MODELING BICYCLE ACCIDENT INJURIES IN NORTH CAROLINA

by

TAMMIE LEAH-ASKEW SIMMONS

A thesis submitted to the Graduate Faculty of
Elizabeth City State University
in partial fulfillment of the
requirements for the Degree of
Master of Science in Mathematics.

May

2020

APPROVED BY

Kenneth L. Jones, Ph.D. Committee Member	Mohammed H. Talukder, Ph.D. Committee Member
Shatoya Covert, M.S. Committee Member	Kuldeep Rawat, Ph.D. Committee Member

Julian A. D. Allagan, Ph.D.
Committee Chair

ABSTRACT

This thesis explores factors that influence traffic accidents at both signaled and non-signalized intersections in North Carolina. This follows from a recent Masters' thesis [4] on bicyclist injury severity at unsignalized intersections in North Carolina. We revisit the original data for our analysis, increase the number of observations and find that only three injury's levels can be correctly classified (83% accuracy rate) using a neural network machine learning algorithm. With this information, we examine the relation between Ambulance and Injury levels and study the effect of traffic control (or its absence) on injury levels. We conclude that there is an increase in odds in the severity of the injury level in the absence of a traffic control. Further, given the (3) levels of injuries, we find that Road Speed Limit, Driver's Estimated Speed Limit, Light Conditions (daylight or not) and Road Characteristics (curve or straight) are statistically significant factors in a multinomial logistic regression model.

DEDICATION

This thesis work is dedicated to my husband, Brandon, whose consistent encouragement and support during the challenges of graduate school and life, afforded me the opportunity to complete the program. I am truly blessed to have him in my life. This work is also dedicated to my mom, known as Busy Bee, who unconditionally loved and prepared me for such a time as time. Finally, to my three children, Brandon, Breaunna, and Benjamin, who journeyed with me through this math maze.

ACKNOWLEDGEMENT

My sincere and utmost gratitude goes to Dr. Allagan for keeping me grounded and focused whilst I explored this topic. Without his persistent help, the goal of this project would not have been realized.

The physical contribution of the Transportation Research Board is truly appreciated. Without their support and funding, this project could not have reached its goal.

To my other professors, the frequent infusion of mathematical humor and introducing me to the beauty of the various branches mathematics which increased my thirst to want to learn more.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background and Overview

Developed in 2004, North Carolina established a committee for highway safety to support a federal-state association plan with a goal of reducing the fatality rate to 1.0 fatalities per 100 million vehicle miles. Just two years later, the revised plan highlighted 14 emphasis areas, and included in this plan was bicycle and pedestrian safety. The strategic framework implemented resulted in significant progress towards reaching the vision zero goal. Ten years from the inception of the federal and state safety initiative, North Carolina's goal was to decrease the number of fatalities and serious injuries by about a half, from figures gathered in 2013 by the year 2030. This long range goal became the mission and vision that motivated stakeholders to activate an vigorous yet obtainable plan for all transportation users in North Carolina. They defined nine areas of focus and among them was pedestrians and bicyclist. Through coordination with other agencies, North Carolinas' focus area's help to realize the goal especially with the push for alternative means of transportation to help reduce the carbon footprint with congestion, to promote healthier and active lifestyles, and to become more environmentally friend. Pedestrians traffic and bicycling offer a viable alternative; however, this opportunity also raised the vulnerability because of the inherent nature of the size, speed, and overall protection of the pedestrians and cyclists. In some cases, the areas of focus overlapped due the type of incident, which means that the overall quality of improvement in safety touches on several focus areas.

From the mandate established in 2004, North Carolina comprised a committee for highway safety to support a federal-state association plan with a goal of reducing the fatality rate. Although they reached specific milestones of the plan, they continue to

seek the zero tolerance by 2030 [9]. Hence the goal of this research is to research past incidents at both signalized and unsignalized intersection based on the data collected by their study to determine predictors of future incidents.

The sample studied comprised of data provided by UNC Highway Safety Research Center in Chapel Hill, NC, which include records of bicycle-motor vehicle crashes from 2007 to 2015 from the various cities and counties across the state. All city and county police agencies in the state provided the information from their perspective locations. Each incident represented the information collected by the policing official for that area. The qualified agent recorded demographics on the bicyclist, the driver, the direct and indirect influences of the incident to include environmental factors as well as interviews at the scene of the accident. Because of the statewide mandate to improve the transportation issue, the highway safety committee set goals and developed initiatives to accomplish those goals by 2030.

In Chapter 2, we describe some of the variables in the data and their distributions. In Chapter 3, we describe the different transformations that were applied to some of the variables, as they are prepared for analysis which took place in Chapter 4. There, we first introduce the reader to several basic concepts of data analysis. Neural network, binary logistics and multinomial logistics are the modeling techniques used for our research to answer the following five (5) questions:

Question n° 1. *Does the data contain enough information to help predict the need of an ambulance when a bicycle accident occurs?*

Question n° 2. *Does the data contain enough information to help predict the different (5) levels of injuries when a bicycle accident occurs?*

Question n° 3. *Do bikers' injury levels (3) help predict the use of ambulance when accidents occur?*

Question n° 4. *Does the presence of some traffic control (or the lack thereof) help explain the (3) levels of injuries of the bikers?*

Question n° 5. *What factors help explain the (3) levels of injuries of the bikers?* Answers for each question is provide and a model equation with an interpretation is given, where appropriate. We close our thesis with Chapter 5 where we summarize our finding

and offer some recommendations.

1.2 Literature Review

While extensive research on bicycle-motor vehicle crash (BMVC) raised awareness, the need continue preventing bicycle accidents still exist especially with an aggressive approach to reaching the goal of 0% tolerance and an increase in bicycle presence. Earlier research has brought about significant changes with incidents with bicycle, and much of those studies focused on classical techniques which we will also use. Hence, the need centers on the fact of using more complex tools such as deep learning and boosting techniques or even generative models which can be compared to the logistic regression models typically explored to give a good base line for the machine learning algorithm. Because of this, the literature review will explore research on factors associated with bicyclist's injury severity. The literature review is also aimed at examining past literature from the mathematical perspective to provide an insight into how cyclist and transportation research have been affected by the commitment to reduce congestion with recent demands for alternative modes of transportation.

A summary of existing studies comprise of four areas for severity of bicycle crashes(1) design;(2) volume as contributing factors;(3) human intervention as a factor and(4) safety and awareness. The various research areas arise from the data collected which helps recognize variation for analysis. With the severity of injury from bicycle crashes, researchers investigated traffic volume and design as well as a socio-economic contributions to accidents. In comparison to the volume collected from other transportation modes, bicycle data presents challenges especially because of the scarcity. Hence, data collected attempts to use all measures of bicycle exposure. Studies involving different factors in bicyclist incidents capture bicyclist gender and other demographics, environment, roadway infrastructure, vehicle types, and even traffic; yet, despite the decline in the number of incidents over the past 15 years, bicycle incidents still occur [11, 5, 18]. Approximately 1,000 bicyclists are involved in police-reported crashes with motor vehicles. Children and young adults are the most frequent victims [9].In the current review, over 100 sources

containing these factors and key terms were investigate as well as provided statistics for those factors. Intersections are BMVC prone locations [18]. For that reason, the design of this analysis centers around signalized and unsignalized intersections to build predictive models.

1.2.1 Design

Crashes occur for a number of reasons. In 2017, Asgarzedeh et al. researched the role of intersection and street design on severity of bicycle-motor vehicle crashes using geographical information system (GIS) along with data collected from New York's police records on bicycle-motor vehicle crashes (BMVC) accidents. They concluded from the various environmental variables of the area that non-orthogonal intersections and non-intersection street segments are more likely to result in a severe injury to bicyclist than BMVCs at orthogonal intersections [1]. Their study captured street design as it relates to severity of the BMVC as well as the risk ratios using multivariate log-binomial linear regression along with log-binomial regression to model the risk ratios. Others researchers of intersections found that clearance time, bicycle lane exclusivity, and vehicle movement delay contribute to greater chances of BMVC incidents [6]. Madsen and Lahrman's study of the design of intersections and concluded that there should be recessed bicycle track to minimize BMVS [8]. However, they also found that the various volumes of traffic made it difficult to determine which design was safest since the number of incidents were relatively small.

1.2.2 Volume

As innovation progresses toward increase in bicycle traffic concerns that other factors associated with bicycle accidents also sparked interest, which included volume and speed. Their study investigated the impact traffic volumes, facility type and land use on cyclist safety; and develop statistical models to predict the number of cyclist collisions. It is assumed that more crashes occur in population density areas than in location with lower population density . Traditional statistics produced by NHTSA provide cyclist fatality rates by population (per 10,000 residents), but these metrics are not sensitive to the

amount of time or distance that the cyclist is exposed to vehicular traffic Another study also found that land mix as little as an increase of 10% at intersections increases the bicycle traffic by 8%, but without the appropriate interventions results in increase in cyclist injuries [16].

1.2.3 Human Factor

Although people across ethnic groups, genders, age groups, and even socio-economic groups are affected by bicycle accidents, those at a lower risk are males where all else being equal [16] However, with regard to the biker and driver specific data, research on the amount of time spent on the road, which includes distance, as well as speed, found a strong correlation of speed and density where electric bicycle were compared to bicycle [2]. This is particularly important with the increase in electric bicycle usage. Others also saw a direct correlation between the willingness to commute by bicycle in low-stress commutes to and from destinations [3]. The concern for safety was a factor even with regard to the riders age. Schepers et al. investigation of unsignalized intersections found that employing speed reducing mechanisms along with design changes were most effective with a clearance between 2 and 5 m is safer than a cycle lane [14]. Hence, when Tang et al. studied the impact of group behavior on bicycle flow at intersection, it was interesting to note the negative impact on the bicyclists as well as with traffic flow [17].

1.2.4 Safety and Awareness

From the first BMVC in the United States, which occurred in New York City on May 30th 1896 , the range of accidents at intersections is about 50-64% [7], which they assert are influenced by unware of surroundings. Others have studied safety and awareness of bicyclist travels and found that with increase in safety and awareness the number of incidents have increased [10]. Jannet et al research found that driver attention focuses mostly on nearby cars and car within their frontal view and with little regard for peripheral blind spots, which agreed with their previously informal notions [5].

1.3 Conclusion

With so many various studies, there is still a need for understanding and improvement. Bicycling offers an alternative to the issue of traffic congestion as well as other health benefits. However, when is error, safety should always come first. In addition to the height use which is not well develop and has limitations, BMVC studies are a valuable pursuit.

Chapter 2 Data Exploration

2.1 Basic Definitions

1. **Type:** Indicates if the data is Numerical or String (text/alphabet/character)
2. **Label:** Signifies the name or label of the variable. The exact label described in this column will appear in the tables/ graphs.
3. **Values:** The Coded value of the variable when indicated.
4. **Missing:** There could be instances within data that observations are label "n/a", "NULL", ".".
5. **Nominal:** A variable can be treated as nominal (or *categorical or qualitative*) when its values represent categories with no intrinsic ranking. Names or labels of the categories e.g. Gender, hair color, location.
6. **Ordinal:** A variable can be treated as ordinal (or *categorical or qualitative*) when its values represent categories with some intrinsic ranking. Order of the categories: e.g, Levels of injuries, satisfaction.
7. **Scale:** A variable can be treated as scale (or *numerical or quantitative*) when its values represent a meaningful metric. e.g., age in years, and income, test score.

2.2 Data Description

Initially, the data consisted of 66 variables (string, numerical, date) and 8418 observations for each of those variables. It contains variables such as Year of incidents, Crash date, Bikers Age, Gender, Race, Date (Year, month, day) of crash, Time (hour,

minute), Location, Street condition, Level of injury (Ambulance required or not, death, etc.)

See Appendix for details.

2.2.1 Location

The accidents occur throughout the state of North Carolina as shown in Figures 2.1 and 2.2.

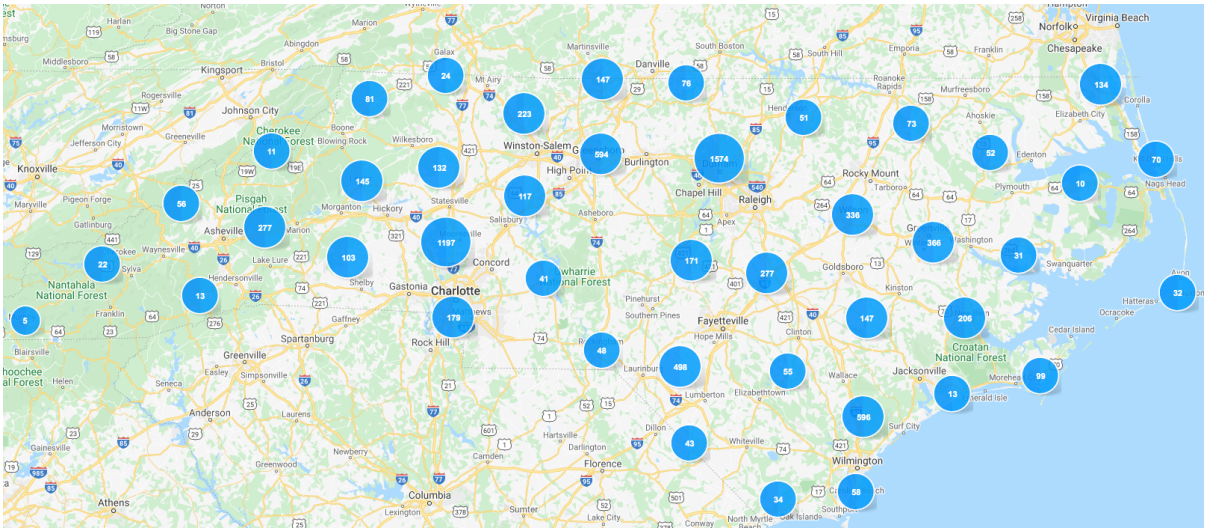

Figure 2.1: Location of Yearly Incidents

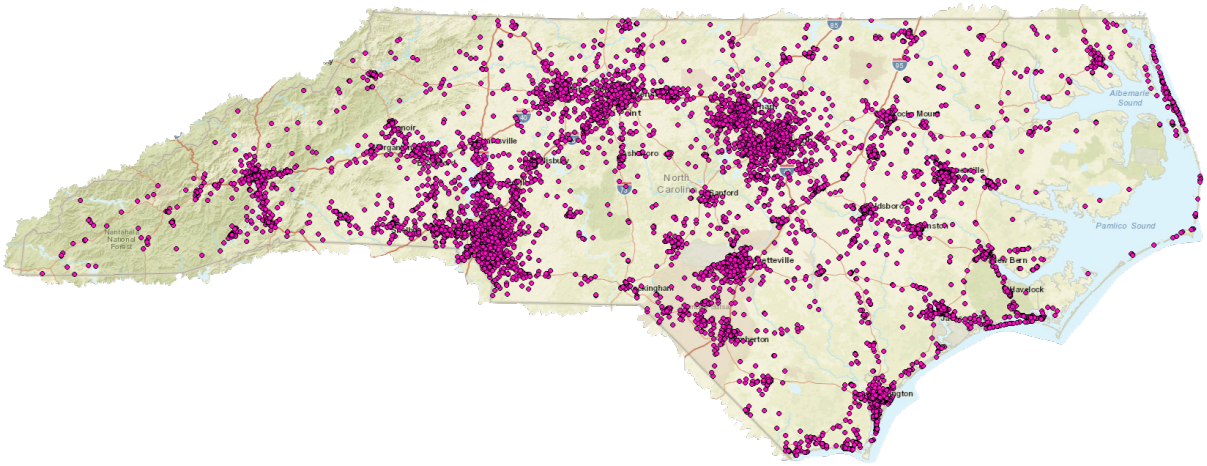

Figure 2.2: Cluster of Incidents: 2007-2018

There are 296 cities that are recorded from 100 counties. One county (Wake county) accounted for almost 14% of the incidents in North Carolina of which one city (Raleigh) within that county accounted for 790 out of the 1157 values captured.

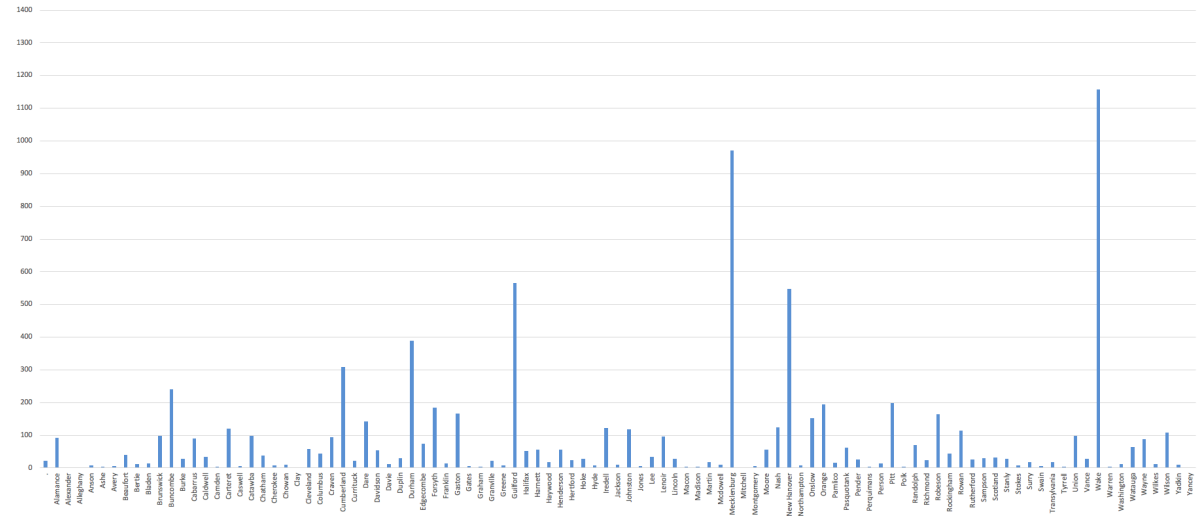

Figure 2.3: Count of Incidents by County

2.2.2 Level of injuries

The biker's injury level (BikeInjury) includes five categories ranging from no injury to deaths. Table 2.1 describes each category in the order of severity. There are 230 unknown cases

Category of Injury	Description of Injury
No Injury	No injury reported
Possible Injury	No visible injury, but complaints of pain annotated, or temporary unconscious experienced.
Evident Injury	Apparent injury at the scene, but not killed or disabling
Disabling Injury	Injury, which prevents the person from performing daily activities for at least one day beyond the day of the incident
Killed	Deaths (occurring within 12 months after the crash) resulting from injuries sustained from incident

Table 2.1: Bicyclist Injury Levels

Throughout the years (2007-2015) the number of recorded accidents has remained consistent (about 10-12%) as shown in Table 2.2. About 80% of these accidents show some evident/possible injuries.

Incidents by Year	Disabling Injury	Evident Injury	Possible Injury	Killed	No Injury	Unknown Injury	Grand Total
2007	68	432	365	17	112	7	1001
	1.92%	12.21%	11.01%	8.81%	12.39%	12.73%	11.89%
2008	50	430	403	29	104	2	1018
	12.11%	12.15%	12.16%	15.03%	11.50%	3.64%	12.09%
2009	45	352	328	14	68	4	811
	10.90%	9.95%	9.90%	7.25%	7.52%	7.27%	9.63%
2010	43	448	341	23	105	8	968
	10.41%	12.66%	10.29%	11.92%	11.62%	14.55%	11.50%
2011	50	354	358	19	111	10	902
	12.11%	10.00%	10.80%	9.84%	12.28%	18.18%	10.72%
2012	39	417	425	27	107	5	1020
	9.44%	11.78%	12.82%	13.99%	11.84%	9.09%	12.12%
2013	40	345	379	23	107	3	897
	9.69%	9.75%	11.44%	11.92%	11.84%	5.45%	10.66%
2014	39	355	349	17	81	9	850
	9.44%	10.03%	10.53%	8.81%	8.96%	16.36%	10.10%
2015	39	406	366	24	109	7	951
	9.44%	11.47%	11.04%	12.44%	12.06%	12.73%	11.30%
Grand Total	413	3539	3314	193	904	55	8418

Table 2.2: Table of Incidents by Year

As we look at the relation between road condition (1=wet, 0=dry) and injury, we see from the graph, Figure 2.4 that there is no significant difference between the number of incidents when the road is wet vs when it is dry.

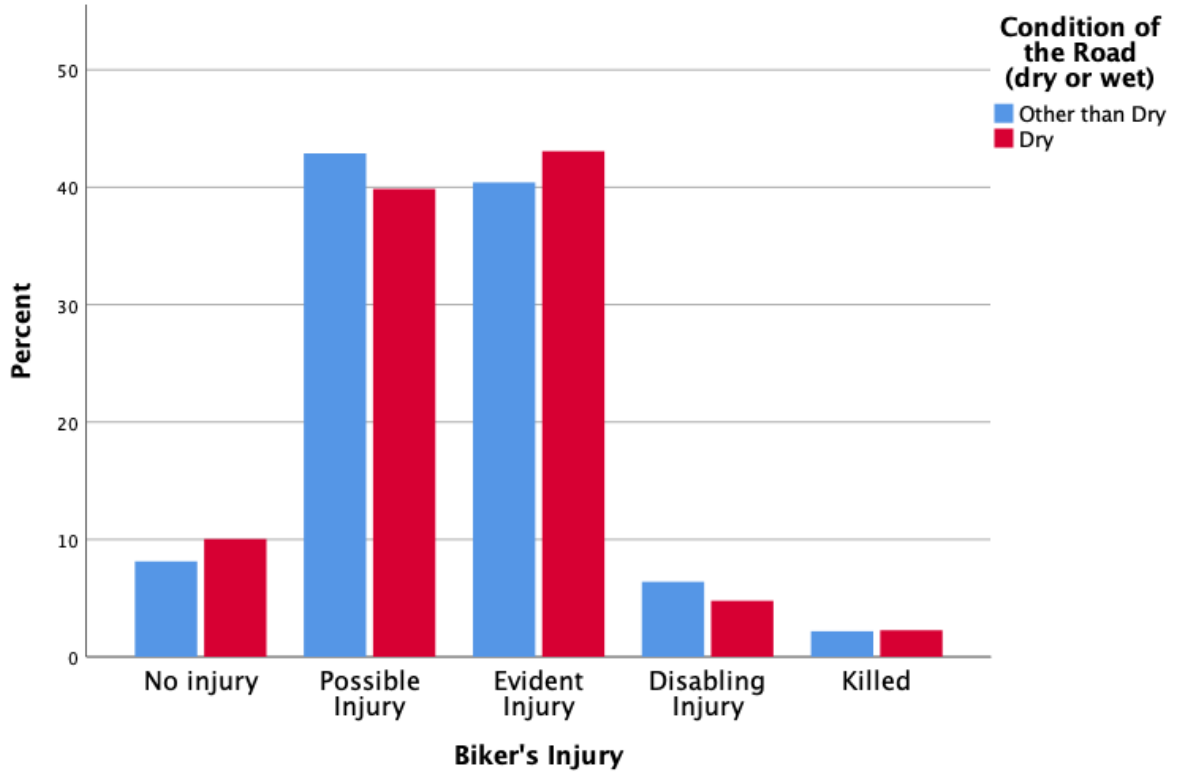

Figure 2.4: Incidents by road condition

Furthermore, as related to work zone, evidence from graph (Figure 2.5) suggests that work zone has no bearing on the level of injuries. In fact, that was reported death not

in a workzone.

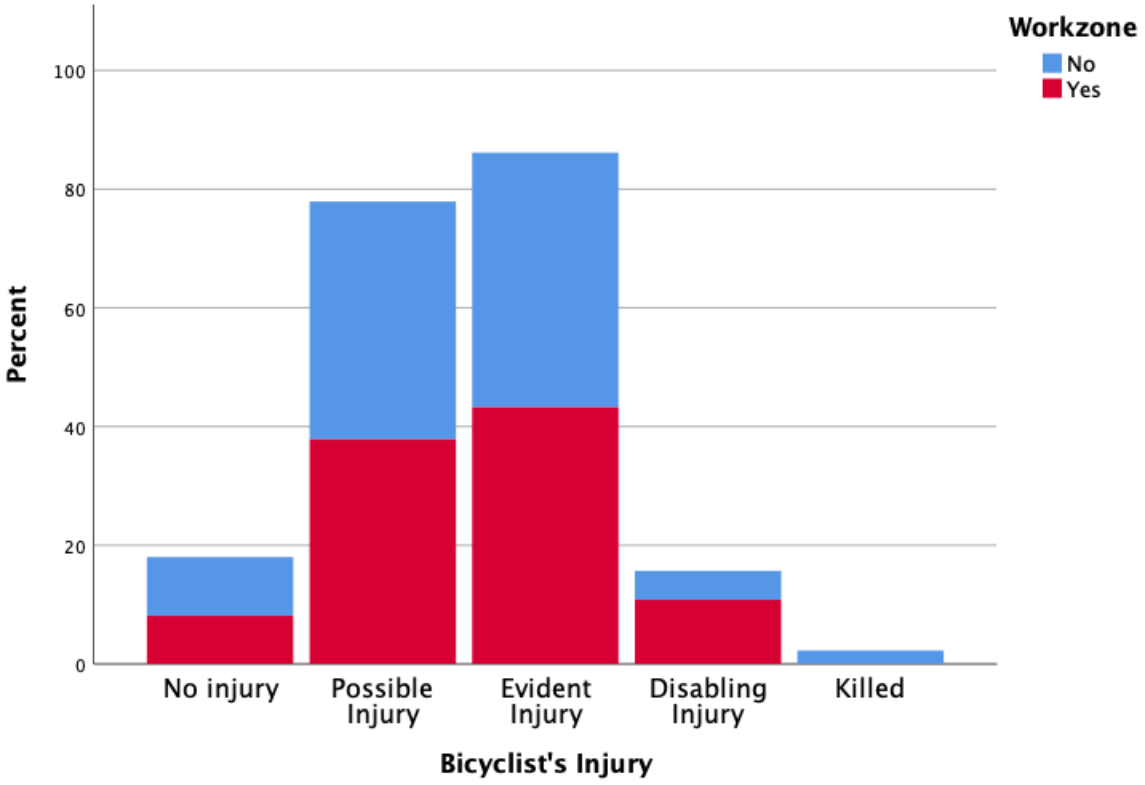

Figure 2.5: Incidents by workzone

Most accidents occur when the drivers's average speed is its lowest as shown in Figure 2.6.

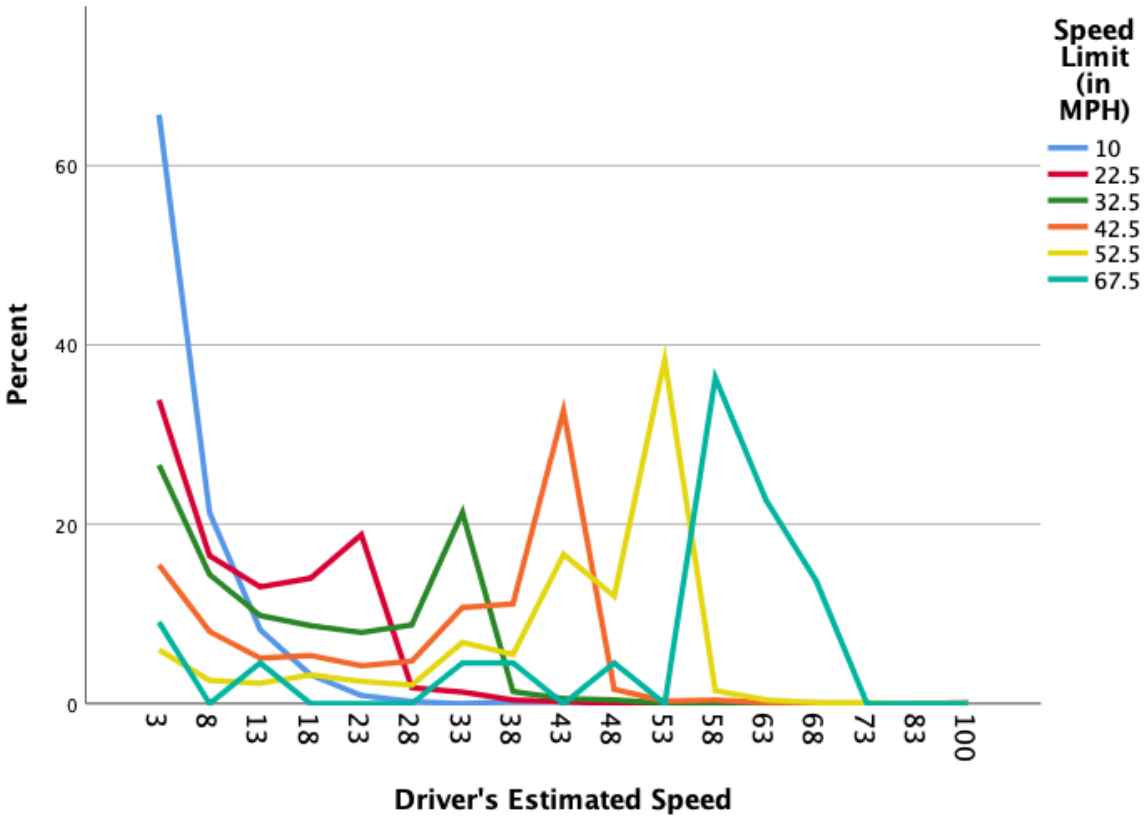

Figure 2.6: Average driver's speed

Furthermore, most drivers were driving within the speed limit as shown in Table 2.3

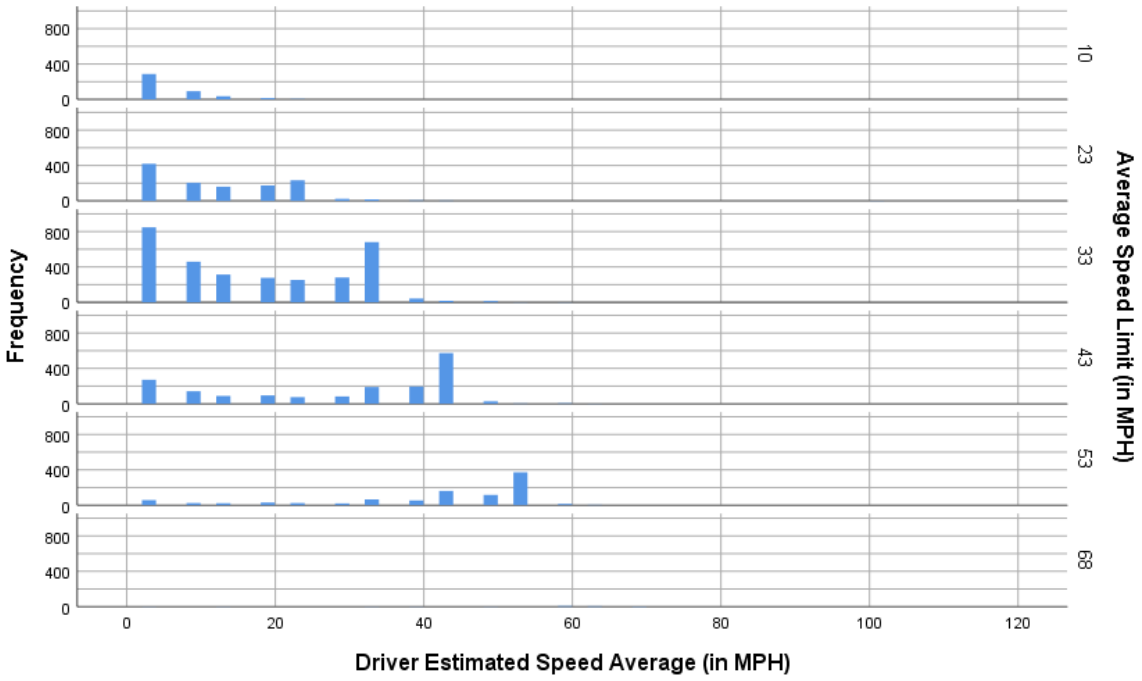

Table 2.3: Drivers' speed vs Actual speed limit

Most accidents occur during the day as shown in Figure 2.7

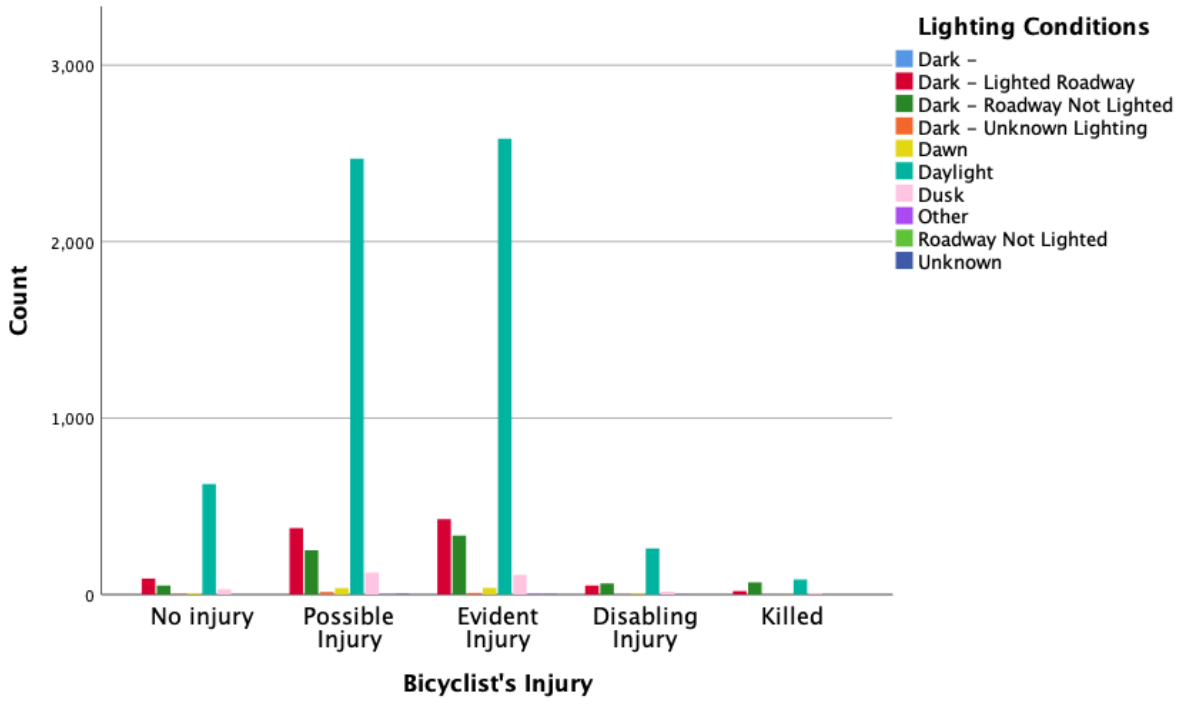

Figure 2.7: Light condition during accidents

Most accidents occur with or without evident of injury where there is no traffic control light or signal as shown in Figure 2.8

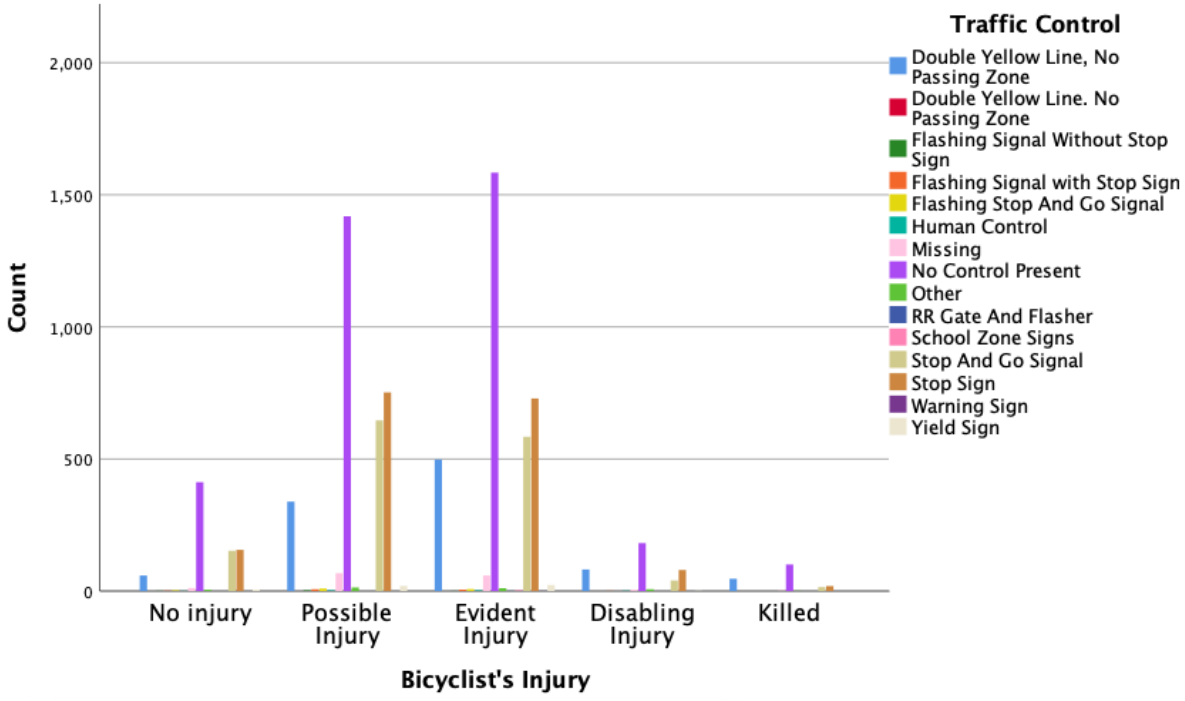

Figure 2.8: Accidents reported at signalized or unsignalized intersections

2.2.3 Substance Involvement

Quite naturally, the data also contains some information on whether or not individuals were under the influence of some substance. Substances include alcohol, alcohol and drug, for the biker and the driver. Table 2.4 and Figure 2.9 show that less than 8% of reported accidents involve some substance use. Moreover, Figure 2.10 indicates that hit-and-run incidents have no effect on the level of injuries.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	7771	92.3	92.4	92.4
	Yes	643	7.6	7.6	100.0
	Total	8414	100.0	100.0	
Missing	System	4	.0		
Total		8418	100.0		

Table 2.4: Record of Substance Use

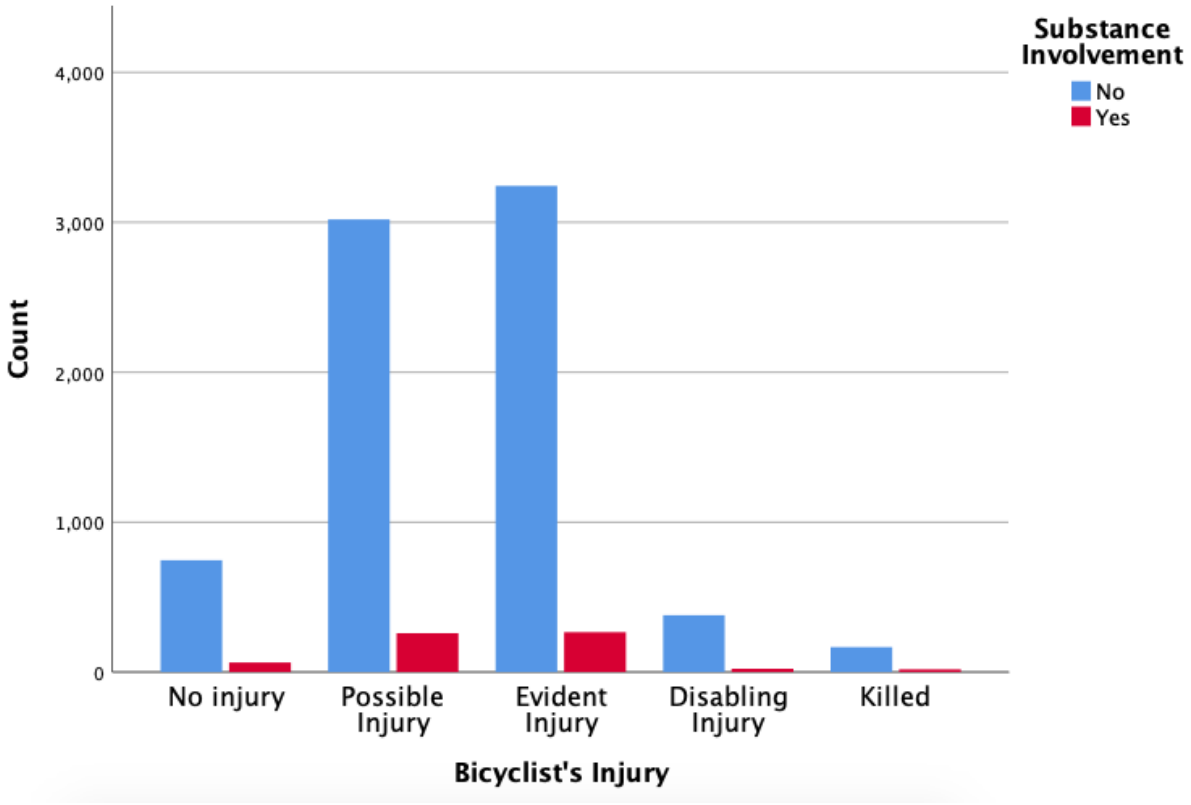

Figure 2.9: Substance use and level of injuries

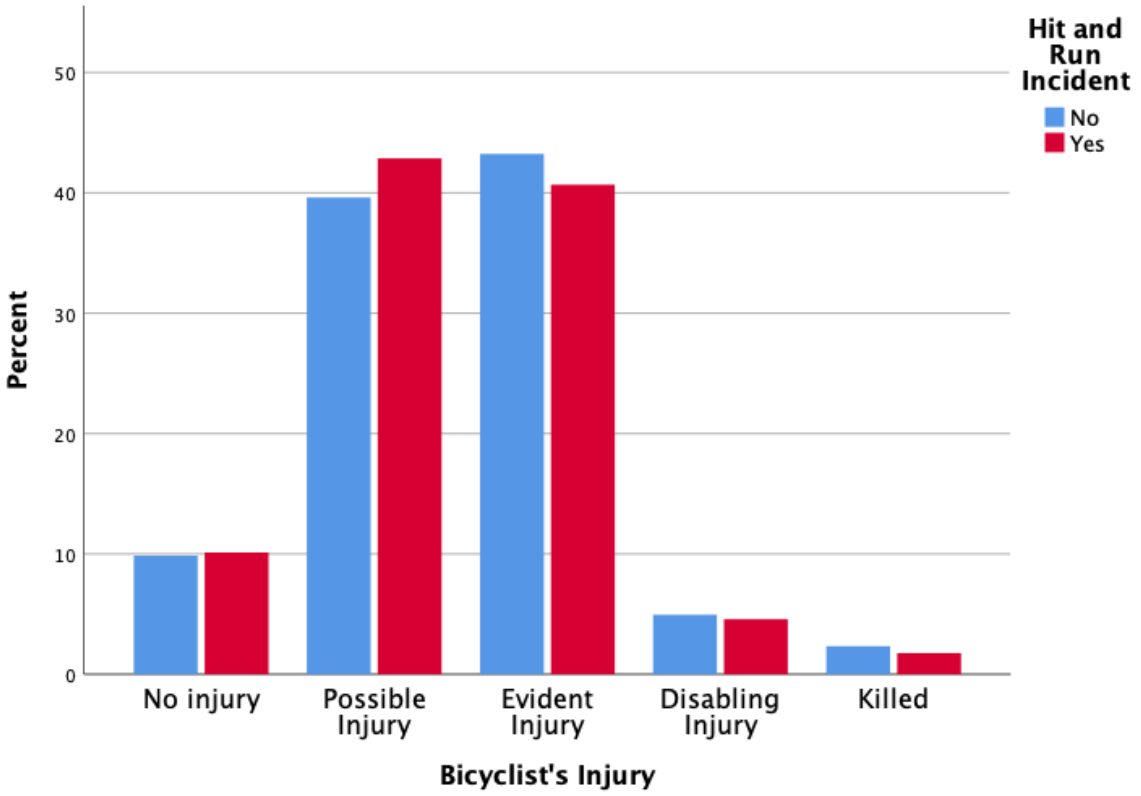

Figure 2.10: Hit and run with reported level of injuries

2.2.4 Ambulance

About 68% of the incidents require an ambulance while 32% do not. Table 2.5 shows such a distribution by levels of injuries.

		Biker's Injury					
		No injury	Possible Injury	Evident Injury	Disabling Injury	Killed	Total
Ambulance Required	No	639	1111	754	26	11	2541
	Yes	171	2170	2757	375	174	5647
	Total	810	3281	3511	401	185	8188

Table 2.5: Distribution of Ambulance requirement

From Table 2.5 and Figure 2.11 , it appears that there are instances where ambulance was not required while there is some evident/possible injury even in the case of death (11 cases).

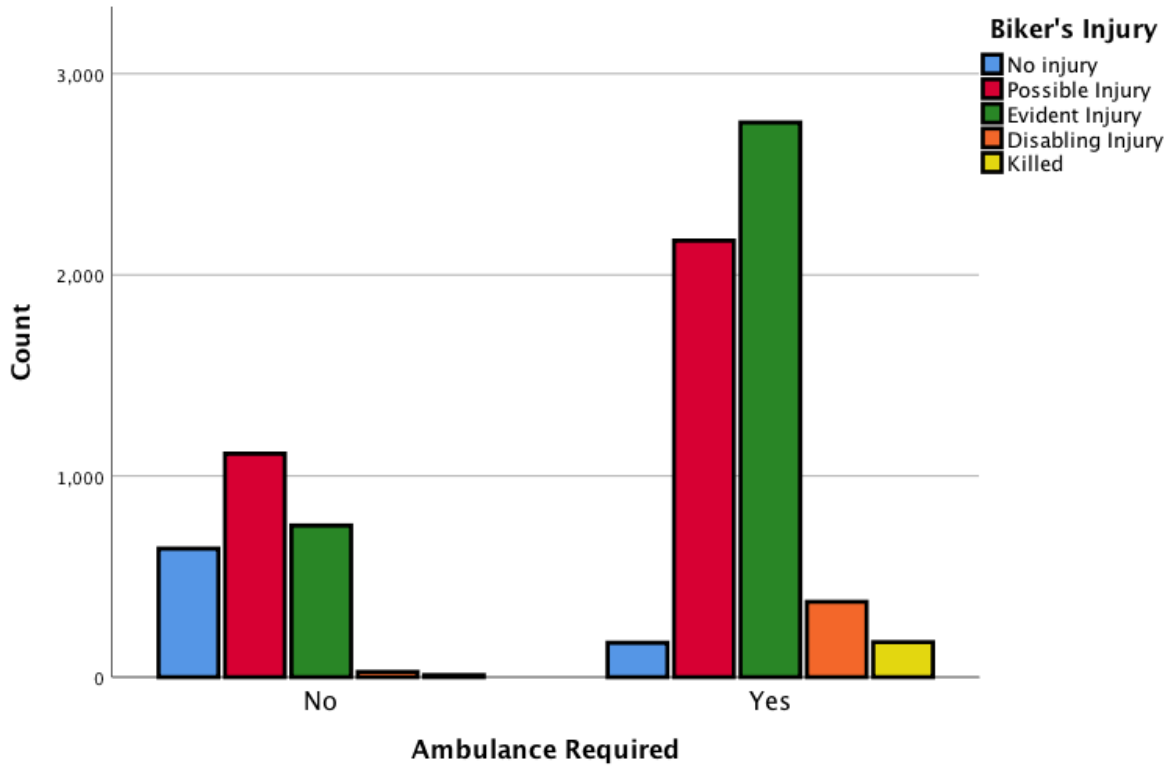

Figure 2.11: Ambulance requirement and level of injuries

2.2.5 Biker Age

The data captures the age of the bicyclist in two different ways. One variable was the numerical age in years which up until 2015, an observation from 2015 forward, expressed the age as a group as well entries and in some cases so the data was transposed to the appropriate field. Out of 8418 observations, 124 are recorded as missing. The mean age recorded is about 32 with 1 infant (0 age) and being 121 individuals being 70 years or older. Details are shown in Figure 2.12 and Table 2.6.

Statistics	Values
Mean	32
Median	28
Std. Deviation	17.4
Minimum	0
Maximum	70

Table 2.6: Bikers Age Statistics

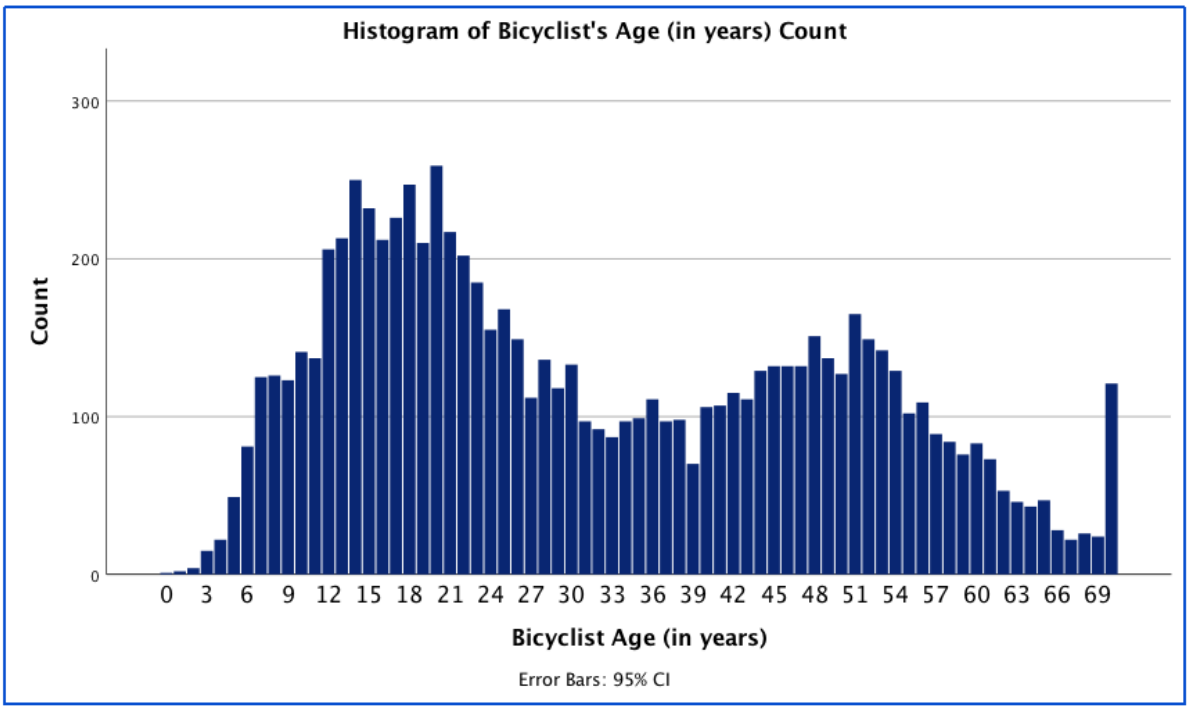

Figure 2.12: Distribution of Bikers Age

2.2.6 Biker's Race

The data contains the race of the bikers involved in an accident. 55% of them are white, 34.5% are black while the remaining 10% accidents involve other races. Figure 2.13 shows such distribution.

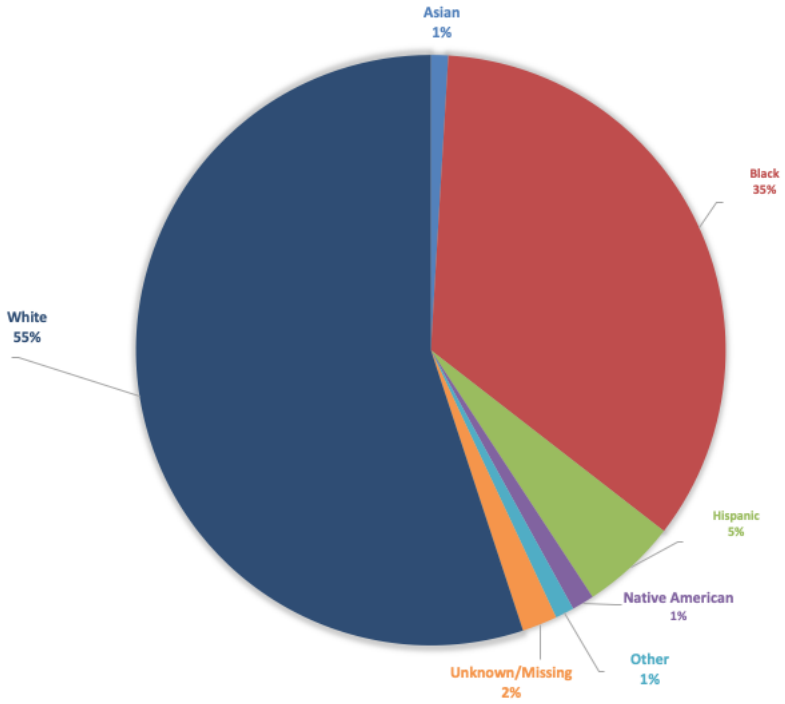

Figure 2.13: Biker's Race

Figure 2.14 shows although most accidents affect all races, Native American are reported to be disproportionately killed even though the account of 1% of the recorded accidents.

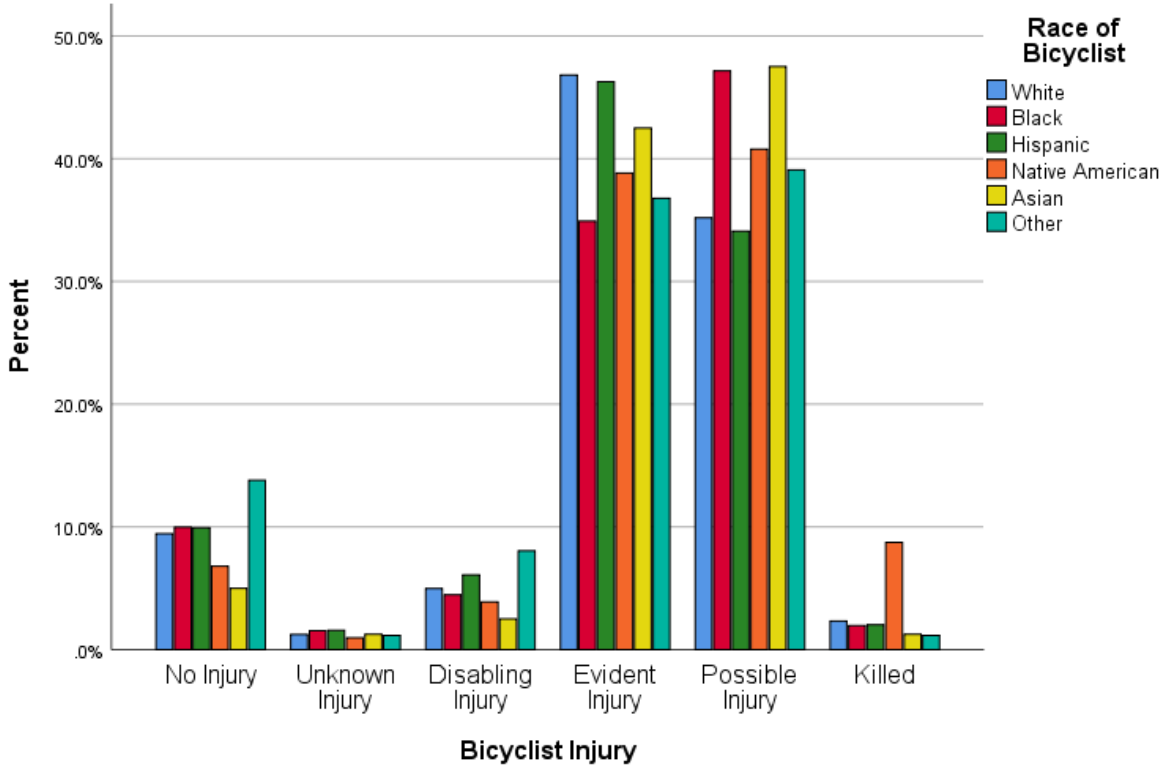

Figure 2.14: Biker's Race and injury level

2.2.7 Biker's Gender

Bicyclist's and Driver's genders also provide some additional background information about the individuals that are involved in the accident. There were 116 missing gender values for the bicyclists compared to the 1138 missing values for the driver. We focus on bicyclists, as it is the focus of this research. As shown in Table 2.7, most accidents (85%) involve males.

Gender of Bicyclist

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	1238	14.7	14.9	14.9
	Male	7064	83.9	85.1	100.0
	Total	8302	98.6	100.0	
Missing	System	116	1.4		
Total		8418	100.0		

Table 2.7: Biker's Gender

Most recorded bikers injuries span across both genders, in which there is no obvious relationship.

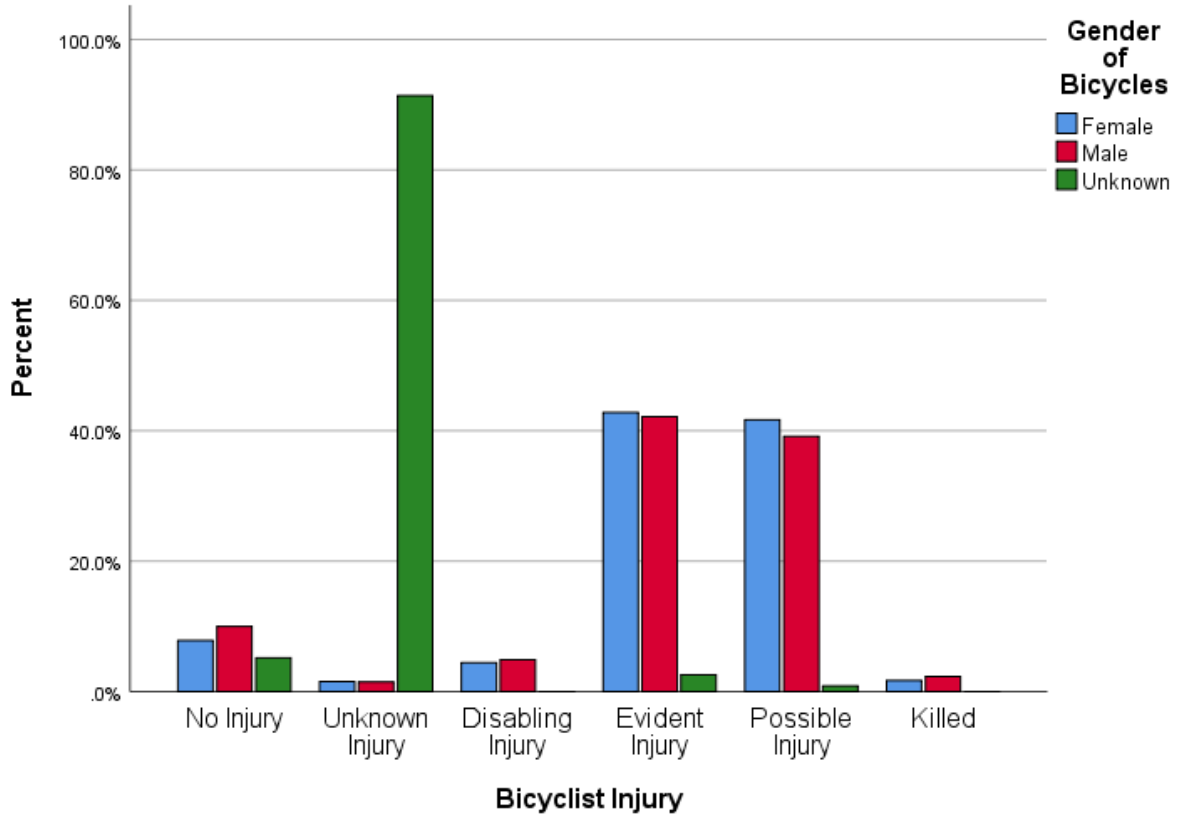

Figure 2.15: Biker's Gender with recorded injury levels

Chapter 3 Data Mining

3.1 Basic definition

1. **Binning:** a grouping of contiguous values of existing variables into a limited number of distinct categories

2. **Data cleaning:** a process of identifying and correcting or even deleting errors and inconsistencies in a data due to incomplete or inaccurate data entry, which requires replacement, modification or even deletion of irrelevant data.

3. **Dichotomous variable:** a variable that contains precisely two distinct values.

4. **Nominal Data or nominal scale:** the simplest for a scale measure but provides no quantitative value.

5. **Ordinal Data:** a data type with a set order or scale to it.

6. **Proxy or Proxy variable:** a variable that serves as an immeasurable variable and is itself indirectly relevant.

7. **Redundancy:** can be data duplication or cross-correlation where two seemingly unrelated variables have some relation or cause that at one time were independent of each variable.

3.2 Data preparation

In the original form, the data is not easily recognizable for machine learning which required recoding of some existing variables and in other cases proxy variables were employed. However, before an analysis can take place, data cleaning, transformation or

integration must occur. This process also checks for irregularities, eliminate duplicated data, detect and correct missing values. In some cases, proxy variable, also called dummy variables are generated to make for a cleaner analysis. Data mining tools can effectively create valid and insightful models only when the information provided is free of nuisance and noise factors. The first step in the procedure describes each of the variables in the data set. The description of the variables ensures continuity of understanding. The original coding represents the data as it was originally presented. The coding employs encoding or continuation of categorical data transformed into their numerical counterparts. The process for encoding is necessary for modeling methods such as linear regression or bivariate correlation to illustrate a continuous variable. The two main types of encoding binary or target-based encoding.

Much of the encoding converting the string data to numerical data types presented a nominal relationship in the table, which meant the various categories of data had no relationship. From the data collected from a profession agent at the scene of the incident expertise in the subject matter for each incident, 34 variables were utilized in the order presented in the data set (see Appendix 5). The data set contained missing value and where necessary some scaling was necessary for the purpose of this analysis. There are occasions when it is necessary to reduce the number of categories in an ordinal or nominal variable by combining ('collapsing') them in order to perform a certain type of analysis.

The initial cleaning resulted in dropping and deleting several variables. Among them were Object Identification, Biker and driver age group, crash time, crash date, Distance mile from, NumBikesAI,NumBikesBI, NumBikesCI, NumBikesKI, NumBikesNo, NumBikesTo, NumBikesUI, RdFrm, OnRoad, TowrdRd. In 2014, data on DistnMiFrm and FrmRd was captured but for this study, is was deleted; however, the location

3.2.1 Location

Throughout the examination, many of the variables appeared to be redundant and among the observations, several observations collected information regarding location. The narrative of the city, county, crash location, development, distance mile from, from road, on

road, region, and toward road variables' tell the same story as does the latitudinal and longitudinal. The latitude and the longitude data provided specifics of the accident's location. Since the goal is to use tools like genetic boosting, location data, city and county variables were considered for location segments analysis.

3.2.2 Level of Injuries

Level of Injury's observational data was ordinal which allowed the data to be coded based on rank and order of the injury from *1* to *5*. If no injury was collected, the string variable was converted to a numerical value of 0. The next value was possible injury which was assigned *2* in rank. The following value, evident injury, was *3* in the rank followed by disabling injury assigned 4. The last of the order values was encoded a *5* for deaths. The level of injuries was further delineated into *3* bins.

3.2.3 Substance Involvement

Variables Biker and driver alcohol only as well as alcohol and drug involvement along with crash alcohol variables were merged into a single variable category called substance use/involvement. If either the biker or the driver had a yes value regardless of impairment, the value of the substance use variable returned a yes value. It appears that the predefined columns were further delineated to reflect a more comprehensive extent of the information captured. The number of yes and no were considered, and the dominant category received the value associated with the category. The missing or null values were calculated also based on the majority category as well. Given the ratio of no to yes, we extend that ratio to randomly impugned values to the null and encoded to keep the same percentage count. Rather than employing the mean (or median) of a certain attribute calculated by looking at all the rows in a database, we limited the calculations to the relevant yes or no response to make the value more relevant to the row in review.The goal was to savage all of the values. Because of the swap, the data was updated to reflect the transposed information. After cleaning the initial values, the encoding reflected 0s for no values and 1s for yes values which in turn transformed this variable to a numerical nominal variable.

3.2.4 Ambulance

As a string variable, recoding the same variable, ambulance required variable consisted of no and yes responses. Hence, this variable transformed into 0s and 1s encoding where 0 represented the no responses and 1 represented the yes. Changing the coding to binary coding makes for decent machine learning as it aids in processing time and accuracy.

3.2.5 Biker Age

Biker Age naturally correlate with Biker Age Group which is another variable in the data. So, we dropped Biker Age group and kept Biker Age which was a numerical and scaled value. Additionally, 124 of the observations are unknown and we used a global constant "." to handle this missing value. There are also some rows that had special operators, such as 70+ that triggered a string type. To handle these cases, the data was grouped and tagged as 70 years in age in order to treat the data as a numerical type. During the examination of the data, it was observed that beginning 2015, the age and age group were swapped, so to keep the data consistent, the two values were transposed for the analysis, which further substantiated the need to evaluate only the one variable. with the changes implemented, the age value code was the numerical age of the biker.

3.2.6 Biker's Race

At first, the biker's race consisted of 7 different race groups, which captured the unknown values as well. Proxy variables were created for each trait and then encoded with a 1 if the attribute was present for that "dummy variable" and all other characteristics was encoded as a 0. Each attribute now reflected a numerical type with a nominal value of 1 or 0. With regard to the various changes in this variables, we deleted this variable as a factor.

3.2.7 Biker Gender

Gender is treated as a dichotomous variable, because at the time of the survey, only two distinct choices were presented. Changing the values from a string type to a numerical

type along with nominal values of 1 or 0, made for easier learning environment. Also with this variable, we deleted this variable.

3.2.8 Crash Group and Crash Type

Data collected also included crash group and crash type which were identified by the professional agent at the scene of the incident. The type of accident help explain the events that led up to the incident. Crash group centralizes the various types of crashes.

Chapter 4 Data Modeling

Predictive analytics generally seek to extract information from the raw data in order to predict trends or indicate certain patterns of behavior. Here we rely on standard statistical data modeling such as logistic regression and a well-known machine learning technique called neural network. Fundamentally, we are trying to capture the relationships between each of the responses Injury and Ambulance, and several predictors such as Drivers speed, Road condition, traffic control, age, gender, etc. The events recorded in the data as they occurred a few years ago (up to 2015) are analyzed to help predict when we outcomes such as death or injuries for a biker that is involved in a road accident. We begin by introducing the reader to some common statistics, models, and technical terms.

4.1 Basic Statistics and Machine Learning

4.1.1 Level of significance

Also known as **alpha level**, this value is used as a probability cutoff for making decisions about the null hypothesis. Its value represents the probability we are willing to place on our test for making an incorrect decision in regards to rejecting the null hypothesis. In other words, it is the level of risk we are willing to take as we reject a possibly correct hypothesis. For example, a significance level of 0.05 indicates a 5% risk of concluding there is a statistically significant result or difference when there is none.

4.1.2 P-value and Confidence Interval

P-values (labelled **Sig.**, in SPSS) are the probability of obtaining an effect or a relationship at least as extreme as the one in the sample data, as we assume the truth of the

null hypothesis. When a p-value is less than or equal to the significance level (typically 0.05), we reject the null hypothesis.

The range of values, for which the p-value exceeds a specified alpha level is called **confidence interval**. In other words, this interval gives a range of values within which lies a true (population) parameter. So, with an estimated parameter at $\alpha = 0.05$, a confidence interval indicates that, with repeated samplings (identical studies in all respects except for random error), we are "confident" that, in spite of margin-of-error (or deviations), 95% of the parameter estimates will lie within this interval. With the margin-of-error we can state that the interval includes the true population parameter.

4.1.3 Correlation

A simple correlation measures the relationship between two (ideally normally distributed) variables. For our thesis we used Pearson's r which measures a linear relationship (or association) between two continuous (numeric) variables without taking into account other variables. For each pair of variables (X_i, X_j) Pearson's correlation coefficient is computed using

$$r = \frac{\sum_{i=1}^{n}(x - \overline{x_i})(y - \overline{y_i})}{\sqrt{\sum_{i=1}^{n}(x - \overline{x_i})^2 \sum_{i=1}^{n}(y - \overline{y_i})^2}}.$$

Its value range between -1 and 1 and $|r| \sim 1$ indicates a strong dependence or correlation and $|r| \sim 0$ indicates a strong independence between the variables.

The objective of any data analysis is to extract information (or accurate estimation) from the original (raw) data. Typically, we seek to determine whether or not there is statistical relationship between a response variable (Y) and explanatory variables (X_i). One way to answer this question is to use some regression analysis in order to *model* its relationship. By modeling we try to predict the outcome (Y) based on values of a set of predictor variables (X_i). There are several types of regression analysis and each type of the regression model depends on the type of the distribution of Y. They are often used to assess the impact of multiple variables (a.k.a. covariates and factors) in the same model. Here, we focus on two of these which we define next.

4.1.4 Linear regression

This is an extension of the simple correlation. In regression, one or more variables X_i (*predictors* or *factors* or *independent variables* or *inputs*) are used to predict an outcome Y_i (*response* or *target* or *criterion* or *dependent variable* or *output*). In practice, a linear regression model or equation returns estimates of the coefficients of a linear equation that involves one or more independent variables that best predict the values of an ouput or the dependent variable which must be quantitative continuous or scale. It is often written as

$$E(Y_i) = \beta_0 + \beta X_i \quad \text{or} \quad Y_i = \beta_0 + \beta X_i + \epsilon_i$$

for each i observation or data point with errors ϵ_i.

Regression coefficients or coefficient estimates β_i represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.

The *p-value* for each term tests the null hypothesis that the coefficient is equal to zero (no effect). Thus, a low p-value (< 0.05) indicates that we can reject the null hypothesis, in which case the corresponding predictor is likely to be a meaningful addition (or is *statistically significant*) to your model. Likewise, a larger (insignificant) p-value suggests that changes in the predictor are not associated with (or do not help explain) changes in the response. Thus, for our analysis, we use the coefficient p-values to determine which variables are useful for our final model.

As it is true for any model, part of the process involves checking to make sure that the data we want to analyze can actually be done using the chosen model. For a linear model it is required that, for each value of the independent variable, the distribution of the dependent variable must be normal. Typically, we plot the errors (residuals) to see if they follow a normal distribution. A QQ- plot is an example of such a residual plot that can be used to reveal biased results more effectively than a simple computation. Further, the variance of the distribution of the dependent variable should be constant for all values of the independent variable. Finally, the relationship between the dependent variable and the independent variables should be linear, and all observations should be

independent. In brief, the residuals of a good model should be normally and randomly distributed.

In the event the response variable takes a form where the residuals look completely different from a normal distribution, it is preferable to consider another class of models known as *generalized linear models (GLM)*; in which case the response variable Y_i follows an exponential family distribution. Logistic regression is an example of a GLM as we define it, next.

4.1.5 Binomial Logistic regression

Binomial Logistic regression which is simply called a *logistic regression* estimates the probability of an occurrence of an event Y_i based on a set of predictors X_i. The basic mathematical concept behind logistic regression is *logit* which is the natural logarithm (ln) of an odds; and odds are ratios of probability "success" p (for instance, an ambulance was needed) to probability "failure" $1-p$ (when no ambulance was needed, for instance). Thus, given a response categorical variable Y and m predictors X_i, we have

$$logit(Y) = log(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{m} \beta_i X_i \tag{4.1}$$

where β_0 is the Y intercept (i.e., mean of Y independent of X_i's) and β_i's are the *regression coefficients* (or *parameter estimates*) for each predictor X_i, for $i = 1, \ldots, m$.

We note that, by taking exponential (or antilog) of both sides of equation 4.1, we derive the equation to predict the probability of the occurrence of an outcome of interest as follows:

$$\begin{aligned} p &= Probability\ (Y = outcome\ of\ interest \mid X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_m = x_m) \\ &= \frac{e^{\beta_0 + \sum_{i=1}^{m} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{m} \beta_i X_i}}, \end{aligned}$$

where $e \sim 2.71828$ is the natural base.

Interpretation :

(i) The sign $(\pm)$ of a coefficient (or slope) β_j gives the direction of the relationship

(negative or positive) between the predictor X_j's and the logit of Y.

(*ii*) The intercept or log average odd $\beta_0 = log(\frac{p}{1-p})$ is an estimate of the model (null model) if we consider no predictor; this is also known as *unconditional log odds* of the response. Thus, the **average odd** is e^{β_0} and the **average probability of success**, p is $\frac{e^{\beta_0}}{1+e^{\beta_0}}$.

(*iii*) The coefficient β_j, for some predictor X_j. Fixing the levels of the remaining predictors X_k, $k \neq j$, this value gives the log(odds) of the effect of X_j on Y (beyond the average) for each unit increase (in a scale variable) or in comparison to a fixed (base) level in X_j. Thus, for a predictor X_j, the **estimated odds** value is e^{β_j} and the **percentage change** in odds (per unit increase or relative to a base level) is

$$(e^{\beta_j} - 1) \times 100\%.$$

As related to inferential statistics, a *null hypothesis* would state that, for some $\beta_j = 0$, $j > 0$, i.e., there is no linear relationship between logit of Y and X_j, in the population. So, rejecting such a null hypothesis would imply that a linear relationship exists between logit of Y and X_j. As indicated earlier for linear regression, we will rely on the p-values and the alpha level of .05, to help make our decision on the significance of the coefficients.

4.1.6 Multinomial Logistic regression

Multinomial logistic regression (or *multinomial regression*) is used to predict a nominal dependent variable (with two or more factors or categories) given one or more independent variables. As such, it is an extension of binomial logistic regression to allow for a dependent variable with more than two categories.

4.1.7 R-squared

Also known as **coefficient of determination**. it is a statistical measure of how close the data are to the fitted regression line. In other words, it is the percentage of the response

variable variation that is explained by a linear model in which case

$$R^2 = \frac{Explained\ variation}{Total\ variation} \times 100$$

0% indicates that the model explains none of the variability of the response data around its mean and 100% indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits your data but there are risks of "overfitting" or bias, which makes the model less adaptable to a different data taken under a similar circumstance.

4.1.8 Pseudo R-squared

As opposed to an R-squared value that is obtain from evaluating a model built on a continuous response, such an indicator does not make sense for models built on an ordinal response where the variance is fixed instead. However, a similar metric (in scale) called a "Pseudo" R-squared is used for models such as logistic regressions. In which case, the higher the value the better model but they are only meaningful when comparing these values for distinct models. There are several such pseudo R-squared values but SPSS software returns the values for Nagelkerke, and Cox & Snell (Pseudo) R-squareds.

4.1.9 Confusion Matrix

In the area of *machine learning* when it comes to statistical classification we often rely on a confusion matrix (or *error matrix*) which gives the performance of a classifier or supervised learning algorithm; neural network, which we define later, is an example of a classifier. The **confusion table** or **confusion matrix** is a 2 matrix with the number of **true positives** (TP; hit) and **true negatives** (TN; correct rejection) on row 1 and the number of **false positives** (FP; false alarm or Type I error) and **false negatives** (FN; miss or Type II error) on row 2, respectively by columns. The performance of a classifier will be measured with the following statistics:

4.1.10 Sensitivity

It is the measure of the proportion of actual positives (TP) that are correctly identified. (e.g., the percentage of injured bikers who are correctly identified as being injured)

Thus, **sensitivity** or true positive rate (TPR) is given by

$$TPR = \frac{TP}{TP + FN}.$$

where

4.1.11 Specificity

It is the measure of the proportion of actual negatives that are correctly identified. (e.g., the percentage of bikers who suffered no accident related injury and who are correctly identified as not injured).

Thus, **specificity** or true negative rate (TNR) is given by

$$TNR = 1 - TPR = \frac{TN}{TN + FP}.$$

4.1.12 Receiver Operating Characteristic (ROC)

This is a plot of the diagnostic ability of the classifier system as we vary its discrimination threshold (or cut-points). Thus, a curve is obtained as we plot the true positive rate (TPR) against the false positive rate (FPR) at various cut points. In general, the closer the curve is to the top left corner in the plane, the better the classification as shown in Figure 4.1.

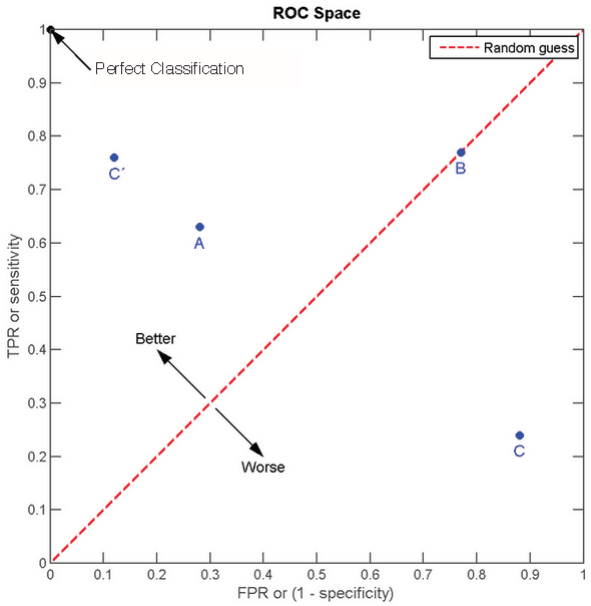

Figure 4.1: An ROC curve space

In order to check the performance of our classifier, we will rely on the **AUC (Area Under Curve)** of the ROC curve; this is a measure of discrimination or diagnostics. As such, a higher AUC, the better the model at distinguishing between say, injured bikers vs non-injured bikers, following an accident. Thus, an excellent classifier has AUC ~ 1 while a poor classifier has AUC ~ 0.

4.1.13 Neural Network

This is a sophisticated classifier that is applied to a data when the nature of the relationship between the predictors and the response is not clear; this relationship is learned through repetitive "training" methods. For example, *gradient methods* such as *gradient descent* (on a loss function) are used to train multilayer networks by updating weights to minimize loss.

Following these definitions, we begin by answering some specific questions that are related to the data using analytical methods in each upcoming sections.

4.2 Neural Network (Multilayer Perceptron)

Question n° 1. *Does the data contain enough information to help predict the need of an ambulance when a bicycle accident occurs?*

Procedure: Using Neural Network, we test the strength or the performance of any classifier built on Ambulance against the predictors in the data.

4.2.1 Dependent: Ambulance

We include all variables as independent in the neural network algorithm except ambulance (0=No, 1=Yes) which is used as dependent.

Classification

Sample	Observed	Predicted 0	Predicted 1	Percent Correct
Training	0	449	1452	23.6%
	1	163	3837	95.9%
	Overall Percent	10.4%	89.6%	72.6%
Testing	0	196	605	24.5%
	1	61	1653	96.4%
	Overall Percent	10.2%	89.8%	73.5%

Figure 4.2: Neural Network Classification Output

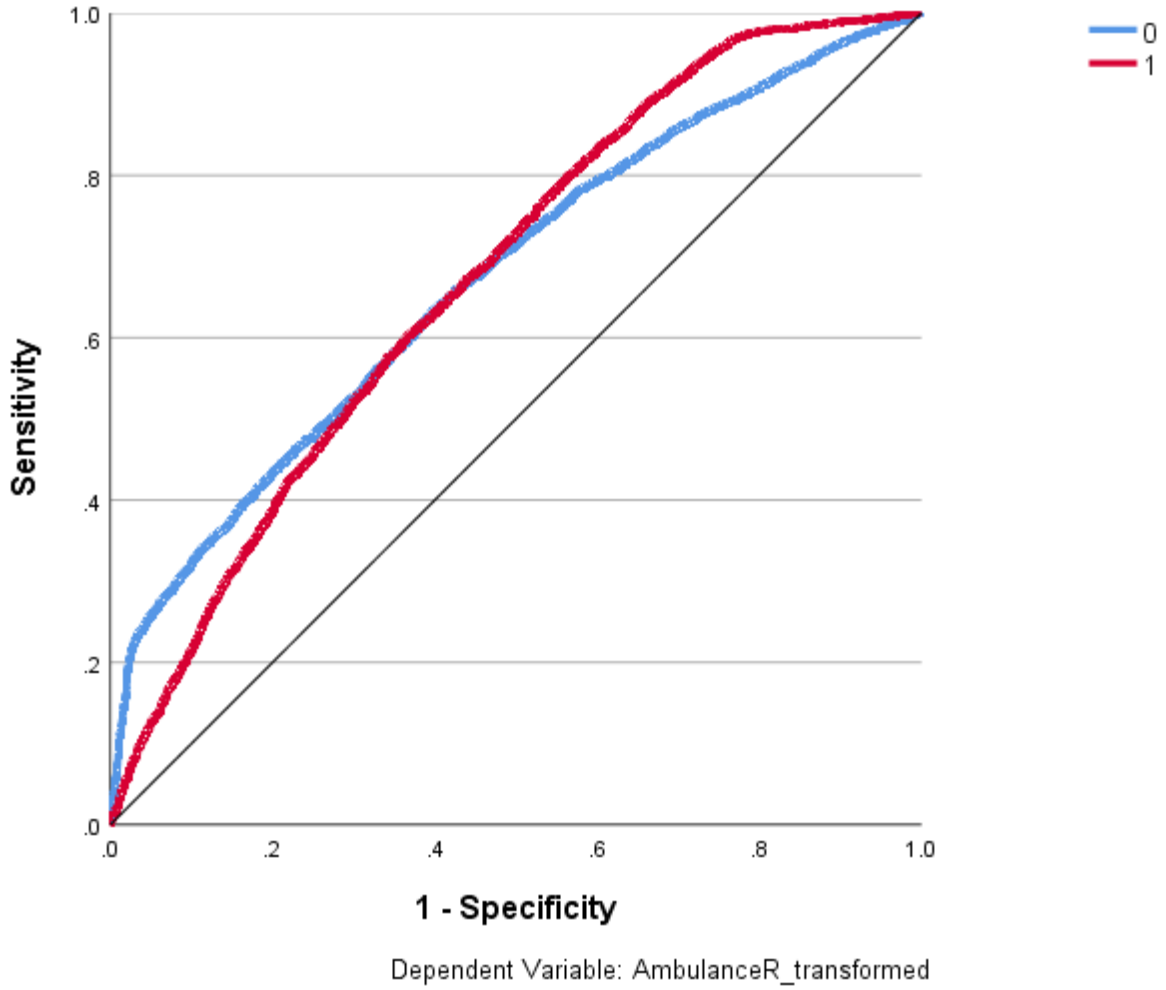

Figure 4.3: ROC curve for classification of ambulance requirement

		Area
AmbulanceR_transformed	0	.673
	1	.673

Figure 4.4: AUC output for ambulance requirement

Conclusion: We found that the variables used in the data can predict the outcome variable correctly 73% of the times. Both the training and testing data agree with this prediction level as shown in Figure 4.2. This means that the performance level would hold if applied to the larger population. Thus, we think that the data is sound, and it is reasonable to consider a classification model on Ambulance against other predictors in

the data.

4.2.2 Dependent: Bike Injuries

Question n° 2. *Does the data contain enough information to help predict the different (5) levels of injuries when a bicycle accident occurs?*

Procedure: Similarly, using Neural Network, we test the performance of any classifier that is built on Injury against any other predictor in the data. We observe two separate results, depending on the number of levels of injuries.

i. Five classes

From the Neural Network output, most levels of injuries cannot be classified. In which case the percent of prediction is below 50% for both training and testing data as shown in Figures 4.5

		Predicted
Sample	Observed	Percent Correct
Training	No injury	0.0%
	Possible Injury	45.0%
	Evident Injury	71.1%
	Disabling Injury	0.0%
	Killed	0.0%
	Overall Percent	48.8%
Testing	No injury	0.0%
	Possible Injury	43.7%
	Evident Injury	67.5%
	Disabling Injury	0.0%
	Killed	0.0%
	Overall Percent	47.7%

Dependent Variable: BikeInjury_transformed

Figure 4.5: Neural Network Classification Output for five injury levels

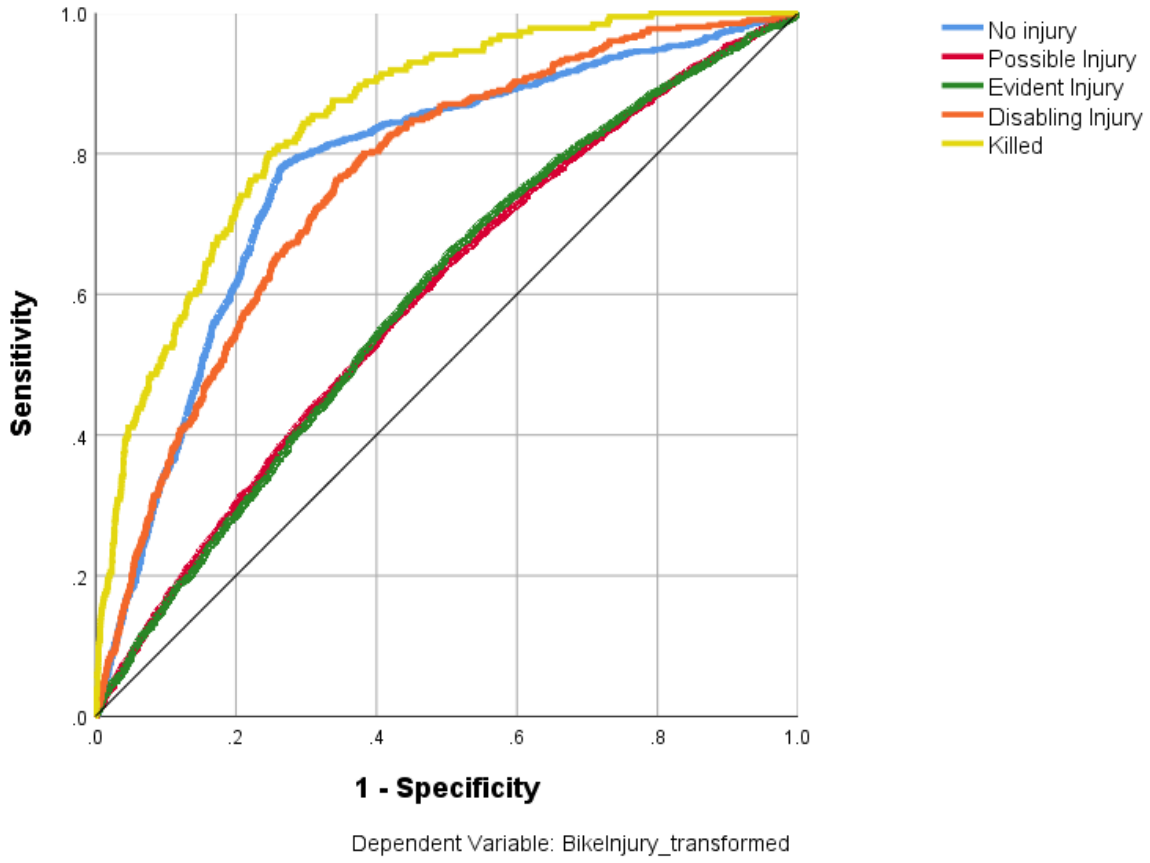

Figure 4.6: ROC curve for classification of five injury levels

		Area
BikeInjury_transformed	No injury	.774
	Possible Injury	.595
	Evident Injury	.596
	Disabling Injury	.760
	Killed	.849

Figure 4.7: AUC output for five injury levels

Conclusion: Given the performance of the model, we can think that either the data does not have sufficient predictors or would need more observations in order to be robust enough for such an analysis.

For this reason, we decided to recode the variable Injury level, down from five classes to three classes as shown in Figure 4.8.

ii. Three classes

		Frequency	Parameter coding (1)	Parameter coding (2)
Bicyclist's Injury	No injury	810	1.000	.000
	Possible or Evident Injury	7022	.000	1.000
	Disabling Injury or Killed	586	.000	.000

Figure 4.8: Injury levels recoded

With the newly recoded variable, the Neural Network output indicates an overall 83% predictive strength of the three injury levels, consistently for both training and testing data, as shown in Figure 4.9. However, as injury or possible injury occurrences are predictable 100% of the time, other levels such as no injury or death are highly unpredictable; this does not come as a surprise since these events are heavily influenced by many other factors, perhaps not recorded or accounted for.

Sample	Observed	Predicted: No injury	Predicted: Possible or Evident Injury	Predicted: Disabling Injury or Killed	Percent Correct
Training	No injury	1	541	0	0.2%
	Possible or Evident Injury	2	4955	1	99.9%
	Disabling Injury or Killed	0	423	1	0.2%
	Overall Percent	0.1%	99.9%	0.0%	83.7%
Testing	No injury	0	268	0	0.0%
	Possible or Evident Injury	4	2057	1	99.8%
	Disabling Injury or Killed	0	162	0	0.0%
	Overall Percent	0.2%	99.8%	0.0%	82.5%

Dependent Variable: Bicyclist's Injury

Figure 4.9: Neural Network Classification Output for three injury levels

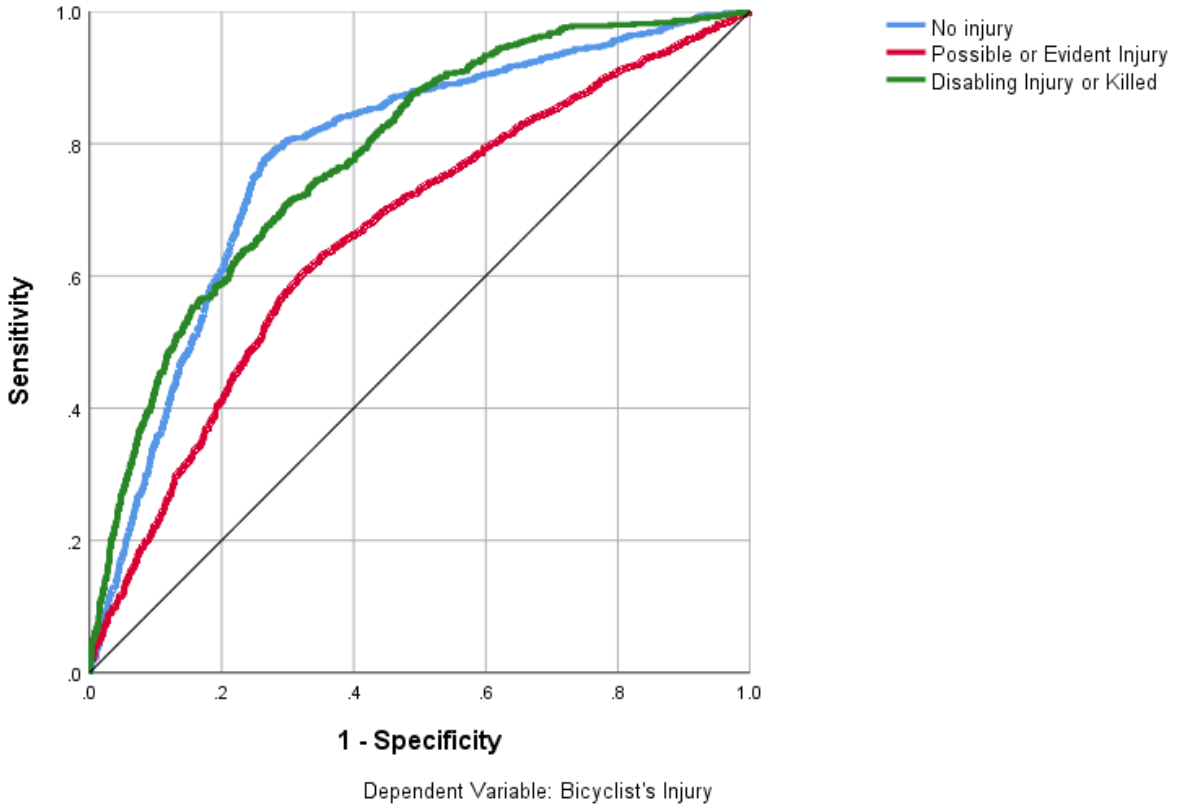

Figure 4.10: ROC curve for classification of three injury levels

		Area
Bicyclist's Injury	No injury	.779
	Possible or Evident Injury	.663
	Disabling Injury or Killed	.781

Figure 4.11: AUC output for three injury levels

Conclusion: Given the overall predictive power (83%) of the classifier relative to the (3) Injury levels, we consider the data sound for analysis.

For each of the previous questions, the response variable is nominal, it makes sense to use binary and multinomial logistic regressions.

4.3 Logistic Regressions

Question n° 3. *Do bikers' injury levels help predict the use of ambulance when accidents occur?*

4.3.1 Ambulance vs Injury levels

The first model in the output (Figure 4.12) is a null model, that is, a model with no predictors.

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.749	.023	1028.741	1	.000	2.114

Figure 4.12: Logistic Regression on Ambulance - Null Model

Figure 4.13 gives the overall test for the model that includes the predictors which are significant with a p-value ~ 0.000. This means our model as a whole fits significantly better than the null model. The logistic regression coefficients give the change in the log odds of the outcome Ambulance for biker's injury levels 1 and 2 compared to biker's injury level 3.

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Bicyclist's Injury			741.705	2	.000	
	Bicyclist's Injury(1)	-4.015	.190	444.649	1	.000	.018
	Bicyclist's Injury(2)	-1.795	.172	109.102	1	.000	.166
	Constant	2.697	.170	252.172	1	.000	14.838

Figure 4.13: Logistic Regression on Ambulance - Variables Output

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	9550.779[a]	.114	.159

Figure 4.14: Logistic Regression on Ambulance - Model Performance Summary

Model Equation:.

$$\ln(y) = 2.7 - 4.02x_1 - 1.8x_2$$

where y :=*Predicted odds of Ambulance*

x_1:=Injury-level1(No Injury)

x_2:= Injury-level2(Possible or Evident Injury)

Interpretations:

Having been involved in a bicycle's accident, the odds of getting an ambulance with a level 2 injury compared to injury level 3 decreases by $(1 - .2) \times 100\% = 80\%$ while the odds of getting an ambulance with a level 1 injury decreases by $(1 - .02) \times 100\% = 98\%$ compared to injury level 3.

Conclusion:

Because the odds of getting an ambulance increase significantly as the level of injury increases, we conclude that there is a linear relationship between Ambulance use and level of injuries.

Further, it is reasonable to assume that Ambulance is very related to some level of injury, for the remaining analysis, we use Injury levels as our response variable and drop Ambulance from the model.

4.3.2 Injury levels vs Traffic Controls

Question n° 4. *Does the presence of some traffic control (or the lack thereof) help explain the (3) levels of injuries of the bikers?*

With this question, we seek to understand the effect of traffic control on the levels of injuries as we disregard other factors.

This is a multinomial logistic regression table that details a likelihood ratio test with traffic control presence of a bicycle accident (Figure 4.15), which indicates statistical significance.

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	30.299	.000	0	.
Traffic Control Presents	44.075	13.776	2	.001

Figure 4.15: Multinomial Logistic Regression on Traffic Control-Likihood Ratio

Cox and Snell	.147
Nagelkerke	.217
McFadden	.141

Figure 4.16: Multinomial Logistic Regression on Traffic Control-Pseudo R^2

Bicyclist's Injury[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B) Lower Bound	Upper Bound
No injury	Intercept	.265	.077	11.807	1	.001			
	[Traffic Control Presents=0]	.110	.109	1.016	1	.313	1.116	.901	1.383
	[Traffic Control Presents=1]	0[b]	.	.	0	.	.	.	.
Possible or Evident Injury	Intercept	2.546	.060	1778.509	1	.000			
	[Traffic Control Presents=0]	-.147	.087	2.894	1	.089	.863	.728	1.023
	[Traffic Control Presents=1]	0[b]	.	.	0	.	.	.	.

Figure 4.17: Multinomial Logistic Regression on Traffic Control-Variable Output

Model Equation:

$$\ln(y_1) = .3 + .11x_1$$

$$\ln(y_2) = 2.55 - .15x_1$$

where,

y_1 :=*Predicted odds of Level 1 injury*

y_2 :=*Predicted odds of Level 2 injury*

x_1=Injury-level1

x_2= Injury-level2

Interpretations:

An odds ratio > 1 indicates that the risk of the outcome falling in the comparison group relative to the risk of the outcome falling in the referent group increases as the variable increases. In other words, the comparison outcome is more likely. An odds ratio < 1 indicates that the risk of the outcome falling in the comparison group relative to the risk of the outcome falling in the referent group decreases as the variable increases. In general, if the odds ratio < 1, the outcome is more likely to be in the referent group. Given the other variables in the model are held constant with bicycle accidents, the odds of the incidents with no injury occurring at intersections without traffic controls increases by .110 relative to level 3 injuries. In other words, level *1* injury compared to injury level *3* increases by $(1.11 - 1) \times 100\% = 11\%$ while the odds of a level 2 injury decreases by $(1 - .86) \times 100\% = 14\%$ compared to injury level 3 when there are no traffic controls present. Restated, the odds of bicycle injury where traffic controls exist suggest both an increase and a decreases, and that some other factors or covariates affect the level of injuries. Because the p-value$> .05$ for each variable in the model, we fail to reject the the null hypothesis at a 5% level of risk.

Conclusion:

There is no linear relationship between being injured (or not) and traffic control at a standard 5% level of risk. However, there may be some relationship between traffic control and the level of severity of an injury, especially when other factors are consider. This is why we considered the next analysis.

4.3.3 Injury levels vs Other factors

With regard for the other factors, this is a multinomial logistic regression which considers all other factors for the model built on (3) Injury levels. Neither number of lanes, gender, race and age of bicyclists and driver, road conditions, substance involvement, weather conditions, work zone, were statistically significant at explaining the different levels of injuries for bicyclists who crashed with a motor vehicle. Figure 4.18 is a list of the factors that are significant in our model. (See Appendix 5.2 for the full list of factors and estimates).

Bicyclist's Injury[a]		B	Std. Error	Wald	df	Sig.	Exp(B)
No injury	Intercept	-3.746	.464	65.130	1	.000	
	Bicyclist's Age	-.271	.059	21.185	1	.000	.762
	Driver's Estimated Speed (in MPH)	.210	.020	110.398	1	.000	1.234
	Speed Limit Average (in MPH)	.136	.047	8.469	1	.004	1.145
	[Lighting Conditions=Not Daylight]	-.499	.134	13.966	1	.000	.607
	[Lighting Conditions=Daylight]	0	.	.	0	.	.
	[Road Characteristics=Curve]	-.524	.242	4.667	1	.031	.592
	[Road Characteristics=Straight]	0	.	.	0	.	.
	[Traffic Control Presents=No Control Present]	.244	.119	4.195	1	.041	1.277
	[Traffic Control Presents=Control Present]	0	.	.	0	.	.
Possible or Evident Injury	Intercept	.112	.342	.107	1	.743	
	Bicyclist's Age	-.218	.044	24.628	1	.000	.804
	Driver's Estimated Speed (in MPH)	.174	.014	150.352	1	.000	1.190
	Speed Limit Average (in MPH)	.147	.036	16.567	1	.000	1.158
	[Lighting Conditions=Not Daylight]	-.484	.096	25.333	1	.000	.617
	[Lighting Conditions=Daylight]	0	.	.	0	.	.
	[Road Characteristics=Curve]	-.476	.160	8.833	1	.003	.621
	[Road Characteristics=Straight]	0	.	.	0	.	.
	[Traffic Control Presents=No Control Present]	-.027	.091	.088	1	.767	.973
	[Traffic Control Presents=Control Present]	0	.	.	0	.	.

a. The reference category is: Disabling Injury or Killed.

Figure 4.18: Multinomial Logistic Regression on Traffic Control-Variable Output

Question n° 5. *What factors help explain the levels of injuries of the bikers?*

Following from Figure 4.18, we have obtained the following.

Model Equation:

$$\ln(y_1) = -4 + .27x_1 + .21x_2 + .14x_3 - .5x_4 - .52x_5 + .24x_6$$

$$\ln(y_2) = .11 - .22x_1 + .17x_2 + .15x_3 - .48x_4 - .48x_5 - .03x_6$$

where,

y_1 :=*Predicted odds of Level 1 injury (No injury)*

y_2 :=*Predicted odds of Level 2 injury (Possible or Evident Injury)*

x_1= Biker's Age

x_2= Driver's Estimated Speed

x_3= Average Speed Limit

x_4= Lighting Conditions

x_5= Road Characteristics

x_6= Traffic Controls

Interpretations:

The odds of being injury level 1 vs injury level 3 decreases by 27% as the age of the biker increases.

The odds of being injury level 2 vs injury level 3 decreases by 22% as the age of the biker increases.

The odds of injury level 1 vs injury level 3 increases by 21% for each 1 unit increase in driver's estimated speed.

The odds of being injury level 2 vs injury level 3 increases by 17% for each 1 unit increase in driver's estimated speed.

The odds of injury level 1 vs injury level 3 increases by 14% for each 1 unit increase in average speed limit.

The odds of being injury level 2 vs injury level 3 increases by 15% for each 1 unit increase in average speed limit.

The odds of being injury level 1 vs injury level 3 decreases by 50% when changing from having lighting to having no lighting.

The odds of being injury level 2 vs injury level 3 decreases by 48% when changing from having lighting to having no lighting.

The odds of being injury level 1 vs injury level 3 decreases by 52% when changing from a straight to a road with curves.

The odds of being injury level 2 vs injury level 3 decreases by 48% when changing from a straight to a road with curves.

Lastly, the odds of the incidents with no injury occurring at intersections without traffic controls increases by 28% relative to level 3 injuries while the odds of a level 2 injury decreases by $(1 - .97) \times 100\% = 3\%$ compared to injury level 3 when there are no

traffic controls present.

Conclusion: The average posted speed limit shows that incidents occur when the average speed of 32.5 accounts for 45% of the accidents total. Lighting conditions and injury severity of BMVC have a negative correlation injury even though more incidents occur during the day. Driver's wariness and prudence are plausible justifications where there is low to no light source present. This may also be the case where there is a straight road versus a curve road involved.

Chapter 5 Conclusion and Recommendations

This thesis was inspired by a recent work by Estime [4] as she studied factors that influence bicyclists injury severity levels at unsignalized intersections in North Carolina. In her results, based on 1273 observations, Estime found that light conditions, age (55 or older), driver's speed, road features, type of day and time of year are factors. We explore the bicyclists' injuries, we found a relationship between ambulance being properly used or sent with the level of severity of the injury. We fix the data which helps to increase the total observations to 8418. This significant increase in the sample size had allowed us to run a neural network algorithm on the data to access its predictive strength. With 5 levels of injuries, the data is not suitable for prediction, but with 3 levels of injuries, we found an overall predictive power of 83% accuracy, given the variables in the data. With this information, we were able to address some analysis questions as outlined in Chapter 1, Section 1.1. From these questions, we found that the odds of getting an ambulance increase with the levels of injuries. So, we decided not to keep Ambulance in any model built on levels of injuries. Traffic control alone was not significant at predicting when an accident will result in an injury or not. Thus, it became obvious that other factors were involved in predicting the levels of injuries. We found Biker's Age, Speed Limit, Driver's Estimated Speed, Light Condition (daylight compared to no daylight), and Road Characteristics (straight compared to curve) are significant factors at explaining the levels of injuries. As with Estime's analysis biker's age and lighting are contributing factors; however, road features (intersections, etc..), time of the year play no role. It is clear from our final analysis that drivers ought to slow down, drive within the speed limit, especially on curved roads. These preventive measures will likely reduce the level of severity of the accidents and possibly save lives.

Bibliography

[1] Asgarzadeh, M., Verma, S., Mekary, R. A., Courtney, T. K., & Christiani, D. C. (2017). The role of intersection and street design on severity of bicycle-motor vehicle crashes. Injury Prevention, 23(3), 179-185. doi:10.1136/injuryprev-2016-042045

[2] Cao, Z. M., Shen, J. J., & Wang, Q. (2015). Correlation model between speed and density of electric bicycles at signalized intersections. Applied Mechanics and Materials, 744-746, 1803-1807. doi:10.4028/www.scientific.net/AMM.744-746.1803

[3] Chen C., Anderson J.C., Wang H., Wang Y., Vogt R., & Hernandez S. (2017) How bicycle level of traffic stress correlate with reported cyclist accidents injury severities: A geospatial and mixed logit analysis Accident Analysis and Prevention, 108 , pp. 234-244.

[4] Estime, S. (2019). Advocating Safety for Bicyclists at Intersections: Investigating Factors that Influence Bicyclist Injury Severity at Unsignalized Intersections in North Carolina, Masters Thesis, Elizabeth City State University, North Carolina.

[5] Jannat, Mafruhatul, Hurwitz, D. Monsere, C. & Funk, K. "The Role of Driver's Situational Awareness on Right-hook Bicycle-motor Vehicle Crashes." Safety Science 110 (2018): 92-101. Web.

[6] Jiang, L., Jiang, H., Ma, Y., Chen, G., & Wang, D. (2018). Evaluation of the dispersion effect in through movement bicycles at signalized intersection via cellular automata simulation. Physica A: Statistical Mechanics and its Applications, 498, 138-147. doi:10.1016/j.physa.2017.12.130

[7] Kim, M., Kim, E., Oh, J., & Jun, J. (2012). Critical factors associated with bicycle accidents at 4-legged signalized urban intersections in South Korea. KSCE Journal of Civil Engineering, 16(4), 627-632. doi:10.1007/s12205-012-1055-1

[8] Madsen, T. K. O., & Lahrmann, H. (2017). Comparison of five bicycle facility designs in signalized intersections using traffic conflict studies. Transportation Research Part F: Psychology and Behaviour, 46, 438-450. doi:10.1016/j.trf.2016.05.008

[9] 2014 North Carolina Strategic Highway Safety Plan. Retrieved from `http://www.ncshsp.org/wp-content/themes/SHSP_Custom/pdfs/SHSP_Complete.pdf`

[10] Ohlms, P. B., & Kweon, Y. (2018). Facilitating bicycle travel using innovative intersection pavement markings. Journal of Safety Research, 67, 173-182. doi:10.1016/j.jsr.2018.10.007

[11] Ou, H., Tang, T., Rui, Y., & Zhou, J. (2018). Electric bicycle management and control at a signalized intersection. Physica A: Statistical Mechanics and its Applications, 512, 1000-1008. doi:10.1016/j.physa.2018.06.116

[12] Phillips, R. O., Bjørnskau, T., Hagman, R., & Sagberg, F. (2011). Reduction in car–bicycle conflict at a road–cycle path intersection: Evidence of road user adaptation? Transportation Research Part F: Psychology and Behaviour, 14(2), 87-95. doi:10.1016/j.trf.2010.11.003

[13] Portilla, C., Valencia, F., Espinosa, J., Núñez, A., & De Schutter, B. (2016). Model-based predictive control for bicycling in urban intersections. Transportation Research Part C, 70, 27-41. doi:10.1016/j.trc.2015.11.016

[14] Schepers, J. P., Kroeze, P. A., Sweers, W., & Wüst, J. C. (2011). Road factors and bicycle–motor vehicle crashes at unsignalized priority intersections. Accident Analysis and Prevention, 43(3), 853-861. doi:10.1016/j.aap.2010.11.005

[15] Stipancic, J., Zangenehpour, S., Miranda-Moreno, L., Saunier, N., & Granié, M. (2016). Investigating the gender differences on bicycle-vehicle conflicts at urban intersections using an ordered logit methodology. Accident Analysis and Prevention, 97, 19-27. doi:10.1016/j.aap.2016.07.033

[16] Strauss, J., & Miranda-Moreno, L. F. (2013). Spatial modeling of bicycle activity at signalized intersections. Journal of Transport and Land use, 6(2), 47-58. doi:10.5198/jtlu.v6i2.296

[17] Tang, T., Rui, Y., Zhang, J., & Wang, T. (2018). Impacts of group behavior on bicycle flow at a signalized intersection. Physica A: Statistical Mechanics and its Applications, 512, 1205-1215. doi:10.1016/j.physa.2018.08.022

[18] Wang, K., & Akar, G. (2018). The perceptions of bicycling intersection safety by four types of bicyclists. Transportation Research Part F: Psychology and Behaviour, 59, 67-80. doi:10.1016/j.trf.2018.08.014

[19] Wang, Y., & Nihan, N. L. (2004). Estimating the risk of collisions between bicycles and motor vehicles at signalized intersections. Accident Analysis and Prevention, 36(3), 313-321. doi:10.1016/S0001-4575(03)00009-5

[20] Warner, J., Hurwitz, D. S., Monsere, C. M., & Fleskes, K. (2017). A simulator-based analysis of engineering treatments for right-hook bicycle crashes at signalized intersections. Accident Analysis and Prevention, 104, 46-57. doi:10.1016/j.aap.2017.04.021

[21] Zhao, J., Yan, J., & Wang, J. (2019). Analysis of alternative treatments for left turn bicycles at tandem intersections. Transportation Research Part A, 126, 314-328. doi:10.1016/j.tra.2019.06.020

Appendix

Name	Type
OBJECTID	Numeric
AmbulanceR	String
BikeAge	String
BikeAgeGrp	String
BikeAlcDrg	String
BikeAlcoho	String
BikeDir	String
BikeInjury	String
BikePos	String
BikeRace	String
BikeSex	String
City	String
County	String
CrashAlcoh	String
CrashDate	String
CrashDay	String
CrashGrp	String
CrashHour	Numeric
CrashLoc	String
CrashMonth	String
CrashSevr	String
CrashTime	String
CrashType	String
CrashYear	Numeric
Developmen	String
DrvrAge	String
DrvrAgeGrp	String
DrvrAlcDrg	String
DrvrAlcoho	String
DrvrEstSpd	String
DrvrInjury	String
DrvrRace	String
DrvrSex	String

Name	Type
DrvrVehTyp	String
DstncMiFrm	String
FrmRd	String
HitRun	String
Latitude	Numeric
LightCond	String
Locality	String
Longitude	Numeric
NumBikesAl	String
NumBikesBl	String
NumBikesCi	String
NumBikesKi	String
NumBikes...	String
NumBikesTo	String
NumBikesUl	String
NumLanes	String
NumUnits	Numeric
OnRoad	String
RdCharacte	String
RdClass	String
RdConditio	String
RdConfig	String
RdDefects	String
RdFeature	String
RdSurface	String
Region	String
RteInvdCd	String
RuralUrban	String
SpeedLimit	String
Towrd_Rd	String
TraffCntrl	String
Weather	String
Workzone	String

Figure 5.1: Original Variables with Original Type

Ambulance vs.	**Correlation**	**P-value**	**Written Results**
Bike Injury	.34	.000	There is a significant positive relationship between ambulanceR and Bike Injury, r(8186)=.34, p = .000
Biker's Age	.046	.000	There is a significant positive relationship between ambulanceR and Biker's Age, r(8186)=.046, p = .000
Crash Severity	.361	.000	There is a significant positive relationship between ambulanceR and Crash Severity, r(8186)=.34, p = .000
Driver's Age	-.019	.076	There is a weak negative relationship between ambulanceR and Driver's Age r(8416)=-.019, p = .076
Driver's Estimated Speed	-.111	.000	There is a significant negative relationship between ambulanceR and Driver's Estimated Speed, r(8416)=-.111, p = .000
Hit and Run	.081	.000	There is a some positive relationship between ambulanceR and Bike Hit and Run r(8416)=..081, p = .000
Lighting Conditions	-.061	.000	There is a weak negative relationship between ambulanceR and Lighting conditions r(8416)=-.061, p = .000
Number of Lanes	.004	.691	There is no relationship between ambulanceR and the Number of Lanes r(8416)=.004, p = .691
Road Characteristics	-.023	.037	There is a weak negative relationship between ambulanceR and Road Characteristics, r(8370)=.34, p = .000
Road Conditions	-.044	.000	There is a weak negative relationship between ambulanceR and Road Conditions, r(8416)=-.044, p = .000
Substance Involvement	-.018	.105	There is no significant relationship between ambulanceR and Substance Involvement, r(8416)=-0.018, p = .105
Speed Limit	-.017	.113	There is a significant positive relationship between ambulanceR and Speed Limit, r(8416)=-.017, p = .113
Traffic Control Present	.008	.491	There is no relationship between ambulanceR and Traffic Control Presence, r(8263)=.008, p = .491
Weather	-.020	.068	There is no significance between ambulanceR and Weather, r(8416)=-.020, p = .068
Workzone	-015	.171	There is a significant positive relationship between ambulanceR and Workzone, r(8416)=-.015, p = .171

Bicyclist's Injury[a]		B	Std. Error	Wald	df	Sig.	Exp(B)
No injury	Intercept	-2.697	.439	37.794	1	.000	
	Driver's Estimated Speed (in MPH)	.248	.019	171.364	1	.000	1.282
	Bicyclist's Age	-.310	.056	30.810	1	.000	.733
	Speed Limit Average (in MPH)	.129	.044	8.471	1	.004	1.138
	[Lighting Conditions=Not Daylight]	-.647	.127	26.094	1	.000	.524
	[Lighting Conditions=Daylight]	0	.	.	0	.	.
	[Road Characteristics=Curve]	-.609	.228	7.119	1	.008	.544
	[Road Characteristics=Straight]	0	.	.	0	.	.
	[Traffic Control Presents=No Control Present]	.229	.114	4.069	1	.044	1.257
	[Traffic Control Presents=Control Present]	0	.	.	0	.	.
Possible or Evident Injury	Intercept	.225	.337	.446	1	.504	
	Driver's Estimated Speed (in MPH)	.184	.014	173.922	1	.000	1.202
	Bicyclist's Age	-.224	.044	26.211	1	.000	.800
	Speed Limit Average (in MPH)	.150	.035	17.805	1	.000	1.161
	[Lighting Conditions=Not Daylight]	-.522	.095	30.271	1	.000	.593
	[Lighting Conditions=Daylight]	0	.	.	0	.	.
	[Road Characteristics=Curve]	-.514	.158	10.571	1	.001	.598
	[Road Characteristics=Straight]	0	.	.	0	.	.
	[Traffic Control Presents=No Control Present]	-.037	.090	.165	1	.684	.964
	[Traffic Control Presents=Control Present]	0	.	.	0	.	.

a. The reference category is: Disabling Injury or Killed.

Figure 5.2: Multinomial Logistic Regression on Other Factors - Variable Output