

**DEVELOPMENT OF A VEHICLE CRASH PREDICTION MODEL BASED ON  
DRIVER CHARACTERISTICS AND TRAFFIC VIOLATIONS HISTORY**

Prepared by:

Armando González-Bonilla

Project submitted in partial fulfillment of the requirements for the degree of  
Master of Engineering in Civil Engineering

University of Puerto Rico at Mayagüez

2017

Approved by:

\_\_\_\_\_  
Ivette Cruzado-Vélez, PhD  
President, Graduate Committee

\_\_\_\_\_  
Date

\_\_\_\_\_  
Enrique González-Vélez, PhD  
Member, Graduate Committee

\_\_\_\_\_  
Date

\_\_\_\_\_  
Pedro A. Torres-Saavedra, PhD  
Member, Graduate Committee

\_\_\_\_\_  
Date

\_\_\_\_\_  
Ismael Pagán-Trinidad, MSCE  
Director of Civil Engineering and Surveying Department

\_\_\_\_\_  
Date

\_\_\_\_\_  
Hilton Alers-Valentín, PhD  
Graduate School Representative

\_\_\_\_\_  
Date

## Table of Contents

List of Figures .....	4
List of Tables .....	5
Acknowledgements.....	8
1. Introduction .....	9
1.1 Background .....	9
1.2 Goals and Objectives.....	10
1.3 Benefits of the Study.....	10
2. Literature Review .....	11
3. Description of data.....	23
3.1 Data Collection.....	23
3.2 Raw Dataset Results.....	25
3.2.1 General Information .....	26
3.2.2 Traffic Violations History .....	29
3.2.3 Traffic Crash History.....	31
3.3 Database Development.....	34
3.4 Descriptive Statistics .....	36
3.4.1 General Information .....	36
3.4.2 Traffic Violations .....	40
3.4.3 Traffic Crashes .....	45
4. Analysis Methodology.....	47
4.1 Preliminary Analyses .....	47
4.1.1 Contingency Tables .....	48
4.1.2 Chi-Square Test of Independence.....	48
4.1.3 Simple Logistic Regression .....	51

4.2 Model Selection.....	54
4.3 Model Assessment.....	55
5 Analysis Results .....	58
5.1 Preliminary Analysis.....	58
5.1.1 Contingency Tables .....	58
5.1.2 Chi-Square Test of Independence.....	68
5.1.3 Simple Logistic Regression .....	73
5.2 Model Development.....	77
5.3 Model Assessment.....	87
6. Conclusions and Recommendations .....	90
6.1 Conclusions.....	90
6.2 Recommendations .....	98
References.....	99
Appendix.....	101
A.1 Committee for Protection of Human Rights in Research Approval Document .....	101
A.2 Example of Survey .....	102
A.3 Chi-Square Test of Independence Results.....	104
A.4 Simple Logistic Regression Raw Output .....	108

## LIST OF FIGURES

Figure 1: Distribution of Age in Raw Data	26
Figure 2: Distribution of Sex in Raw Data	27
Figure 3: Distribution of Years driving a vehicle a Vehicle in Raw Data	28
Figure 4: Distribution of Daily Hours Spent Driving Reported by Participants	29
Figure 5: Reception of Traffic Violations Among Participants	30
Figure 6: Distribution of Traffic Violation Type among Participants	31
Figure 7: Distribution of Crash Involvement Reported by Participants	31
Figure 8: Distribution of Age in Crashes Reported by Participants	32
Figure 9: Distribution of Crash Severity among Crashes Reported	33
Figure 10: Sample Distribution of Drivers Based on their Age	37
Figure 11: Sample Distribution of Drivers Based on Sex	37
Figure 12: Sample distribution of Females Based on their Age	38
Figure 13: Sample Distribution of Males Based on their Age	38
Figure 14: Sample Distribution Based on Years driving a vehicle	39
Figure 15: Sample Distribution Based on Drivers Daily Hours Spent Driving	40
Figure 16: Distribution of Drivers Based on whether they Received Traffic Violations or Not	40
Figure 17: Sample Distribution of Traffic Violations	44
Figure 18: Distribution of Total Traffic Violations Based on Age and Sex	44
Figure 19: Sample Distribution of Crash Involvement Among Participants	45
Figure 20: Sample Distribution of Crash Severity among Crashes Reported	46
Figure 21: Distribution of Total Crashes Based on Age and Sex of Participants	46
Figure 22: Odds ratios for Reckless/Maneuvering Violations	83
Figure 23: Odds ratios for Non-Reckless/Maneuvering Violations	84
Figure 24: Odds ratios for Reckless/Maneuvering Violations (Model Without Variable of Sex)	87
Figure 25: Odds ratios for Non-Reckless/Maneuvering Violations (Model Without Variable of Sex)	87
Figure 26: ROC Curve for the Selected Model (Obtained from Minitab)	88
Figure 27: Page 1 of 2 from the Developed Survey	102
Figure 28: Page 2 of 2 from the Developed Survey	103

## LIST OF TABLES

Table 1: Variables Considered in Study	12
Table 2: Significant Variables Found in Literature Review	21
Table 3: Significant Variables Found in Literature Review (Continued)	22
Table 4: Distribution of Age in Raw Data	26
Table 5: Distribution of Sex in Raw Data	27
Table 6: Distribution of Years driving a vehicle a Vehicle in Raw Data	28
Table 7: Distribution of Daily Hours Spent Driving Reported by Participants	29
Table 8: Reception of Traffic Violations among Participants	30
Table 9: Distribution of Traffic Violation Type among Participants	30
Table 10: Distribution of Crash Involvement Reported by Participants	31
Table 11: Distribution of Age in Crashes Reported by Participants	32
Table 12: Distribution of Crash Severity among Crashes Reported	33
Table 13: Categorical Variables, Table 14: Categorical Variables	35
Table 15: Descriptive Statistics for Age	36
Table 16: Descriptive Statistics for Sex	37
Table 17: Descriptive Statistics for Years driving a vehicle	39
Table 18: Descriptive Statistics for Daily Hours Spent Driving	39
Table 19: Descriptive Statistics for Whether Participants Were Involved in Traffic violations or Not	40
Table 20: Descriptive Statistics for Driving Over the Speed Limit	41
Table 21: Descriptive Statistics for Driving Under the Influence of Alcohol or Drugs	41
Table 22: Descriptive Statistics for Ignoring Traffic Signals and Signs	42
Table 23: Descriptive Statistics for Driving too Close to Front Vehicle	42
Table 24: Descriptive Statistics for Illegal Turning	42
Table 25: Descriptive Statistics for Illegal Parking	42
Table 26: Descriptive Statistics for Illegal Lane Switch	43
Table 27: Descriptive Statistics for Not Using Seatbelt While Driving	43
Table 28: Descriptive Statistics for Using Cellphone While Driving	43
Table 29: Descriptive Statistics for Other Traffic violations	43
Table 30: Distribution of Participants Based on Crash Involvement	45

Table 31: Sample Distribution of Crashes Reported Based on Severity	45
Table 32: Example of a Contingency Table	48
Table 33: Classification Table for ROC Curve	56
Table 34: Contingency Table Age vs Crash Involvement	59
Table 35: Contingency Table Sex vs Crash Involvement	59
Table 36: Contingency Table for Driving over the Speed Limit Violations vs Crash Involvement	60
Table 37: Contingency Table for DUI Violations vs Crash Involvement	60
Table 38: Contingency Table for Ignoring Traffic Signs and/or Signals Violations vs Crash Involvement	61
Table 39: Contingency Table for Driving Too Close to Front Vehicle Violations vs Crash Involvement	61
Table 40: Contingency Table for Illegal Parking Violations vs Crash Involvement	62
Table 41: Contingency Table for Illegal Turn Violations vs Crash Involvement	62
Table 42: Contingency Table for Reckless Lane Switch Violations vs Crash Involvement	63
Table 43: Contingency Table for No Seatbelt Used Violations vs Crash Involvement	63
Table 44: Contingency Table for Cellphone Use Violations vs Crash Involvement	64
Table 45: Contingency Table for Other Violations vs Crash Involvement	64
Table 46: Classification of Traffic Violations	66
Table 47: Contingency Table for Age vs Crash Involvement (After Merging Categories)	67
Table 48: Contingency Table for Reckless/maneuvering violations vs Crash Involvement	67
Table 49: Contingency Table for non-Reckless/maneuvering violations vs Crash Involvement	67
Table 50: Results of Chi-Square Test of Independence for Age	69
Table 51: Results of Chi-Square Test of Independence for Sex	70
Table 52: Results of Chi-Square Test of Independence for Reckless/maneuvering violations	71
Table 53: Results of Chi-Square Test of Independence for Non-Reckless/maneuvering violations	73
Table 54: Results of Simple Logistic Regression Analysis	75
Table 55: Results for Stepwise Backwards Elimination Procedure	78
Table 56: Results of Final Estimation Model	80
Table 57: Calculated Odds Ratios for Terms in Final Estimation Model	83

Table 58: Results for Stepwise Backwards Elimination Procedure (Without the Variable of Sex)	85
Table 59: Results of Final Estimation Model (Without the Variable of Sex)	86
Table 61: Calculated Odds Ratios for Terms in Final Estimation Model (Without the Variable of Sex)	86
Table 61: Observed and Expected Frequencies for Hosmer-Lemeshow Test	88

## **ACKNOWLEDGEMENTS**

The author would like to thank the Transportation Informatics University Transportation Center for the funding provided for the development and exposure of this project. Also, the authors would like to acknowledge Maria Torres-Rodriguez and Ivette Cruzado-Velez for their collaboration in collecting the data for this project.



## **1. INTRODUCTION**

Transportation systems should be designed to move people and goods in an efficient and safe manner. Safety on roads and highways can be measured in terms of the number of traffic crashes that occur in a time period. Although there are many factors that could contribute to the occurrence of a vehicle crash, human factors are usually the most associated with a vehicle crash. Some of the common causes related to human factors that are associated with vehicle crash occurrence are distraction, reckless driving and driving under the influence of alcohol. Identifying drivers whose behavior and characteristics make them prone to being involved in vehicle crashes may be helpful in reducing the number of injuries and fatalities along roads and highways.

### **1.1 Background**

Highway safety has been identified as a top priority in the United States and all around the world. Road traffic crashes are one of the global leading causes of death and injuries. According to the World Health Organization (WHO), in the year 2012, road traffic crashes were the leading cause of death for people between the ages of 15-29 years old. In addition to the undesirable effects that traffic crashes have on highway safety, economy is also affected. Road traffic crashes cost countries approximately 3% of their gross national product; this figure can rise to 5% in some low and middle-income countries (WHO, 2015). Several program initiatives all over the world have taken place to reduce the number of road traffic crashes. The Highway Safety Improvement Program is one example of the continuous effort that countries all over the world are making to improve road safety. Since the year 2007, the number of road traffic deaths has plateaued despite the increase in population, motorization and the predicted rise in deaths (WHO, 2015). This suggests that the efforts made to improve road safety have revealed satisfactory results.

It is commonly acknowledged that factors such as human factors, vehicle characteristics, road design and environmental factors can contribute to the occurrence of vehicle crashes. Since human factors usually have a major influence on vehicle crash occurrence, studies normally focus on the effect that some driver characteristics such as age, sex, alcohol usage and driving have on the occurrence of a vehicle crash. One of the topics that these types of studies explore is the effect that a driver's traffic violations and crash history has on the same driver being involved in a future traffic crash. Several studies have shown that there is a positive correlation between previous traffic violations and crashes and vehicle crash involvement (Gebers, 1999). Thus, the purpose of

this research project is to develop a statistical model that could be used to estimate the likelihood of being involved in a vehicle crash based on a series of human factors such as age and sex as well as traffic violations and crash history. The research approach includes the collection and study of existing traffic violation and crash records databases (if possible), identification of possible variables that could be used for the development of the model, development and assessment of the proposed model it provides a good represent for the phenomena under study.

## **1.2 Goals and Objectives**

The main goal of this project was to develop a statistical model that incorporates several factors such as traffic violation and crash data to estimate the likelihood of involvement in a future vehicle crash. To reach this goal, several specific objectives were identified:

- Perform a review of past studies with the purpose of exploring significant factors and methodologies commonly used in crash prediction models,
- Collect traffic violations and crash data for the driving population of Puerto Rico,
- Develop a new database using the previously collected driver records and crash databases,
- Identification of possible variables for estimating the likelihood of crash involvement for drivers in Puerto Rico,
- Develop a statistical model using on the newly created database from which the likelihood of a driver being involved in a traffic crash could be estimated, and
- Assess the developed model using appropriate statistical tests and procedures.

## **1.3 Benefits of the Study**

Studying factors and characteristics of drivers and their behavior that could affect the likelihood of being involved in a vehicle crash can help in identifying high-risk drivers as well as developing measures to attend the situations caused by this type of drivers. The model that is going to be developed could potentially serve as a monitoring tool to identify possible trends in the future for driver characteristics in Puerto Rico. This study should provide the reader with an insight of how human factors affect the likelihood of future crash involvement for drivers in Puerto Rico.

## 2. LITERATURE REVIEW

Factors associated with likelihood of road traffic crashes are normally grouped into four categories: human, roadway, vehicle, and environment. Engineers are constantly working towards making roads and vehicles safer to reduce the amount and severity of road traffic crashes. Although there is undergoing research regarding every one of the aforementioned categories, studies normally focus on driver characteristics such as age, sex and driver behavior, since human factors are the most significant factors that influences the occurrence of traffic crashes. Most of the traffic crashes and violations committed by drivers are associated with reckless or negligent driving and not obeying traffic rules. The following literature review seeks to identify and understand which factors regarding human characteristics and behavior are most commonly associated with traffic violations and crashes. In addition to exploring common factors, this literature review also has the purpose of identifying common methodologies used for studying the relationship that these factors have on the occurrence of traffic crashes.

A study conducted by Michael Gebers (Gebers, 1999) in California consisted of developing crash prediction models to determine the crash probability of high risk/crash prone drivers. The data used for this study consisted of driving records of a sample of approximately 140,000 licensed drivers. The information used consisted of variables such as age, sex, total crashes and total citations for each driver of the sample. Multiple logistic regressions were used to develop the prediction models used for analyzing the data. The dependent variable for the models was the probability of one outcome, such as crash involvement, while the independent variables consisted of different combinations of driver characteristics. Goodness-of-fit tests were used to select the model that had the better prediction capability while using the fewest independent variables. A total of 17 models with different combinations of independent variables were identified for comparison (Gebers, 1999). The results from the study indicated that models where previous total crashes were considered performed better than models that did not consider this variable. Additionally, models that used demographic information, such as age and sex, licenses class and various combinations of traffic citations and crashes, performed better than others that did not considered these factors. The two models that produced the best fit were then selected to determine which factors were most influential on crash occurrence. Several variables, such as being young, being male, holding a commercial driver's license, and increased prior citation and crash frequency, were identified as being of significance in the determination of crash occurrence

probability (Gebers, 1999). The models were compared in terms of their classification and prediction accuracy.

A study performed by the Department of Motor Vehicles of California aimed to explore the viability of predicting future crashes for the general driving population of California based on equations developed previously for estimating future traffic convictions (Gebers and Peck, 2000). The data used for this study corresponded to driver records for a 1% sample of licensed drivers in California (246,000 drivers) extracted from the 1992 driver license database. The variables used in this study were classified into license or driver record variables or territorial variables inside the ZIP-Code of residence. Table 1 provides the variables that were considered:

*Table 1: Variables Considered in Study*

Licensing and driver record variables	Territorial variables within ZIP-Code of residence
<ul style="list-style-type: none"> <li>• Sex (0 = Man; 1 = Woman)</li> <li>• Age (at the beginning of the criterion period)</li> <li>• Prior 3-year total accidents as defined below</li> <li>• Prior 3-year total citations as defined below</li> <li>• Possession of a commercial driver license (0 = no; 1 = yes)</li> <li>• Presence of a physical or mental condition on record (0 = no; 1 = yes)</li> <li>• Presence of a driver license restriction on record (0 = no; 1 = yes)</li> </ul>	<ul style="list-style-type: none"> <li>• % Black</li> <li>• % Hispanic</li> <li>• % on public assistance</li> <li>• % Unemployed</li> <li>• % age 55 or older</li> <li>• Median annual household income (\$)</li> <li>• 3-year (1989-91) mean ZIP-Code total citations</li> <li>• 3-year (1989-91) mean ZIP-Code total accidents</li> </ul>

Total accidents and total citations were the dependent variables being estimated. Multiple linear regression analyses as well as canonical correlations analyses were performed to determine the combination of predictors that would provide the best estimation of future accidents and future citations. A confidence level of 90% was used to determine if a predictor should have been included or not in the model. A cross-validations procedure was performed to validate the models for estimating each of the dependent variables. The results of the multiple linear regression and canonical correlations analysis presented in the study indicated that the following predictors were significant (Gebers and Peck, 2000):

- Increased prior citation frequency,
- Increased prior accident frequency,
- Having a commercial driver license (which is mostly held by high-mileage professional drivers),
- Being young,
- Being male,
- Having a commercial driver license,
- A higher percentage of Blacks residing within a ZIP-Code area,
- A higher percentage of Hispanics residing within a ZIP-Code area,
- A higher median income within a ZIP-Code area,
- Having one or more PandM conditions on record, and
- Having one or more driver license restrictions on record.

The researchers concluded that the risk level of a traffic crash for a group of drivers can be better predicted than for a simple driver. Additionally, the researchers also concluded that any of the methodologies presented can be used to identify crash prone drivers, although the canonical correlation analysis resulted to be better.

In Quebec, a study was performed to describe the most common type of accidents in the elderly population (65 years or older) that lives in Quebec. Driver records for 426,408 elder drivers were analyzed for the time period of 1992 to 1997. The data was initially analyzed using descriptive statistics followed by a linear regressions analysis to determine the association between previous crashes and convictions and the likelihood of future crashes.

The results presented in this study for the descriptive statistics analysis indicated that elder drivers are characterized for crashes that involve more than one vehicle. Results also indicated that

right-angle and left turn crashes on intersection increase for driver older than 65 years while single vehicle crashes decrease (Daigneault, et al., 2002). The researchers concluded that crashes involving elder drivers were not related to difficult climate conditions or risk taking, instead, most of the crashes occurs because of situations that required perception of various details and the processing of complicated information (Daigneault, et al., 2002). The results for the linear regression analyses performed indicated that there was a significant association between convictions and previous accidents and the risk of future accidents as drivers get older, however, the amount of traffic convictions decreases as drivers get older (Daigneault, et al., 2002). According to the researchers, this can be attributed the driving habits of elder drivers such as; not driving at night or difficult climate conditions and driving at a lower speed than younger drivers.

A study took place in Australia to examine the relationship between the culpability of drivers and their fatal crash and violations history. The data sample used consisted of 388 vehicle and motorcycle drivers in South Australia who were involved in multiple fatal crashes for the years 1999 to 2002 (Wundersitz et al., 2004). The average age of drivers was 41.7 years with 299 being male and 89 females (Wundersitz et al., 2004). A total of 36 motorcycle drivers were included. Traffic violations were obtained by matching a driver's license number with the driver license and traffic violations database (Wundersitz et al., 2004). Traffic violations regarding illegal parking and speeding were not considered. The number and type of previous crashes and violations for culpable and non-culpable drivers involved in multiple crashes were compared to determine any relationship between (Wundersitz et al., 2004). The results presented in this study indicated that drivers who were responsible for a fatal crash were more likely to be younger than 25 years of age and older than 75 years. At the moment of a crash, drivers who were responsible for a crash were more likely to be under the effect of alcohol than non-responsible drivers. Being involved in previous crashes, regardless of culpability, resulted not to be associated with future multiple fatal crashes (Wundersitz et al., 2004). Violation for driving under the influence of alcohol was the only type of traffic violations associated with the culpability of multiple fatal crashes. Culpable drivers had more than triple of traffic violations for driving under the influence of alcohol than non-culpable drivers, although the sample of drivers who committed this violation was small (Wundersitz et al., 2004).

Similar to the study conducted by Michael Gebers in 1999, Chandraratna and Stamatiadis developed a logistic regression model to predict the likelihood that a driver will be responsible of

a future crash. Data used for this study consisted of driver license and crash databases of the state of Kentucky for the years 1995 to 2002. Only drivers who had at least two crash involvements were selected for the sample analyzed. Variables such as age, sex, at fault or not at fault crashes, and speeding, were selected for the development of the regression model. The dependent variable of the model was the probability of being at fault in a crash while all other variables were considered as independent variables. The model developed by Chandraratna and Stamatiadis was validated by means of a holdout procedure where a set of the data obtained was not used for development of the model in order to be used later to validate the model. Significance at the 0.05 level was used to test the variables used in the model. The results of the study indicated that being at-fault and not at-fault in a crash, traffic school attendance, driver license suspensions, non-speeding violations, time between last two crashes, driver age, sex, and crash type were significant at the 0.05 alpha level (Chandraratna and Stamatiadis, 2004).

The American Transportation Research Institute (ATRI) published a study in which the relationship between the behavior of truck drivers who had certain driving records and future crash involvement was explored by means of a logistic regression model (Murray, 2006). The dependent variable for this model was crash involvement while the independent variables were driver specific performance indicators. The data used in this study consisted of driver information regarding their past vehicle inspections, crashes, and convictions. The data was initially explored with descriptive statistics to determine possible subsets of data to be analyzed. After creating an initial model with the selected subsets of data, a stepwise regression procedure was performed to determine the most significant variables for the overall model (Murray, 2006). The results of the study yielded that significant variables such as reckless driving and serious speeding violations cause an increase in likelihood of future crash involvement. Additionally, drivers who had a past crash experience also had a significant increase in their likelihood of future crash involvement.

In Japan, a research performed by Yasushi Nishida took place with the goal of studying the relationship between crash and violation experiences and crash involvement rates (Nishida, 2009). The crash involvement rates by crash and violation experience were determined using driver information such as prefecture (county), age, sex, and type of traffic violations of all drivers. A crosstab analysis was performed to determine the crash involvement rate based on the recorded crashes and traffic violations of previous years. The results of the study indicated that, as the number of crash or violation records increases, crash involvement rate increases at the same rate

(Nishida, 2009). According to the authors, driving behavior and frequency of driving are the two factors that significantly contribute to increased crash involvement rate. Results from the aforementioned studies indicate that traffic violations and road traffic crash involvement have a positive correlation with crash involvement rate. Additionally, demographic information such as age and sex and driving experience are also important factors that should be considered.

A study performed in Eskisehir, Turkey aimed to identify if there is a relationship between traffic crashes and the age and sex of drivers. The data use in the analysis was classified into: age, sex, fundamental faults and minor faults. Sex was divided into males and females while age was divided into the following categories:

- 20 or younger,
- 21-30,
- 31-40,
- 41-50,
- 51-60,
- 61-70, and
- 71 or older.

Fundamental faults corresponded to fault resulting from violations of traffic laws while minor faults corresponded to reckless behavior while driving. The crashes that were examined in this study occurred between January 1 and December 31 of 2009. A preliminary analysis was initially performed using tables to explore the distribution of frequencies for the different faults based on the age and sex of drivers. Subsequently, chi-square tests were performed to determine if there was a significant relationship between traffic faults, age and sex. The results of this study indicated the following:

- Males were involved in more crashes than females.
- The most common fundamental fault in males was “violation of passing priorities in intersections” while females committed more “crashing into the back of a vehicle” violations.
- Males and females mainly committed the minor fault regarding “failure to decrease vehicle speed while approaching junctions, curves, hilltops, pedestrian crossings, level crossings, tunnels, narrow bridges and vents/outlets and while entering into construction and maintenance areas.”



- Drivers between the ages of 21 and 30 were the most who committed fundamental and minor faults while drivers 71 years and older were the ones who committed the least amount of minor and fundamental faults.

The researches of this study concluded that there is a significant relationship between age, sex and traffic faults. Specifically, young drivers caused more accidents than middle age and elderly drivers. Young drivers also caused more crashes because of driving under the influence of alcohol and disobeying speed limits.

Zhang et.al performed a study that aimed to determine the risk factors associated with traffic violations and crash severity. The research team analyzed traffic crash data for the period 2006-2010 in the Guangdong Province in China. Tables were constructed to associate risk factors with traffic violations and injury severity. Chi-square tests of independence were conducted to evaluate the significance of the various variables at the 0.05 alpha level of significance. Similar to the studies mentioned previously, logistic regression analyses were used to estimate the effect of different predictor variables on the likelihood of traffic violations and injury severity. The dependent variables were identified as whether there was any traffic violation and whether the crash resulted in a fatality or a serious injury. Initially, all factors were included in the model and insignificant factors were subsequently removed by an iteration process. The results of this study indicated that traffic violations have a positive relationship with crash severity (Zhang, 2013). This conclusion was determined by comparing the proportion of drivers who were involved in a fatal/serious injury crash and a traffic violation with drivers who were involved in a fatal/serious injury crash but were not involved in a traffic violation (Zhang, 2013). Additionally, the results showed that driver sex is one of the most important factors associated with traffic violations. Males, unfit safety status, overload in a vehicle, no street lightning at night, bad visibility, and weekends were variables that also resulted in an increased likelihood of being associated with traffic violations and crashes (Zhang, 2013). The study also indicated that novice drivers (drivers with less than two years driving a vehicle) have a significantly higher risk of performing traffic violations. Moreover, drivers with three to five years driving a vehicle, passenger vehicles, bad weather, and morning rush hour also seemed to display an increased risk of traffic violations.

The Puerto Rico Department of Transportation and Public Works (DTOP for its meaning in Spanish) in collaboration with the Puerto Rico Highway Authority have developed the Strategic Highway Safety Plane of Puerto Rico. This plan seeks to reduce the fatalities and injuries

associated with highway crashes. As part of this plan, an analysis of data regarding the causes of these crash is presented. The results indicate that the following causes were the most associated with crashes in Puerto Rico for the years 2007 to 2009, and 2012:

- Aggressive driving,
- Vulnerable road users,
- Driving under the influence of alcohol,
- Runoff road crashes,
- Young drivers (18 to 24 years old), and
- Intersections.

A study performed in Louisiana aimed to study the impact that drivers who are more prone to traffic crashes have on safety and to estimate how the traffic crash history of these drivers affect future crash involvement. The data used was obtained from the grouping of various Microsoft Access tables from which characteristics regarding crashes, roads and vehicles was obtained for a period of eight years (2004-2011). Drivers were divided into the following four categories for analysis purposes:

- Not at-fault crash prone drivers,
- At-fault crash prone drivers,
- Not at-fault non-crash prone drivers, and
- At-fault non-crash prone drivers.

Drivers who experimented multiple crashes were defined as crash prone while drivers who were responsible for the crash were defined as at-fault. Drivers who experimented one crash only were defined as non-crash prone. Researchers selected some of the variables contained in the crash database with the purpose of selecting non-redundant ones. A linear regression was performed to further remove non-significant variables. An exhaustive search was performed using the software R to find the subset of variables that could best estimate the dependent variable. The selection of variables that were used to develop the model were:

- Driver culpability,
- Alcohol,
- Road alignment,
- Road lightning,
- Crash Severity,

- Crash type,
- Sex,
- Age,
- Driver distraction, and
- Drugs.

Data regarding these variables was analyzed using descriptive statistics. The results of the descriptive statistics analysis indicated that men between the ages of 15 and 24 years of age are more prone to crash involvement. Road lightning was a problem for crash prone drivers. The rate of fatalities and sever injuries were higher for at-fault crash prone drivers. Alcohol and drugs were seen with more frequency in crash prone drivers. Crashes where the vehicle ended off road were more frequent in at-fault drivers, crash prone or not. At-fault crash prone drivers were involved in more crashes on curved alignments. After the descriptive statistics analysis, a logistic regression model was developed using the variables mentioned a previously as predictors and culpability of crash prone drivers as the dependent variable. In addition to the logistic regression model, an analysis of variance (ANOVA) was performed. The Akaike Information Criterion was used to compare different models. The Receiving Operator Characteristic (ROC) curve was used to evaluate the success of the best model. The area under the ROC curve showed a value of 0.77, which according to the authors, is generally accepted. The resulting model could have been used to correctly identify as many as 62.40% of the incidence of at-fault drivers in the following year of data.

A study conducted in Abu Dhabi, capital of the United Arab Emirates, aimed to predict the probability of a high-risk driver to be involved in future recurrent crashes based on their driver record between the years 2008 -2015. The data used on this study was taken from four different databases of the traffic division from the Abu-Dhabi traffic police. These four databases correspond to:

- Traffic violations,
- Property Damage crashes,
- Severe crashes, and
- Driver License.

Driver data from these four databases were integrated into a single database using a unique code provided to each driver when they receive their license. In 2015, the total number of driver's

licenses was 1,234,009. Data filtration processes were carried out to select the data sample that would be used in this study. The final data sample included 324,644 drivers, with a total of 4,116,149 traffic violations, 578,619 property damage crashes and 7,676 severe crashes. This database shows detailed history of each driver as well as groups of certain traffic violations.

The analysis of this study was divided into two parts; the first one consisted on developing the relationship between severe crash frequency based on the type of traffic violation, demographic characteristics and frequency of the type of violation and crashes, while the second part consisted on estimating a model for determining the best variables or predictors that could be used to identify high risk drivers. Traffic violations were classified in two categories, the first category included speed related violations in which there were six types of violations. The second category included violations related to risky behavior in which there were 14 types of violation.

After the analysis performed in the first part of the methodology, the researchers developed the model that would be used to identify the most significant variables to determine high risk drivers. The model used in this study was a negative binomial regression. Eight models of different combinations of variables were developed and compared using the Akaike Information Criterion. The results of this study indicated that there is a strong relationship between frequency of severe crashes and their history of collisions of property damage and traffic violations. Other factors that were found to be related to an elevated risk of severe crashes were; women, young drivers, local drivers and few years of experience. The model that was selected indicated that the following traffic violations proved to be the best variables to identify high risk drivers; exceeding speed limit by more than 60 km/h, exceed speed limit by values between 50 and 60 km/h, dangerous driving behavior, use of alcohol, use of cell phone, driving near front vehicle, entering the taxiway suddenly, not wearing a protective helmet, and violations related to passing other vehicles.

Most of the studies included in this literature review used driver record databases for collecting the data and information of their unit of analysis, mainly drivers and vehicles, although questionnaires can also be used. The databases used in these studies were most likely created from the driver records that law enforcement officials obtain from traffic crashes and violations. Table 2 and table 3 summarize the variables that were found to be significant on the studies included in this literature review. Results show that several of the studies agreed on the fact that sex, age and prior traffic citations and crashes are significant factors for estimating of future traffic crashes.

Other significant factors that were included in some of these studies are driving behavior and type of license.

Table 2: Significant Variables Found in Literature Review

Study	Variables Found to be Significant
Gebers, 1999	Previous Total Crashes Age Sex Being young Being Male Holding a Commercial Driver's License Increased prior citation and crash frequency
Gebers and Peck, 2000	Increased prior citation frequency Increased prior accident frequency Having a commercial driver license Being young Being male Having a commercial driver license A higher percentage of Blacks residing within a ZIP-Code area A higher percentage of Hispanics residing within a ZIP-Code area A higher median income within a ZIP-Code area Having one or more P&M conditions on record Having one or more driver license restrictions on record
Daigneault, et al., 2002	Previous crashes
Wundersitz et al., 2004	Drivers younger than 25 years of age Drivers older than 75 years Driving under the influence of alcohol
Chandraratna and Stamatiadis, 2004	Being At-fault in a Crash Being not At-fault in a Crash Traffic School Attendance Driver License Suspension Non-Speeding Violations Time Between Last Two Crashes Age Sex Crash Type
Murray, 2006	Reckless Driving Speeding Violations Past Crash Experience
Nishida, 2009	Traffic Crashes Traffic Violations Driving Behavior Frequency of Driving

Table 3: Significant Variables Found in Literature Review (Continued)

Study	Variables Found to be Significant
Zhang,2013	Traffic Violations Males Unfit safety status Overload in a vehicle No Street Lightning at Night Bad Visibility Weekends
SHSP, 2014	Age Sex Traffic faults
Subasish D., et al, 2015	Driver culpability Alcohol Road alignment Road lightning Crash Severity Crash type Sex Age Driver distraction Drugs
Shawky and Al-Ghafli, 2016	Exceeding Speed Limit by More than 60 kph, Exceed Speed Limit by Values Between 50 and 60 kph, Dangerous Driving Behavior, Use of Alcohol, Use of Cell Phone, Driving Near Front Vehicle, Entering the Taxiway Suddenly, Not Wearing a Protective Helmet Violations Related to Passing Other Vehicles

### 3. DESCRIPTION OF DATA

This chapter provides information that describes the data used for the development of the proposed models of this study. A description of the data collection process used to obtain a sample from the population of licensed drivers in Puerto Rico is described. The data collection process includes development of the survey created to obtain information from licensed drivers in Puerto Rico in addition to indicating how the distribution of the survey took place. The resulting dataset obtained from the survey results is described afterwards, first, a dataset of raw data is described on which a data filtration process was performed in order to create a uniform dataset. Following the data filtration process, the final dataset that was used for analysis and model development is presented with the respective descriptive statistics to give an overview of the results obtained from the survey responses.

#### 3.1 Data Collection

Most of the studies revised in the literature review section used driver records and crash databases as the main source of information for their analysis. Unfortunately, lack of access to driver records made difficult the acquiring of information regarding traffic violations of drivers and thus prevented the use of this type of database. Given this issue, a survey was performed on a sample of the driving population of Puerto Rico to obtain data regarding history of traffic violations and crashes. The only requirement for participants of this survey was to have experience driving motorized vehicles. The questions included on this survey were selected based on the findings of the literature review regarding significant variables that contribute to increased likelihood of future crash involvement.

The survey used in this study had two versions; electronic and paper. The electronic version of the survey was created using *SurveyMonkey*, a web based tool created for developing surveys and questionnaires. This tool provides the user with various outlets to distribute the surveys or questionnaires such as links to social media sites (Face, Twitter, etc.) and email. The paper version was developed to have another tool for collecting information from subjects that would not necessarily be reached via social media and emails, mainly subjects of advanced age. The paper version was developed to be an identical copy of the electronic version and was distributed by personal interactions on several locations in Puerto Rico. Responses that were collected using the

paper based form were then manually imported into the electronic version of the survey into to have all the responses in one single database.

The survey initially included a brief introduction regarding the purpose of the survey as well as informing participants of the responsibilities and conditions of participating. During the literature review process, several variables such as age, sex, being young (i.e being inexperienced), frequency of driving, traffic violation and crash history were found to be significant variables for estimating likelihood of future crashes. Thus, questions where participants could provide information related to these variables were included in the survey. The survey was categorized in three parts; general information, traffic violations history and traffic crash history. The first part of the survey included questions regarding the following information:

- Age,
- Sex as indicated on the driver license,
- Years of experience driving a private motor vehicle, and
- Daily hours spent driving a motor vehicle.

The question for “Age” was categorized in different intervals that range from 16 to 89 years of age while the questions for “Sex” was categorized in two levels: Males and Females. The reason for categorizing the answers to these questions was to maintain the survey as controlled and user-friendly as possible.

The second part of the survey included questions regarding a participant’s traffic violations history. A list of traffic violations was provided on the survey and participants indicated the amount of violations received on the respective type of traffic violation. The following traffic violations were considered:

- Driving over the speed limit,
- Driving under the influence of drugs and alcohol,
- Ignoring traffic signals and signs,
- Not using safety belt,
- Driving too close to front vehicle,
- Illegal parking,
- Illegal turn,
- Reckless lane switching, and
- Using cellphone while driving.



An additional space was provided so participants could indicate any traffic violation that they received but were not included in the above list. The traffic violations included in this question were consulted with various officers of the Puerto Rico Police Department to have a more detailed list of the most common traffic violations they encounter when on duty. Since the survey was meant to be as controlled as possible, a list where participants would select the choice that better applied to them seemed like a more attractive approach than letting the question open to freely writing an answer.

The third and final part of the survey included questions regarding a participant's history of traffic crashes. In this part, the total amount of crashes the participant has been involved in as a driver were determined in addition to his or her age, severity and responsibility in each of the crashes. For each vehicle crash, the participant had to indicate the following:

- Age at the moment of the crash
- Severity of the crash; participant had to select among the following:
  - Property Damage Only (PDO) - Nobody was injured, only damage to the vehicle or other property.
  - Light (L) - At least one person was injured but no hospitalization was required.
  - Severe (S) - At least one person was hospitalized as a result of injuries from the traffic crash.
  - Fatal (F) - At least one person died because of the traffic crash.
- Responsibility; participant had to select one of the two following options:
  - Responsible - The traffic crash occurred as a result of the participant's actions.
  - Not Responsible - The traffic crash occurred because of actions beyond the participant's control.

An example of the survey used to collect data is provided in Appendix A.1. Once the data collection period was finished, the set of raw data was exported in to a *Microsoft Excel* spreadsheet from the *SurveyMonkey* database as well as the manually collected surveys. Descriptive statistics were developed for this raw dataset and are presented in the following section.

### **3.2 Raw Dataset Results**

Once the data collection period was finished, the responses collected were exported to a Microsoft Excel spreadsheet to begin the process of analyzing the data. The sample obtained with

the survey contains information about drivers ranging from the age of 16 to 89 years old as well as considering both Male and Female drivers. A total of 1005 responses were collected throughout the data collection period. The survey had a completion rate of 95%, meaning that 95% of the participants completed the survey in its entirety whereas 5% did not answer at least one of the questions in the survey. Descriptive statistics for the information obtained from the raw dataset are provided in this section. The information presented was categorized into general information, traffic violations and traffic crashes, corresponding to the three parts of the survey.

### 3.2.1 General Information

The general information section provided information regarding the age, sex, years driving a vehicle and daily hours spent driving of participants. Table 4 and Figure 1 provide information regarding the distribution of age among participants of the survey.

Table 4: Distribution of Age in Raw Data

Age	Response Count	Percent
16-20	219	21.9%
21-30	394	39.4%
31-40	104	10.4%
41-50	108	10.8%
51-60	114	11.4%
61-70	43	4.3%
71-80	15	1.5%
81-89	2	0.2%
<b>Total</b>	<b>999</b>	<b>100%</b>

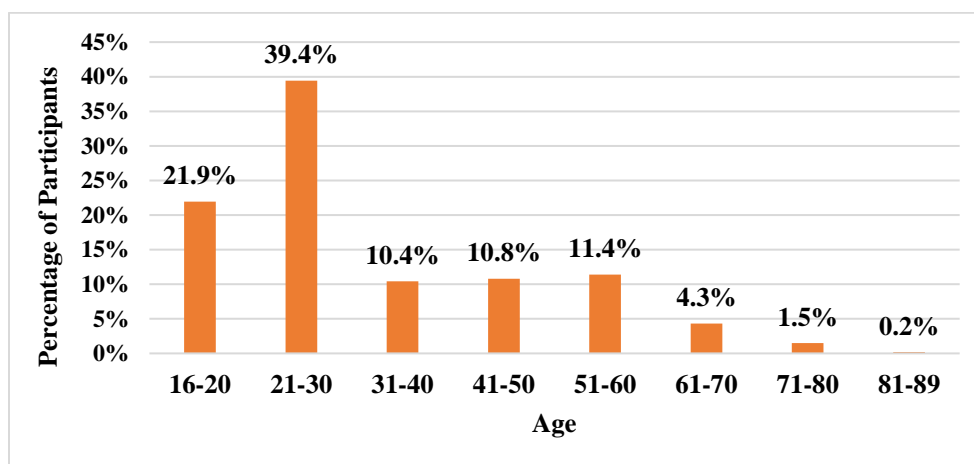
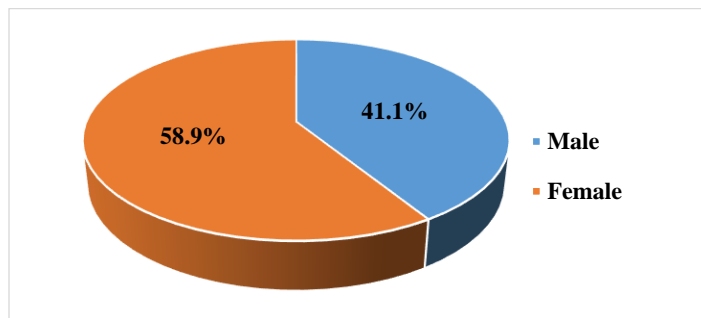


Figure 1: Distribution of Age in Raw Data

Results for the distribution of age indicate that most of the responses collected from the survey correspond to drivers in the age ranges of 16-20 and 21-30 years old. This can be attributed to the fact that most of the responses were collected using social media outlets as well as email for which drivers on this age ranges are more likely to be involved with. The total response count is lower than the reported number of responses collected because several participants skipped this question. Table 5 provides information regarding the distribution of sex among responses in the raw data set. Results from the distribution of sex show that the majority of responses collected corresponded to female participants with a 58.9% percent of responses (compared to the 41.1% of male participants).

*Table 5: Distribution of Sex in Raw Data*

<b>Sex</b>	<b>Count</b>	<b>Percent</b>
<b>Male</b>	409	41.1%
<b>Female</b>	587	58.9%



*Figure 2: Distribution of Sex in Raw Data*

In addition to factors such as age and sex, information regarding years driving a vehicle a vehicle and daily hours spent driving was also considered. Table 6 provides the results for the distribution of years driving a vehicle a vehicle among participants. Most participants reported to have been driving a vehicle for 0 to 10 years. These results were expected since most of the participants of the survey were between the ages of 16 and 30 years of age.

Table 6: Distribution of Years driving a vehicle a Vehicle in Raw Data

Years driving a vehicle	Count	Percent
0-10	556	55.7%
10-20	152	15.2%
20-30	101	10.1%
30-40	108	10.8%
40-50	57	5.7%
50-60	14	1.4%
60-70	3	0.3%
Other	7	0.7%
<b>Total</b>	<b>998</b>	<b>100%</b>

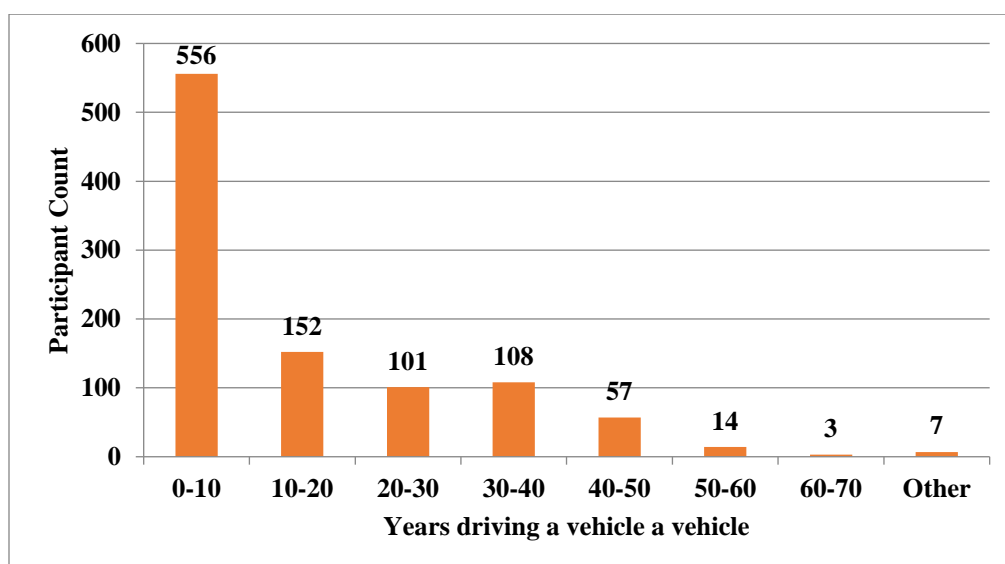


Figure 3: Distribution of Years driving a vehicle a Vehicle in Raw Data

Daily hours spent driving was another factor considered in the general information category. Table 7 provides the results for the distribution of daily hours spent driving reported by participants. According to the results, 62.7% of participants reported to drive an average between 0 and 2 hours on any usual day, while 26.2% reported to drive an average between 2 and 4 hours per day.

Table 7: Distribution of Daily Hours Spent Driving Reported by Participants

Daily Hours Spent Driving	Count	Percent
0-2	626	62.7%
2-4	261	26.2%
4-6	67	6.7%
6-8	11	1.1%
8-10	5	0.5%
More	28	2.8%
<b>Total</b>	<b>998</b>	<b>100%</b>

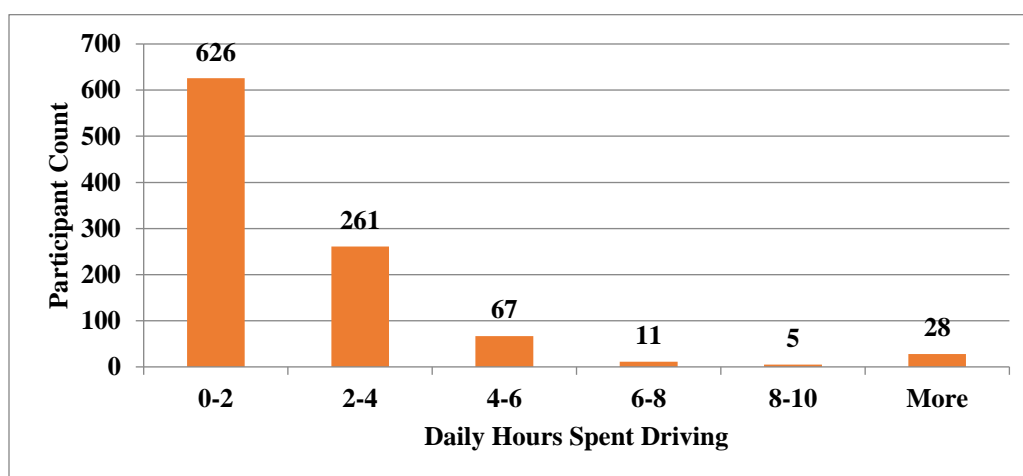


Figure 4: Distribution of Daily Hours Spent Driving Reported by Participants

### 3.2.2 Traffic Violations History

In addition to general information such as age and sex, traffic violations history information was also obtained from the answers provided by participants. Table 8 provides the information regarding response of participants to whether or not they had received traffic violations. The results indicate that 70.5% of participants reported to have received traffic violations in their experience as drivers, Figure 5 provides an illustration of these results.

Table 8: Reception of Traffic Violations among Participants

Reception of Traffic Violations Reported by Participants	Percent	Response Count
Yes	70.5%	690
No	29.5%	289

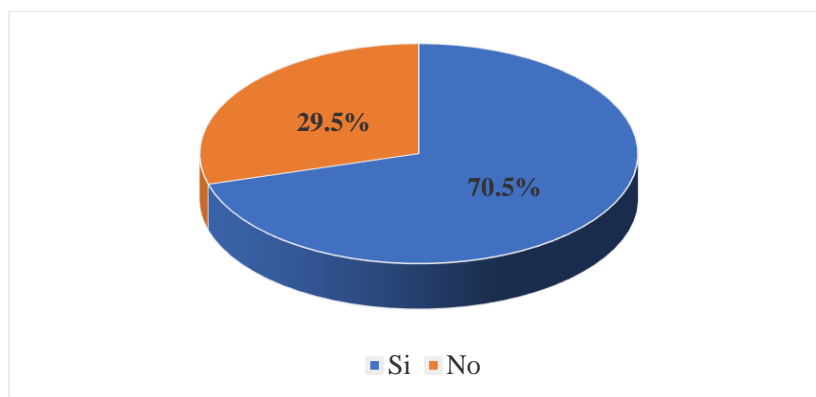


Figure 5: Reception of Traffic Violations Among Participants

The distribution of the type of traffic violations received by participants is provided in Table 9 and Figure 6. According to the information provided by participants, driving over the speed limit is the most common type of traffic violation received (14% of responses), followed by illegal parking (12.9% of responses).

Table 9: Distribution of Traffic Violation Type among Participants

Traffic Violation	Response Count of Traffic Violations Received						Total	Percent
	0	1	2	3	4	5 or more		
Driving Over Speed Limit	116	217	123	43	19	59	577	14.0%
Driving Under the Influence of Alcohol or Drugs	376	8	2	0	0	2	388	9.4%
Ignore Traffic Signals and/or Signs	276	152	16	7	3	4	458	11.1%
Driving too Close to Front Vehicle	364	18	2	0	0	3	387	9.4%
Illegal Parking	156	199	86	42	12	39	534	12.9%
Illegal Turn	344	41	1	2	1	2	391	9.5%
Illegal Lane Switch	337	47	8	1	1	3	397	9.6%
Not Using Seatbelt	291	105	32	3	1	10	442	10.7%
Using Cellphone	319	85	6	2	0	4	416	10.1%
Other	*	*	*	*	*	*	135	3.3%
<b>Total</b>	<b>2579</b>	<b>872</b>	<b>276</b>	<b>100</b>	<b>37</b>	<b>126</b>	<b>4125</b>	<b>100%</b>

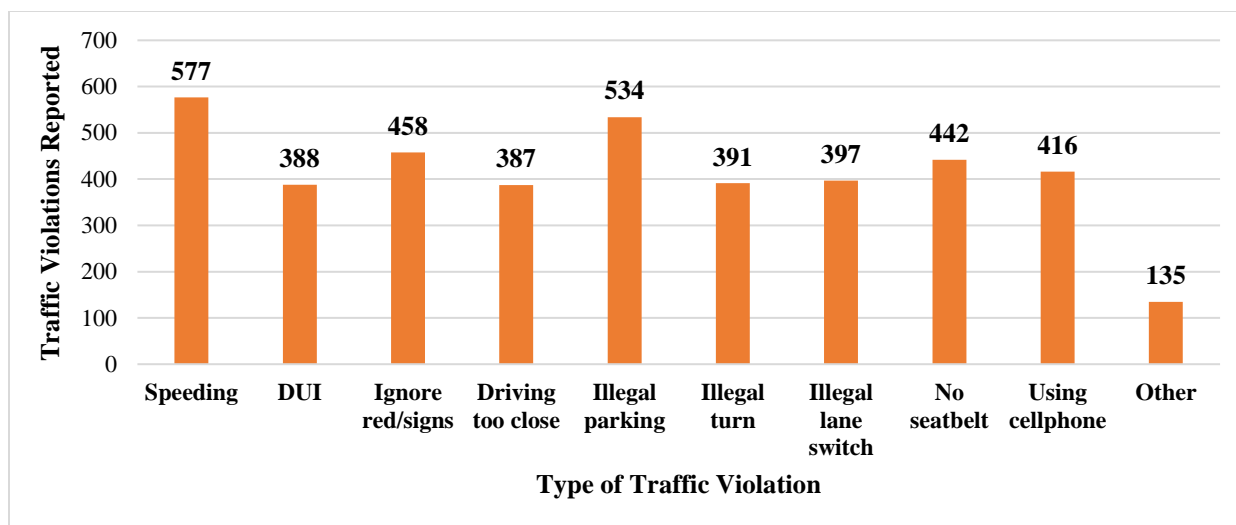


Figure 6: Distribution of Traffic Violation Type among Participants

### 3.2.3 Traffic Crash History

The final category of the raw dataset obtained corresponds to the participant's vehicle crash history. The distribution of participants on whether they had been involved in a crash or not is displayed in Table 11 and Figure 8. Results show that 65.2% of participants indicated that they had been involved in a traffic crash as a driver.

Table 10: Distribution of Crash Involvement Reported by Participants

Crash Involvement Reported by Participants	Response Count	Percent
Yes	627	65.2%
No	335	34.8%

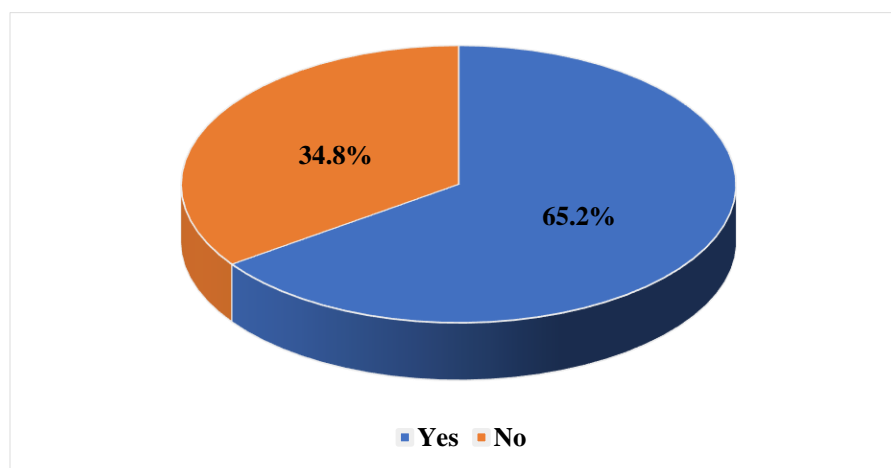
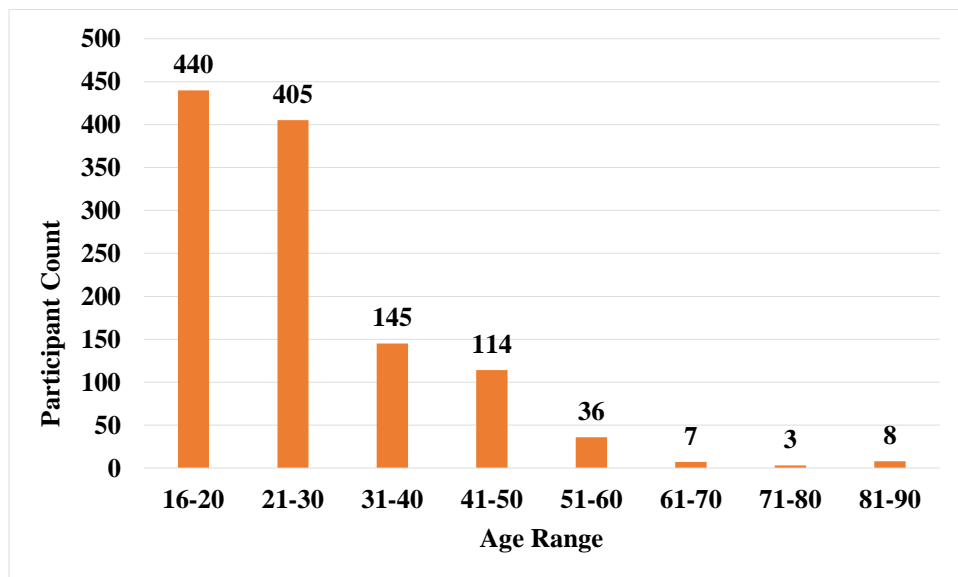


Figure 7: Distribution of Crash Involvement Reported by Participants

Furthermore, the distribution of age among the crashes reported in the survey is provided in Table 12 and Figure 9. According to results provided in this table, most crashes reported in the survey correspond to participants between the ages 16 and 30 years of age. These results indicate that young drivers are the most involved in traffic crashes, which complies with some of the studies included in the literature review which indicated that being young is a significant factor in likelihood of future traffic crashes.

*Table 11: Distribution of Age in Crashes Reported by Participants*

Age	Response Count	Percent
<b>16-20</b>	440	38.0%
<b>21-30</b>	405	35.0%
<b>31-40</b>	145	12.5%
<b>41-50</b>	114	9.8%
<b>51-60</b>	36	3.1%
<b>61-70</b>	7	0.6%
<b>71-80</b>	3	0.3%
<b>81-90</b>	8	0.7%
<b>Total</b>	1158	100%



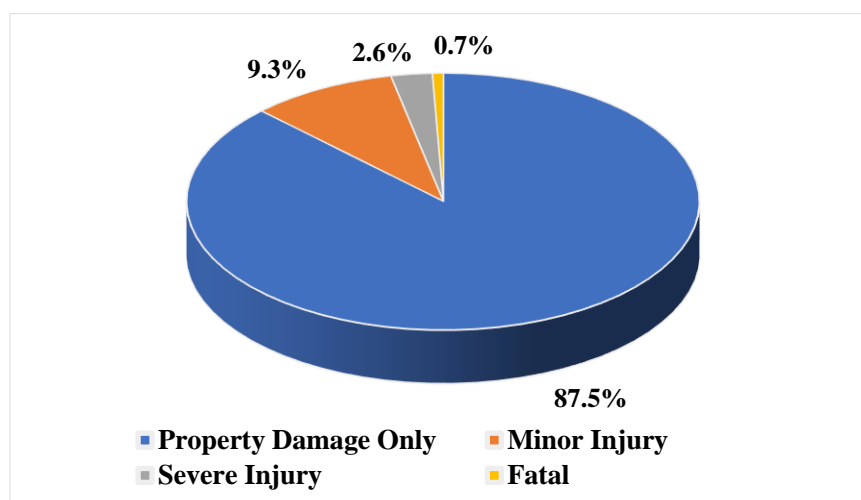
*Figure 8: Distribution of Age in Crashes Reported by Participants*



Participants who were involved in traffic crashes also had to indicate the severity and responsibility at the moment of the crash. Table 13 and Figure 10 provides information regarding the distribution of crash severity among crashes reported. Property damage only (PDO) crashes corresponded to 87.5% of all crashes reported, making it the severity in the majority of crashes. Meanwhile, 43.3% of drivers acknowledge to be responsible of traffic crashes.

*Table 12: Distribution of Crash Severity among Crashes Reported*

<b>Crash Severity</b>	<b>Response Count</b>	<b>Percent</b>
<b>Property Damage Only</b>	1011	87.5%
<b>Minor Injury</b>	107	9.3%
<b>Severe Injury</b>	30	2.6%
<b>Fatal</b>	8	0.7%
<b>Total</b>	<b>1156</b>	<b>100%</b>



*Figure 9: Distribution of Crash Severity among Crashes Reported*

Many of the responses received contained skipped questions as well as answers that were not completely clear and required some type of assumption for interpretation, therefore it was necessary to filter the responses and create a database of the data that was going to be considered in the development of models.

### 3.3 Database Development

Once the raw dataset was obtained and analyzed, a data filtering process was performed to organize and edit the data so there would be a sense of uniformity between the answers that were provided when creating the database. Several responses were deleted because they were incomplete and did not include enough information to be considered for the development of the proposed model. The following criteria points were used as a base to deleting these responses:

- Responses where the participant accepted the informed consent but did not answer any more questions of the survey.
- Responses where the participant did not indicate if he or she received traffic violations and involvement in traffic crashes.
- Responses where the participant did not indicate that he or she was involved in traffic crashes but did acknowledged to receiving or not receiving traffic violations.
- Responses where the participant did not indicate that he or she received traffic violations but did acknowledged any involvement in traffic crashes.
- Responses where the participant acknowledged to receiving traffic violations but did not specified which ones.
- Responses where the participant indicated to be involved in a traffic crash but did not provide any more information regarding the crash.

A total of 952 survey responses remained after the data filtering process was finished, this was the sample of data used for development of the proposed models. The created database contains the predictor variables that were going to be initially considered for the models. These variables are of both continuous and categorical type, Tables 13 and 14 provide the categorical variables that were considered for this study while the following list provides the continuous variables initially selected.

- Years driving a vehicle,
- Daily Hours Driven,
- Total Traffic Crashes,
- PDO Crashes,
- Minor Injury Crashes, and
- Severe Injury Crashes.

Table 13: Categorical Variables

Variable	Categories
<b>Age</b>	16-20
	21-30
	31-40
	41-50
	51-60
	61-70
	71-80
	81-89
<b>Sex</b>	Male
	Female
<b>Driving Over Speed Limit</b>	1
	2
	3
	4
	5 or more
<b>DUI</b>	1
	2
	3
	4
	5 or more
<b>Ignoring Traffic Signals or Signs</b>	1
	2
	3
	4
	5 or more
<b>Driving too Close to Vehicle</b>	1
	2
	3
	4
	5 or more

Table 14: Categorical Variables

Variable	Categories
<b>Illegal Parking</b>	1
	2
	3
	4
	5 or more
<b>Illegal Turn</b>	1
	2
	3
	4
	5 or more
<b>Illegal Lane Switch</b>	1
	2
	3
	4
	5 or more
<b>No seatbelt use</b>	1
	2
	3
	4
	5 or more
<b>Use of cellphone while driving</b>	1
	2
	3
	4
	5 or more
<b>Other</b>	1
	2
	3
	4
	5 or more

As indicated before, only PDO, minor injury, and severe injury crashes were included in the database. There were no fatal crashes considered for this study as the responses for these in the survey were filtered and eliminated from the database. Eight people responded to have been involved in a fatal crash, but the same eight people did not complete the survey. Therefore it was determined that these could not be considered as there was missing information for these eight observations.

### 3.4 Descriptive Statistics

Once the database was created, descriptive statistics were calculated for the variables identified in the previous section. Descriptive statistics were performed to obtain an initial understanding of the data presented in the sample. The information provided for each variable consists of the categories of the variable (for categorical variables), response count, percent and the mean. When a response variable is of a dichotomous type (i.e it has two outcomes) the mean of a categorical predictor corresponds to the proportion that achieves one of the outcomes, in this case, the outcome being achieved is whether the participant was involved in a traffic crash. Similar to how the survey was composed, the descriptive statistics shown below were divided in three parts; general information, traffic violations and traffic crashes.

#### 3.4.1 General Information

The results of the survey are presented in this section for each of the three parts of the survey. Results for the variable of age are provided and illustrated in Table 15 and Figure 10. From this information, it can be seen that most of the responses collected from the survey correspond to drivers in the age ranges of 16-20 and 21-30 years old. This can be attributed to the fact that most of the responses were collected using social media outlets as well as email for which drivers on this age ranges are more likely to be involved with. Table 16 and Figure 11 provide information regarding the distribution of sex in the sample. Results indicate that the percentage of responses obtained from females is larger than that of males.

*Table 15: Descriptive Statistics for Age*

<b>Age</b>	<b>Count</b>	<b>Percent</b>	<b>Mean</b>
<b>16-20</b>	210	22%	0.429
<b>21-30</b>	374	39%	0.655
<b>31-40</b>	99	10%	0.778
<b>41-50</b>	103	11%	0.796
<b>51-60</b>	110	12%	0.782
<b>61-70</b>	42	4%	0.738
<b>71-80</b>	13	1%	0.769
<b>81-89</b>	1	0%	1
<b>Total</b>	<b>952</b>	<b>100%</b>	<b>*</b>

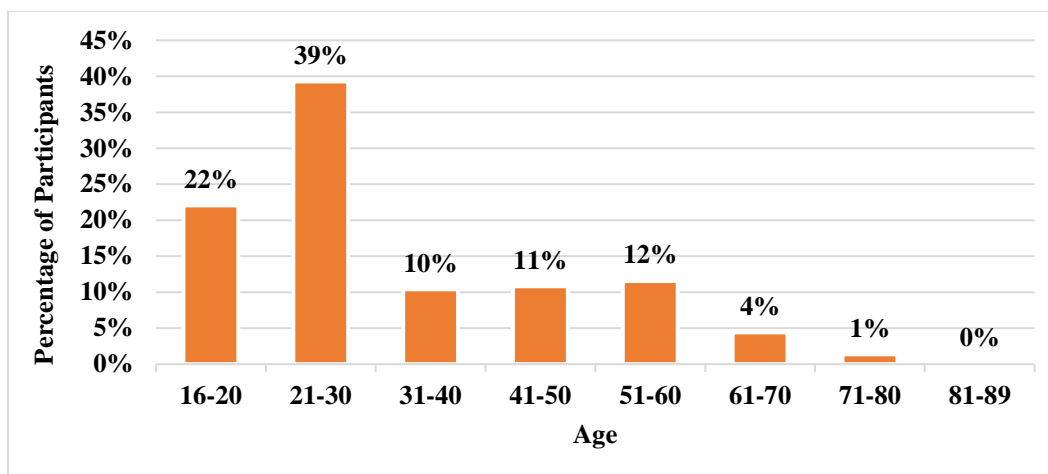


Figure 10: Sample Distribution of Drivers Based on their Age

Table 16: Descriptive Statistics for Sex

Sex	Count	Percent	Mean
Female	564	59%	0.644
Male	388	41%	0.668
Total	952	100%	*

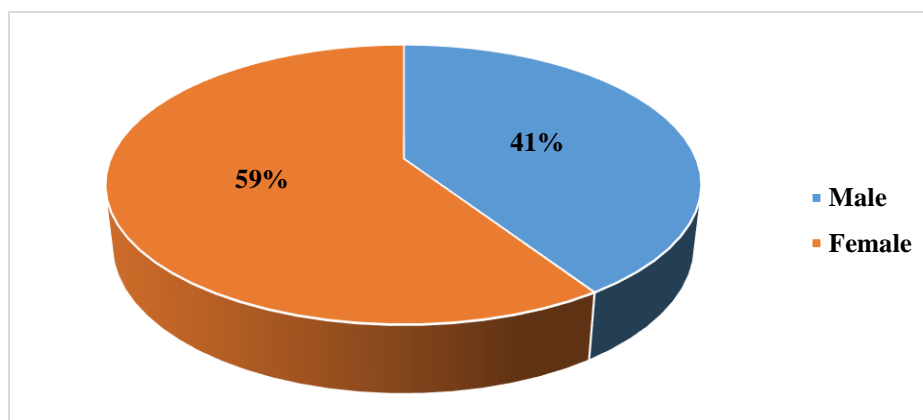


Figure 11: Sample Distribution of Drivers Based on Sex

Information regarding the distribution of age in females is presented and illustrated in Figure 12 while information regarding the distribution of age in males is illustrated in Figure 13. When comparing the data for females and males separately, results show that the proportion of females between the ages of 16-20 (25%) is larger than that of males (18%). This is also the case for the age interval of 21-30 years, where 41% of females correspond to this age interval while

males have a 36%. On the other hand, males comprised a higher percentage of responses than females from 31 to 89 years of age.

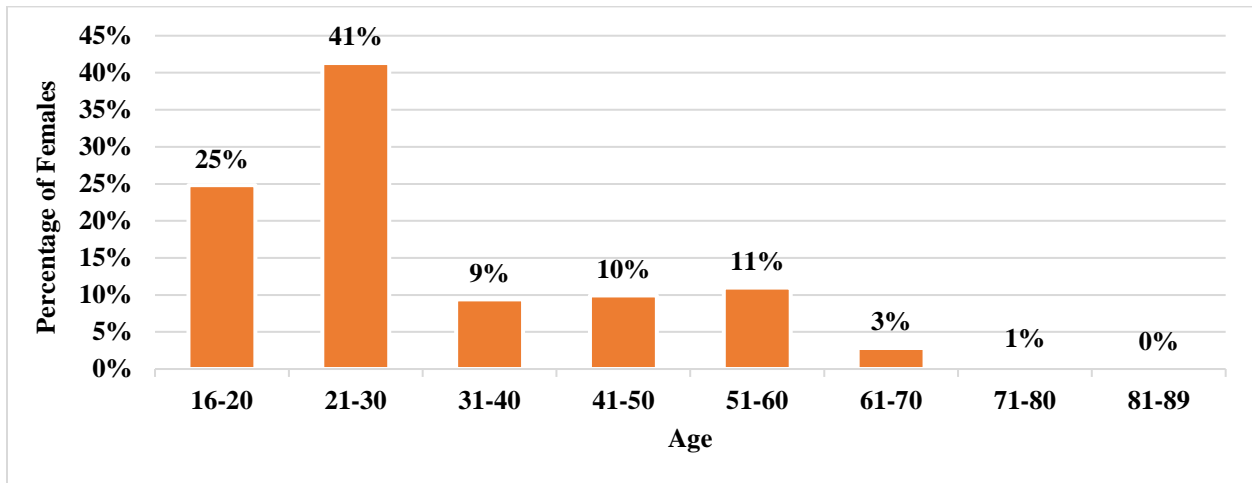


Figure 12: Sample distribution of Females Based on their Age

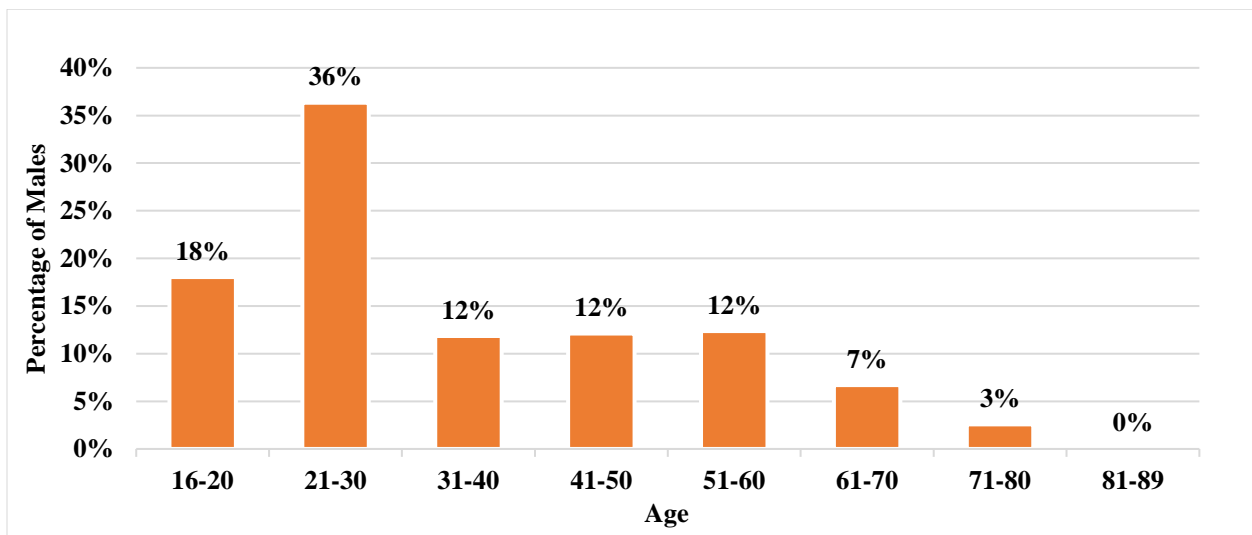


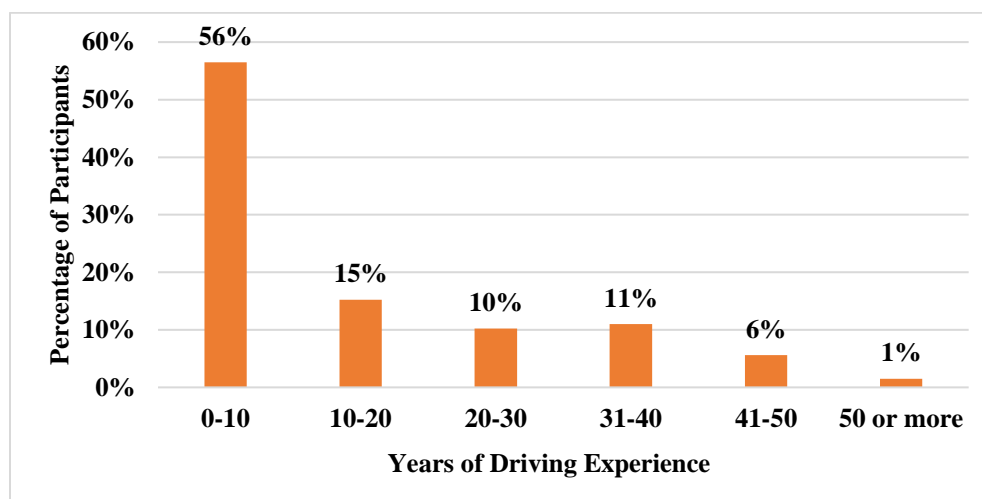
Figure 13: Sample Distribution of Males Based on their Age

Information regarding years driving a vehicle from the participants in the sample is presented in Table 17 and Figure 14. This question was left as an open answer in the survey but was categorized for illustration purposes. Most of the responses obtained corresponds to drivers who have a driving experience of 10 years or less. This was expected since most of the responses obtained corresponded to drivers between the ages of 16 and 30. Whenever a participant indicated that he or she had less than a year of driving experience, a value of 0.5 was designated in the

database. This was done to represent the average of the answers provided of less than one year of driving experience.

*Table 17: Descriptive Statistics for Years driving a vehicle*

<b>Variable</b>	<b>Total Count</b>	<b>Percent</b>	<b>Mean</b>
<b>Years of Driving Experience</b>	952	99.5	15.2



*Figure 14: Sample Distribution Based on Years driving a vehicle*

Another factor that was included in the survey was the number of hours spent driving in a day. Like years driving a vehicle, this question was provided as an open answer in the survey but was categorized for the purpose of reporting the information. Results provided in Table 18 for daily hours spent driving indicate that most of the participants spend from zero to two hours a day driving.

*Table 18: Descriptive Statistics for Daily Hours Spent Driving*

<b>Variable</b>	<b>Total Count</b>	<b>Percent</b>	<b>Mean</b>
<b>Daily Hours</b>	952	94.5	2.5

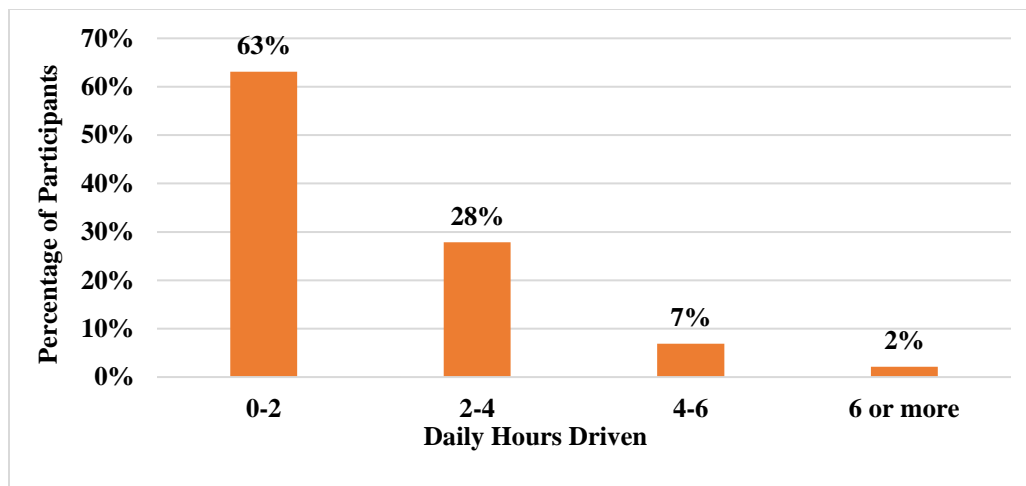


Figure 15: Sample Distribution Based on Drivers Daily Hours Spent Driving

### 3.4.2 Traffic Violations

Table 19 and Figure 16 show the distribution of the responses from the survey based on whether participants received traffic violations or not. Results indicate that 70% of participants have received traffic violations. The most common of these traffic violations are speeding and illegal parking with 36 and 28% of the responses as shown in Figure 20.

Table 19: Descriptive Statistics for Whether Participants Were Involved in Traffic violations or Not

Traffic Violations Received	Count	Percent	Mean
Yes	671	30%	0.445
No	281	70%	0.741
<b>Total</b>	952	100%	-

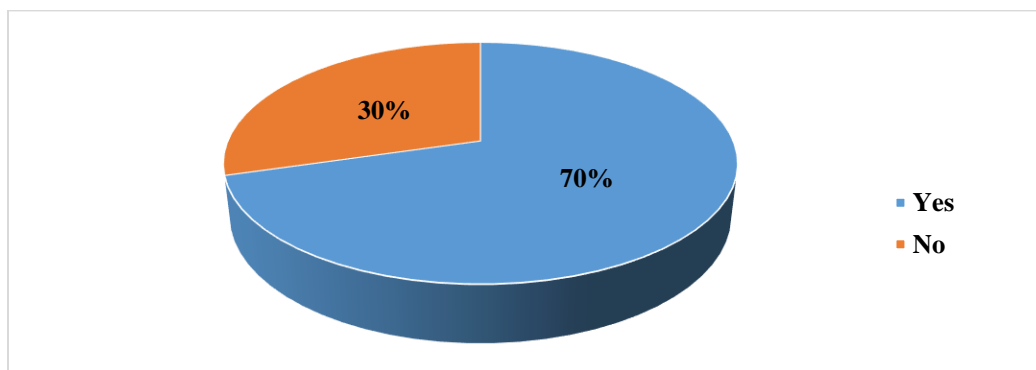


Figure 16: Distribution of Drivers Based on whether they Received Traffic Violations or Not



Descriptive statistics for the different type of traffic violations are presented throughout tables 20 to 29. The information displayed in these tables corresponds to the participants responses regarding the amount of traffic violations received for each type. As mentioned previously, the mean corresponds to the percentage of participants in each respective category that indicated to be involved in a vehicle crash. Figure 17 displays the distribution of traffic violations, it can be seen that the most common traffic violations among the responses provided were “driving over the speed limit” and “illegal parking”.

*Table 20: Descriptive Statistics for Driving Over the Speed Limit*

<b>Driving Over Speed Limit</b>	<b>Count</b>	<b>Percent</b>	<b>Mean</b>
<b>0</b>	504	53%	0.542
<b>1</b>	211	22%	0.725
<b>2</b>	123	13%	0.756
<b>3</b>	42	4%	0.881
<b>4</b>	16	2%	0.938
<b>5 or more</b>	56	6%	0.911
<b>Total</b>	<b>952</b>	<b>100%</b>	<b>*</b>

*Table 21: Descriptive Statistics for Driving Under the Influence of Alcohol or Drugs*

<b>Driving Under the Influence</b>	<b>Count</b>	<b>Percent</b>	<b>Mean</b>
<b>0</b>	942	98.9%	0.650
<b>1</b>	7	0.7%	1
<b>2</b>	2	0.2%	1
<b>3</b>	0	0	0
<b>4</b>	0	0	0
<b>5 or more</b>	1	0.1%	1
<b>Total</b>	<b>952</b>	<b>100.0%</b>	<b>*</b>

Table 22: Descriptive Statistics for Ignoring Traffic Signals and Signs

<b>Ignoring Traffic Signals and Signs</b>	<b>Count</b>	<b>Percent</b>	<b>Mean</b>
<b>0</b>	771	81.0%	0.627
<b>1</b>	152	16.0%	0.763
<b>2</b>	16	1.7%	0.625
<b>3</b>	7	0.7%	1
<b>4</b>	3	0.3%	1
<b>5 or more</b>	3	0.3%	1
<b>Total</b>	<b>952</b>	<b>100.0%</b>	<b>*</b>

Table 23: Descriptive Statistics for Driving too Close to Front Vehicle

<b>Driving Too Close to Front Vehicle</b>	<b>Count</b>	<b>Percent</b>	<b>Mean</b>
<b>0</b>	930	97.7%	0.648
<b>1</b>	18	1.9%	0.833
<b>2</b>	2	0.2%	1
<b>3</b>	0	0	0
<b>4</b>	0	0	0
<b>5 or more</b>	2	0.2%	1
<b>Total</b>	<b>952</b>	<b>100.0%</b>	<b>*</b>

Table 24: Descriptive Statistics for Illegal Turning

<b>Illegal Turn</b>	<b>Count</b>	<b>Percent</b>	<b>Mean</b>
<b>0</b>	907	95.3%	0.646
<b>1</b>	40	4.2%	0.775
<b>2</b>	1	0.1%	1
<b>3</b>	2	0.2%	1
<b>4</b>	1	0.1%	1
<b>5 or more</b>	1	0.1%	1
<b>Total</b>	<b>952</b>	<b>100.0%</b>	<b>*</b>

Table 25: Descriptive Statistics for Illegal Parking

<b>Illegal Parking</b>	<b>Count</b>	<b>Percent</b>	<b>Mean</b>
<b>0</b>	585	61.4%	0.557
<b>1</b>	194	20.4%	0.747
<b>2</b>	83	8.7%	0.880
<b>3</b>	41	4.3%	0.805
<b>4</b>	12	1.3%	1
<b>5</b>	37	3.9%	0.892
<b>Total</b>	<b>952</b>	<b>100.0%</b>	<b>*</b>

Table 26: Descriptive Statistics for Illegal Lane Switch

<b>Illegal Line Switch</b>	<b>Count</b>	<b>Percent</b>	<b>Mean</b>
<b>0</b>	894	93.9%	0.641
<b>1</b>	46	4.8%	0.870
<b>2</b>	8	0.8%	0.625
<b>3</b>	1	0.1%	1
<b>4</b>	1	0.1%	1
<b>5 or more</b>	2	0.2%	1
<b>Total</b>	<b>952</b>	<b>100.0%</b>	<b>*</b>

Table 27: Descriptive Statistics for Not Using Seatbelt While Driving

<b>Not Using Seatbelt</b>	<b>Count</b>	<b>Percent</b>	<b>Mean</b>
<b>0</b>	805	84.6%	0.620
<b>1</b>	104	10.9%	0.808
<b>2</b>	30	3.2%	0.867
<b>3</b>	3	0.3%	1
<b>4</b>	1	0.1%	1
<b>5 or more</b>	9	0.9%	1
<b>Total</b>	<b>952</b>	<b>100.0%</b>	<b>*</b>

Table 28: Descriptive Statistics for Using Cellphone While Driving

<b>Using Cellphone</b>	<b>Count</b>	<b>Percent</b>	<b>Mean</b>
<b>0</b>	857	90.0%	0.645
<b>1</b>	84	8.8%	0.726
<b>2</b>	6	0.6%	0.667
<b>3</b>	2	0.2%	1
<b>5 or more</b>	3	0.3%	0.667
<b>Total</b>	<b>952</b>	<b>100.0%</b>	<b>*</b>

Table 29: Descriptive Statistics for Other Traffic violations

<b>Other</b>	<b>Count</b>	<b>Percent</b>	<b>Mean</b>
<b>0</b>	863	90.7%	0.641
<b>1</b>	53	5.6%	0.793
<b>2</b>	19	2.0%	0.790
<b>3</b>	8	0.8%	0.5
<b>4</b>	3	0.3%	0.667
<b>5 or more</b>	6	0.6%	1

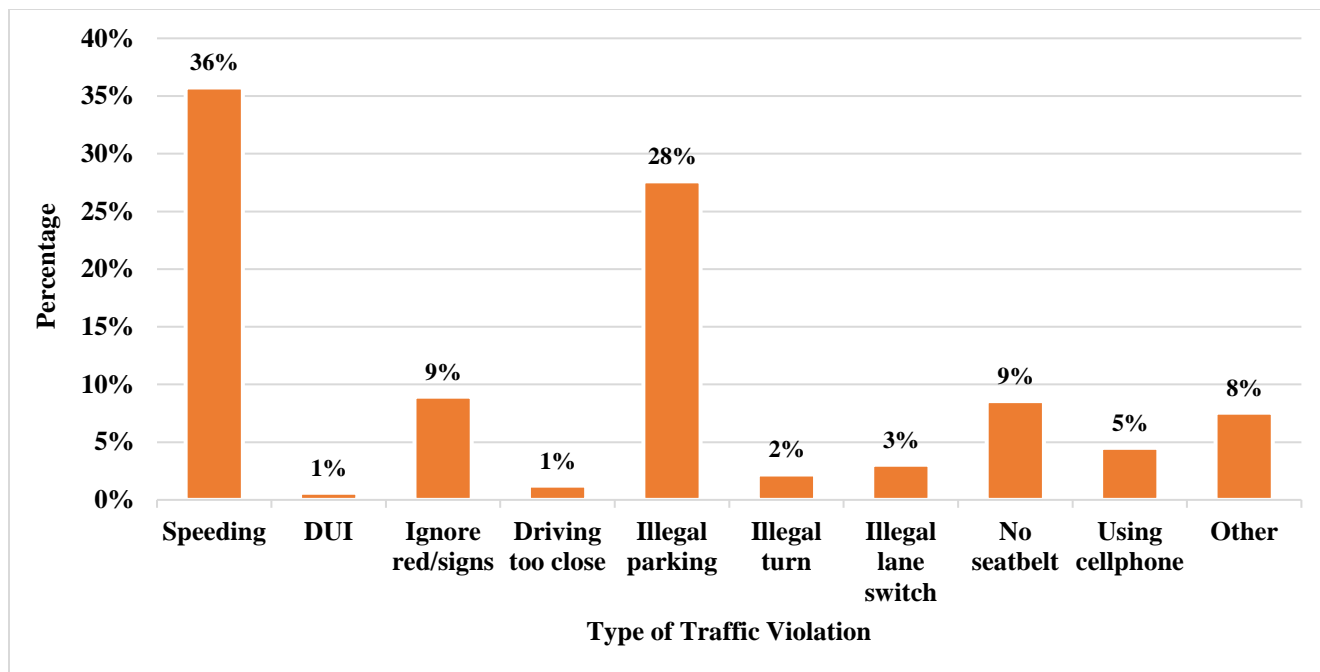


Figure 17: Sample Distribution of Traffic Violations

Figure 18 displays the distribution of the total traffic violations reported in the survey based on participants' age and sex. Results show that most traffic violations corresponded to males and females between the ages of 21 and 30. Additionally it can be seen that female participants received more traffic violations between the ages of 16 and 30 while males received more traffic violations between the ages of 31 and 70.

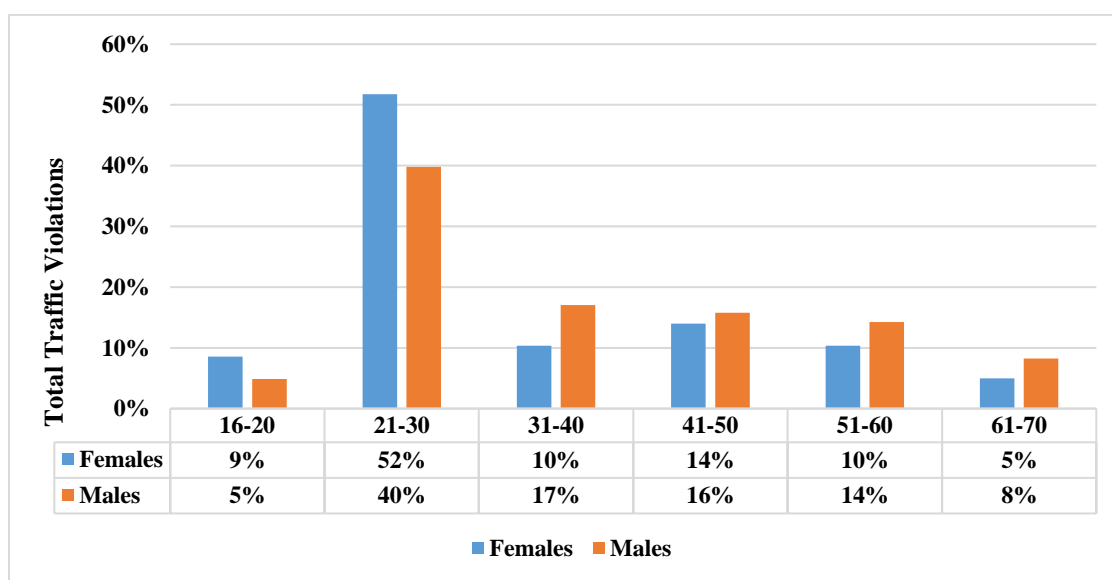


Figure 18: Distribution of Total Traffic Violations Based on Age and Sex

### 3.4.3 Traffic Crashes

Table 30 provides the distribution of participants with respect to whether they were involved or not in a traffic crash. From these results, it can be seen that 65% of participants indicated they were involved in a traffic crash as a driver while 35% have never been involved in a traffic crash. Moreover, table 32 provides the distribution of traffic crashes with respect to crash severity. From this table, it can be seen that the majority of the crashes that participants indicated they were involved in had a severity of property damage only (PDO).

Table 30: Distribution of Participants Based on Crash Involvement

Crash Involvement as a Driver	Response Percent	Response Count
Yes	65.3%	622
No	34.7%	330
<b>Total</b>	<b>100%</b>	<b>952</b>

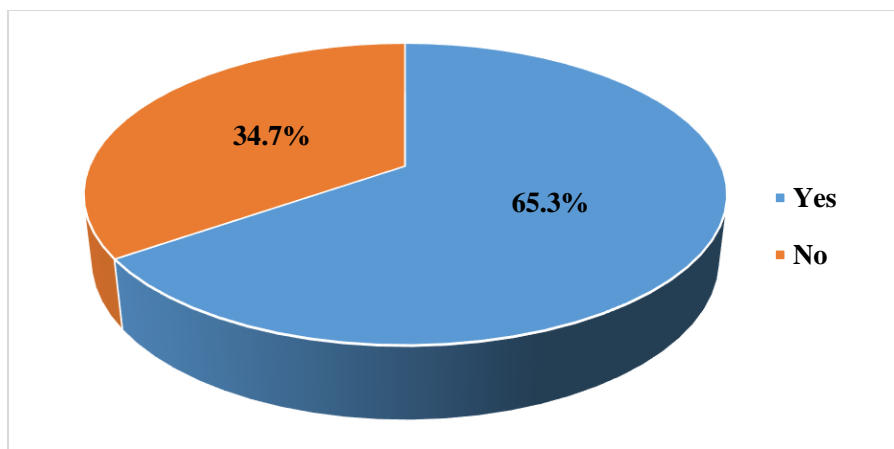


Figure 19: Sample Distribution of Crash Involvement Among Participants

Table 31: Sample Distribution of Crashes Reported Based on Severity

Severity	Response Percent	Response Count
<b>PDO</b>	88%	1010
<b>Minor Injury</b>	9%	107
<b>Severe Injury</b>	2%	28
<b>Total</b>	<b>100%</b>	<b>1145</b>

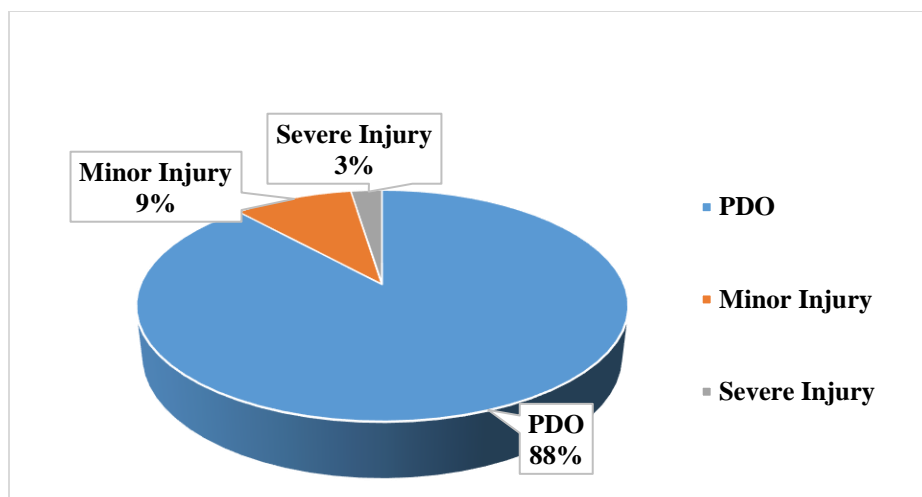


Figure 20: Sample Distribution of Crash Severity among Crashes Reported

Figure 21 displays the distribution of total traffic crashes based on participants age and sex. Results show that most of the crashes reported correspond to participants between the ages of 21 and 30. It can also be seen that female participants were the most involved in crashes between the ages of 16 and 30 while male participants were the most involved in crashes between the ages of 31 and 70.

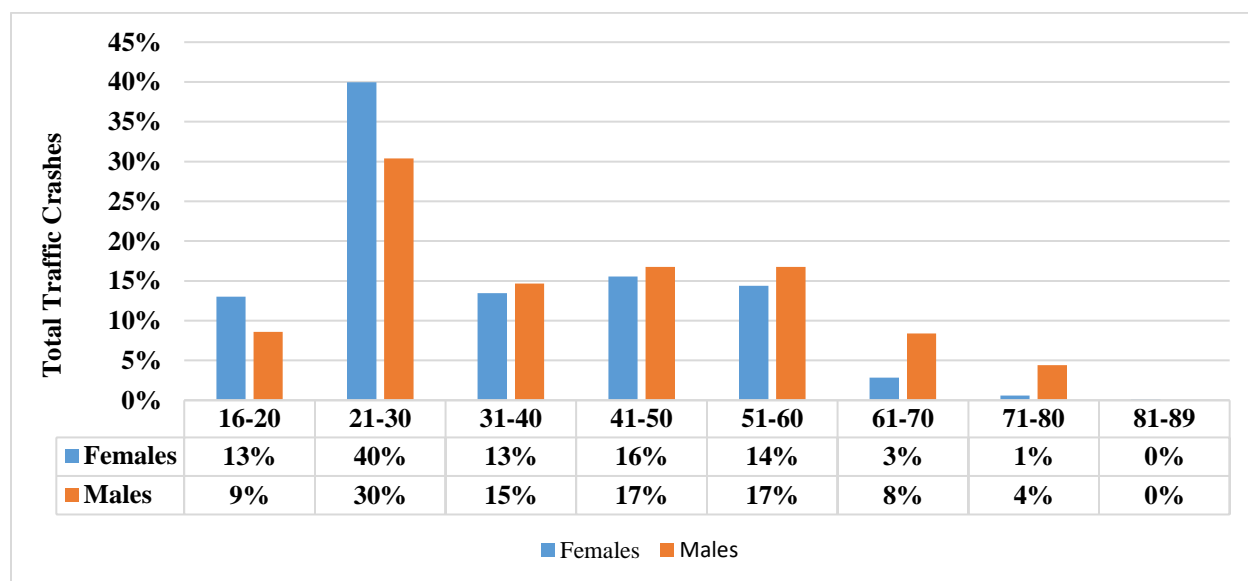


Figure 21: Distribution of Total Crashes Based on Age and Sex of Participants

## 4 ANALYSIS METHODOLOGY

This chapter describes the methodology that took place in order to analyze the data and develop the estimation models. The first part of the methodology corresponds to preliminary analysis that were performed to determine the association between each of the independent variable/predictors identified in the database and the dependent variable of crash involvement for participants. The procedures included in the preliminary analysis were subject to the type of independent variable that was being analyzed, for categorical variables chi-square tests of independence were performed while simple logistic regression was used on continuous variables. Both of these are used when the dependent variable is of a dichotomous type, i.e. it has two possible outcomes. Once the preliminary analysis was finished, a stepwise multiple logistic regression procedure was performed in order to select the model with the best subset of independent variables. On a multiple logistic regression model, a dependent dichotomous variable is being compared to multiple independent variables (2 or more) while a simple logistic regression, such as the one performed in the preliminary analysis compares a dependent dichotomous variable with a single independent variable.

### 4.1 Preliminary Analyses

The purpose of performing these analyses before starting development of the model was to understand how each of the different independent variables that were identified in section 3.4 of this study affect whether a participant was involved in a traffic crash or not. Several steps were followed to help in the selection of independent variables for a logistic regression model. The process of variable selection should begin with a detailed analysis for each variable independently. For categorical variables, Hosmer and Lemeshow suggested that an analysis of contingency tables is should be performed between the response variable and its two outcomes versus the independent variables and its different levels (Hosmer and Lemeshow, 2000). Additionally, Hosmer and Lemeshow also suggested that it is a good idea to estimate the odds ratios for variables that show a moderate level of association using one of the levels as reference. For continuous variables, the most desirable analysis consists of performing logistic regression models for each variable independently so the coefficient, standard error of the coefficient, likelihood ratio test for the significance of the coefficient and the Wald statistic could be estimated (Hosmer and Lemeshow, 2000). Thus, before development of the proposed models started, a series of bivariate analyses

took place to explore the association between the different predictor variables identified in the database independently and the response variable. The analyses that were performed in this section depended on the type of independent variable that was going to be compared. When a categorical variable was being analyzed for its association with the response variable, chi-square tests of independence were performed, on the other hand, simple logistic regression analysis were used to analyze continuous predictor variables.

#### 4.1.1 Contingency Tables

Contingency tables are a mean of displaying the frequencies or proportions between the categories of two categorical variables. Table 32 shows an example of a 2 x 2 contingency table. The rows of the table correspond to the categories of one of the variables, say X, while the columns correspond to the categories of the remaining variable, Y. If X and Y are categorical variables with I and J categories respectively, then the cells of a contingency table represent the joint frequency counts of X and Y (Agresti, 2002). The sum of these outcomes for each row and column are referred to as the marginal totals. The grand total, which is displayed on bottom left cell, is the sum of the marginal total for the rows or columns.

Table 32: Example of a Contingency Table

Variable X	Variable Y		Row Totals
	J <sub>1</sub>	J <sub>2</sub>	
I <sub>1</sub>	I <sub>1</sub> J <sub>1</sub>	I <sub>1</sub> J <sub>2</sub>	Marginal Totals = I <sub>1</sub> J <sub>1+</sub> I <sub>1</sub> J <sub>2</sub>
I <sub>2</sub>	I <sub>2</sub> J <sub>1</sub>	I <sub>2</sub> J <sub>2</sub>	Marginal Totals = I <sub>2</sub> J <sub>1+</sub> I <sub>2</sub> J <sub>2</sub>
Column Totals	Marginal Totals = I <sub>1</sub> J <sub>1+</sub> I <sub>2</sub> J <sub>1</sub>	Marginal Totals = I <sub>1</sub> J <sub>2+</sub> I <sub>2</sub> J <sub>2</sub>	Grand Total

Displaying data in this manner helps in identifying how the frequencies between two categorical variables are distributed along each of their respective categories. Several contingency tables were developed in this study to compare the association between being involved or not in a traffic crash and the other categorical variables identified in the previous sections. Once these tables were developed, chi-square tests of independence were performed for each categorical independent variable.

#### 4.1.2 Chi-Square Test of Independence

The chi-square test of independence or chi-square test of association, is a non-parametric statistical test used to determine if two categorical variables in a sample are independent of each



other. If two independent variables are independent of each other, by consequence, there would not be an association between them. A non-parametric test means that the data is not required to fit a normal distribution. Several assumptions are included to perform the chi-square test of independence:

- Data in the contingency should be in frequency or counts rather than percentages.
- The categories of the variables being compared must be mutually exclusive.
- Each subject may contribute data to only one cell of the contingency table.
- Study groups must be independent.
- The two variables being analyzed must be categorical.
- The expected value of a cell should be 5 or more on 80% of the cells, and no cell should have an expected value of less than one.

The chi-square test of independence was performed using the statistical software *Minitab*. Information for each cell in the contingency table is provided in the Minitab output results and are displayed as follows:

- Observed Cell Count/Frequency,
- Expected Cell Count/Frequency.
- Adjusted Standardized Residual. and
- Chi-Square statistic for the respective cell.

The chi-square test seeks to compare the observed frequencies, or cell counts, for the cells presented in a contingency table with a set of expected frequencies for the same cells. The cell count/frequency corresponds to the count obtained directly from the survey and that was presented on the contingency tables. The expected cell count/frequency is the frequency value that would be present in a cell if both variables were completely independent of each other, i.e. there would not be any association between them.

Analyses of the adjusted residuals were performed to further understand the association stated by the probability value. The analysis of residuals also presented with an opportunity to study the association between the categories of the independent and dependent variable respectively. The standardized adjusted residuals correspond to values that follow a normal distribution, meaning that the residuals can be associated to the Z values of a normal distribution (Agresti, 2002). For the case of this study, the confidence interval was stated as 95% which has upper and lower bounds of +1.96 and -1.96 respectively and thus any value larger than these

bounds is statistically significantly different from  $H_0$ , meaning that there is a significant association with the response variables (Agresti, 2002). Finally, the chi-square statistic determines how variables affect each other, with values equal or lower than one indicating that the observed and expected frequencies are approximately equal (i.e. one variable does not have an effect on the other for the particular category that the respective cell is included in) (Agresti, 2002). However, the main result that was considered for analyzing purposes was the probability values associated with the Pearson and likelihood ratio chi-square statistics. These probability values were analyzed using the following statistical hypotheses:

- $H_0$ : Both variables are independent of each other.
- $H_1$ : There is not sufficient evidence to state that both variables are independent.

Since the chi-square tests of independence presented in this section were performed at a confidence level of 95% ( $\alpha = 0.05$ ), if a probability value was lower than 0.05,  $H_0$  would be rejected, meaning that both variables are not independent of each other (there is a significant association).

In addition to the determining the association using the probability values, an assessment of this association was performed using the following goodness of fit measures:

- Cramer's V-Square,
- Pearson's R, and
- Spearman's Rho.

Cramers V-squared is to measure the strength of association between two categorical variables. The values for this measure range from 0 to 1, 0 being there is not any association between the variables and 1 being both variables have a perfect association. In addition to Cramer's V-square statistic, values for Pearson's R and Spearman's Rho statistics were also determined, which, in similar fashion to Cramer's V-squared, also seek to measure the strength of association between two categorical variables. The values for these measures range from -1 to +1, the closer the absolute value is to 1 the stronger the association between the variables.

As it was mentioned, the chi-square of independence is a statistical test used to determine the association between two categorical variables. Since the independent variables included in the sample also contain continuous variables, such as years driving a vehicle and daily hours driven, this test cannot be used for such variables. Thus, simple logistic regressions analyses were

performed to study the association that these independent variables have with being involved or not in a traffic crash.

#### ***4.1.3 Simple Logistic Regression***

Before discussing the methodology for the simple logistic regression analyses performed in this section, a description of the concept of logistic regressions is presented. Logistic regression is a form of regression analysis used when the dependent variable is dichotomous, meaning it can have one of two outcomes. The main objective when doing a logistic regression analysis is to find a model with the best fit that could describe the relationship between a response variable and a set of independent variables. When only one independent variable is included, the model referred to is called a simple logistic regression while a multiple logistic regression model considers the use of more than one predictor variable (Hosmer and Lemeshow, 2000). Logistic regressions are useful when one or more independent variables are important and there is no linearity between dependent and independent variables (Chandraratna and Stamatiadis, 2004). Moreover, logistic regressions do not require normally distributed variables and overall are less strict than linear regressions. One of the differences between a logistic regression analysis and a linear regression is reflected when the choice of using a parametric model and the assumptions for each type of regression are considered. When this difference is cleared, logistic regression analyses use the same principles considered for linear regression. Another important difference between logistic regression and linear regression is the nature of the relationship between the independent variable and the response variable. In any regression model, the key value is the mean value of the dependent variable given a value of the independent variable (Hosmer and Lemeshow, 2000). This value, called the conditional mean, is expressed as  $E(Y/x)$  or the expected value of Y (response variable) given a value of x (independent variable). In linear regression analyses, it is assumed that this quantity may be expressed as the following linear equation:

$$E(Y|x) = B_0 + B_1x$$

*Equation 1: Expression of the Conditional Mean for Linear Regression*

In this equation,  $B_0$  and  $B_1$  correspond to the coefficients of the slope of the equation and the first independent variable respectively. This expression implies that the value of  $E(Y/x)$  can take any value when the range of value for x is between  $-\infty$  y  $\infty$ . For logistic regression, the conditional mean must be equal to or larger than zero and equal to or smaller than 1 since the

values of the response variable for a logistic regression model ranges between one and zero. The specific form of representing a logistic regression model is as follows:

$$\pi(x) = \frac{e^{B_0 + B_1x}}{1 + e^{B_0 + B_1x}}$$

*Equation 2: Conditional Mean for a Logistic Regression Model*

To simplify the notation,  $E(Y/x)$  was substituted for  $\pi(x)$  to represent the conditional mean of Y given a value of x. An important part of the logistic regression model is the logit transformation, which can be defined as:

$$g(x) = \ln \left[ \frac{\pi(x)}{1 - \pi(x)} \right] = B_0 + B_1x$$

*Equation 3: Logit Transformation of the Conditional Mean for Logistic Regression Models*

In the above equation  $g(x)$  is defined as the probability of the outcome of the response variable,  $x$  is defined as the independent variable used to represent the different factors that affect the dependent variable,  $\beta_0$  is the coefficient associated with the slope of the equation and  $\beta_1$  is the coefficient associated with the independent variable. The expected outcome  $g(x)$  of the logistic regression model for this study is established as the likelihood that a driver will be involved in a traffic crash versus the likelihood of a driver not being involved in a future traffic crash. The outcome will depend on the relationship between the different independent variables used in the model. The importance of the logit transformation is the fact that  $g(x)$  takes on many desirable properties of a linear regression model (Hosmer and Lemeshow, 2000). The parameters of the logit transformation of  $g(x)$  are linear, can be continuous and can take any value from  $-\infty$  to  $\infty$ , depending on the range of values of  $x$ .

The studies included in the literature review indicate that the most common approach for estimating a driver's likelihood of future crashes based on a series of different factors is the use of logistic regression analyses. For the case of estimating likelihood of future traffic crash occurrence, the values can be established as likelihood of traffic crash occurrence vs likelihood of no traffic crash occurrence. Some of the findings of the literature review indicate that factors such as age, sex and previous traffic violations and crashes are commonly used as independent variables of logistic regression models that seek to estimate future crash occurrence. Simple logistic regression is a non-parametric analysis, similar to the chi-square test of independence, in which the data is not required to have a normal distribution. Unlike the chi-square test of independence, simple

logistic regression analyses were used to study the association between a single continuous independent variable and one of the outcomes of the dependent variable in addition to also be able to study the association of two categorical variables. For the simple logistic regression analyses performed, the outcome selected from the dependent variables was that participants were involved in a traffic crash. The analyses of simple logistic regression were also performed using the statistical software Minitab. The following information was analyzed from the output results provided by Minitab:

- Coefficients,
- Odds Ratios, and
- Goodness of Fit Tests.

A probability value, based on the chi-squared distribution, was used to indicate if the terms included in the regression are statistically significant or not. A confidence interval of 95% was used with an alpha value of 0.05 to test for significance, similar to the chi-square test of independence. The null hypothesis for this test was that the coefficient being analyzed was equal to zero, thus any value lower than the stipulated alpha would reject the null hypothesis that the value of the coefficient was zero, this would indicate that there is a significant association between the independent variable and the response outcome. The coefficients section provides information regarding the coefficient value of the predictor variable, which describes the size and direction of the relationship with the response outcome and how significant is this relationship by means of the probability value. The odds ratio (OR) section provides information regarding the odds of the independent variable associated with achieving one of the outcomes of the response variable. For logistic regression, the odds ratios can be defined as the odds of one of the outcomes (Y) of the response variable occurring versus the odds that the outcome does not occur (1-Y). The odds ratios obtained in this study correspond to the odds of being involved in a traffic crash divided the odds of not being involved in a traffic. For continuous variables, the odds ratio represents the odds that the selected continuous variable has of achieving one of the outcomes of the response variable. Odds ratios for continuous predictors were also performed based on the outcome that a participant was involved in a traffic crash, meanwhile, the odds ratio for categorical variables are interpreted quite different. For each categorical variable, a category was selected as the base or reference category, this reference category was identified by the row with the zero values and \*, then, the odds ratio for a categorical variable were interpreted as the odds that one category has of achieving

the selected outcome of the response variable based on the odds the reference category has of achieving the same outcome. It must be noted that the odds ratio for a logistic regression model containing only one single predictor is considered to be unadjusted because there are no other variables whose influence must be adjusted or subtracted out (Stoltzfus, 2011). Values for the odds ratios range from  $-\infty$  to  $\infty$  and can be interpreted as follows:

- If  $OR = 1$ , predictor does not affect the outcome.
- If  $OR > 1$ , predictor is associated with higher odds of outcome.
- If  $OR < 1$ , predictor is associated with lower odds of outcome.

Finally, a goodness of fit section is provided to display information on how well the predicted probabilities deviate from the observed probabilities

## 4.2 Model Selection

Once the preliminary analyses were performed, a multiple logistic regression analysis took place to determine the model that best estimates the likelihood of being involved in a traffic crash based on a series of independent variables. The problem with simple logistic regression analyses, as Hosmer and Lemeshow stated, is the fact that “it ignores the possibility that a collection of variables can become an important predictor of the outcome when taken together” (Hosmer-Lemeshow, 2000). The process of multiple logistic regression analyses was identical to the simple regression analysis discussed in the previous section, the only difference between both analyses is the number of predictors included in the model. The selection of the best model depends on the relationship between the predictor variables chosen to be included in the model and the response variable. Ideally, a model should include as many independent variables as possible in order to have a sense of completeness. On the other hand, if every independent variable is considered, the model would suffer of overfitting as a result of irrelevant independent variables being included (Agresti, 2002). To determine the model with best subset of independent variables, a stepwise logistic regression was performed. In a stepwise logistic regression, independent variables are included or excluded in an iterative process that stops when a model that contains only significant independent variables is obtained. The significance of the independent variables was determined using the probability value obtained from the *Minitab* output results presented in the following chapter. To compare the models developed in each of the steps where an independent variable was included or excluded, criterion based methods are applied. Criterion based methods are useful to

determine which combination of predictor variables are most effective at estimating the dependent variables outcome. The goal is to find a balance between a complex model with many independent variables and simple model with few independent variables (Agresti, 2002). The criterion base method that was used to select the best model was the Akaike Information Criterion (AIC), which judges a model based on how close its fitted values tend to be to the true value in terms of a certain expected value (Agresti, 2002). The AIC allows for comparison of models even if they do not have the same number of predictor variables. For logistic regression, a lower AIC value indicates that the model has a better fit. After selecting the model that best fits the data, an assessment of the fit for the selected model was performed.

### **4.3 Model Assessment**

An assessment of the selected model was performed to know how effective the model is at describing the outcome of the dependent variable. This is referred to as the model's goodness of fit (Hosmer and Lemeshow, 2000). The Minitab output provides goodness of fit measures for a logistic regression model. The measures provided by the output results include: deviance and Pearson chi-square and Hosmer-Lemeshow tests. When fitting (developing) a logistic regression model in the software Minitab, the user is asked to select in which format the data is going to be entered. The two options provided are binary response/frequency and event/trial. The format used for this study was the binary response/frequency format since each row corresponds to one trail (subject). When choosing the binary response/frequency format, the deviance and Pearson chi-square test measures of goodness of fit are usually not trustworthy because they depend on the amount of trials per row that are presented in the dataset; as the trials per row decrease so does the probability value that indicates if a model represents a good fit. Since the binary response format includes only one trial per row, contrary to the event/trial format which can have multiple trials per row, the deviance and Pearson measures of goodness of fit should not be considered. When the probability value is less than the stated alpha, the model is considered to provide a bad fit. Thus, for this study, the goodness of fit test used to assess the selected model was the Hosmer-Lemeshow measure since it does not depend on the format used to enter the data in the software.

The Hosmer-Lemeshow measure is obtained from the observed and expected frequencies presented in the two-way table provided by the output results of Minitab. This table groups the observed and expected frequencies for the two outcomes of the response variable of being involved

or not in a traffic crash based on the probability estimated from the model. The estimated probabilities for each subject were grouped in ten intervals or groups (the first interval starts with a probability of 0 while the last one ends with a probability of 0.99) and each of these intervals has a number of observed and expected frequencies that correspond to the probabilities included in each interval. The outcomes of the dependent variable are displayed in the columns while the rows correspond to the estimated probability intervals. The table of observed and expected frequencies provides the opportunity to assess whether the frequency of the estimated probabilities for the selected model are similar to the observed ones. If the Hosmer-Lemeshow probability value presented in the results is larger than the stated alpha, there is not enough evidence to say that it does not provide a good fit.

Another procedure used to assess the predictive ability of the model was the development of a Receiver Operating Characteristic (ROC) curve. The ROC curve is a graph which provides a measure of the model's ability to discriminate between subjects who experience the outcome of interest versus those who not (Hosmer and Lemeshow, 2000). For this study, the ROC curve indicates if the model classifies participants who were involved or not in a crash correctly. This is done by assigning a value of one or zero to the estimated probability of the model depending if it is greater or lesser than a specified cutoff value. The cutoff value for this study was 0.5, if the estimated probability of the model is greater than or equal to 0.5, the model classifies the subject's predicted probability as one (being involved in a crash). On the other hand, if the estimated probability is less than 0.5, the model classifies the subject's predicted probability as zero (as not being involved in a traffic crash). A classification table is used to display this procedure and the ROC curve is the output of the information provided. Table 33 provides an example of the classification table used to develop the ROC curve. Values in the true positive and true negative cells represent the subjects that the model classified correctly as being involved in a traffic crash or not.

*Table 33: Classification Table for ROC Curve*

<b>Predicted Crash Involvement</b>	<b>Observed Crash Involvement</b>	
	<b>Yes</b>	<b>No</b>
<b>Yes</b>	True Positive	False Negative
<b>No</b>	False Positive	True Negative



Using the true positive and true negative values, the sensitivity and specificity of the model can be calculated. The sensitivity indicates the number of subjects who were involved in a traffic crash and were predicted to be involved in a traffic crash by the estimated probability while the specificity indicates the same for subjects who were not involved in a traffic crash. The ROC curve is a plot of sensitivity versus 1-specificity. To plot the ROC curve, several classification tables have to be developed using different cutoff values to calculate the values of sensitivity and 1-specificity. The area under the ROC curve provides a value which indicates if the model has a good predictive ability. The value for the area under the ROC curve range from 0.5 to 1 and can be interpreted as follows:

- If  $ROC = 0.5$ : No predictive ability.
- If  $0.7 \leq ROC \leq 0.8$ : Acceptable predictive ability.
- If  $0.8 \leq ROC \leq 0.9$ : Excellent predictive ability.
- If  $ROC \geq 0.9$ : Outstanding predictive ability.

## **5 ANALYSIS RESULTS**

In this chapter, the results of the methodologies discussed in chapter 4 are presented. First, the results for the preliminary analyses are presented. These preliminary analyses were performed to study the association between each the independent variables used in this study and the dependent variable of being involved or not in a traffic crash. Chi-square tests of independence were performed for independent categorical variables while simple logistic regression analyses were performed for continuous and categorical variables. Following the discussion of the results for the preliminary analyses, the model selection results are presented. Stepwise multiple logistic regression analyses were performed to develop and select the final model. Finally, goodness of fit and model diagnostics is discussed to assess how the results from the model describe the dependent variable outcome.

### **5.1 Preliminary Analysis**

The preliminary analyses results presented in this section start with chi-square tests of independence. As mentioned in section 4.2, chi-square tests of independence were performed to tests the association of the categorical predictor variables that were going to be included in the proposed model and the response variable. In order to perform this analysis, the data had to be rearranged in contingency tables.

#### ***5.1.1 Contingency Tables***

The resulting contingency tables developed to be used with the chi-square test of independence are provided below. The first value in each cell corresponds to the observed frequency while the second value corresponds to the expected frequency.

Table 34: Contingency Table Age vs Crash Involvement

Age Range	Crash Involvement		Total
	Yes	No	
<b>16-20</b>	90 137.2	120 72.8	210
<b>21-30</b>	245 244.4	129 129.6	374
<b>31-40</b>	77 64.6	22 34.3	99
<b>41-50</b>	82 67.3	21 35.7	103
<b>51-60</b>	86 71.9	24 38.1	110
<b>61-70</b>	31 27.4	11 14.6	42
<b>71-80</b>	10 8.5	3 4.5	13
<b>81-89</b>	1 0.7	0 0.3	1
<b>Total</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

Table 35: Contingency Table Sex vs Crash Involvement

Sex	Crash Involvement		Total
	Yes	No	
<b>Females</b>	363 368.5	201 195.5	564
<b>Males</b>	259 253.5	129 134.5	388
<b>Totals</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

When analyzing these contingency tables, it can be seen that the resulting table for sex complies with the following assumptions stated for using the chi-square tests of independence; data is displayed in frequencies, the categories of both variables are independent since each of the frequencies corresponds to only one of the 952 responses collected and there are no cells with observed frequencies values of less than three. However, this was not the case for the contingency table of "Age" since the cells corresponding to the age range of 81-89 contains cells with frequency values lower than three. Similarly, the following tables display the frequencies of the different

traffic violations received by participants of the survey based on whether they were involved or not in a traffic crash.

Table 36: Contingency Table for Driving over the Speed Limit Violations vs Crash Involvement

Driving Over the Speed Limit	Crash Involvement		Total
	Yes	No	
<b>0</b>	273 329.3	231 174.7	504
<b>1</b>	153 137.9	58 73.1	211
<b>2</b>	93 80.4	30 42.6	123
<b>3</b>	37 27.4	5 14.6	42
<b>4</b>	15 10.5	1 5.6	16
<b>5 or more</b>	51 36.6	5 19.4	56
<b>Total</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

Table 37: Contingency Table for DUI Violations vs Crash Involvement

Driving Under the Influence	Crash Involvement		Total
	Yes	No	
<b>0</b>	612 615.5	330 326.5	942
<b>1</b>	7 4.6	0 2.4	7
<b>2</b>	2 1.3	0 0.7	2
<b>3</b>	0 0	0 0	0
<b>4</b>	0 0	0 0	0
<b>5 or more</b>	1 0.7	0 0.3	1
<b>Total</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

Table 38: Contingency Table for Ignoring Traffic Signs and/or Signals Violations vs Crash Involvement

Ignoring Traffic Signal/Signs	Crash Involvement		Total
	Yes	No	
<b>0</b>	483 503.7	288 267.3	771
<b>1</b>	116 99.3	36 52.7	152
<b>2</b>	10 10.5	6 5.5	16
<b>3</b>	7 4.6	0 2.4	7
<b>4</b>	3 2	0 1	3
<b>5 or more</b>	3 2	0 1	3
<b>Total</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

Table 39: Contingency Table for Driving Too Close to Front Vehicle Violations vs Crash Involvement

Driving Too Close to Front Vehicle	Crash Involvement		Total
	Yes	No	
<b>0</b>	603 607.6	327 322.4	930
<b>1</b>	15 11.8	3 6.2	18
<b>2</b>	2 1.3	0 0.7	2
<b>3</b>	0 0	0 0	0
<b>4</b>	0 0	0 0	0
<b>5 or more</b>	2 1.3	0 0.7	2
<b>Total</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

Table 40: Contingency Table for Illegal Parking Violations vs Crash Involvement

Illegal Parking	Crash Involvement		Total
	Yes	No	
<b>0</b>	326 382.2	259 202.8	585
<b>1</b>	145 126.8	49 67.2	194
<b>2</b>	73 54.2	10 28.8	83
<b>3</b>	33 26.8	8 14.2	41
<b>4</b>	12 7.8	0 4.2	12
<b>5 or more</b>	33 24.2	4 12.8	37
<b>Total</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

Table 41: Contingency Table for Illegal Turn Violations vs Crash Involvement

Illegal Turn Violations	Crash Involvement		Total
	Yes	No	
<b>0</b>	586 592.6	321 314.4	907
<b>1</b>	31 26.1	9 13.9	40
<b>2</b>	1 0.7	0 0.3	1
<b>3</b>	2 1.3	0 0.7	2
<b>4</b>	1 0.7	0 0.3	1
<b>5 or more</b>	1 0.7	0 0.3	1
<b>Total</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

Table 42: Contingency Table for Reckless Lane Switch Violations vs Crash Involvement

Reckless Lane Switch Violations	Crash Involvement		Total
	Yes	No	
<b>0</b>	573 584.1	321 309.9	894
<b>1</b>	40 30.1	6 15.9	46
<b>2</b>	5 5.2	3 2.8	8
<b>3</b>	1 0.7	0 0.3	1
<b>4</b>	1 0.7	0 0.3	1
<b>5 or more</b>	2 1.3	0 0.7	2
<b>Total</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

Table 43: Contingency Table for No Seatbelt Used Violations vs Crash Involvement

No Seatbelt Used Violations	Crash Involvement		Total
	Yes	No	
<b>0</b>	499 526	306 279	805
<b>1</b>	84 67.9	20 36.1	104
<b>2</b>	26 19.6	4 10.4	30
<b>3</b>	3 2	0 1	3
<b>4</b>	1 0.7	0 0.3	1
<b>5 or more</b>	9 5.9	0 3.1	9
<b>Total</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

Table 44: Contingency Table for Cellphone Use Violations vs Crash Involvement

Minimum Number of Cellphone Use Violations	Crash Involvement		Total
	Yes	No	
<b>0</b>	553 559.9	304 297.1	857
<b>1</b>	61 54.9	23 29.1	84
<b>2</b>	4 3.9	2 2.1	6
<b>3</b>	2 1.3	0 0.7	2
<b>4</b>	0 0	0 0	0
<b>5 or more</b>	2 2	1 1	3
<b>Total</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

Table 45: Contingency Table for Other Violations vs Crash Involvement

Other Violations	Crash Involvement		Total
	Yes	No	
<b>0</b>	553 563.9	310 299.1	863
<b>1</b>	42 34.6	11 18.4	53
<b>2</b>	15 12.4	4 6.6	19
<b>3</b>	4 5.2	4 2.8	8
<b>4</b>	2 2	1 1	3
<b>5 or more</b>	6 3.9	0 2.1	6
<b>Total</b>	<b>622</b>	<b>330</b>	<b>952</b>
Observed Value Expected Value			

Recalling the assumption for the use of chi-square tests of independence which stated that no cell should have an expected frequency of three or less, it can be seen that there are expected frequencies lower than one in the contingency tables of age and the following traffic violations:

- Driving Under the Influence,



- Driving too Close to Front Vehicle,
- Illegal Turn,
- Reckless Lane Switch,
- Not Using Seatbelt While Driving, and
- Using Cellphone While Driving.

Because of this, the chi-square test of independence could not be performed for these independent variables. To solve this issue, a merge of categories and variables was performed.

For the variable of age, the categories of 71-80 and 81-89 years of age were merged into a single category. Similarly, the traffic violations mentioned above were classified into one of two categories, reckless/maneuvering violations and non-reckless/maneuvering violations. This was done in order to have a more complete distribution of frequencies in the contingency tables and thus reduce the number of cells that had frequencies of less than three. The study presented by Chandrata and Stiamadis in 2004 proposed a similar approach; every traffic violation that was considered to be indicative of risky behavior was categorized into any of these four groups: lapses (LAPSES), errors (ERORRS), non-speeding violations (VIOLATE) or speeding violations (SPEEDING). On the other hand, traffic violations that were not indicative of risky behavior were classified as no-risk citations (NORISK). The downside of this is the fact that the number of variables that are going to be included in the proposed model was reduced greatly since all of the traffic violations were merged into these two classifications. One could argue that by doing this, the model would not be as robust as it would be if every traffic violation was included separately. Traffic violations that involve a reckless behavior or some type of maneuver while driving, such as driving over the speed limit or unsafe lane switch are considered reckless/maneuvering violations. Non-reckless/maneuvering violations correspond to traffic violations that don't show a reckless behavior or some type of maneuver while driving, such as illegal parking. In the case of the traffic violation for not wearing a seatbelt, it was considered as a non-manuevering violation for the purpose of this study because this violation is not necessarily indicative of a reckless behavior that could affect the safety of other drivers. Table 46 shows the traffic violations obtained from the survey classified by reckless/maneuvering or non-manuevering violations.

Table 46: Classification of Traffic Violations

<b>Reckless/maneuvering violations</b>	<b>Non-Reckless/maneuvering violations</b>
Driving Over the Speed Limit	Illegal Parking
Driving Under the Influence of Alcohol or Drugs	Not Using Seatbelt
Ignoring Traffic Signals and Signs	Window Tints
Driving too close to front vehicle	Not Carrying Driver License
Illegal Turn	Expired Car License
Unsafe Lane Switch	EZ-Pass
Using Cellphone While Driving	Traffic Lights Turned Off
Using Shoulder Lane	Long Traffic Lights
Driving in Wrong Way (Against Incoming Traffic)	Damaged Taillights
Street Racing	HID Lights
Reckless Driving	Expired Park Meter
Overtake Yellow Traffic Signal	Illegal Car Exhaust
Driving Between Lanes	Disturbing the Peace
	Mechanical Malfunction
	Damaged Signal Light

The survey that was developed for this study provided a list of commonly issued traffic violations in Puerto Rico so participants could select whichever one applied to them. In addition to this, a space was provided so participants could write in traffic violations they had received and that were not included in the survey while also indicating the quantity of these. Once the traffic violations were classified into reckless/maneuvering or non-reckless/maneuvering violations, new contingency tables were created for these classifications. The contingency tables for age, reckless/maneuvering violations and non-reckless/maneuvering violations are shown below in tables 47, 48 and 49 respectively.

Table 47: Contingency Table for Age vs Crash Involvement (After Merging Categories)

Age	Traffic Crash Involvement		Total
	Yes	No	
<b>16-20</b>	90	120	210
<b>21-30</b>	245	129	374
<b>31-40</b>	77	22	99
<b>41-50</b>	82	21	103
<b>51-60</b>	86	24	110
<b>61-89</b>	42	14	56
<b>All</b>	<b>622</b>	<b>330</b>	<b>952</b>

Table 48: Contingency Table for Reckless/maneuvering violations vs Crash Involvement

Reckless/Maneuvering violations	No	Yes	Total
<b>0</b>	195	199	394
<b>1</b>	69	151	220
<b>2</b>	37	105	142
<b>3</b>	14	60	74
<b>4</b>	8	36	44
<b>5 or more</b>	7	71	78
<b>Total</b>	<b>330</b>	<b>622</b>	<b>952</b>

Table 49: Contingency Table for non-Reckless/maneuvering violations vs Crash Involvement

Non-Reckless/Maneuvering violations	No	Yes	Total
<b>0</b>	232	268	500
<b>1</b>	56	135	191
<b>2</b>	22	96	118
<b>3</b>	11	43	54
<b>4</b>	4	22	26
<b>5 or more</b>	5	58	63
<b>Total</b>	<b>330</b>	<b>622</b>	<b>952</b>

Comparing these contingency tables with those provided previously regarding the different traffic violations and considering the assumptions regarding the frequencies in a contingency table, it can be seen that there are no cells with frequencies less than one. Additionally, the contingency tables for reckless/maneuvering violations and non-reckless/maneuvering violations comply with every other of the assumptions stated in section 4.1. Once all the data was correctly arranged into contingency tables, chi-square tests of independence were performed.

### ***5.1.2 Chi-Square Test of Independence***

The results of the chi-square test of independence are provided in this section, starting with the independent variable of age. The results displayed in Table 50 indicate that there appears to be an association between the variable of age and being involved in a vehicle crash since the probability value is less than 0.05, meaning that the null hypothesis of independence is rejected. Although the probability value indicates that there is an association between two variables, it lacks the power to quantify how the strength of this association or the cause it. From the results provided in Table 51, it can be seen that the hypothesis of independence has a good fit for drivers in the age ranges of 21-30, 61-70 and 71-89 since the adjusted residuals for these categories lie outside the confidence interval of 95% stated for this study, making these categories independent from the dependent variable of being involved in a traffic crash. Recall from section 4.2.2, on which the methodology for chi-square tests of independence was discussed, that values between  $-1.96$  and  $+1.96$  are associated with a fit to the hypothesis of independence, i.e. these values are not associated with the dependent variable. The fit of the independence hypothesis can also be determined when one compares the observed and expected frequency values of the cells. For participants in the age categories of 21-30, 61-70 and 61-89, the observed and expected values are relatively close, meaning that being involved in a traffic crash is independent of the fact that a participant is in the age ranges of 21-30, 61-70 and 71-89. Additionally, results regarding goodness of fit measures obtained for the variable of age are also shown. The Cramer's V-square statistic yielded a value of 0.077019, which indicates that, although there is a positive association between age and being involved in a traffic crash, it is not a strong one since its value is close to zero. However, this hypothesis cannot be inferred from the results for the Pearson R and Spearman's Rho statistics since values of  $-0.221240$  and  $-0.257256$  were obtained respectively, which indicate a negative but close to zero association between the variables considered. The values provided in each cell of the tables presented in this section correspond to:

- Observed Cell Count/Frequency,
- Expected Cell Count/Frequency,
- Adjusted Standardized Residual and
- Chi-Square statistic for the respective cell.

Table 50: Results of Chi-Square Test of Independence for Age

Age	Traffic Crash Involvement		Row Marginal totals
	Yes	No	
16-20	90	120	210
	137.21	72.79	
	-7.75	7.75	
	16.24	30.61	
21-30	245	129	374
	244.36	129.64	
	0.09	-0.09	
	0.002	0.003	
31-40	77	22	99
	64.68	34.32	
	2.748	-2.748	
	2.34	4.42	
41-50	82	21	103
	67.30	35.70	
	3.22	-3.22	
	3.21	6.05	
51-60	86	24	110
	71.87	38.13	
	3.01	-3.01	
	2.78	5.24	
61-70	31	11	42
	27.44	14.56	
	1.18	-1.18	
	0.46	0.87	
71-89	11	3	14
	9.15	4.85	
	1.05	-1.05	
	0.38	0.71	
<b>Column Marginal Totals</b>	622	330	<b>952</b>
Pearson Chi-Square = 73.322 Degrees of Freedom = 6 P-Value < 0.001	Likelihood Ratio Chi-Square = 72.686 Degrees of Freedom = 6 P-Value < 0.001	Cramer's V-square = 0.077019 Pearson's R = -0.221240 Spearman's Rho = -0.257256	

A similar analysis was conducted for the variables of sex, reckless/maneuvering violations and non-reckless/maneuvering violations. Table 51 displays the results for males and females with respect to being involved in a traffic crash. The results provided for the variable of sex indicate that there is not a significant association between sex and being involved in a traffic crash (the hypothesis for fit of independence is rejected since the probability values are larger than 0.05). This can also be assessed by taking a closer look to the adjusted residuals obtained for the cells, each of these values lies in the range of -1.96 and +1.96 which indicate that both variables are

independent of each other. Moreover, the expected frequency values obtained for each cell resulted to be very close to those from the observed frequencies, leading to the same conclusion that both variables are independent of each other. The values provided in each cell of the tables presented in this section correspond to:

- Observed Cell Count/Frequency,
- Expected Cell Count/Frequency,
- Adjusted Standardized Residual, and
- Chi-Square statistic for the respective cell.

*Table 51: Results of Chi-Square Test of Independence for Sex*

Sex	Traffic Crash Involvement		Row Marginal Totals
	Yes	No	
Female	363	201	563
	368.5	195.5	
	-0.762	0.762	
	0.082	0.154	
Male	259	129	388
	253.5	134.5	
	0.762	-0.762	
	0.119	0.225	
<b>Column Marginal Totals</b>	<b>330</b>	<b>622</b>	<b>952</b>
Pearson Chi-Square = 0.580 Degrees of Freedom = 1 P-Value = 0.446	Likelihood Ratio Chi-Square = 0.581 Degrees of Freedom = 1 P-Value = 0.446		Cramer's V-square = 0.0006094 Pearson's R = 0.0246865 Spearman's Rho = 0.0246865

On the other hand, reckless/maneuvering violations resulted to have a significant association with being involved in a traffic crash as shown in Table 52. The probability values obtained from the Pearson and likelihood ratio chi-squares statistics for number of reckless/maneuvering violations resulted to be lower than 0.001, which indicates that the null hypothesis of both variables being independent of each other can be rejected. According to the results in this table, there is a significant association between being involved in a traffic crash and having received 0, 2, 3, 4 and 5 or more reckless/maneuvering violations. Additionally, the adjusted residuals for these categories are located beyond the range of -1.96 and +1.96 which assesses the fact that there is an association between number of reckless/maneuvering violations and being involved in a traffic crash. The goodness of fit statistics presented for this variable also comply with this hypothesis, however, the association is not a strong one since the values obtained

for Cramer's V-square, Pearson's R and Spearman's Rho are 0.084099, 0.27488 and 0.287869 respectively, all which are close to zero which indicates a relatively weak association. The values provided in each cell of the tables presented in this section correspond to:

- Observed Cell Count/Frequency,
- Expected Cell Count/Frequency,
- Adjusted Standardized Residual, and
- Chi-Square statistic for the respective cell.

Table 52: Results of Chi-Square Test of Independence for Reckless/maneuvering violations

Reckless/Maneuvering Violations	Traffic Crash Involvement		Row Marginal Totals
	Yes	No	
<b>0</b>	199	195	<b>394</b>
	257.42	136.58	
	-8.079	8.08	
	13.26	24.99	
<b>1</b>	151	69	<b>220</b>
	143.74	76.26	
	1.173	-1.173	
	0.367	0.691	
<b>2</b>	105	37	<b>142</b>
	92.78	49.220	
	2.337	-2.337	
	1.610	3.035	
<b>3</b>	60	14	<b>74</b>
	48.35	25.650	
	2.964	-2.964	
	2.808	5.292	
<b>4</b>	36.00	8	<b>44</b>
	28.75	15.250	
	2.352	-2.352	
	1.829	3.448	
<b>5 or more</b>	71	7	<b>78</b>
	50.96	27.040	
	4.976	-4.976	
	7.879	14.850	
<b>Column Marginal Totals</b>	<b>622</b>	<b>330</b>	<b>952</b>
Pearson Chi-Square = 80.062 Degrees of Freedom = 5 P-Value < 0.001	Likelihood Ratio Chi-Square = 83.369 Degrees of Freedom = 5 P-Value < 0.001	Cramer's V-square = 0.084099 Pearson's R = 0.274884 Spearman's Rho = 0.287869	

Similar to the reckless/maneuvering violations, non-reckless/maneuvering violations were also found to be associated with being involved in a traffic crash. Table 53 provides the results for

non-reckless/maneuvering violations and its different categories. The probability values obtained from the Pearson and likelihood ratio chi-squares statistics were found to be lower than 0.001, which indicates that the null hypothesis of both variables being independent of each other can be rejected. The categories of 0, 2, 3, 4 and 5 or more non-reckless/maneuvering violations were found to be associated with being involved in a traffic crash since the adjusted residuals for these categories were less than -1.96 or greater than +1.96 which assesses the fact that there is an association between number of reckless/maneuvering violations and being involved in a traffic crash. The goodness of fit statistics presented for this variable also indicate that there is a positive correlation with the response variable, however, in a similar fashion to reckless/maneuvering violations, is not a strong one since the values obtained for the different goodness of fit statistics were found to be close to zero.

From the results presented in this section it is concluded that age, reckless/maneuvering violations and non-reckless/maneuvering violations were associated with being involved in a traffic crash. The probability value obtained from the Pearson and likelihood ratio chi-square statistics were used as the main results for determining if a predictor variable was associated with the response variable. Additionally, the adjusted residuals obtained from the Minitab output were used to further compliment the conclusion obtained from the probability value as well as to also analyze the various categories included in each predictor to see which are associated with being involved in a traffic crash and which are not. Finally, several goodness of fit statistics were used to assess the strength of the relationship between the two variables being analyzed. The values provided in each cell of the tables presented in this section correspond to:

- Observed Cell Count/Frequency,
- Expected Cell Count/Frequency,
- Adjusted Standardized Residual, and
- Chi-Square statistic for the respective cell.



Table 53: Results of Chi-Square Test of Independence for Non-Reckless/maneuvering violations

Non-Reckless/maneuvering violations	Traffic Crash Involvement		Row Marginal Totals
	Yes	No	
<b>0</b>	268	232	<b>500</b>
	326.68	173.32	
	-8.003	8.00	
	10.541	19.87	
<b>1</b>	135	56	<b>191</b>
	124.79	66.21	
	1.736	-1.736	
	0.835	1.574	
<b>2</b>	96	22	<b>118</b>
	77.1	40.900	
	3.907	-3.907	
	4.635	8.736	
<b>3</b>	43	11	<b>54</b>
	35.28	16.720	
	2.272	-2.272	
	1.689	3.183	
<b>4</b>	22.00	4	<b>26</b>
	16.99	9.010	
	2.094	-2.094	
	1.479	2.788	
<b>5 or more</b>	58	5	<b>63</b>
	41.16	21.840	
	4.613	-4.613	
	6.888	12.983	
<b>Column Marginal Totals</b>	<b>622</b>	<b>330</b>	<b>952</b>
Pearson Chi-Square = 75.197 Degrees of Freedom = 5 P-Value < 0.001	Likelihood Ratio Chi-Square = 81.703 Degrees of Freedom = 5 P-Value < 0.001	Cramer's V-square = 0.078989 Pearson's R = 0.262395 Spearman's Rho = 0.279033	

### 5.1.3 Simple Logistic Regression

As mentioned in the previous chapter, simple logistic regression analyses were performed to determine the association between independent and dependent variable. Table 54 shows the results of the simple logistic regression analysis. The response event being analyzed is being involved in a traffic crash, the response variable has two outcome events: being involved in a traffic crash and not being involved in a traffic crash. A 95% confidence level was used for the significance tests being considered in this analysis. Results for the continuous predictors are initially discussed followed by a discussion on the results for categorical predictors.

Initially, the total number of vehicle crashes reported by participants was going to be included in the logistic regression analyses but complete separation prevented the development of

the model. The phenomenon of complete separation occurs when one of the independent variables is associated with only one of the outcomes of the dependent variables. For this study, the outcomes of the dependent variable were being involved and not being involved in a vehicle crash. Because of the nature of the questions that were asked to collect information for vehicle crashes, every participant who indicated to be involved in a crash had at least one crash while participants who indicated to not be involved in a crash had zero crashes only. Because of this phenomenon, the software could not fit a model using the variables of total crashes since there is no diversity in the way it is associated with the dependent variable. It was decided that this independent variable was not going to be considered in the subsequent analyses.

The results show that years driving a vehicle is a significant predictor of being involved in a traffic crash since the probability value obtained was lower than 0.001, thus the null hypothesis was rejected. On the other hand, results for daily hours spent driving indicate that it is not a significant predictor since it yielded a probability value of 0.389, which is greater than 0.05, thus not rejecting the null hypothesis. The results obtained for the continuous predictors in this section yielded similar conclusions to those obtained from the chi-square test of independence analysis. The odds ratios for years driving a vehicle and daily hours spent driving resulted with value of 1.04 and 0.97 respectively, indicating that there is a positive correlation between years driving a vehicle and being involved in a traffic crash but not for daily hours spent driving since the odds ratio obtained was lower than one. The odds ratios for the years driving a vehicle predictor can be interpreted as follows; the odds being involved in a traffic crash increase by 4% for year of driving experience. On the other hand; the odds ratio of being involved in a traffic can be interpreted as; the odds of being involved in a traffic crash decrease by 3% for every hour spent driving. After discussing the odds ratios for the continuous predictors included, it can be said that although a positive and negative correlation were associated with years driving a vehicle and daily hours spent driving respectively, it is not a relatively significant one since the odds ratios obtained for these predictors were very close to 1. It is important to remember that an odds ratio of one indicates that the predictor is not correlated with the response variable.

Table 54: Results of Simple Logistic Regression Analysis

Predictor	Coefficient	SE	Z-Value	P-value	Odds Ratio
<b>Years driving a vehicle</b>	0.04289	0.006	7.15	<0.001	1.04
<b>Daily Hours Spent Driving</b>	-0.0332	0.0385	-0.86	0.389	0.97
<b>Age</b>					
Age (16-20)	0	0	*	*	Reference
Age (21-30)	0.929	0.177	5.25	<0.001	2.53
Age (31-40)	1.54	0.279	5.52	<0.001	4.67
Age (41-50)	1.65	0.282	5.86	<0.001	5.21
Age (51-60)	1.56	0.270	5.80	<0.001	4.78
Age (61-70)	1.32	0.378	3.51	<0.001	3.76
Age (71-89)	1.59	0.666	2.38	0.017	4.89
<b>Sex</b>					
Female	0	0	*	*	Reference
Male	0.106	0.139	0.76	0.446	1.11
<b>Reckless/Maneuvering violations</b>					
0	0	0	*	*	Reference
1	0.768	0.177	4.34	<0.001	2.16
2	1.001	0.215	4.67	<0.001	2.72
3	1.531	0.322	4.76	<0.001	4.62
4	1.489	0.404	3.69	<0.001	4.43
5 or more	2.302	0.409	5.63	<0.001	9.99
<b>Non-Reckless/Maneuvering violations</b>					
0	0	0	*	*	Reference
1	0.736	0.182	4.03	<0.001	2.09
2	1.329	0.253	5.26	<0.001	3.78
3	1.242	0.349	3.56	<0.001	3.46
4	1.514	0.553	2.74	0.006	4.54
5 or more	2.307	0.475	4.86	<0.001	10.04

\*Indicates that is not available since it was used as reference

The results for the predictor of age and its categories are also presented in table 54. The category of 16-20 years of age was established as the reference category. When inspecting the probability value for the various categories of age, results show that it is a significant predictor of being involved in a vehicle crash since the probability value for all the categories is lower than 0.001. The null hypothesis that the subset of coefficients for this predictor is equal to zero can be rejected. A similar conclusion was also established when the results of the chi-square test of independence for age also indicated that there is a significant association with being involved in a traffic crash. Results for the odds ratio of the various categories for the predictor of age indicate that there is a positive correlation with the response variable. Inspecting the odds ratio column for the age predictor, it can be seen that there is an increase in the odds ratio from the age of 21 to the

age of 50, then the odds for participants between the ages of 51 to 70 decrease and lastly the odds increase again for participants between the ages of 71 to 89. Thus, results of the odds ratio analysis indicate that as the age of participants increases, the odds of being involved in a traffic crash also increase for certain age ranges while they decrease for others. For drivers between the ages of 21-30, the odds of being involved in a traffic crash are 2.53 times more than the odds of drivers between the ages of 16-20 while drivers between the ages of 71-89 are approximately five times more likely to be involved in a traffic crash than drivers between the ages of 16-20.

The results for the predictor of sex show that there is not sufficient evidence to suggest that the subset of coefficient for this predictor is different from zero since the probability value obtained was 0.446 which is greater than 0.05. Inspecting the odds ratios for this predictor, results show that for male drivers, the odds of being involved in a traffic crash are 1.11 times more than the odds of female drivers being involved in a traffic crash or males are 11% more likely to be involved in a crash than females. Although the variable of sex resulted to be a non-significant independent variable, analyzing the odds ratios can complement the significance results obtained using the probability value of this independent variable. The category of female participants was used as the reference category. Although there is a positive correlation between sex and being involved in a traffic crash, it can be inferred it is not a significant one since the odds ratio obtained is almost equal to one, which would indicate that the predictor does not affect the odds of being involved in a traffic crash.

For reckless/maneuvering violations, the probability values obtained for each of the categories were lower than 0.001, which indicates that number of reckless/maneuvering violations received is a significant predictor of being involved in a traffic crash. Similarly, non-reckless/maneuvering violations also resulted to be a significant predictor of the outcome of being involved in a traffic crash since the probability values obtained were also lower than 0.001. The odds ratios for reckless/maneuvering and non-reckless/maneuvering violations indicate that, as the number of violations increase, the odds of being involved in a traffic crash also increase. No traffic violations received was chosen as the reference category for discussing odds ratios. For participants who received two reckless/maneuvering violations, the odds of being involved in a traffic crash are 2.16 times more for drivers who received two reckless/maneuvering violations than for drivers who did not receive any traffic violations. Meanwhile, participants who received

5 or more reckless/maneuvering violations are approximately ten times more likely to be involved in a traffic crash than participants who did not receive any traffic violations.

The simple logistic regression analyses performed in this section revealed several conclusions regarding the significance of the predictor variables included in this study and being involved in a traffic crash. Results of this analysis indicated that years driving a vehicle, age, reckless/maneuvering violations and non-reckless/maneuvering violations are significant predictors of being involved in a traffic crash while sex and daily hours spent driving were not significant. Past studies have shown that previous traffic violations, age, sex and being young have a significant association with being involved in a traffic crash, which is consistent with the results obtained from the simple logistic regression analyses performed in this section. However, one has to consider that every predictor was analyzed without the interaction of other predictors. When additional predictors are present in a logistic regression analysis, the effect of a predictor single predictor can vary, as discussed in the following section.

## **5.2 Model Development**

Unlike the simple logistic regression analyses presented in the previous section, multiple logistic regression examines the relationship between two or more predictor variables and a dichotomous response variable. Examining multiple variables is generally more informative because it reveals the unique contribution of each variable after adjusting for the others (Stoltzfus, 2011). However, including too many variables in a model would provide results that are not realistic since there may be insignificant factors included. Although variables display a certain behavior when compared solely to the response variable, this may not be the case when other predictor variables are also included.

When selecting the best subset of variables in a logistic regression, Minitab provides options to perform stepwise regression procedures, which seek to determine the best model based on an iterative process of inputting and/or removing independent variables. The process of inputting of removing variables is based on statistical algorithms that check for the importance of variables based on the statistical significance of their coefficient (Hosmer and Lemeshow, 2000). A backwards elimination stepwise procedure was performed in this study in order to obtain the subset of variables that would provide the best fitting model. The procedure starts by fitting the full model, which in this case includes all six predictor variables being considered. Subsequently,

predictors that are determined not to be significant are removed iteratively. This process stops when the remaining predictors in the model are significant at the specified confidence interval. The Akaike Information Criterion (AIC) is provided and was used to compare the different models that were developed. The AIC score indicates how well a model fits the sample data by balancing the under-fitting of models with few variables and over-fitting models with many variables, low scores of AIC indicate that the model has a better fit. Table 55 displays the results of the backwards elimination procedure with the respective iterative steps.

*Table 55: Results for Stepwise Backwards Elimination Procedure*

Term	Step 1		Step 2		Step 3	
	Coefficient	P-Value	Coefficient	P-Value	Coefficient	P-Value
<b>Constant</b>	-0.496		-0.625		-0.416	
<b>Years of Driving Experience</b>	0.055	0.001	0.055	0.001	0.034	<0.001
<b>Daily Hours Driving</b>	-0.051	0.249				
<b>Age</b>	-1.153	0.199	-1.192	0.176		
<b>Sex</b>	-0.332	0.040	-0.322	0.046	-0.323	0.043
<b>Reckless/maneuvering violations</b>	1.345	0.004	1.334	0.005	1.471	0.001
<b>Non-Reckless/maneuvering violations</b>	1.572	<0.001	1.554	<0.001	1.679	<0.001
<b>AIC</b>	1066.20		1065.55		1062.63	

Information regarding the coefficient and probability value for each predictor being included is displayed in addition to the AIC value the resulting model in each step. The first step of this procedure consisted of fitting the full model which included predictors for years driving a vehicle, daily hours spent driving, age, sex, reckless/maneuvering violations and non-reckless/maneuvering violations. An inspection of the coefficients obtained for the model in first step indicated that the predictors for daily hours spent driving, age and sex are negatively correlated with the outcome of being involved in a traffic crash since their coefficient value is less than zero. When inspecting the probability values (P-Values) for each predictor, hours spent driving and age resulted to be non-significant predictors since their p-value was larger than 0.05.

For the second step of this iterative process, the least significant of the predictors, in this case daily hours spent driving was removed and the model with the remaining predictors was subsequently fitted. The results for the model fitted in the second step show that the coefficient

and most of the probability values obtained remained significantly equal to those obtained in the first step. To proceed to the next step, the predictor of age which yielded the highest probability value (0.176) among the predictors included in the model was removed. The fitted model in the third step contains the following variables; years of driving experience, sex, reckless/maneuvering violations and non-reckless/maneuvering violations. Inspection of the coefficient values for the predictors included show a significant increase in these values from the second step which in turn increased and decreased the correlation for predictors with values larger and smaller than zero respectively. Moreover, inspection of the probability values shows that in this step, none of the predictors yielded values larger than the stated alpha of 0.05 thus the backward elimination procedure taking place can safely be stopped since all the predictors included are significantly associated with the dependent variable. Additionally, the AIC value obtained for the fitted model in step 3 is lower than those of steps 1 and 2 thus indicating that the model in step 3 has a better fit for the data.

The resulting model from the backwards elimination process is shown in table 56 displays the coefficient analysis results for this model. The coefficients column provides the magnitude and direction of the coefficients associated with the different predictors included. The magnitude of the coefficient indicates how much the response variable changes with respect to a unit change of the respective predictor while the direction is determined from the sign of the coefficient, a negative sign indicates that the probability for the outcome of the response variable decreases while a positive sign indicates an increase. The standard error of the coefficient column indicates the precision at which the coefficient value for a certain predictor was estimated, lower values indicate a greater precision. The 95% confidence interval column provides a range on which the exact value of the coefficient can be located. The probability value (P-value) column provides information regarding the significance of the predictors included in the model. Recall from the simple regression analysis performed in the previous section that a probability value larger than 0.05 indicates that there is not sufficient evidence to say the coefficient of the variable is different from zero and thus is not significantly associated the response variable. Additionally, odds ratios were also analyzed for both continuous and categorical variables.

A column corresponding to the variance inflation factor values (VIF) is also provided to indicate the level of multicollinearity presented in each predictor variable. Multicollinearity can be defined as correlation between predictors; when predictors are correlated with each other and

not the response variable it creates a phenomenon where a redundant predictor would result to be important because the correlation with other predictors is causing this. This value was not included in the simple regression analysis because the regression models that were developed in that section only had one predictor variable included. Values for the VIF range from 1 to  $\infty$ , with values close to 1 indicating that the predictor has no multicollinearity with other predictors.

Table 56: Results of Final Estimation Model

Term	Coefficient	Standard Error of Coefficient	Z-Value	P-Value	VIF
<b>Constant</b>	-0.416	0.137	-3.03	0.002	
<b>Years Driving a Vehicle</b>	0.034	0.006	5.39	<0.001	1.08
<b>Sex</b>					
<b>Female</b>	*	*	*	*	*
<b>Male</b>	-0.323	0.160	-2.02	0.043	1.07
<b>Reckless/Maneuvering Violations</b>					
0	*	*	*	*	*
1	0.501	0.193	2.59	0.010	1.20
2	0.541	0.234	2.31	0.021	1.19
3	1.007	0.344	2.92	0.003	1.11
4	0.831	0.428	1.94	0.052	1.10
5 or more	1.471	0.441	3.34	0.001	1.14
<b>Non-Reckless/Maneuvering Violations</b>					
0	*	*	*	*	*
1	0.427	0.197	2.16	0.031	1.12
2	1.020	0.266	3.84	<0.001	1.07
3	0.929	0.364	2.55	0.011	1.05
4	0.787	0.582	1.35	0.176	1.05
5 or more	1.679	0.500	3.36	0.001	1.08

$$\begin{aligned}
 Y = & -0.416 + 0.034 \text{ Years of Experience} - 0.323 \text{ Males} + 0.501 \text{ Moving Violations}_1 \\
 & + 0.541 \text{ Moving Violations}_2 + 1.007 \text{ Moving Violations}_3 \\
 & + 0.831 \text{ Moving Violations}_4 + 1.471 \text{ Moving Violations}_{5 \text{ or more}} \\
 & + 0.427 \text{ Non - Moving Violations}_1 + 1.020 \text{ Non - Moving Violations}_2 \\
 & + 0.929 \text{ Non - Moving Violations}_3 + 0.787 \text{ Non - Moving Violations}_4 \\
 & + 1.679 \text{ Non - Moving Violations}_{5 \text{ or more}}
 \end{aligned}$$

Equation 4: Model Equation



The continuous predictor for years driving a vehicle resulted with a coefficient value of 0.034, which is larger than zero and thus indicates there is a positive correlation with the outcome of being involved in a traffic crash. This coefficient value also yielded a standard error of 0.006 which indicates that the coefficient value was estimated with a prominent level of precision. When inspecting the 95% confidence interval, it can be seen that a coefficient value of one is not included thus it can safely be said that this predictor will maintain its positive correlation with the response outcome. Like it was shown in the backward elimination results, the probability value for this predictor resulted to be less than the established value of 0.05, which indicates that the coefficient is significantly different from zero and thus is a significant predictor of being involved in a traffic crash. The VIF for this predictor resulted with a value of 1.08, indicating there is no significant collinearity with the other predictors. Table 57 displays the results of the calculated odds ratios for continuous and categorical predictors. According to these results, the predictor for years driving a vehicle yielded an odds ratio value of 1.034 indicating that the odds of being involved in a traffic crash increase by 1.034 for each year of driving experience. This statement makes sense based on the hypothesis that as a person grows older, his or her experience while driving would improve the awareness needed to drive safely. Although the odds ratio shows an increase in odds of being involved in a traffic crash, it is not a significant one since the value obtained is close to one which would indicate that the odds of the response outcome are not affected by the predictor being analyzed. The coefficient value obtained for years driving a vehicle in the coefficients table also corroborates this inference since that value obtained is very close to zero, however, the significance tests performed for this predictor indicate that the coefficient value is significantly different from zero.

Discussion of the categorical predictors for the model starts with the predictor of sex. The sex variable has two categories, males and females with females being the reference group. The probability value obtained for sex was 0.043, indicating there is significance with the response variable, however, it must be noticed that this value is very close to the stated alpha value of 0.05. The VIF value obtained for sex was 1.07 which indicates that there is little multicollinearity with other predictors. Results of the odds ratios obtained for sex show that the odds of male participants being involved in a traffic crash are 0.72 times more than female drivers, indicating that males have decreased odds of being involved in a traffic crash than females since the odds ratio obtained was less than one. The 95% confidence interval for the odds ratios of sex show that the odds ratio

will remain lower than one thus the relationship with being involved in a traffic crash will remain the same for this model.

When analyzing the coefficients of reckless/maneuvering violations and non-reckless/maneuvering violations, results do not show a consistent trend that as the number of violations increase. The 95% confidence interval for the coefficients show a positive association for all categories of both type of variables, except the category of four violations which has the possibility of having a negative correlation. The probability values obtained for both types of violations indicate a significant association with the response variable since all the values obtained are less than 0.05, with the exception of the category corresponding to four violations. The fact that the confidence interval indicated that this category could change from a positive to a negative correlation corroborates the non-significance of this category for both types of violations. The odds ratios for reckless/maneuvering violations and non-reckless/maneuvering violations indicate that participants who received traffic violation have increased odds of being involved in a traffic crash when compared to participants who did not receive traffic violations, which can be concluded from the fact that the odds ratios increase as the number of violations increases. This complies with the results obtained from the simple logistic regression as well as the literature review studies which indicate that there is a positive correlation between traffic violations received and traffic crash involvement. Figure 22 displays the odds ratios for reckless/maneuvering violations while figure 23 displays the same behavior for non-reckless/maneuvering violations.

Table 57: Calculated Odds Ratios for Terms in Final Estimation Model

Term		Odds ratio	95% CI
Years of Driving Experience		1.035	(1.022, 1.047)
Sex			
Level A	Level B	Odds ratio	95% CI
Male	Female	0.730	(0.534, 0.998)
Reckless/maneuvering violations			
Level A	Level B	Odds ratio	95% CI
1	0	1.650	(1.130, 2.408)
2	0	1.736	(1.094, 2.756)
3	0	2.435	(1.2586, 4.709)
4	0	2.271	(0.982, 5.250)
5 or more	0	4.307	(1.813, 10.234)
Non-Reckless/maneuvering violations			
Level A	Level B	Odds ratio	95% CI
1	0	1.536	(1.044, 2.262)
2	0	2.792	(1.659, 4.696)
3	0	2.470	(1.209, 5.047)
4	0	2.329	(0.748, 7.255)
5 or more	0	5.378	(2.019, 14.316)

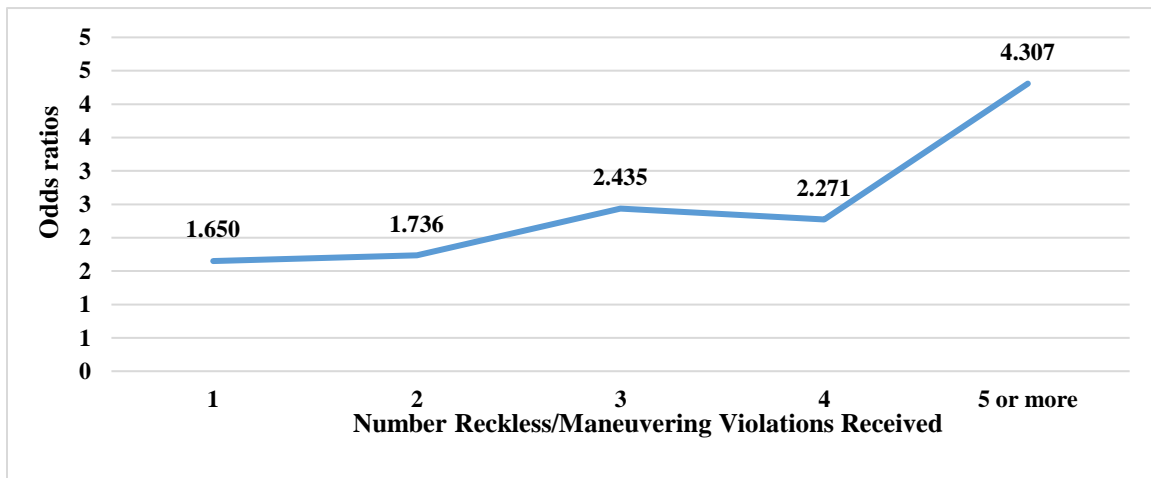


Figure 22: Odds ratios for Reckless/Maneuvering Violations

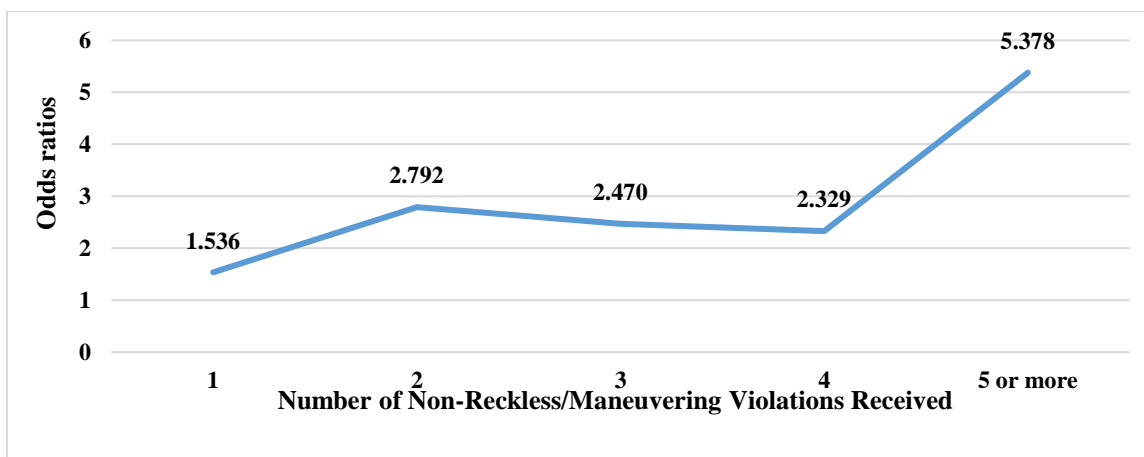


Figure 23: Odds ratios for Non-Reckless/Maneuvering Violations

The final estimation model indicated that the variables of years driving a vehicle, sex, reckless/maneuvering violations and non-reckless/maneuvering violations have a significant association with the dependent variables when analyzed at the same time. However, the variable sex showed some inconsistencies during the multiple regression analysis. Preliminary analyses indicated that sex was not associated with the dependent variable (as shown in the results of the chi-square tests and the simple logistic regression). Moreover, the preliminary analyses indicated that the association between sex (male) and the dependent variables of being involved in a vehicle crash was positive while the results of the multiple regressions analysis indicated that this association is negative. Due to these inconsistencies, the variable of sex was determined to not be included as an independent variable. Therefore, a new model was developed without the variable of sex using the same backwards elimination procedure. The results are shown in Table 58.

Table 58: Results for Stepwise Backwards Elimination Procedure (Without the Variable of Sex)

Term	Step 1		Step 2		Step 3	
	Coefficient	P-Value	Coefficient	P-Value	Coefficient	P-Value
<b>Constant</b>	-0.598		-0.713		-0.502	
<b>Years of Driving Experience</b>	0.052	0.002	0.053	0.002	0.033	< 0.001
<b>Daily Hours Driving</b>	-0.046	0.294				
<b>Age</b>	-1.148	0.187	-1.184	0.168		
<b>Reckless/Maneuvering Violations</b>	1.258	0.010	1.249	0.011	1.379	0.002
<b>Non-Reckless/Maneuvering Violations</b>	1.524	< 0.001	1.509	<0.001	1.636	< 0.001
<b>AIC</b>	1068.43		1067.55		1064.74	

The results show that the final model, as indicated in the step 3 column, contains the variables of years driving a vehicle and both reckless/maneuvering and non-reckless/maneuvering violations, similar to the previous estimation model. When comparing these results with the ones for the estimation model with the variable of sex, it can be seen that the removal of this variable did not cause any significant changes in the coefficients and probability values of the other independent variables. Although the results of the previous estimation model indicated that the variable of sex had a significant association with the dependent variable, the results of table 59 show that it was in fact a redundant variable that did not have a significant effect on the rest of the independent variables being considered. Table 59 shows the results of the final estimation model without including the variable of sex.

The results shown in Table 59, as mentioned before, indicate that the removal of the variable of sex did not cause significant changes in the probability values of the categories for the variables of reckless/maneuvering violations and non-reckless/maneuvering violations. Additionally, the VIF values of all variable being considered are still close to one, similar to the results of the previous multiple regression model that included the variable of sex. This adds to the conclusion that the variable of sex was a redundant predictor of being involved in a crash for this study. Table 60 provides results for the odds ratios for the model without the variable of sex. Results show the trend of the odds ratios did not change because of the removal of the variable of sex. It also shows that there is still an ascending behavior from zero to three traffic violations; the

odds descend for the category of four violations, and finally the odds ascend again when there were five violations or more.

Table 59: Results of Final Estimation Model (Without the Variable of Sex)

Term	Coefficient	Standard Error of Coefficient	Z-Value	P-Value	VIF
<b>Constant</b>	-0.502	0.130	-3.85	< 0.001	
<b>Years Driving a Vehicle</b>	0.033	0.006	5.22	<0.001	1.07
<b>Reckless/Maneuvering Violations</b>					
0	*	*	*	*	*
1	0.476	0.192	2.47	0.013	1.19
2	0.507	0.233	2.18	0.029	1.18
3	0.931	0.341	2.73	0.006	1.10
4	0.746	0.425	1.76	0.079	1.08
5 or more	1.379	0.437	3.15	0.002	1.12
<b>Non-Reckless/Maneuvering Violations</b>					
0	*	*	*	*	*
1	0.438	0.197	2.23	0.026	1.12
2	1.025	0.265	3.87	<0.001	1.07
3	0.882	0.362	2.44	0.015	1.04
4	0.821	0.582	1.41	0.159	1.05
5 or more	1.636	0.498	3.29	0.001	1.08

Table 60: Calculated Odds Ratios for Terms in Final Estimation Model (Without the Variable of Sex)

Term		Odds ratio	95% CI
<b>Years of Driving Experience</b>		1.033	(1.021, 1.046)
<b>Reckless/Maneuvering Violations</b>		<b>Odds ratio</b>	<b>95% CI</b>
<b>Level A</b>	<b>Level B</b>		
1	0	1.609	(1.104, 2.345)
2	0	1.661	(1.053, 2.621)
3	0	2.537	(1.302, 4.946)
4	0	2.109	(0.917, 4.848)
5 or more	0	4.307	(1.685, 9.354)
<b>Non-Reckless/Maneuvering Violations</b>		<b>Odds ratio</b>	<b>95% CI</b>
<b>Level A</b>	<b>Level B</b>		
1	0	1.550	(1.054, 2.280)
2	0	2.787	(1.658, 4.684)
3	0	2.416	(1.188, 4.915)
4	0	2.272	(0.726, 7.110)
5 or more	0	5.136	(1.935, 13.633)

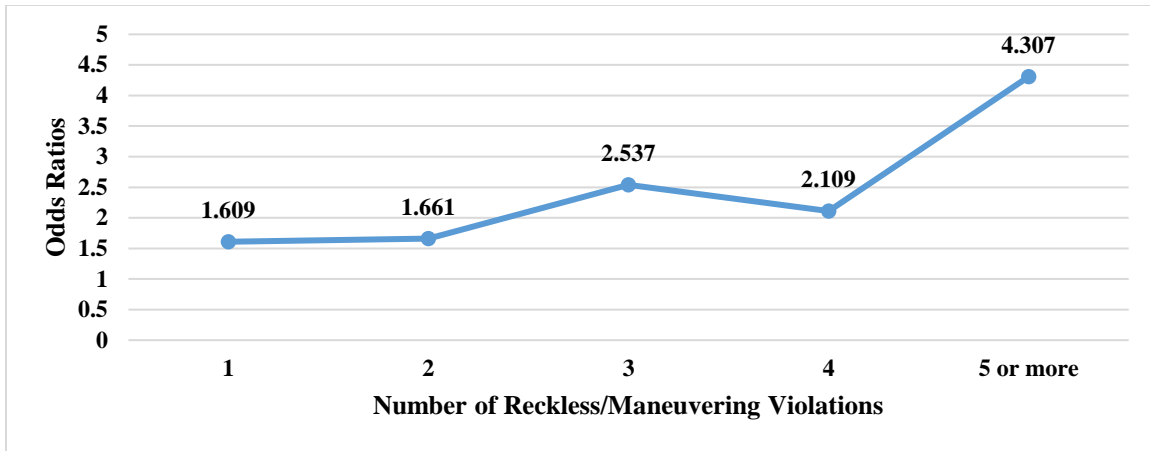


Figure 24: Odds ratios for Reckless/Maneuvering Violations (Model Without Variable of Sex)

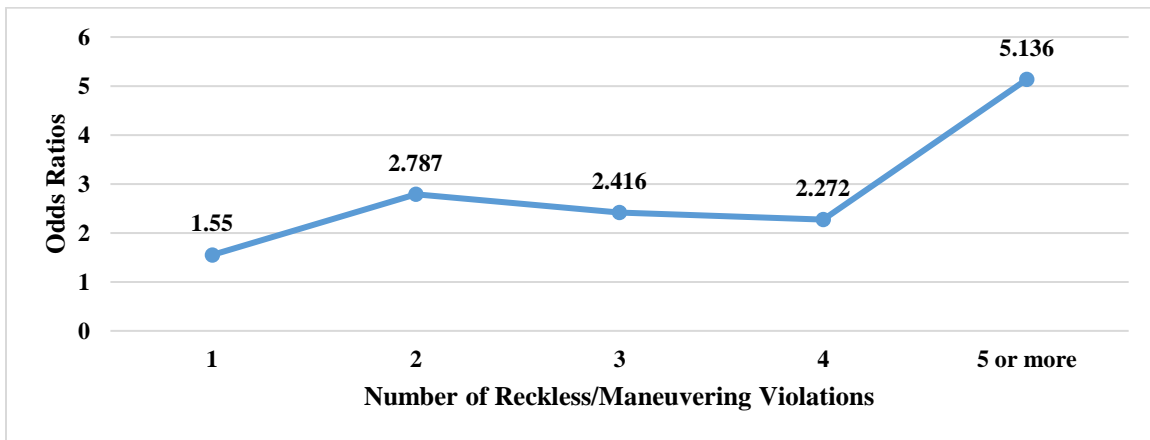


Figure 25: Odds ratios for Non-Reckless/Maneuvering Violations (Model without Variable of Sex)

### 5.3 Model Assessment

An assessment of the final estimation model was performed to determine how effective it is at describing the outcome variable. The Hosmer-Lemeshow test was used to analyze the goodness of fit of the selected model. A probability value of 0.863 was obtained for the Hosmer-Lemeshow test, indicating that the model provides a good fit of the data. Additionally, the results for the observed and expected frequencies obtained for the Hosmer-Lemeshow test are displayed in table 61. It can be seen that the observed and expected frequencies for each group are similar to each other which further assesses the goodness of fit of the model.

Table 61: Observed and Expected Frequencies for Hosmer-Lemeshow Test

Group	Event Probability Range	Crash Involvement			
		Yes		No	
		Observed	Expected	Observed	Expected
1	(0.000, 0.414)	40	34.8	54	59.2
2	(0.414, 0.447)	39	39.3	53	52.7
3	(0.447, 0.536)	41	46.1	53	47.9
4	(0.536, 0.605)	48	52.5	44	39.5
5	(0.605, 0.679)	62	59.6	30	32.4
6	(0.679, 0.728)	67	64.4	25	27.6
7	(0.728, 0.783)	70	69.6	22	22.4
8	(0.783, 0.844)	75	74.8	17	17.2
9	(0.844, 0.904)	80	80.3	12	11.7
10	(0.904, 0.989)	84	84.6	6	5.4

In addition to assessing the model using the Hosmer-Lemeshow test, a Receiving Operating Characteristic (ROC) curve was developed. The area under the resulting ROC curve provides a value which indicates the model's predictive ability. Figure 24 shows the ROC curve for the selected model provided by the Minitab output. The area under the ROC curve resulted with a value of approximately 0.73, which indicates that the model has an acceptable predictive ability.

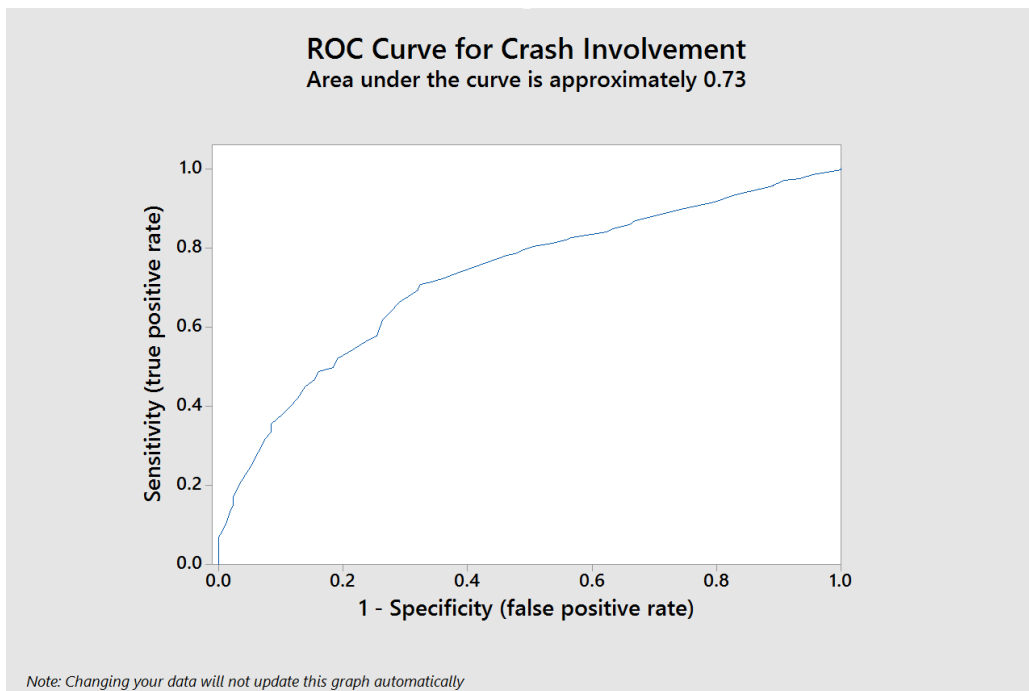


Figure 26: ROC Curve for the Selected Model (Obtained from Minitab)



The multiple logistic regression analysis presented in this section provided a fitted model that contained the most significant predictors of the outcome of being involved in a traffic crash. The significance of predictors was determined by the probability value obtained from the Minitab output results, with values lower than 0.05 being considered as significant predictors. The resulting model contained the following predictors:

- Years driving a vehicle,
- Number of Reckless/Maneuvering Violations Received, and
- Number of Non-Reckless/Maneuvering Violations Received.

Of all the predictors that were initially considered, these three predictors resulted to have the most significant association with the response outcome when analyzed along other independent variables. The set of predictors obtained in this chapter are similar to some that were identified as significant factors in the results of the literature review. One must mention that there were also some predictors that resulted to be significant in other studies of similar nature but for this research they were not significant predictors. Obviously, several factors such as the database and type of subject being analyzed can be associated with this issue. However, other studies have shown that age is not necessarily a significant predictor. It is the author's opinion that the predictor of age can be associated with years driving a vehicle in the sense that it is not necessarily the age that affects the likelihood of being involved in a traffic crash, but rather the experience that a person has for driving a vehicle for a longer period. The variable of sex resulted to be a significant independent variable along the others mentioned above but was ultimately left out of the model because of inconsistencies between the preliminary analyses and the multiple regression analysis. Because of this, the model that was developed in this section was fitted again using the backwards elimination procedure but without including the variable of sex. Results showed that the same variables (years driving a vehicle and both type of traffic violations) resulted to be the best subset of independent variables similar to the multiple regression model that included the variable of sex. This indicates that the removal of the variable of sex does not have a significant effect on the others variable that resulted to be significant predictors of being involved in a vehicle crash.

## 6. CONCLUSIONS AND RECOMMENDATIONS

The purpose of this study was to estimate the likelihood that a driver has of being involved in a traffic crash based on several factors such as traffic violations and crash history, among others. A literature review was performed to identify and understand which factors regarding human characteristics and behavior are most commonly associated with traffic crash involvement. Several studies found that factors such as age, sex, type of license, traffic school attendance, previous traffic violations and crashes, driving behavior and frequency of driving are significantly related with traffic crash involvement. In addition to exploring common factors, this literature review also served as a way of identifying common methodologies used for studying the relationship that these factors have on the likelihood of traffic crashes involvement

### 6.1 Conclusions

According to these studies, the most common approach for estimating whether a driver will be involved in a traffic crash or not based on a set of factor or variables is the use of multiple logistic regression procedures. Logistic regression is a type of regression analysis where the dependent variable is binary or dichotomous, meaning it can have one of two possible outcomes. The main objective of a logistic regression analysis is to find a model with the best fit that could describe the relationship between a response binary variable and a set of independent variables or predictors. For this study, the outcomes of the response variables were established as being involved in a traffic crash or not being involved in a traffic crash. In contrast to linear regression, logistic regression models do not require the data to follow a certain distribution and are overall less stringent than linear regression models.

In order to develop the proposed logistic regression model, information regarding the driving population of Puerto Rico was required. To obtain a sample of the population of licensed drivers in Puerto Rico for the development of the proposed model, a survey was developed to obtain information regarding demographics as well as traffic violation and crash history. The reason for developing this survey was the lack of access and availability of driver records that could provide detailed information regarding the traffic violations history of a sample of licensed drivers. The only requirement for participants of this survey was to have a driver's license and be at least 16 years old. The survey was developed and deployed using the web tool *SurveyMonkey*, which provides the user the opportunity to develop different types of questionnaires as well as

outlets for distributing the survey. This proved to be convenient since the survey was deployed using outlets such as e-mail and social media (*Facebook*).

The questions included in this survey were prepared based on the information obtained from the literature review results regarding the significant factors related to traffic crash involvement. The composition of the survey consisted of three parts; general information, traffic violations history and traffic crash history. In the general information part, information such as age, sex years driving a vehicle and daily hours spent driving were inquired about. Age and sex are commonly used factors in any type of study of this nature while years driving a vehicle and daily hours driving were also considered important factors of traffic crash involvement by the author of this study as well the studies included in the literature review. The next part of the survey consisted of information regarding traffic violations history, where participants were asked to indicate which type traffic violations they have received as well as the number of violations received. A list of traffic violation was provided to participants so they could choose which ones they had received as well as the quantity. Additionally, participants were provided a space to include traffic violations that were not included in the list. The types of traffic violations provided in this part were chosen based on the ones included in the studies of the literature review and the ones recommended by police officers of the Puerto Rico Police Department.

The last part of the survey consisted of the traffic crash history of participants, where participants were asked to indicate in how many traffic crashes they have participated as drivers as well as indicating the age at the time of the crash, the severity and whether they were responsible or not for the crash. Crash severity was identified as property damage only (PDO), minor injury, severe injury and fatal. A total of 1005 responses were obtained from the survey where 409 (41.1%) of responses corresponded to male participants while 587 (58.9%) corresponded to female participants. Most responses corresponded to drivers between the ages of 16 and 30 years of age.

After creating the database using the results obtained from the survey, a data filtering process took place to remove certain observations that were not answered completely or observations where the response did not make sense or were answered wrongfully; a total of 952 responses remained after this data filtering process. Several variables were identified in the created database and were categorized into two groups: continuous and categorical variables. Continuous variables were comprised of years driving a vehicle as well as daily hours spent driving and total traffic crashes which are variables that consist of numerical values provided by the participants.

On the other hand, categorical variables such as sex and age are variables based on questions where the participants were asked to choose from a list of options or categories. These variables do not consist on any numerical number provided by participants but rather of categories that a participant thinks applied to him or her.

Once the database was created and filtered, descriptive statistics were obtained from the final sample of data. The process of descriptive statistics was performed to have an initial understanding of the data included in the sample without doing complicated statistical analyses. The following was concluded from the information in the sample:

- 51% of participants corresponded to people between the ages of 16 and 30.
- The majority of the responses of the survey corresponded to females (59%).
- Most female and male participants corresponded to ages between 16 and 30 years (66% and 54% respectively).
- The mean of years driving a vehicle is 15 years.
- The mean of daily hours spent driving according to participants is 2.5 hours/day.
- 70% of participants indicated they had received traffic violations.
- The most common traffic violations received in the sample corresponded to driving over the speed limit and illegal parking violations (36% and 28%, respectively).
- 65% of participants indicated to have been involved in a traffic crash as a driver,
- 88% of crashes reported in the survey corresponded to PDO crashes.
- Between the ages of 31-60, males had a larger percentage of traffic crashes compared to females (49% vs. 43%).

After the descriptive statistics analyses were finished, bivariate preliminary analyses were performed. These analyses consist of comparing two variables, a response and a predictor variable, with the purpose of analyzing the significance that the predictor variable has on the response variable. The purpose of these preliminary analyses was to study the relationship between the variables identified in the sample database and traffic crash involvement. Information for response variable chosen was obtained from the responses of the question where participants were asked if they had been involved in a traffic crash or not. The predictor variables compared consisted of the information obtained from the variables mentioned previously such as age, sex, traffic violations, and traffic crashes. These analyses were performed using the statistical software *Minitab* and consisted of chi-square tests of independence and simple logistic regression analyses.

These analyses depended on the type of predictor variable being compared; when a predictor was categorical, chi-square tests of independence were performed while simple logistic regression whereas used for continuous predictors as recommended by Hosmer and Lemeshow in 1999.

The chi-square test of independence is a parametric test, which means that it does not require a specific distribution for the data; it is used to determine if two categorical variables are independent of each other. To perform this test, data has to be rearranged into contingency tables which are a mean of displaying the joint frequencies of two categorical variables. In the case of this study, the frequencies corresponded to the joint responses obtained from the survey. Several assumptions regarding the distribution of values in a contingency table had to be met in order to perform the chi-square test of independence, the most important one being that no cell in the table should have an expected value lower than one. This was an issue since several of the tables corresponding to traffic violations had cells with expected values lower than one. For the traffic violations presented in this study, participants had to indicate in the survey the number of violations receive by selecting one of the 5 choices or categories provided which corresponded to 1, 2, 3, 4 and 5 or more traffic violations received. This was done so the responses of the survey would be maintained as controlled as possible. Because of this, many of the categories for number of traffic violations received did not apply to participants. For instance, almost all of the participants indicated to never have received traffic violations for driving under the influence, thus the frequencies for the categories of 1, 2 3 and 4 driving under the influence of alcohol traffic violations would be equal to zero. This was the case for other traffic violations, with the exception of driving over the speed limit and illegal parking since these were the most commonly received responses as indicated by the descriptive statistics analysis. Because of this, traffic violations were categorized into reckless/maneuvering and non-reckless/maneuvering violations.

Reckless/maneuvering violations concern traffic violations where the participant showed a reckless behavior or performed an illegal maneuver whereas non-reckless/maneuvering violations corresponded to traffic violations that were not indicative of reckless behavior or any type of illegal maneuver. After the various traffic violations were compacted into these two categories, contingency tables were developed again so they could comply with the requirement of frequency values in the cells.

Once the data concerning to the respective categorical variables was rearranged into contingency tables, chi-square tests of independence analyses were performed using the *Minitab*

software. The main result used to determine independence between the two variables being compared was the probability value associated with the Pearson and likelihood ratio chi-square statistics using the following hypotheses:

- If P-Value  $\leq 0.05$ , there is a significant association between both variables at the 95% confidence level
- If P-Value  $> 0.05$ , there is not enough information to say that there is a significant association.

Additionally, several goodness of fit tests were used to assess this association. Results for the chi-square tests of independence analyses performed indicated that age, reckless/maneuvering violations, and non-reckless/maneuvering violations have a significant association with being involved in a traffic crash; the variable sex was not found to have a significant association with the response variable.

In addition to chi-square tests of independence, simple logistic regression analyses were also performed to study the relationship between the different predictor variables identified and being involved or not in a traffic crash. Simple logistic regression consists of logistic regression model where only one predictor variable is being compared to the response variable. The difference between simple and the multiple logistic regression procedure mentioned previously and in the literature review is the number of predictors included. Whenever more than one predictor is being compared to the response variable it becomes a multiple logistic regression model rather than a simple logistic regression. The purpose of performing a simple logistic regression is because this analysis provides the opportunity of comparing a continuous variable with a response binary variable, unlike chi-square tests of independence. However, categorical variables were also analyzed using simple logistic regression to compare the results with the ones obtained from the chi-square tests of independence.

When starting the simple logistic regression analyses in Minitab, several statements have to be established such as, the outcome event chosen for the response and the confidence interval for the significance tests. The outcome chosen for these analyses was being involved in a traffic crash while a 95% confidence interval was chosen for the level of significance. The variables that were initially considered were years driving a vehicle, daily hours spent driving, total traffic crashes, PDO crashes, minor injury crashes and severe injury crashes. Unfortunately, the simple logistic regression models obtained for data regarding traffic crashes suffered from complete

separation which occurs when a linear combination of predictor variables provide a perfect prediction of the outcomes of the response variable, in this case being involved or not in a traffic crash. Consider that the frequencies or counts being used in these analyses correspond to participants, the number of participants that indicated to have been involved in a traffic crash or not is the same regardless of the predictor that is being used for comparing. When analyzing predictors such as age and years driving a vehicle, complete separation does not occur because participants that had received traffic violations did not have to necessarily be involved in a traffic crash and vice versa. When analyzing total vehicle crashes, complete separation occurs because participants who indicated to be involved in a traffic crash also had a number of total traffic crashes while participants who were not involved in a vehicle crash had zero total vehicle crashes. Since this was a problem that was created from the data collection process, it was decided that predictors concerning to vehicle crashes were going to be omitted from further analyses.

The output results provided by Minitab included information regarding the following; coefficients, odds ratios and goodness of fit tests. The coefficients information indicates the directions of the correlation between the predictor and response variable as well as the magnitude of these correlations. The significance of the predictors was determined using the probability value column. The odds ratio column provides information regarding the odds of achieving the outcome event based on the odds of the predictor variable. For continuous predictors, the odds ratio indicate how much the odds of achieving the response outcome increase or decrease for a unit change in the predictor coefficient. On the other hand, the odds ratio for a categorical variable can be interpreted as the odds that one of the categories of the predictor has of achieving the response event outcome based on the odds of the reference category. For each categorical predictor, a category had to be chosen as the reference or control category. Results for the simple logistic regression analyses indicated the following;

- Variables for sex and daily hours spent driving resulted to be non-significant predictors of being involved in a traffic crash.
- Variables for years driving a vehicle, age, reckless/maneuvering violations and non-reckless/maneuvering violations resulted to be significant predictors of being involved in a traffic crash.

- An increase in years driving a vehicle indicated an increase in the odds of being involved in a traffic crash while an increase in daily hours spent driving showed a decrease in the odds of being involved in a traffic crash.
- Older participants were shown to have increased odds of being involved in a vehicle crash when compared to younger drivers.
- Male participants have decreased odds of being involved in a traffic crash than females
- Participants that indicated to have committed at least one reckless/maneuvering violation showed increased odds of being involved in a traffic crash than participants who indicated to not have committed traffic violations.
- Participants that indicated to have committed at least one non-reckless/maneuvering violation also showed increased odds of being involved in a traffic crash than participants who indicated to not have committed traffic violations.

Once the preliminary analyses were finished and an idea of the association between each predictor considered and being involved in a traffic crash was obtained, multiple logistic regression analyses were performed.

The process of multiple logistic regression analyses started with the fitting of a logistic regression that contained all six predictor variables being considered. Results for the significance of the coefficients in this model indicated that daily hours spent driving and age were non-significant predictors of being involved in a traffic crash when being analyzed with other independent variables in the same model. The results obtained for the independent variable of daily hours spent driving was the same as the one obtained in the simple logistic regression analysis; in both analyses this predictor was non-significant, however, this is not the case for age in participants. The results for the chi-square tests of independence and simple logistic regression analyses indicated that age has a significant association with being involved in a traffic crash, but this was not the case when other predictor variables were included in a logistic regression model. Results for years driving a vehicle, sex, reckless/maneuvering violations and non-reckless/maneuvering violations indicated that these variables are significant predictors of being involved in a traffic crash.

In order to compare other models that could have a better fit than the full model including all six variables, a backwards elimination stepwise procedure was performed. Comparison of models was determined using the Akaike Information Criterion (AIC) which indicates how well a



model fits the data regardless of the number of predictors included; lower values of AIC indicate a better fit. In this stepwise procedure, predictors that result to be non-significant are removed in an iterative process that stops when a model that contains only significant predictors remains. For the full model obtained in this analysis, the first step was to remove the predictor that resulted to have the most non-significance for the response outcome; in this case, daily hours spent driving was removed and the remaining model was fitted again. The remaining model provided a better fit since the AIC value obtained was lower. When inspecting the significance of the remaining predictors, the predictor that showed the least significance in the model was age with a p-value of 0.11 (which is larger than 0.05). Thus, this predictor was removed and the remaining predictors were fitted in another model. In this third step, the resulting model had an even lower AIC value and also showed that every predictor included was significant at the 95% confidence interval. The remaining model contained the following predictors:

- Years driving a vehicle,
- Sex,
- Reckless/maneuvering violation, and
- Non-reckless/maneuvering violations.

The results obtained from this multiple logistic regression analysis regarding which predictors can be considered significant when predicting traffic crash involvement are similar to the results shown by previous studies in the literature review while also being consistent with the results obtained in the preliminary analyses. Also, the resulting model makes sense when thinking about it from a point of view of experience: younger drivers (16-20) can be more likely to be involved in a traffic crash since they have almost no experience and usually have a more immature mentality than older driver. However, the variable of sex showed inconsistencies between the preliminary and multiple regression analyses. Because of these inconsistencies and the fact that a model fitted without the variable of sex yielded almost equal results to one where this variable was included, it was ultimately decided to remove it from the final model. Finally, traffic violations, in the form of reckless/maneuvering and non-reckless/maneuvering violations, can be considered significant if one considers that a traffic violation history that includes many traffic violations committed might be associated with a pattern of reckless behavior when driving and not obeying traffic laws.

## 6.2 Recommendations

Although surveys and questionnaires can be helpful for the fact that they can be used to obtain information directly from the study subjects, this is not necessarily an ideal thing since the type of information that is being collected can affect whether participants want to complete the survey. The experience when collecting data for this dissertation was that not every person that was approached to complete the survey accepted to participate, especially when performing on-site surveys where participants were completing the survey in the presence of the person conducting the survey. The studies included in the literature review usually indicated that a database that consisted of crash and driver records were used. Unfortunately for this study, this type of database for the population of Puerto Rico was not accessible. Additional benefits of using this type of database are:

- Larger sample size can be achieved,
- Data can be obtained for certain time periods,
- Increased number of variables can be considered, and
- Better assessment and validation of models can be achieved.

If the purpose of a future study requires a large sample of data to be obtained through survey or questionnaires, it is recommended that a group of people should help the researcher in collecting the required data. Electronic tools such as *SurveyMonkey* and others can facilitate the deployment and distribution of surveys by using outlets such as e-mail and social media. However, a paper version of the survey was developed to obtain responses from senior participants that do not necessarily use such outlets. The problem with having such a wide distribution of ages among the target population was the fact that most of the responses that were collected were obtained from social media outlets and e-mails which senior participants are not necessarily familiarized with, this is the main reason of why there was a small number of senior participants included in the final sample of data. Another factor that could increase the amount of responses obtained for a survey or questionnaire is offering a reward to participants who complete the survey of questionnaire. The problem with this approach is the fact that it requires an increased economic influx into the research project if the sample that is wished to be obtained needs to be large.



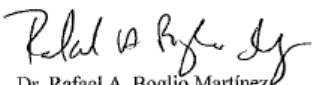
## REFERENCES

- Agresti, A. *Categorical Data Analysis, Second Edition*. John Wiley & Sons Inc., New Jersey, 2002.
- Allison, P. Measures of Fit for Logistic Regression. SAS Global forum, Washington, D.C., 2014.
- Chandraratna, S., and N Stamatiadis. Evaluation of the Characteristics of Drivers with Multiple Crashes. University of Kentucky, Southeast Transportation Center, 2004.
- Daigneault G., et al. Previous Convictions or Accidents and the Risk of Subsequent Accidents for Older Drivers. *Accident Analysis and Prevention*, Vol. 34, 2002, pp. 257–261.
- Gebers, M. *Strategies for Estimating Driver Accident Risk in Relation to California's Negligent-Operator Point System*. California Department of Motor Vehicles – Research and Development Branch. Technical Monograph 183, 1999.
- Gebers M., and R. Peck. Using Traffic Conviction Correlates to Identify High Accident-Risk Drivers. *Accident analysis and Prevention*, No. 35, 2003, pp. 903–912. Guangnan, Z., et al. C. Risk Factors Associated with Traffic Violations and Accident Severity in China. *Accident Analysis and Prevention* 59, 18-25, 2013.
- Hosmer and Lemeshow. *Applied Logistic Regression: Second Edition*. John Wiley & Sons Inc., New Jersey, 2000.
- Karacasu M., and E. Arzu. An Analysis on Distribution of Traffic Faults in Accidents, Based on Driver's Age and Sex: Eskisehir Case. *Procedia Social and Behavioral Sciences*, Vol. 20, 2001, pp. 776–785.
- Shawky M., and A. Al-Ghafli. Risk Factors Analysis for Drivers with Multiple Crashes. *International Journal of Engineering and Applied Sciences (IJEAS)*, Vol. 3, No. 11, 2016.
- Murray, D., et al. Predicting Truck Crash Involvement: Developing a Commercial Driver Behavior Model and Requisite Enforcement Countermeasures. 47<sup>th</sup> Annual Transportation Research Forum, 2006.
- Nishida, Y. Road Traffic Accident Involvement Rate by Accident and Violation Records: New Methodology for Driver Education based on Integrated Road Traffic Accident Database. National Research Institute of Police Science, Japan. 99-106, 2009.

- Peden, M., et al. World Health Organization. World Report on Road Traffic Injury Prevention. Geneva, Switzerland. *World Health Organization Catalogue*, 2004.
- Subasish D., et al. Estimating likelihood of future crashes for crash-prone drivers. *Journal of Traffic and Transportation Engineering*, Vol. 2, No. 3, 2015, pp. 145–157.
- Stoltzfus, J.C. Logistic Regression: A Brief Primer. Society of Academic Emergency Medicine. *Academic Emergency Medicine* 18, 1099-1104, 2011.
- World Health Organization. Global Status on Road Safety 2015. Geneva, Switzerland. *World Health Organization Catalogue*, 2015.
- Wundersitz LN, and NR Burns. Relationships Between Prior Driving Record, Driver Culpability, and Fatal Crash Involvement.

## APPENDIX

## A.1 Committee for Protection of Human Rights in Research Approval Document

	<p><b>Comité para la Protección de los Seres Humanos en la Investigación</b>  <b>CPSHI/IRB 00002053</b>          Universidad de Puerto Rico – Recinto Universitario de Mayagüez          Decanato de Asuntos Académicos          Call Box 9000          Mayagüez, PR 00681-9000</p>	
<p>25 de febrero de 2016</p> <p>Armando González Bonilla          Ingeniería Civil          RUM</p> <p>Estimado estudiante:</p> <p>El Comité para la Protección de los Seres Humanos en la Investigación (CPSHI) ha considerado el proyecto titulado <i>Desarrollo de un modelo estadístico para identificar conductores de alto riesgo</i> (# Protocolo 20160219). Luego de evaluar el mismo hemos certificado que este cumple con todos los requisitos para ser aprobado como exento bajo la Categoría 2 del 45 CFR 46.101(b). La determinación de exención implica que este proyecto no requiere ser re-evaluado ni re-autorizado por nuestro comité. Le recordamos que la aprobación emitida por nuestro comité no lo exime de cumplir con cualquier otro requisito institucional o gubernamental relacionado al tema o fuente de financiamiento de su proyecto.</p> <p>Cualquier cambio al protocolo o a la metodología que altere los criterios de exención deberá ser revisado y aprobado por el CPSHI ANTES de su implantación, excepto en casos en que el cambio sea necesario para eliminar algún riesgo inmediato para los/as participantes. El CPSHI deberá ser notificado de dichos cambios tan pronto le sea posible al/ a la investigador/a. Igualmente, el CPSHI deberá ser informado de inmediato de cualquier efecto adverso o problema inesperado que surgiera con relación al riesgo de los seres humanos, de cualquier queja sobre la conducción de esta investigación y de cualquier violación a la confidencialidad de los participantes.</p> <p>Atentamente,</p> <p style="text-align: center;">           Dr. Rafael A. Boglio Martínez          Presidente          CPSHI/IRB          UPR - RUM       </p> <p style="text-align: center; font-size: small;">         Teléfono: (787) 832 - 4040 x 6277, 3807, 3808 – Fax: (787) 831-2085 – Página Web: <a href="http://www.uprm.edu/cpshi">www.uprm.edu/cpshi</a>          Email: <a href="mailto:cpshi@uprm.edu">cpshi@uprm.edu</a> </p>		

## A.2 Example of Survey

<b>HISTORY OF TRAFFIC VIOLATIONS AND CRASHES</b>	
<p><b>Instructions:</b> The purpose of the following questionnaire is the collection of data regarding traffic violations and crashes of drivers in Puerto Rico. Participating in this questionnaire is voluntary, if you refuse to participate or feel uncomfortable at any time, feel free to stop. No reward will be given for participating in this questionnaire. For more information about this questionnaire and the project for which it was created, please contact Dr. Ivette Cruzado to <a href="mailto:ivette.cruzado@upr.edu">ivette.cruzado@upr.edu</a>.</p>	
<b>General Information</b>	
1. Age	
2. Sex (As indicated on driver license)	
3. How many years of experience do you have driving a motor vehicle?	
4. On a regular day, how many hours do you drive your motor vehicle?	
<b>History of Traffic Violations</b>	
5. Have you received traffic violations?	
Please indicate the number of traffic violations received next to the corresponding traffic violation:	
Number	Traffic violation
	Speeding
	Driving under the influence of drugs and alcohol
	Ignoring traffic signals and signs
	Not using safety belt
	Driving too close to front vehicle
	Illegal parking
	Illegal turn
	Reckless lane switching
	Using cellphone while driving
Indicate any traffic violation that is not indicated on the previous table	

Figure 27: Page 1 of 2 from the Developed Survey

**History of traffic crashes**

6. Have you been involved in a traffic crash as a driver?

If you have been involved in any traffic crash as a driver, please indicate your age at the time of the crash, the severity of the crash and whether you were responsible for the crash or not

Age	Severity	Responsibility

Severity of the crash can be one of the following:

Property damage (PDO): Nobody was injured, only damage to the vehicle or other property.

Light (L): At least one person was injured but no hospitalization was required.

Severe (S): At least one person was hospitalized as a result of injuries from the traffic crash.

Fatal (F): At least one person died as a result of the traffic crash.

Responsibility can be one of the following:

Responsible (R): The traffic crash occurred as a result of your actions.

Not responsible (NR): The traffic crash occurred as a result of actions beyond your control.

---

Here ends this questionnaire, thank you for your participation.

Figure 28: Page 2 of 2 from the Developed Survey

### A.3 Chi-Square Test of Independence Results

#### Chi-Square Test Results for Age vs Crash Involvement

Rows: Age Columns: Crash Involvement

	No	Yes	All
16-20	120	90	210
	72.79	137.21	
	7.753	-7.753	
	30.612	16.241	
21-30	129	245	374
	129.64	244.36	
	-0.090	0.090	
	0.003	0.002	
31-40	22	77	99
	34.32	64.68	
	-2.748	2.748	
	4.421	2.346	
41-50	21	82	103
	35.70	67.30	
	-3.224	3.224	
	6.055	3.213	
51-60	24	86	110
	38.13	71.87	
	-3.010	3.010	
	5.236	2.778	
61-89	14	42	56
	19.41	36.59	
	-1.566	1.566	
	1.509	0.800	
All	330	622	952

Cell Contents: Count  
Expected count  
Adjusted residual  
Contribution to Chi-square

Pearson Chi-Square = 73.217, DF = 5, P-Value = 0.000  
Likelihood Ratio Chi-Square = 72.556, DF = 5, P-Value = 0.000

Cramer's V-square 0.076908  
Pearsons r 0.223915  
Spearmans rho 0.257155



## Chi-Square Test Results for Sex vs Crash Involvement

Rows: Sex Columns: Crash Involvement

	No	Yes	All
Female	201	363	564
	195.5	368.5	
	0.7617	-0.7617	
	0.1545	0.0820	
Male	129	259	388
	134.5	253.5	
	-0.7617	0.7617	
	0.2246	0.1191	
All	330	622	952

Cell Contents:           Count  
                               Expected count  
                               Adjusted residual  
                               Contribution to Chi-square

Pearson Chi-Square = 0.580, DF = 1, P-Value = 0.446  
 Likelihood Ratio Chi-Square = 0.581, DF = 1, P-Value = 0.446

Fisher's exact test: P-Value = 0.488357

Cramer's V-square   0.0006094  
 Pearsons r           0.0246865  
 Spearmans rho       0.0246865

### Chi-Square Test Results for Reckless/maneuvering violations vs Crash Involvement

Rows: Reckless/maneuvering violations    Columns: Crash Involvement

	No	Yes	All
0	195 136.58 8.079 24.993	199 257.42 -8.079 13.260	394
1	69 76.26 -1.173 0.691	151 143.74 1.173 0.367	220
2	37 49.22 -2.337 3.035	105 92.78 2.337 1.610	142
3	14 25.65 -2.964 5.292	60 48.35 2.964 2.808	74
4	8 15.25 -2.352 3.448	36 28.75 2.352 1.829	44
5	7 27.04 -4.976 14.850	71 50.96 4.976 7.879	78
All	330	622	952

Cell Contents:            Count  
                               Expected count  
                               Adjusted residual  
                               Contribution to Chi-square

Pearson Chi-Square = 80.062, DF = 5, P-Value = 0.000  
 Likelihood Ratio Chi-Square = 85.369, DF = 5, P-Value = 0.000

Cramer's V-square    0.084099  
 Pearsons r            0.274884  
 Spearmans rho        0.287869

## Chi-Square Test Results for Non-Reckless/maneuvering violations vs Crash Involvement

Rows: Non-Reckless/maneuvering violations    Columns: Crash Involvement

	No	Yes	All
0	232	268	500
	173.32	326.68	
	8.003	-8.003	
	19.867	10.541	
1	56	135	191
	66.21	124.79	
	-1.736	1.736	
	1.574	0.835	
2	22	96	118
	40.90	77.10	
	-3.907	3.907	
	8.736	4.635	
3	11	43	54
	18.72	35.28	
	-2.272	2.272	
	3.183	1.689	
4	4	22	26
	9.01	16.99	
	-2.094	2.094	
	2.788	1.479	
5	5	58	63
	21.84	41.16	
	-4.613	4.613	
	12.983	6.888	
All	330	622	952

Cell Contents:            Count  
                                   Expected count  
                                   Adjusted residual  
                                   Contribution to Chi-square

Pearson Chi-Square = 75.197, DF = 5, P-Value = 0.000

Likelihood Ratio Chi-Square = 81.703, DF = 5, P-Value = 0.000

Cramer's V-square    0.078989

Pearsons r            0.262395

Spearmans rho        0.279033

## A.4 Simple Logistic Regression Raw Output

### Simple Logistic Regression: Years of Experience vs Crash Involvement

#### Deviance Table

Source	DF	Seq Dev	Contribution	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	1	61.31	5.02%	61.31	61.314	61.31	0.000
Years of Exp	1	61.31	5.02%	61.31	61.314	61.31	0.000
Error	945	1160.62	94.98%	1160.62	1.228		
Total	946	1221.94	100.00%				

#### Model Summary

Deviance	Deviance	
R-Sq	R-Sq(adj)	AIC
5.02%	4.94%	1164.62

#### Coefficients

Term	Coef	SE Coef	95% CI	Z-Value	P-Value	VIF
Constant	0.050	0.101	(-0.148, 0.248)	0.49	0.622	
Years of Exp.	0.04289	0.00600	(0.03112, 0.05465)	7.15	0.000	1.00

#### Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Years of Exp	1.0438	(1.0316, 1.0562)

#### Regression Equation

$$P(S_i) = \exp(Y') / (1 + \exp(Y'))$$

$$Y' = 0.050 + 0.04289 \text{ Years of Exp}$$

#### Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	945	1160.62	0.000
Pearson	945	967.22	0.301
Hosmer-Lemeshow	7	31.92	0.000

## Simple Logistic Regression: Crash Involvement vs Daily Hours Spent Driving

### Deviance Table

Source	DF	Seq Dev	Contribution	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	1	0.74	0.06%	0.74	0.7355	0.74	0.391
Daily Hours	1	0.74	0.06%	0.74	0.7355	0.74	0.391
Error	925	1191.45	99.94%	1191.45	1.2881		
Total	926	1192.18	100.00%				

### Model Summary

Deviance	Deviance	
R-Sq	R-Sq(adj)	AIC
0.06%	0.00%	1195.45

### Coefficients

Term	Coef	SE Coef	95% CI	Z-Value	P-Value	VIF
Constant	0.731	0.117	( 0.501, 0.961)	6.23	0.000	
Daily Hours	-0.0332	0.0385	(-0.1088, 0.0423)	-0.86	0.389	1.00

### Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Daily Hours	0.9673	(0.8969, 1.0432)

### Regression Equation

$$P(S_i) = \exp(Y') / (1 + \exp(Y'))$$

$$Y' = 0.731 - 0.0332 \text{ Daily hours}$$

### Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	925	1191.45	0.000
Pearson	925	927.22	0.473
Hosmer-Lemeshow	3	3.52	0.318

## Simple Logistic Regression: Crash Involvement vs Age

### Deviance Table

Source	DF	Seq Dev	Contribution	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	5	72.56	5.90%	72.56	14.511	72.56	0.000
Age	5	72.56	5.90%	72.56	14.511	72.56	0.000
Error	946	1156.17	94.10%	1156.17	1.222		
Total	951	1228.73	100.00%				

### Model Summary

Deviance	Deviance	
R-Sq	R-Sq(adj)	AIC
5.90%	5.50%	1168.17

### Coefficients

Term	Coef	SE Coef	95% CI	Z-Value	P-Value	VIF
Constant	-0.288	0.139	( -0.561, -0.014)	-2.06	0.039	
Age						
16-20	0.000000	0.000000	(0.000000, 0.000000)	*	*	*
21-30	0.929	0.177	( 0.582, 1.276)	5.25	0.000	1.52
31-40	1.540	0.279	( 0.993, 2.087)	5.52	0.000	1.22
41-50	1.650	0.282	( 1.098, 2.202)	5.86	0.000	1.21
51-60	1.564	0.270	( 1.035, 2.093)	5.80	0.000	1.24
61-89	1.386	0.339	( 0.723, 2.050)	4.09	0.000	1.14

### Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
Age			
21-30	16-20	2.5323	(1.7905, 3.5814)
31-40	16-20	4.6667	(2.7006, 8.0642)
41-50	16-20	5.2063	(2.9984, 9.0401)
51-60	16-20	4.7778	(2.8161, 8.1058)
61-89	16-20	4.0000	(2.0597, 7.7681)

Odds ratio for level A relative to level B

### Regression Equation

$$P(S_i) = \exp(Y') / (1 + \exp(Y'))$$

$$Y' = -0.288 + 0.0 \text{ Age}_{16-20} + 0.929 \text{ Age}_{21-30} + 1.540 \text{ Age}_{31-40} + 1.650 \text{ Age}_{41-50} + 1.564 \text{ Age}_{51-60} + 1.386 \text{ Age}_{61-89}$$

### Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	946	1156.17	0.000
Pearson	946	952.00	0.439
Hosmer-Lemeshow	3	0.00	1.000

## Simple Logistic Regression: Crash Involvement vs Sex

### Deviance Table

Source	DF	Seq Dev	Contribution	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	1	0.58	0.05%	0.58	0.5814	0.58	0.446
Sex	1	0.58	0.05%	0.58	0.5814	0.58	0.446
Error	950	1228.15	99.95%	1228.15	1.2928		
Total	951	1228.73	100.00%				

### Model Summary

Deviance	Deviance	
R-Sq	R-Sq(adj)	AIC
0.05%	0.00%	1232.15

### Coefficients

Term	Coef	SE Coef	95% CI	Z-Value	P-Value	VIF
Constant	0.5911	0.0879	( 0.4188, 0.7634)	6.72	0.000	
Sex						
Females	0.000000	0.000000	(0.000000, 0.000000)	*	*	*
Males	0.106	0.139	( -0.167, 0.379)	0.76	0.446	1.00

### Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
Sex			
Males	Females	1.1117	(0.8465, 1.4601)

Odds ratio for level A relative to level B

### Regression Equation

$$P(S_i) = \exp(Y') / (1 + \exp(Y'))$$

$$Y' = 0.5911 + 0.0 \text{ Females} + 0.106 \text{ Males}$$

### Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	950	1228.15	0.000
Pearson	950	952.00	0.476
Hosmer-Lemeshow	0	0.00	*

## Simple Logistic Regression: Crash Involvement vs Reckless/maneuvering violations

### Deviance Table

Source	DF	Seq Dev	Contribution	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	5	87.07	7.09%	87.07	17.414	87.07	0.000
Moving	5	87.07	7.09%	87.07	17.414	87.07	0.000
Error	946	1141.66	92.91%	1141.66	1.207		
Total	951	1228.73	100.00%				

### Model Summary

Deviance	Deviance	AIC
R-Sq	R-Sq(adj)	
7.09%	6.68%	1153.66

### Coefficients

Term	Coef	SE Coef	95% CI	Z-Value	P-Value	VIF
Constant	0.015	0.101	( -0.182, 0.213)	0.15	0.880	
Moving						
0	0.000000	0.000000	(0.000000, 0.000000)	*	*	*
1	0.768	0.177	( 0.421, 1.115)	4.34	0.000	1.13
2	1.001	0.215	( 0.581, 1.422)	4.67	0.000	1.10
3	1.531	0.322	( 0.900, 2.161)	4.76	0.000	1.05
4	1.489	0.404	( 0.698, 2.280)	3.69	0.000	1.03
5	2.302	0.409	( 1.500, 3.103)	5.63	0.000	1.03

### Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
Moving			
1	0	2.1552	(1.5238, 3.0484)
2	0	2.7213	(1.7872, 4.1436)
3	0	4.6212	(2.4599, 8.6815)
4	0	4.4318	(2.0090, 9.7767)
5	0	9.9892	(4.4829, 22.2589)

Odds ratio for level A relative to level B

### Regression Equation

$$P(S_i) = \exp(Y') / (1 + \exp(Y'))$$

$$Y' = 0.015 + 0.0 \text{ Total Moving (Category)}_0 + 0.768 \text{ Total Moving (Category)}_1 + 1.001 \text{ Total Moving (Category)}_2 + 1.531 \text{ Total Moving (Category)}_3 + 1.489 \text{ Total Moving (Category)}_4 + 2.302 \text{ Total Moving (Category)}_5$$

### Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	946	1141.66	0.000
Pearson	946	952.00	0.439
Hosmer-Lemeshow	3	0.00	1.000



## Simple Logistic Regression: Crash Involvement vs Non-Reckless/maneuvering violations

### Deviance Table

Source	DF	Seq Dev	Contribution	Adj Dev	Adj Mean	Chi-Square	P-value
Regression	5	81.59	6.64%	81.59	16.319	81.59	0.000
Non Moving	5	81.59	6.64%	81.59	16.319	81.59	0.000
Error	946	1147.14	93.36%	1147.14	1.213		
Total	951	1228.73	100.00%				

### Model Summary

Deviance	Deviance	
R-Sq	R-Sq(adj)	AIC
6.64%	6.23%	1159.14

### Coefficients

Term	Coef	SE Coef	95% CI	Z-Value	P-Value	VIF
Constant	0.1442	0.0897	( -0.0315, 0.3200)	1.61	0.108	
Non Moving						
0	0.000000	0.000000	(0.000000, 0.000000)	*	*	*
1	0.736	0.182	( 0.378, 1.093)	4.03	0.000	1.06
2	1.329	0.253	( 0.834, 1.825)	5.26	0.000	1.04
3	1.242	0.349	( 0.558, 1.926)	3.56	0.000	1.02
4	1.514	0.553	( 0.430, 2.598)	2.74	0.006	1.01
5	2.307	0.475	( 1.376, 3.237)	4.86	0.000	1.01

### Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
Non-Moving			
1	0	2.0869	(1.4593, 2.9843)
2	0	3.7775	(2.3015, 6.2000)
3	0	3.4627	(1.7478, 6.8601)
4	0	4.5448	(1.5378, 13.4311)
5	0	10.0418	(3.9609, 25.4581)

Odds ratio for level A relative to level B

### Regression Equation

$$P(S_i) = \exp(Y') / (1 + \exp(Y'))$$

$$Y' = 0.1442 + 0.0 \text{ Total Non Moving (Category)}_0 + 0.736 \text{ Total Non Moving (Category)}_1 + 1.329 \text{ Total Non Moving (Category)}_2 + 1.242 \text{ Total Non Moving (Category)}_3 + 1.514 \text{ Total Non Moving (Category)}_4 + 2.307 \text{ Total Non Moving (Category)}_5$$

### Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	946	1147.14	0.000
Pearson	946	952.00	0.439
Hosmer-Lemeshow	2	0.00	1.000