TransInfo
Tier I University Transportation Center

FINAL REPORT

Development of a Prediction Model for Crash Occurrence by Analyzing Traffic Crash and Citation Data

Date of report: April 2017

Enrique Gonzalez-Velez, Ph.D., Assistant Professor, University of Puerto Rico at Mayagüez
Armando Gonzalez-Bonilla, BSCE, Research Assistant, University of Puerto Rico at Mayagüez

Prepared by:
Department of Civil Engineering and Surveying
University of Puerto Rico at Mayagüez
PO Box 9000
Mayagüez PR 00681-9000

Prepared for:
Transportation Informatics Tier I University Transportation Center
204 Ketter Hall
University at Buffalo
Buffalo, NY 14260

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
		5. Report Date April 30, 2017
		6. Performing Organization Code
7. Author(s) Enrique Gonzalez-Velez, Ph.D. Armando Gonzalez-Bonilla, BSCE		**8. Performing Organization Report No.**
		10. Work Unit No. (TRAIS
		11. Contract or Grant No. DTRT13-G-UTC48
		13. Type of Report and Period Covered Final
		14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

It is commonly acknowledged that factors such as human factors, vehicle characteristics, road design and environmental factors highly contribute to the occurrence of traffic crashes (WHO, 2004). Since human factors usually have the most significant influence on traffic crash occurrence, studies normally focus on the effect that some driver characteristics have on the occurrence of a traffic crash, such as age, gender, alcohol usage and driving. 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 vehicle crash. This research project aims to estimate the likelihood of a driver being involved or not in a vehicle crash by performing stepwise multiple logistic regression analyses. The data used was obtained by performing a survey on a sample of the driving population of Puerto Rico. Information such as age, gender, years of driving experience, daily hours spent driving and traffic violation and crash history were determined for a sample of the driving population of Puerto Rico. Results indicate that years of driving experience, gender and traffic violations history are significantly associated with being involved in a vehicle crash.

17. Key Words	18. Distribution Statement
Crash likelihood estimation, logistic regression	No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	67	

DEVELOPMENT OF A PREDICTION MODEL FOR CRASH OCCURRENCE BY ANALYZING TRAFFIC CRASH AND CITATION DATA

Enrique Gonzalez-Velez, Ph.D.

Armando Gonzalez-Bonilla, BSCE

Transportation Informatics University Transportation Center

University of Puerto Rico at Mayaguez

April 2017

Acknowledgements

The authors 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 Dr. Ivette Cruzado-Velez for their collaboration in collecting the data for this project.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. PROBLEM

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. 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 good results.

It is commonly acknowledged that factors such as human factors, vehicle characteristics, road design and environmental factors highly contribute to the occurrence of traffic crashes (WHO, 2004). Since human factors usually have the most significant influence on traffic crash occurrence, studies normally focus on the effect that some driver characteristics have on the occurrence of a traffic crash, such as age, gender, alcohol usage and driving. 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 traffic 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 traffic crash based on a series of human factors such as age and gender 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.

1.1 Goals and Objectives

The main goal of this research project is to develop a statistical model that could be used to estimate the likelihood of being involved in a traffic crash based on a series of human factors such as age and gender as well as traffic violations and crash history. The research approach includes the following objectives:

- 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.
- Assess the developed model using appropriate statistical tests and procedures.

2. APPROACH AND METHODOLOGY

This section describes the methodology that took place in order to analyze the data and develop the estimation models. The approach taken is outlined as follows:

- Literature review
- Description of data
- Preliminary analyses
- Model development
- Model assessment
- Conclusions and recommendations

2.1 Literature Review

The literature review performed in this study sought to identify and understand which factors regarding human characteristics and behavior are most commonly associated with future vehicle crash involvement. 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. Most of the studies included in the literature review used driver record databases for collecting the data and information of their unit of analysis, mainly drivers and vehicles. The databases used in most these studies were created from the driver records that law enforcement officials obtain from traffic crashes and violations. Tables 1, 2 and 3 summarize the variables that were found to be significant on the studies included in this literature review. Results show that gender, age and prior traffic citations and crashes as well as the type of citation and crash are significant factors for estimating of future traffic crashes among many of the studies reviewed. Other significant factors that were included in some of these studies are driving behavior and type of license. Additionally, the most common methodology for estimating the likelihood of future crashes based on a series of factors is the use of multiple logistic regression analyses.

Table 1: Significant Variables Found in Literature Review

Study	Variables Found to be Significant
"Strategies for Estimating Driver Accident Risk in Relation to California's Negligent-Operator Point System" (Gebers, 1999)	Previous Total Crashes Age Gender Being young Being Male Holding a Commercial Driver's License Increased prior citation and crash frequency
"Using Traffic Conviction Correlates to Identify High Accident-Risk Drivers" (Gebers & 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

Table 2: Significant Variables Found in Literature Review (Continued)

Study	Variables Found to be Significant
"Previous Convictions or Accidents and the Risk of Subsequent Accidents for Older Drivers" (Daigneault, et al., 2002)	Previous crashes
"Relationships Between Prior Driving Record, Driver Culpability, and Fatal Crash Involvement" (Wundersitz et al., 2004)	Drivers younger than 25 years of age Drivers older than 75 years Driving under the influence of alcohol 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 Gender Crash Type
"Evaluation of the Characteristics of Drivers with Multiple Crashes" (Chandraratna & Stamatiadis, 2004)	
"Predicting Truck Crash Involvement: Developing a Commercial Driver Behavior Model and Requisite Enforcement Countermeasures" (Murray, 2006)	Reckless Driving Speeding Violations Past Crash Experience
"Road Traffic Accident Involvement Rate by Accident and Violation Records: New Methodology for Driver Education based on Integrated Road Traffic Accident Database" Nishida, 2009	Traffic Crashes Traffic Violations Driving Behavior Frequency of Driving
"Risk Factors Associated with Traffic Violations and Accident Severity in China" Zhang,2013	Traffic Violations Males Unfit safety status Overload in a vehicle No Street Lightning at Night Bad Visibility Weekends
"Strategic Highway Safety Plan of Puerto Rico" SHSP, 2014	Age Gender Traffic faults

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

Study	Variables Found to be Significant
"Estimating likelihood of future crashes for crash-prone drivers" Subasish D., et al, 2015	Driver culpability Alcohol Road alignment Road lightning Crash Severity Crash type Gender Age Driver distraction Drugs
"Risk Factors Analysis for Drivers with Multiple Crashes" Shawky & 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

2.2 Description of Data

The data used for this research project consisted of information regarding demographics as well as history of traffic violations and crashes for a sample of the population of licensed drivers in Puerto Rico. 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, for this project, 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 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.

Survey Development and Composition

The survey was distributed using online based platforms such as Facebook and email as well as a paper based platform. 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. A paper based form was also 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 based form 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 as well as information to subjects of the responsibilities and conditions of participating. During the literature review process, several variables such as age, gender, 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,
- Gender as indicated on the driver license,
- Years of experience driving a private motor vehicle, and
- Daily hours spent driving a motor vehicle.

The variable "Age" was categorized in different intervals that range from 16 to 89 years of age while the variable "Gender" was also categorized in two levels: Males and Females. The reason for categorizing the answers to the questions provided in the survey was ease of development and completion of the survey.

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
- 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.

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. First off, 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 4 and 5 provide the categorical variables that were considered for this research. The following list provides the continuous variables initially selected.

- Years of Driving Experience
- Daily Hours Driven

- Total Traffic Crashes
- PDO Crashes
- Minor Injury Crashes
- Severe Injury Crashes

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 overview of the data obtained from the sample. The categories of the variable (for categorical variables), response count, percent and the mean were provided for each variable. 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. The statistical software Minitab was used to determine the descriptive statistics of the data used in this research project as well as all other subsequent analyses.

Table 4: Categorical Variables

Variable	Categories
	16-20
	21-30
	31-40
	41-50
	51-60
	61-70
	71-80
	81-89
	Male
	Female
	1
	2
	3
	4
	5 or more
	1
	2
	3
	4
	5 or more
	1
	2
	3
	4
	5 or more
	1
	2
	3
	4
	5 or more

Table 5: Categorical Variables

Variable	Categories
	1
	2
	3
	4
	5 or more
	1
	2
	3
	4
	5 or more
	1
	2
	3
	4
	5 or more
	1
	2
	3
	4
	5 or more
	1
	2
	3
	4
	5 or more
	1
	2
	3
	4
	5 or more

2.3 Preliminary Analyses

Preliminary analyses were performed to determine the association between each of the independent variable/predictors identified in the database and the dependent variable of crash involvement. The purpose of performing these analyses before starting development of the estimation model was to understand how each of the different independent variables that were identified affect whether a participant was involved in a traffic crash or not. For categorical variables, it is suggested that an analysis of contingency tables should be performed between the response variable and its two outcomes versus the independent variables and its different levels (Hosmer & Lemeshow, 2000). Contingency tables were developed to compare the categorical variables identified previously with the vehicle crash involvement of participants. Contingency tables are a mean of displaying the frequencies or proportions between the categories of two categorical variables. Table 6 shows an example of a 2 x 2 contingency table. The rows of the table correspond to the categories of one variable, 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 6: Contingency Table Example

Variable X	**Variable Y**		**Row Totals**
	J_1	**J_2**	
I_1	I_1J_1	I_1J_2	Marginal Totals = I_1J_{1+} I_1J_2
I_2	I_2J_1	I_2J_2	Marginal Totals = I_2J_{1+} I_2J_2
Column Totals	**Marginal Totals = I_1J_{1+} I_2J_1**	**Marginal Totals = I_1J_{2+} I_2J_2**	**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.

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 value of a cell should be 5 or more on 80% of the cells, and no cell should have a value of less than 3.

The statistical software *Minitab* was used to perform this test. 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 allowed to study the association between the categories of the independent and dependent variable respectively. The standardized adjusted residuals values 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). The main result that was considered for analyzes purposes was the probability values associated with the Pearson and likelihood ratio chi-square statistics. The following statistical hypotheses were considered:

- 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), a probability value lower than 0.05 indicates that 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
- Spearman's Rho

Cramer's 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 of driving experience 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.

Simple Logistic Regression

Results of the literature review indicated that one of the most common approaches for estimating a driver's likelihood of future crashes based on a series of different factors is the use of logistic regression analyses. Some of the findings of the literature review indicate that factors such as age, gender 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
- 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. In addition to the probability values, an analysis of the coefficients of the predictor variable was also performed. The coefficients describe 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 odd 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 odd 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. Odd ratios for continuous predictors were 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 "*". 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 odd 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 odd ratios range from $-\infty$ to ∞ and can be interpreted as follows:

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

2.4 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 vehicle crash based on a series of independent variables. 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. To compare the models that resulted whenever a variable was included or excluded, the Akaike Information Criterion (AIC) was used. The 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). Moreover, this value 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.

2.5 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 & Lemeshow, 2000). For this study, the goodness of fit test used to assess the selected model was the Hosmer-Lemeshow measure. If the Hosmer-Lemeshow probability value presented in the results is larger than the alpha value for a 95% confidence interval ($\alpha = 0.05$), there is not enough evidence to say that it does not provide a good fit.

Another assessment 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-Lemeshow, 1999). For this study, the ROC curve indicated if the model classified participants who were involved or not in a vehicle crash correctly. This was 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 classified the predicted probability as one (being involved in a crash). On the other hand, if the estimated probability was less than 0.5, the model classified the subject's predicted probability as zero (as not being involved in a traffic crash). 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 ranges from 0.5 to 1 and can be interpreted as follows:

- ROC = 0.5: No predictive ability
- $0.7 \leq \text{ROC} \leq 0.8$: Acceptable predictive ability
- $0.8 \leq \text{ROC} \leq 0.9$: Excellent predictive ability
- $\text{ROC} \geq 0.9$: Outstanding predictive ability

3. FINDINGS

Results of the data gathered using the survey as well as the descriptive statistics performed to describe them are presented in this section. Following the results of the survey, 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.

3.1 Documentation of Data Gathered

Results of for the survey responses as well as the descriptive statistics of the variables identified are presented in this section. Results for the variable of age are provided and illustrated in Table 7 and Figure 1. 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 8 and Figure 2 provide information regarding the gender distribution. Results indicate that the percentage of responses obtained from females is larger than males.

Table 7: Descriptive Statistics for Age

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

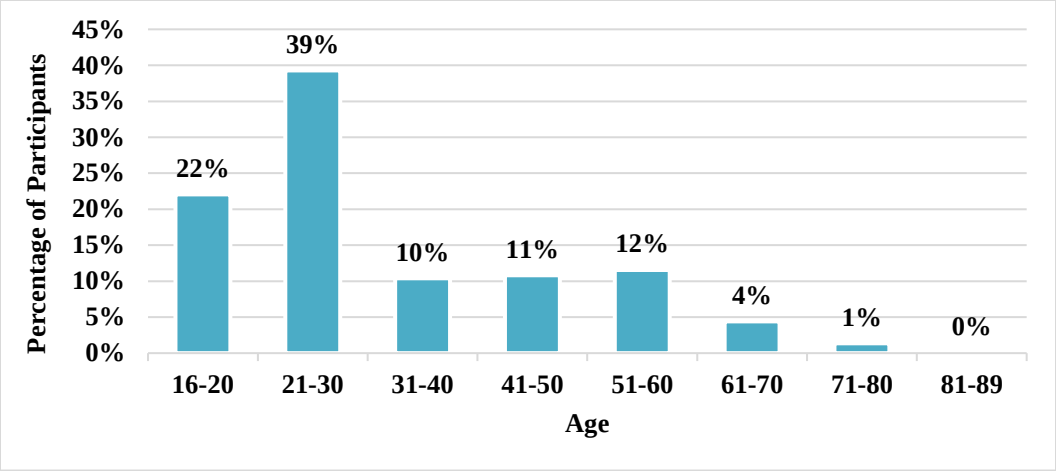


Figure 1: Sample Distribution of Drivers Based on their Age

Table 8: Descriptive Statistics for Gender

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

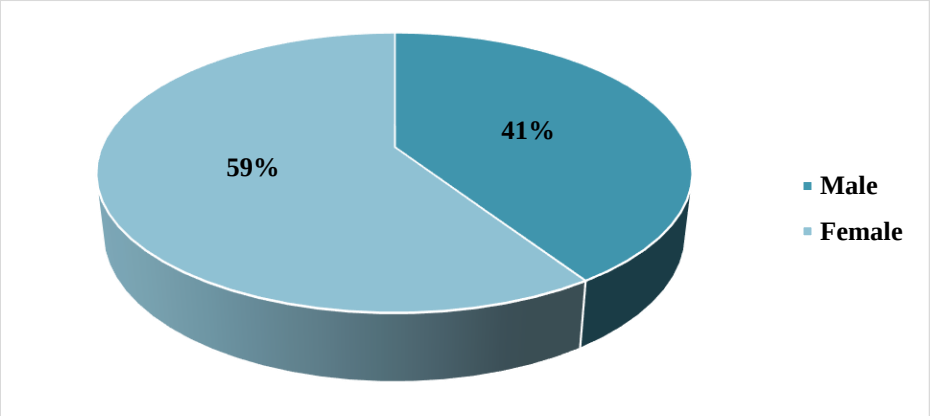


Figure 2: Sample distribution of Drivers Based on Gender

Information regarding the distribution of age in females is illustrated in Figure 3 while information regarding the distribution of age in males is illustrated in Figure 4. 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 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.

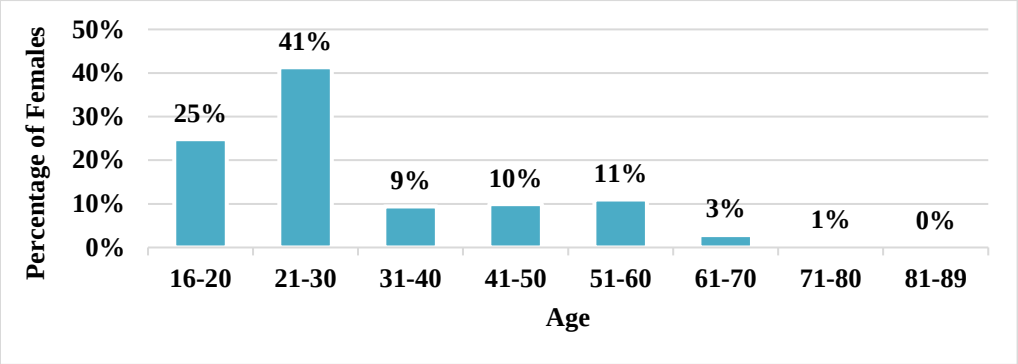


Figure 3: Sample distribution of Females Based on their Age

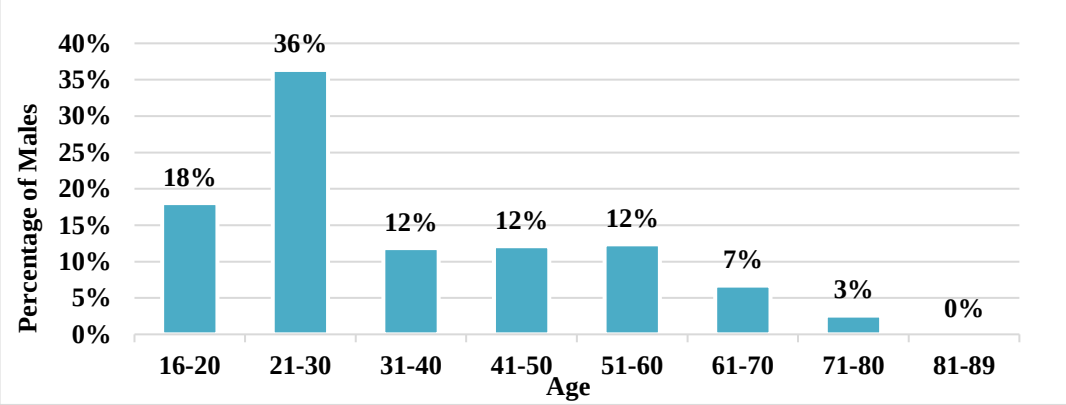


Figure 4: Sample Distribution of Males Based on their Age

Information regarding the years of driving experience from the participants in the sample is presented in Table 9 and Figure 5. 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 9: Descriptive Statistics for Years of Driving Experience

Variable	Total Count	Percent	Mean
Years of Experience	952	99.5	15.2

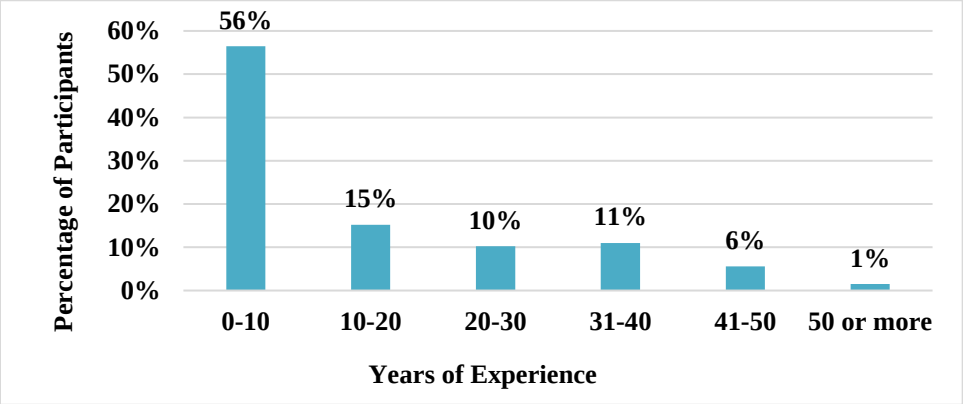


Figure 5: Sample Distribution Based on Years of Driving Experience

Another factor that was included in the survey was the number of hours spent driving in a day. Like years of driving experience, 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 10 for daily hours spent driving indicate that most of the participants spend from zero to two hours a day driving.

Table 10: Descriptive Statistics for Daily Hours Spent Driving

Variable	Total Count	Percent	Mean
Daily Hours	952	94.5	2.5

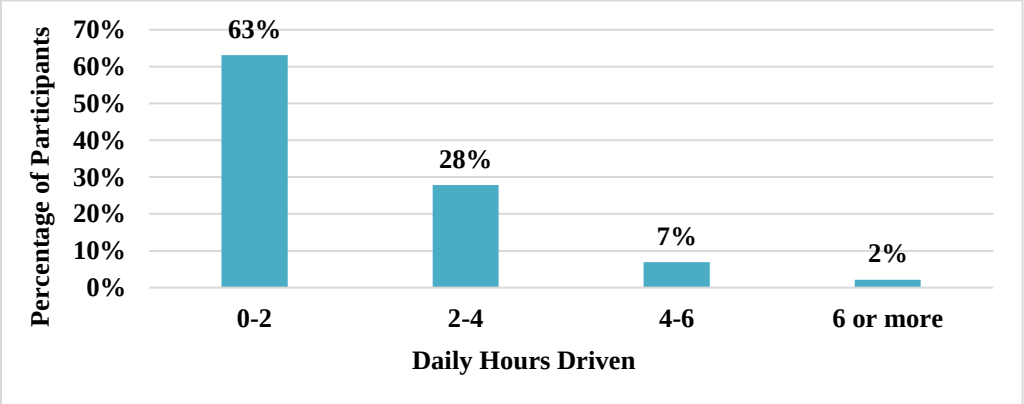


Figure 6: Sample Distribution Based on Drivers Daily Hours Spent Driving

Table 11 and Figure 7 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 11: 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
Total	952	100%	-

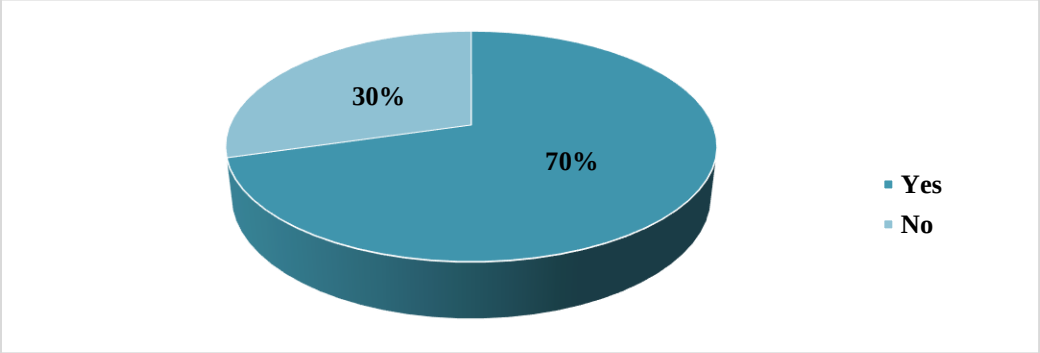


Figure 7: Distribution of Drivers Based on whether they Received Traffic Violations or Not

Descriptive statistics for the different type of traffic violations are presented throughout tables 12 to 21. 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 8 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 12: Descriptive Statistics for Driving Over the Speed Limit

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

Table 13: Descriptive Statistics for Driving Under the Influence of Alcohol or Drugs

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

Table 14: Descriptive Statistics for Ignoring Traffic Signals and Signs

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

Table 15: Descriptive Statistics for Driving too Close to Front Vehicle

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

Table 16: Descriptive Statistics for Illegal Turning

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

Table 17: Descriptive Statistics for Illegal Parking

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

Table 18: Descriptive Statistics for Illegal Lane Switch

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

Table 19: Descriptive Statistics for Not Using Seatbelt While Driving

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

Table 20: Descriptive Statistics for Using Cellphone While Driving

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

Table 21: Descriptive Statistics for Other Traffic violations

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

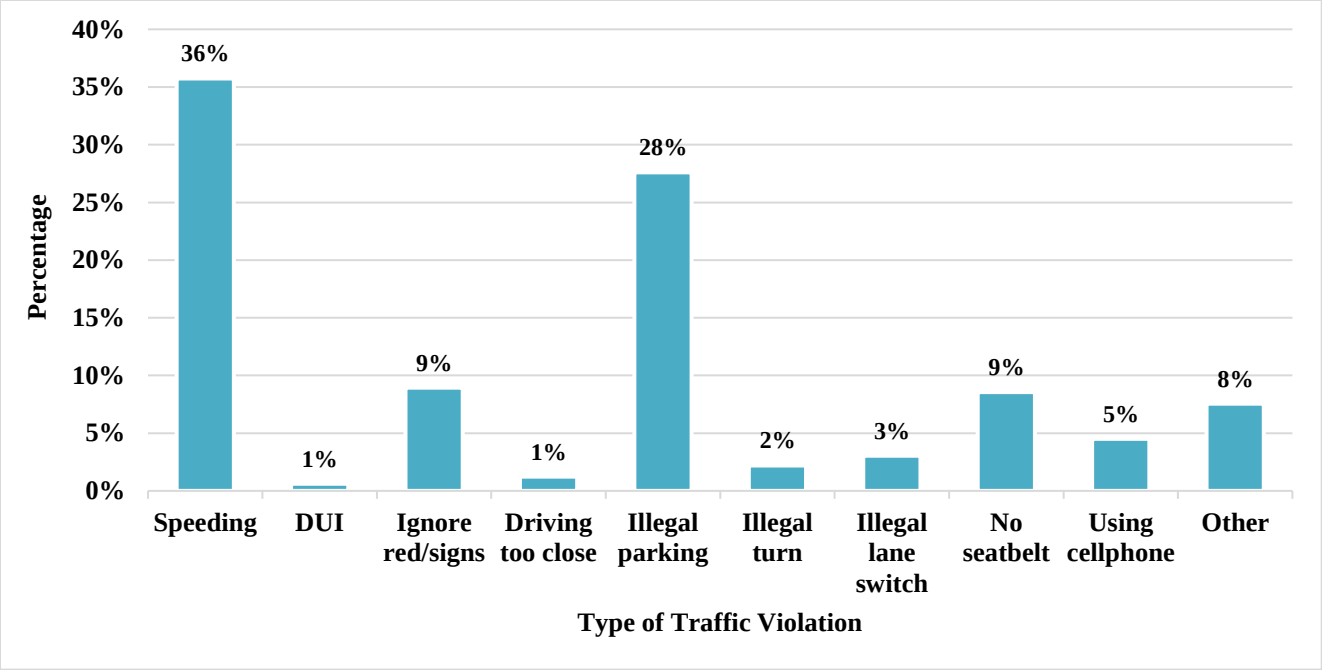


Figure 8: Sample Distribution of Traffic Violations

Table 22 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 30 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 22: Distribution of Participants Based on Crash Involvement

Crash Involvement as a Driver	Response Percent	Response Count
Yes	65.3%	622
No	34.7%	330
Total	**100%**	**952**

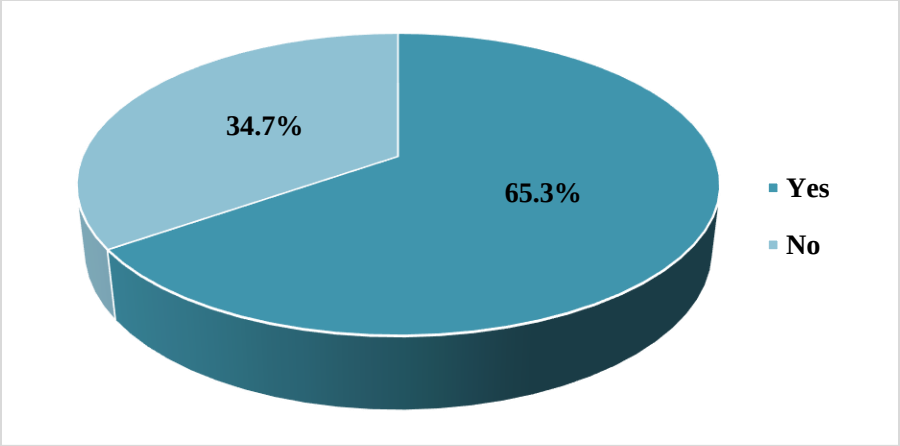


Figure 9: Sample Distribution of Crash Involvement among Participants

Table 23: Sample Distribution of Crashes Reported Based on Severity

Severity	Response Percent	Response Count
PDO	88%	1010
Minor Injury	9%	107
Severe Injury	2%	28
Total	**100%**	**1145**

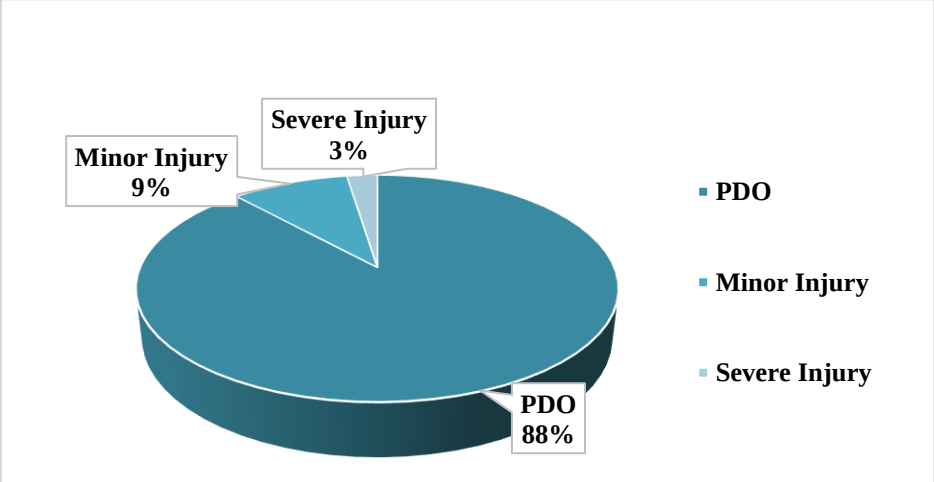


Figure 10: Sample Distribution of Crash Severity among Crashes Reported

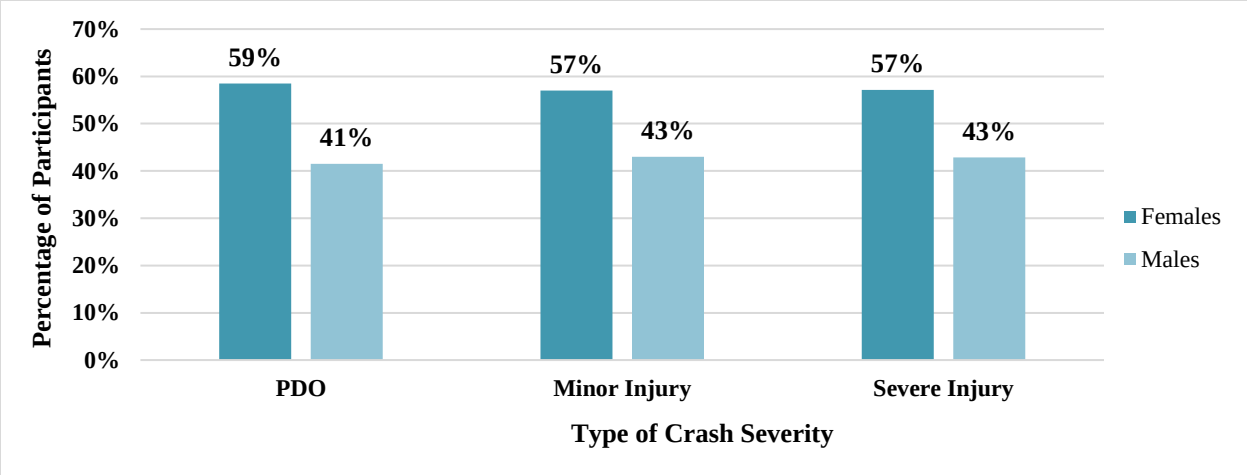
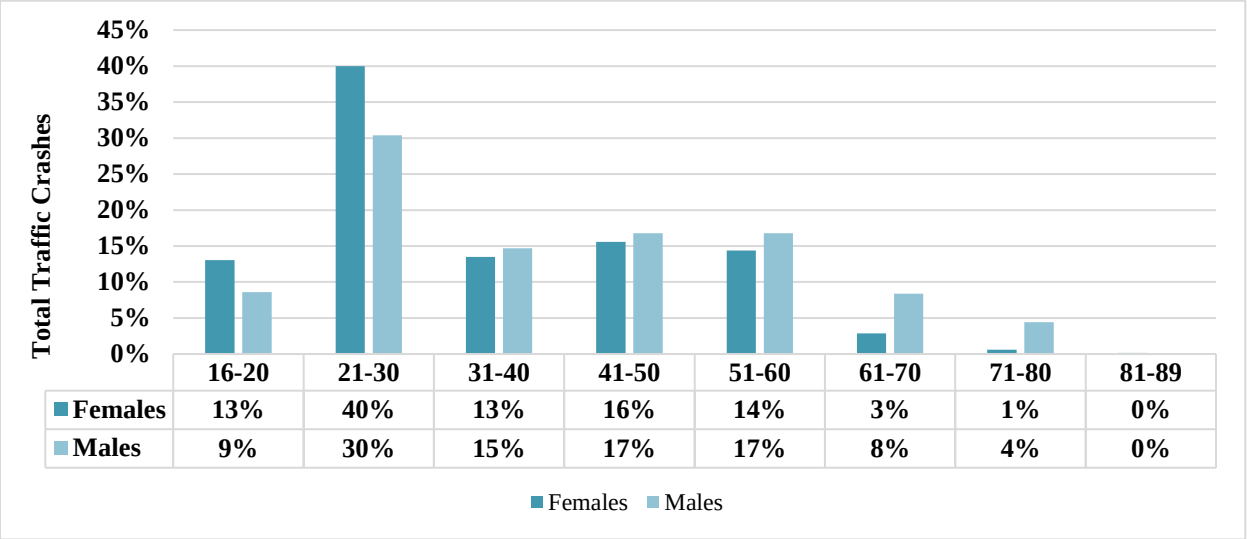


Figure 11: Distribution of Crash Severity Based on Gender of Participants



	16-20	21-30	31-40	41-50	51-60	61-70	71-80	81-89
Females	13%	40%	13%	16%	14%	3%	1%	0%
Males	9%	30%	15%	17%	17%	8%	4%	0%

Figure 12: Distribution of Total Crashes Based on Age and Gender of Participants

3.2 Analysis and Results

Before development of the logistic regression model started, preliminary analyses were performed to have an understanding of the significance of the association between each of the independent variables being considered and the dependent variable of vehicle crash involvement. As mentioned in chapter 2, when studying the association between two categorical variables, chi-square tests of independence were performed. However, in order to perform this test, the data had to be rearranged into contingency tables.

Contingency Tables

The resulting contingency tables developed to be used with the chi-square test of independence are provided below. When inspecting these contingency tables, it can be seen that the table for gender complies with the following assumptions 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 have 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 24: Contingency Table Age vs Crash Involvement

Age Range	Crash Involvement		Total
	Yes	No	
16-20	90	120	210
21-30	245	129	374
31-40	77	22	99
41-50	82	21	103
51-60	86	24	110
61-70	31	11	42
71-80	10	3	13
81-89	1	0	1
Total	**622**	**330**	**952**

Table 25: Contingency Table Gender vs Crash Involvement

Gender	Crash Involvement		Total
	Yes	No	
Females	363	201	564
Males	259	129	388
Totals	**622**	**330**	**952**

Table 26: Contingency Table for Driving over the Speed Limit Violations vs Crash Involvement

Driving Over the Speed Limit	Crash Involvement		Total
	Yes	No	
0	273	231	504
1	153	58	211
2	93	30	123
3	37	5	42
4	15	1	16
5 or more	51	5	56
Total	**622**	**330**	**952**

Table 27: Contingency Table for DUI Violations vs Crash Involvement

Driving Under the Influence Violations	Crash Involvement		Total
	Yes	No	
0	612	330	942
1	7	0	7
2	2	0	2
3	0	0	0
4	0	0	0
5 or more	1	0	1
Total	**622**	**330**	**952**

Table 28: Contingency Table for Ignoring Traffic Signs and/or Signals Violations vs Crash Involvement

Ignoring Traffic Signal/Signs Violations	Crash Involvement		Total
	Yes	No	
0	483	288	771
1	116	36	152
2	10	6	16
3	7	0	7
4	3	0	3
5 or more	3	0	3
Total	**622**	**330**	**952**

Table 29: Contingency Table for Driving Too Close to Front Vehicle Violations vs Crash Involvement

Driving Too Close to Front Vehicle	Crash Involvement		Total
	Yes	No	
0	603	327	930
1	15	3	18
2	2	0	2
3	0	0	0
4	0	0	0
5 or more	2	0	2
Total	**622**	**330**	**952**

Table 30: Contingency Table for Illegal Parking Violations vs Crash Involvement

Minimum Number of Illegal Parking Violations	Crash Involvement		Total
	Yes	No	
0	326	259	585
1	145	49	194
2	73	10	83
3	33	8	41
4	12	0	12
5+	33	4	37
Total	**622**	**330**	**952**

Table 31: Contingency Table for Illegal Turn Violations vs Crash Involvement

Illegal Turn Violations	Crash Involvement		Total
	Yes	No	
0	586	321	907
1	31	9	40
2	1	0	1
3	2	0	2
4	1	0	1
5 or more	1	0	1
Total	**622**	**330**	**952**

Table 32: Contingency Table for Reckless Lane Switch Violations vs Crash Involvement

Reckless Lane Switch Violations	Crash Involvement		Total
	Yes	No	
0	573	321	894
1	40	6	46
2	5	3	8
3	1	0	1
4	1	0	1
5 or more	2	0	2
Total	**622**	**330**	**952**

Table 33: Contingency Table for No Seatbelt Used Violations vs Crash Involvement

No Seatbelt Used Violations	Crash Involvement		Total
	Yes	No	
0	499	306	805
1	84	20	104
2	26	4	30
3	3	0	3
4	1	0	1
5 or more	9	0	9
Total	**622**	**330**	**952**

Table 34: Contingency Table for Cellphone Use Violations vs Crash Involvement

Minimum Number of Cellphone Use Violations	Crash Involvement		Total
	Yes	No	
0	553	304	857
1	61	23	84
2	4	2	6
3	2	0	2
4	0	0	0
5 or more	2	1	3
Total	**622**	**330**	**952**

Table 35: Contingency Table for Other Violations vs Crash Involvement

Other Violations	Crash Involvement		Total
	Yes	No	
0	553	310	863
1	42	11	53
2	15	4	19
3	4	4	8
4	2	1	3
5 or more	6	0	6
Total	**622**	**330**	**952**

Recall the assumption for the use of chi-square tests of independence which stated that no cell should have a frequency of three or less; it can be seen that there are cells in the contingency tables of age and each of the different traffic violations that contained cells with frequencies lower than three, similar to the category of age for participants between 81 and 89 years of age. Because of this, it was not appropriate to perform the chi-square test of independence for these independent variables. If these tests were to be performed ignoring the assumption that no cell should have a value of less than three, the approximation of the chi-square statistic could be invalid and thus results would be unreliable. To solve this issue, a merge of categories and variables was performed.

For the variable of age, the categories of 61-70, 71-80 and 81-89 years of age were merged into a single category. Similarly, the traffic violations included in this study were classified into one of two categories, moving violations and non-moving 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). Traffic violations that involve a driver on a moving vehicle, such as driving over the speed limit or unsafe lane switch are considered moving violations. Non-moving violations correspond to traffic violations that were carried out when the vehicle was not moving, such as illegal parking. In the case of the traffic violation for not wearing a seatbelt, it was

considered as a non-moving violation for the purpose of this study. Table 36 shows the traffic violations obtained from the survey classified by moving or non-moving violations.

Table 36: Classification of Traffic Violations

Moving Violations	**Non-Moving Violations**
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 moving or non-moving violations, new contingency tables were created for these classifications. The contingency tables for age, moving violations and non-moving violations are shown below in tables 37, 38 and 39 respectively.

Table 37: Contingency Tabl7e for Age vs Crash Involvement (After Merging Categories)

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

Table 38: Contingency Table for Moving Violations vs Crash Involvement

Moving Violations	No	Yes	Total
0	195	199	394
1	69	151	220
2	37	105	142
3	14	60	74
4	8	36	44
5 or more	7	71	78
Total	**330**	**622**	**952**

Table 39: Contingency Table for non-Moving Violations vs Crash Involvement

Non-Moving Violations	No	Yes	Total
0	232	268	500
1	56	135	191
2	22	96	118
3	11	43	54
4	4	22	26
5 or more	5	58	63
Total	**330**	**622**	**952**

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 were no cells with frequencies less than three. Additionally, the contingency tables for moving and non-moving violations comply with all other assumptions. Once all the data was correctly arranged into contingency tables, chi-square tests of independence were performed.

Chi-Square Test of Independence

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 40 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 40, it can be seen that the hypothesis of independence has a good fit for drivers between the ages of 61 and 89 years of age since the adjusted residuals for this category lie outside the confidence interval of 95% stated for this study, making this category independent from the dependent variable of being involved in a traffic crash. Recalling the discussion of the methodology for the chi-square tests of independence, 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. Similarly, residuals for drivers between the ages of 21 and 30 years of age and being involved in a traffic crash also seem to indicate a good fit for independence. 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 and 61-89, the observed and expected values are almost equal, meaning that being involved in a traffic crash is independent of the fact that a participant is between the ages of 21-30 or 61-89. Table 52 provides results regarding goodness of fit measures obtained for the variable of age. The Cramer's V-square statistic yielded a value of 0.076908, 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. This hypothesis can also be inferred from the results for the Pearson R and Likelihood Ration chi-square statistics, which yielded values of 0.223915 and 0.257155 respectively, which are close to zero and indicate a weak 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
- Chi-Square statistic for the respective cell

Table 40: Results of Chi-Square Test of Independence for Age

Age	Traffic Crash Involvement		Row Marginal totals
	Yes	**No**	
	90	120	**210**
	137.210	72.790	
	-7.750	7.750	
	16.240	30.610	
	245	129	**374**
	244.36	129.64	
	0.09	-0.09	
	0.002	0.003	
	77	22	**99**
	64.680	34.320	
	2.748	-2.748	
	2.346	4.421	
	82	21	**103**
	67.300	35.700	
	3.224	-3.224	
	3.213	6.055	
	86	24	**110**
	71.870	38.130	
	3.010	-3.010	
	2.778	5.236	
	42	14	**56**
	36.590	19.410	
	1.566	-1.566	
	0.800	1.509	
Column Marginal Totals	622	330	**952**
Pearson Chi-Square = 73.217 Degrees of Freedom = 5 P-Value < 0.001	Likelihood Ratio Chi-Square = 72.556 Degrees of Freedom = 5 P-Value < 0.001		Cramer's V-square = 0.0769 Pearson's R = 0.223915 Spearman's Rho = 0.257155

A similar analysis was conducted for the variables of gender, moving violations and non-moving violations. Table 41 displays the results for males and females with respect to being involved in a traffic crash. The results provided for the variable of gender indicate that there is not a significant association between gender 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
- Chi-Square statistic for the respective cell

Table 41: Results of Chi-Square Test of Independence for Gender

Gender	**Traffic Crash Involvement**		**Row Marginal Totals**
	Yes	No	
	363	201	**563**
	368.5	195.5	
	-0.762	0.762	
	0.082	0.154	
	259	129	**388**
	253.5	134.5	
	0.762	-0.762	
	0.119	0.225	
Column Marginal Totals	**330**	**622**	**952**
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, moving violations resulted to have a significant association with being involved in a traffic crash as shown in Table 42. The probability values obtained from the Pearson and likelihood ratio chi-squares statistics for number of moving 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 moving 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 moving 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
- Chi-Square statistic for the respective cell

Table 42: Results of Chi-Square Test of Independence for Moving Violations

Moving Violations	**Traffic Crash Involvement**		**Row Marginal Totals**
	Yes	**No**	
	199	195	394
	257.42	136.58	
	-8.079	8.08	
	13.26	24.99	
	151	69	220
	143.74	76.26	
	1.173	-1.173	
	0.367	0.691	
	105	37	142
	92.78	49.220	
	2.337	-2.337	
	1.610	3.035	
	60	14	74
	48.35	25.650	
	2.964	-2.964	
	2.808	5.292	
	36.00	8	44
	28.75	15.250	
	2.352	-2.352	
	1.829	3.448	
	71	7	78
	50.96	27.040	
	4.976	-4.976	
	7.879	14.850	
Column Marginal Totals	**622**	**330**	**952**
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 = 0.0841 Pearson's R = 0.274884 Spearman's Rho = 0.287869

Similar to moving violations, non-moving violations were also found to be associated with being involved in a traffic crash. Table 43 provides the results for non-moving 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-moving 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 moving 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 moving 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, moving violations and non-moving 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 was 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
- Chi-Square statistic for the respective cell

Table 43: Results of Chi-Square Test of Independence for Non-Moving Violations

Non-Moving Violations	Traffic Crash Involvement		Row Marginal Totals
	Yes	No	
	268	232	500
	326.68	173.32	
	-8.003	8.00	
	10.541	19.87	
	135	56	191
	124.79	66.21	
	1.736	-1.736	
	0.835	1.574	
	96	22	118
	77.1	40.900	
	3.907	-3.907	
	4.635	8.736	
	43	11	54
	35.28	16.720	
	2.272	-2.272	
	1.689	3.183	
	22.00	4	26
	16.99	9.010	
	2.094	-2.094	
	1.479	2.788	
	58	5	63
	41.16	21.840	
	4.613	-4.613	
	6.888	12.983	
Column Marginal Totals	**622**	**330**	**952**
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

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 44 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 no 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 of driving experience 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 odd ratios for years of driving experience and daily hours spent driving resulted with value of 1.04 and 0.97 respectively, indicating that there is a positive correlation between years of driving experience and being involved in a traffic crash but not for daily hours spent driving since the odd ratio obtained was lower than one. The odd ratios for the years of driving experience 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 odd 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 odd ratios for the continuous predictors included, it can be said that although a positive and negative correlation were associated with years of driving experience and daily hours spent driving respectively, it is not a relatively significant one since the odd ratios obtained for these predictors were very close to 1. It is important to remember that an odd ratio of one indicates that the predictor is not correlated with the response variable.

Table 44: Results of Simple Logistic Regression Analysis

Predictor	Coefficient	SE	Z-Value	P-value	Odds Ratio
Years of Driving Experience	0.04289	0.006	7.15	<0.001	1.04
Daily Hours Spent Driving	-0.0332	0.0385	-0.86	0.389	0.97
Age					
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)	1.564	0.27	5.8	<0.001	4.78
Age (61-89)	1.386	0.339	4.09	<0.001	4.00
Gender					
Female	0	0	*	*	Reference
Male	0.106	0.139	0.76	0.446	1.11
Moving Violations					
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
Non-Moving Violations					
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 44. 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 traffic 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 odd ratios of the various categories for the predictor of age indicate that there is a positive correlation with the response variable. Inspecting the odd ratios column for the age predictor, it can be seen that the odd ratio values range from 2.53 for participants between the ages of 21 and 30 to 4 for driver between the ages of 61-89, showing a steady increase in the

odd ratio values throughout the categories. Thus, results of the odd ratio analysis indicate that as the age of participants increases, the odds of being involved in a traffic crash also increase. 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 61-89 are four times more likely to be involved in a traffic crash than drivers between de ages of 16-20.

The results for the predictor of gender 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 gender resulted to be a non-significant independent variable, analyzing the odd 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 gender and being involved in a traffic crash, it can be inferred it is not a significant one since the odd 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 moving violations, the probability values obtained for each of the categories were lower than 0.001, which indicates that number of moving violations received is a significant predictor of being involved in a traffic crash. Similarly, non-moving 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 odd ratios for moving and non-moving 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 odd ratios. For participants who received two moving violations, the odds of being involved in a traffic crash are 2.16 times more for drivers who received two moving violations than for drivers who did not receive any traffic violations. Meanwhile, participants who received 5 or more moving 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 of driving experience, age, moving violations and non-moving violations are significant predictors of being involved in a traffic crash while gender and daily hours spent driving were not significant. Past studies have shown that previous traffic violations, age, gender 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.

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 & 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 45 displays the results of the backwards elimination procedure with the respective iterative steps.

Table 45: Results for Stepwise Backwards Elimination Procedure

Term	Step 1		Step 2		Step 3	
	Coefficient	P-Value	Coefficient	P-Value	Coefficient	P-Value
Constant	-0.499		-0.625		-0.416	
Years of Experience	0.055	0.001	0.056	0.001	0.034	0.000
Daily Hours Driving	-0.049	0.259				
Age	-1.132	0.127	-1.179	0.11		
Gender	-0.323	0.046	-0.313	0.052	-0.315	0.049
Moving Violations	1.334	0.007	1.322	0.008	1.46	0.001
Non-Moving Violations	1.576	0.000	1.558	0.000	1.682	0.000
AIC	1065.53		1064.82		1063.93	

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 of driving experience, daily hours spent driving, age, gender, moving violations and non-moving violations. An inspection of the coefficients obtained for the model in first step indicated that the predictors for daily hours spent driving, age and gender 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. The only probability value that changed significantly was the probability value for the predictor of age which increased from a value of 0.046 in the first step to a value of 0.052 in the second step, making it a non-significant predictor. However, it is worth noticing that although the value for this value in the first step was below the stated alpha value of 0.05, a value of 0.046 was still very close to this alpha value.

To proceed to the next step, the predictor of age which yielded the highest probability value (0.11) among the predictors included in the model was removed. The fitted model in the third step contains the following variables; years of experience, gender, moving violations and non-moving 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 show 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 46 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 ∞, with values close to 1 indicating that the predictor has no multicollinearity with other predictors.

Table 46: Results of Final Logistic Regression Model

Term	Coefficient	Standard Error of Coefficient	Z-Value	P-Value	VIF
Constant	-0.416	0.137	-3.03	0.002	
Years of Experience	0.034	0.006	5.42	0.000	1.08
Gender					
Female	0.000	0.000	*	*	*
Male	-0.315	0.160	-1.97	0.049	1.07
Moving Violations					
0	0.000	0.000	*	*	*
1	0.501	0.193	2.59	0.009	1.20
2	0.552	0.236	2.34	0.019	1.19
3	0.890	0.337	2.64	0.008	1.12
4	0.820	0.428	1.92	0.055	1.10
5 or more	1.460	0.442	3.31	0.001	1.14
Non-Moving Violations					
0	0.000	0.000	*	*	*
1	0.429	0.197	2.18	0.030	1.12
2	1.027	0.265	3.87	0.000	1.07
3	0.904	0.365	2.48	0.013	1.05
4	0.845	0.580	1.46	0.145	1.05
5 or more	1.682	0.500	3.37	0.001	1.08

$$Y = -0.416 + 0.033\,Years\ of\ Experience - 0.315\,Males + 0.501\,Moving\ Violations_1 + 0.552\,Moving\ Violations_2 + 0.890\,Moving\ Violations_3 + 0.820\,Moving\ Violations_4 + 1.460\,Moving\ Violations_{5\ or\ more} + 0.429\,Non-Moving\ Violations_1 + 1.027\,Non-Moving\ Violations_2 + 0.904\,Non-Moving\ Violations_3 + 0.845\,Non-Moving\ Violations_4 + 1.682\,Non-Moving\ Violations_{5\ or\ more}$$

Equation 1: Model Equation

The continuous predictor for years of driving experience 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 47 displays the results of the calculated odd ratios for continuous and categorical predictors. According to these results, the predictor for years of driving experience yielded an odd ratio value of 1.035 indicating that the odds of being involved in a traffic crash increase by 1.035 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 odd 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 of driving experience 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 gender. The gender variable has two categories, males and females with females being the reference group. The probability value obtained for gender was 0.049, 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 gender was 1.07 which indicates that there is little multicollinearity with other predictors. Results of the odd ratios obtained for gender show that the odds of male participants being involved in a traffic crash are 0.73 times more than female drivers, indicating that males have decreased odds of being involved in a traffic crash than females since the odd ratio obtained was less than one. This can be attributed to the fact the females represent a larger part of the sample population than men. The 95% confidence interval for the odd ratios of gender show that the odd 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 moving and non-moving violations, results show that as the number of violations received increase, the magnitude of the coefficients also increase, indicating that there is a positive correlation between number of traffic violations received and

being involved in a traffic crash. The 95% confidence interval for the moving and non-moving violations coefficients show that only the category of "four moving violations received" can achieve a value of less than zero, which would change the correlation from a positive one to a negative one. The probability values obtained for moving and non-moving violations indicate that it is a significant predictor of the response variables since all the values obtained are less than 0.05, with the exception of the category of four moving and non-moving violations received. 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 predictors. The odd ratios for moving and non-moving 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 odd ratios for moving and non-moving violations increase as the number of violations received 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 13 displays the odd ratios for moving violations while figure 14 displays the same behavior for non-moving violations.

Table 47: Calculated Odd Ratios for Terms in final Model

Term		**Odd Ratio**	**95% CI**
Years of Experience		1.035	(1.022, 1.047)
Gender			
Level A	**Level B**		
Male	Female	0.730	(0.534, 0.998)
Moving Violations			
Level A	**Level B**		
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-Moving Violations			
Level A	**Level B**		
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)

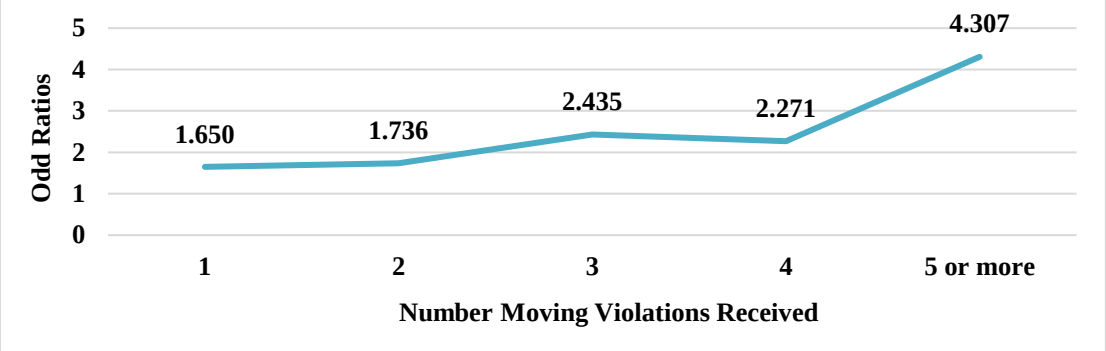


Figure 13:Odd Ratios for Moving Violations

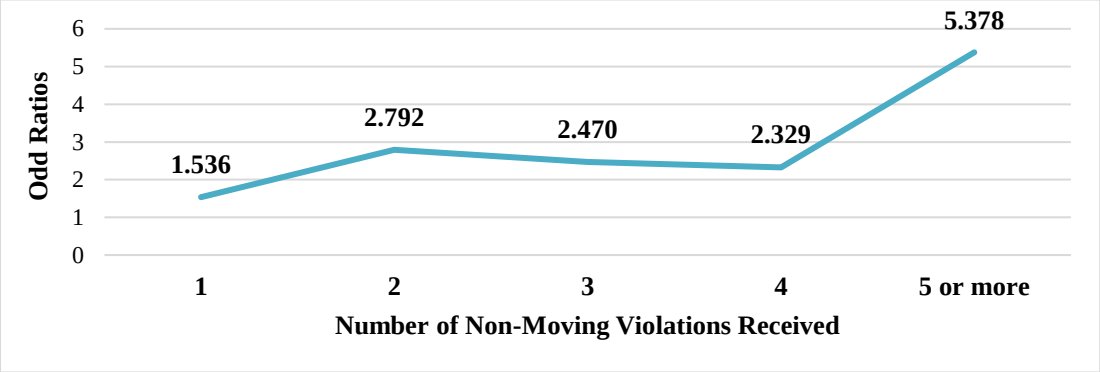


Figure 14: Odd Ratios for Non-Moving Violations

Model Assessment

An assessment of the selected model was performed to determine how effective it is at describing the outcome variable. The Hosmer-Lemeshow measure of association was used to analyze the goodness of fit of the selected model. This measure is obtained from the observed and expected frequencies presented in table 48. 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. A probability value of 0.856 was obtained for the Hosmer-Lemeshow test, indicating that the model provides a good fit of the data. 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 48: 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	35	54	59
2	(0.414, 0.447)	39	39.4	53	52.6
3	(0.447, 0.537)	41	45.1	51	46.9
4	(0.537, 0.601)	48	52.4	44	39.6
5	(0.601, 0.676)	62	59.4	30	32.6
6	(0.676, 0.727)	65	64.4	27	27.6
7	(0.727, 0.781)	68	69.5	24	22.5
8	(0.781, 0.841)	77	74.7	15	17.3
9	(0.841, 0.904)	81	80.1	11	11.9
10	(0.904, 0.990)	85	86.3	7	5.7

In addition to assessing the model using the Hosmer-Lemeshow test, a Receiving Operating Characteristic (ROC) curve was developed. 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-Lemeshow, 1999). 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 49 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 49: Classification Table for ROC Curve

Predicted Crash Involvement	**Observed Crash Involvement**	
	Yes	**No**
Yes	True Positive	False Negative
No	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 area under the resulting ROC curve provides a value which indicates the model's predictive ability. Figure 15 shows the ROC curve for the selected model provided by the Minitab output. The area under the ROC curved resulted with a value of approximately 0.73, which indicates that the model has an acceptable predictive ability.

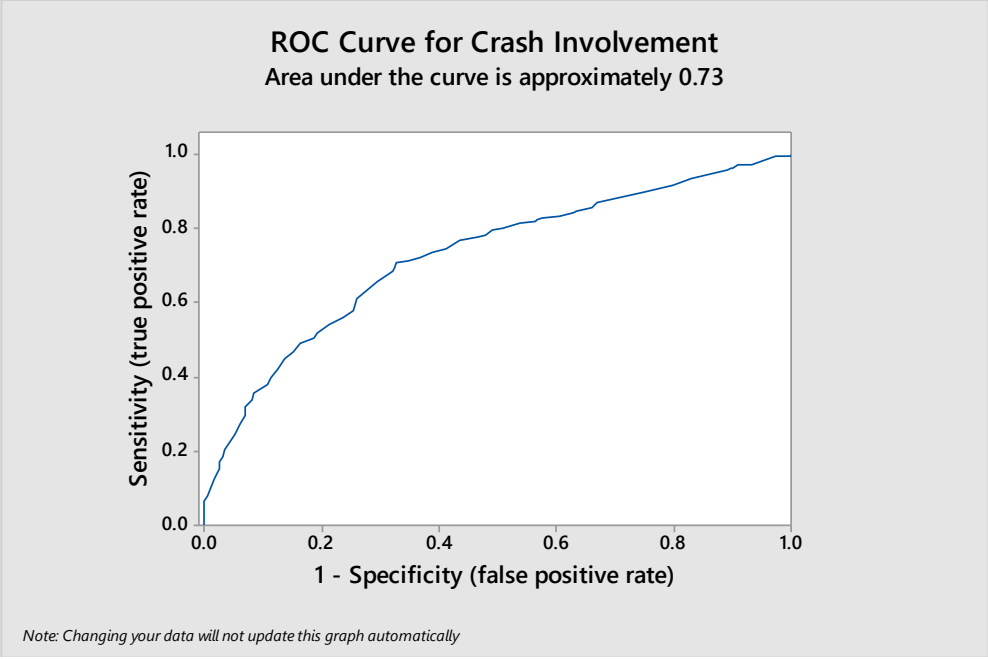


Figure 15: ROC Curve for the Selected Model (Obtained from Minitab)

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 of Driving Experience
- Gender
- Number of Moving Violations Received
- Number of Non-Moving Violations Received

Of all the predictors that were initially considered, these four predictors resulted to have the most significant association with the response outcome. 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 of driving experience 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.

4. CONCLUSIONS

The purpose of this study was to estimate the likelihood of a driver to 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, gender, 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

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 form of linear regression 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, gender years of driving experience and daily hours spent driving were inquired about. Age and gender are commonly used factors in any type of study of this nature while years of driving experience 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 where 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 male 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 of driving experience 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 gender 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 gathered:

- 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 of driving experience 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 where 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, gender, 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 & 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 research, 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 a value of zero. This was an issue since several of the tables corresponding to traffic violations had cells with a value of zero. For the traffic violations presented in this research, 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 example, 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 DUI 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 moving and non-moving violations. Moving violations concern traffic violations where the vehicle was on the move at the moment of the incident whereas non-moving violations corresponded to traffic violation concerning parked vehicles and vehicle equipment. 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 ≤ 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 was used to assess this association. Results for the chi-square tests of independence analyses performed indicated that age, moving violations, and non-moving violations have a significant association with being involved in a traffic crash; the variable gender 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 relations ship 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 of driving experience, 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 of driving experience, 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, odd 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 odd 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;

- Years of driving experience, gender and traffic violations history resulted to be significant predictors of the outcome event of being involved in a traffic crash.
- An increase in years of driving experience 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 moving 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-moving 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 resulted to be non-significant predictors of the outcome of being involved in a traffic crash. The results obtained for the predictor of daily hours spent driving remained 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.

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 remove 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 of driving experience
- Gender
- Moving Violation
- Non-Moving 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 is observed from the 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. Gender does not necessarily have to be considered a significant factor from a common-sense aspect but it is usually included in studies of this nature. Finally, traffic violations, in the form of moving and non-moving 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.

5. 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 research 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
- 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.

6. 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 & 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 Gender: 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. 47th 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 Example of Survey

HISTORY OF TRAFFIC VIOLATIONS AND CRASHES

Instructions: 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 ivette.cruzado@upr.edu.

General Information

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?	

History of Traffic Violations

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 16: 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 17: Page 2 of 2 from the Developed Survey