Nicolas Mavromatis (nima6629@colorado.edu)

CSPB 4502

Bike Accidents in Great Britain-Data Mining Project

Part 3 Project Progress Report

Updated Proposal

In the U.S., Bicycle accidents account for significant deaths, with about 130,000 crashes and 1,000 killed annually. The cost is estimated to exceed $23 billion each year in the U.S. Alarmingly, mortalities from bicycle accidents reached a 28 year high in 2018, perhaps partially due to the increasing popularity of biking [1]. Great Britain, the location of the data collected for this study, has similar issues with serious injuries rising by 26% from 2004 to 2020, and bicyclist traffic growing by 96% in this same timeframe [6]. This indicates the need for an exploration into the causes of such accidents, especially because existing studies are very limited. It would be useful to identify associations between factors that tend to result in accidents, in order to avoid them in the future. Perhaps policy could change after being guided by the results to increase safety. Data can be muddy and difficult to draw conclusions from without a focused discipline; in particular, it can be hard to attribute a cause to a specific attribute when there are many factors involved. Accidents can occur under a great variety of conditions, including weather, road conditions, time of day, and the age and gender of the bicyclist which result in varying accident severity. It is therefore imperative to uncover significant patterns associated with crashes to identify which factors tend to cause severe accidents. Data mining is an excellent approach and powerful tool set to better understand the main causes and associated factors of bicycle accidents and mortalities.

Data mining is exceptional at identifying patterns and relationships in large volumes of data, to extract useful information and make informed predictions. The combination of analytical, mathematical, and statistical tools can result in powerful identification of trends and relationships. This discipline can lead to a deep understanding of the data in order to draw conclusions and deploy effective changes in the future. In this case, a variety of factors interact to contribute to bicycle accidents, confusing the analysis. It will therefore be illuminating to perform a data mining analysis and better understand the main factors that correlate with accidents.

Scant previous work has been done on this topic, but some small studies have provided interesting results. One such study looked at single-bicycle crashes in Denmark, looking at around 350 sample points. This study found that winter road maintenance is crucial to deter accidents, and that in warm weather there is no gender difference associated with accidents. Unsurprisingly, it found that the elderly, above 65 years of age, have a lower number of accidents but a higher chance of injury when they do get in an accident. It found that 18% sustain injuries from car crashes and 82% used a helmet. Significantly, 79% of crashes happen in an urban area, with around half in daylight and half at dusk or dawn. 24% of crashes happen on a dry surface. Also, half of the crashes happened in Winter. Lastly, most crashes happen during rush hour, as may be predicted. The main limitation of this study is that the sample size is very small, and it used self-reporting. It would be illuminating to see if the findings can be replicated, where applicable, by a larger data set. The data used in this project does not contain information on injuries or helmet use, so not all of these findings can be assessed, but there is a lot of the data that can be tested for replication of results [2].

Another study was conducted on crashes in Ohio from 2013-2017 to identify factors associated with injury. The data was collected from Ohio police accident reports and hospital databases. Unfortunately, it did not assess many contributing factors such as weather or road conditions/type. It did find that bicyclists 65 years or older had higher odds of sustaining upper extremity injuries, while those aged 3-24 were more likely to sustain lower extremity injuries. It also found that weather, bicyclist sex, and intersection located crashes were associated with higher rates of injury. Accident rates were lower in cold months, likely due to more protective clothing being worn. Interestingly, more than 80% of bicyclists were male, and there were more observations in warm months. This study did not analyze many attributes that are likely significant, such as road conditions, but the most important finding is that most bicyclists are male, and the elderly experience higher rates of injury [3].

A final study looked at insurance reports from Sweden, with only around 450 data points. It found that more than 75% of accidents were intersection related. It also found the highest number of collisions in May, August, and September, with most happening between 7-8 a.m. and 4-5 p.m. Almost 90% of the crashes happened in Urban areas. In 75% of the crashes, the speed limit was 50 km/h or under. Crashes during daylight were most frequent at 82%, while around 9% were during darkness and 5% during dusk or dawn. The weather was fair in about 73% of the collisions, and was rainy only about 5% of the time. This is perhaps due to the fact that weather was generally good, with dry road conditions about 70% of the time. This study had a different finding, that male and female bicyclist collisions were about 50/50, and that most accidents happened in the age of 35-45 years. This study and the others above provided useful insight that may be corroborated or contested by this project [4].

This data mining project utilizes a larger data set that can be used to replicate and assess the previous findings of the smaller studies, and also to elucidate factors associated with crashes that were not previously explored. The major aims are to analyze the data to characterize patterns associated with accidents and deaths, including age, gender, road conditions, time of day and year, weather, and light. It is crucial to assess whether one age or gender is more likely to bike and be involved in crashes, and if certain times and road conditions result in higher numbers of fatalities. Very limited work with small sample sizes have been performed, so the aim is more so to expand upon findings than replicate the exact limited findings previously available. Other studies have disagreed on the number of male or female bikers and associated accidents. This study could confirm findings such as a lower accident rate in the elderly but higher number of fatal outcomes, and that about half of crashes happen in the Winter with most happening in an intersection during rush hour. Also, most crashes were found to occur at 7-8 a.m. and 4-5 p.m. Other things to confirm are that half of the accidents occur at dusk and dawn, and that about 25% of all crashes happen in dry conditions.

There are several interesting questions to answer. Is there a specific age range or gender most associated with biking accidents? Do certain speed limit zones correlate with high numbers of accidents? Are certain years or time of day most associated with accidents? Could this be explained by brightness? Finally, do certain seasons, and therefore weather and road conditions, associate with accidents? The limited work done on this topic found some interesting results, but with only small sample sizes.

There is a fair amount of work to be done to perform this data mining task. First, the best data set available is split over two excel sheets linked by the ID attribute, so the two excel sheets should be joined. Significant data cleaning will also be necessary, such as converting dates to a workable pandas type and adding in the season attribute to make broader generalizations. There also may need to be accounting for

unknown weather and road conditions, although maybe this data can just be filtered out of the analysis. Features will be extracted by binning data by attribute, such as time, road condition, and age to characterize interesting patterns. Time will need to be intelligently binned, as there is a great variety of times present. Statistical analysis will be crucial in assessing the data, and is well presented graphically in things like scatter and bar plots. In particular, measures such as the average, standard deviation, quartiles, and chi-squared results will be useful to draw conclusions. The chi-square test is particularly useful to compare observed results with the expected to see if any relationships exist between the attributes. Linear regression will also be an important statistical analysis. Other statistical analyses will be used as necessary, but the chi-squared test will likely be the most useful statistical tool. From these findings, the goal is to characterize conditions/factors associated with bicycle accidents and mortalities to recommend safety and policy changes in the future.

The data set utilized for this project is excellent, containing bike accidents from Great Britain from 1970-2018. With over 800,000 data points and a broad date range, this could potentially be the largest data mining project focused on this effort. There are also many useful attributes, including date/time, speed limit, road/weather conditions, accident severity, fatality number, light conditions, gender and age which are crucial in finding patterns associated with bike mortalities and accidents. The data source is: https://www.kaggle.com/datasets/johnharshith/bicycle-accidents-in-great-britain-1979-to-2018 [5].

Evaluation will utilize statistical measures such as the chi-squared test and linear regression to see if any strong relationships exist between the attributes, such as severity and death number and gender. These findings will be presented in tables and graphically to characterize patterns and findings. Other statistics, such as mean and quartile results, will also undoubtedly be useful. Time, in particular, may benefit from being analyzed by quartiles, as there is such a large variety of continuous data. This data will be used to evaluate the current conditions under which bike accidents occur in order to predict and avoid them in the future.

There are several tools that will be instrumental in conducting these analyses. Initially, I considered using python and associated libraries to explore the data. My thinking was this: python is an easy language to use with many well integrated libraries. Pandas is a great tool to merge, clean, and save data like excel sheets. Numpy and scipy are useful for statistical analysis, where scipy builds on numpy with more complex algorithms. Matplotlib is the standard for detailed graphing and would be instrumental in finding patterns and presenting results. However, I quickly realized that using these tools is unwieldly for data exploration which requires a lot of tinkering and faster visualization. I therefore chose to mainly use SPSS which is very easy to use to perform statistical analysis. Chi squared and regression results are available at the click of a button. One is also able to perform data cleaning in this framework, although I split that task between pandas and SPSS. From these results, visualizations are more conveniently done in excel. I am able to do all necessary analyses in python, but there is so much exploration to do that it necessitated the use of the more convenient UI packages of SPSS and excel. I initially did some analyses in python and pandas, but found the visualization process to be much more difficult and time consuming. Lastly, github is used as the industry standard for version control. Other tools will be assessed as necessary.

Milestones Completed/Initial Results

It is important to have milestones to keep progress moving forward. Several milestones have already been completed. Data cleaning was the most difficult part of the project, and one of the most significant. Without clean, cohesive data to start with, downstream analyses are of far lower quality and might not accurately represent patterns in the data. I made significant progress cleaning the data, first by joining the two excel sheets and adding a season attribute. One excel sheet had most of the attributes, while the other had gender and age, so I combined these into one excel dataset by linking accident ID. I also converted the time and date format from an unworkable type to a standard format in pandas. The speed limit data had some errors, including a value of "660 mph" which was changed to 60 mph. There were also values that were clearly noise such as "3 mph". The solution was to bin odd valued speed limits to their closest whole number (nonzero) value, so "27 mph" became 30 mph for example. Road conditions had several attribute values with frequencies so low they would not be appropriate for chi squared analysis, so I chose to use only values of "wet" and "dry" which had high enough frequencies. Other values were simply excluded from analyses. I found that weather overlaps too much with road conditions, so it makes more sense to look at just road conditions and ignore the weather attribute. Road type turned out to be relatively useless, in that it is missing data of interest of accidents that happen in intersections, and 80% of the data occurred on the same road type, so it wouldn't yield interesting results. Most likely weather and road conditions will not be used in final analyses. One of the largest tasks was to bin data by time, as the continuous values of the variable would make it hard to draw conclusions. 15:20 vs. 15:21 would by default be treated as different values, and this is not desirable for finding patterns. It was also in an incorrect format for SPSS analysis, containing semi colons. Time was therefore binned into deciles, with 10% of the accident number placed in each bin to allow for analysis in SPSS. Unknown values of gender were deleted, as there were less than 10 of these values. Age group was recorded from a string to ordinal data, necessary to complete analysis in SPSS. Some of these changes required recoding into a new, cleaned variable in SPSS. There may be other data cleaning steps that are necessary as the project progresses.

I completed significant statistical analysis and exploration in SPSS, and found interesting results. Chi squared tests were run with different attributes and accident severity to uncover hidden patterns. Graphs were produced in excel by dividing the percent of observed data by the percent of expected. It was also illuminating to generate frequency results for some attributes in tandem with chi square to understand the spread of the data. Firstly, The overwhelming majority of accidents have slight severity (~82%) with much less having severe results (~17%) and a minority resulting in fatalities (~1%). When split into gender, about 80% of cyclists getting in accidents are male, and about 20% are female. The chi square results for gender show that male and females have about the same number of slight accidents, but that males have slightly more serious accidents than expected (~101% above expected), compared to females that have less serious accidents than expected (~95% below expected). Males have an even higher observed number of fatalities than expected (~103% above expected), while females have a far lower number of fatalities than expected (~88% below expected). The chi square results of severity by age are also significant. The number of serious and slight accidents from age 6-45 are about at expected or lower, while fatal results are significantly lower than expected. From age 46-75, fatalities are far higher than expected and serious accidents are also higher than expected. From age 56-75, the observed fatalities are much higher than expected. Chi square results of severity by both age and gender showed some interesting results. Females and males both have similar numbers of slight accidents, about the same as

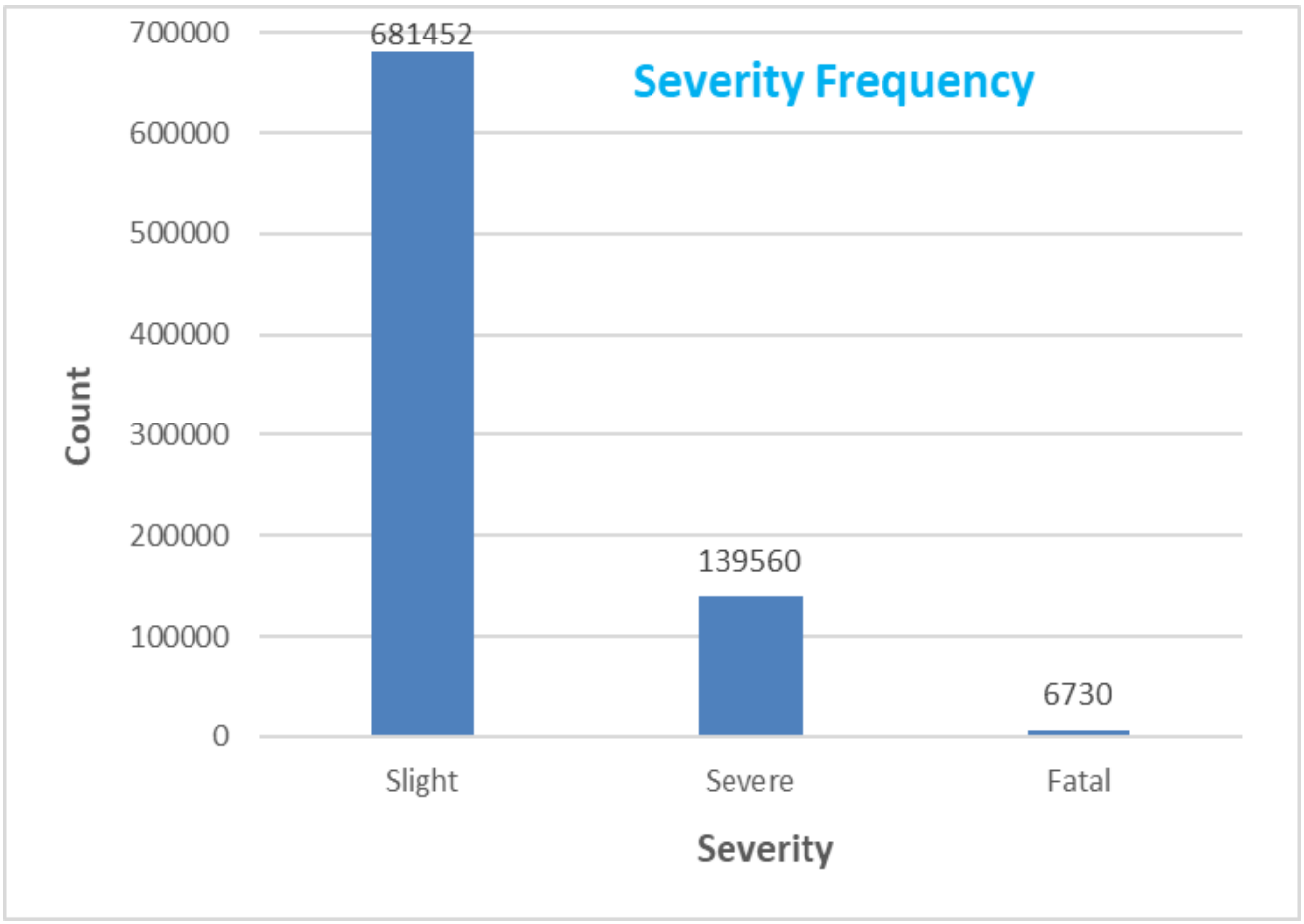
expected. Females tend to have a higher number of serious accidents than males, but both increase with age. Both males and females have a higher number of severe accidents from age 46-75, with males having a much greater number. At age 66-75, females have ~362% above expected fatal results, while males have ~500% above expected fatal results. Clearly, there are significant patterns associated with age and gender of accident severity.

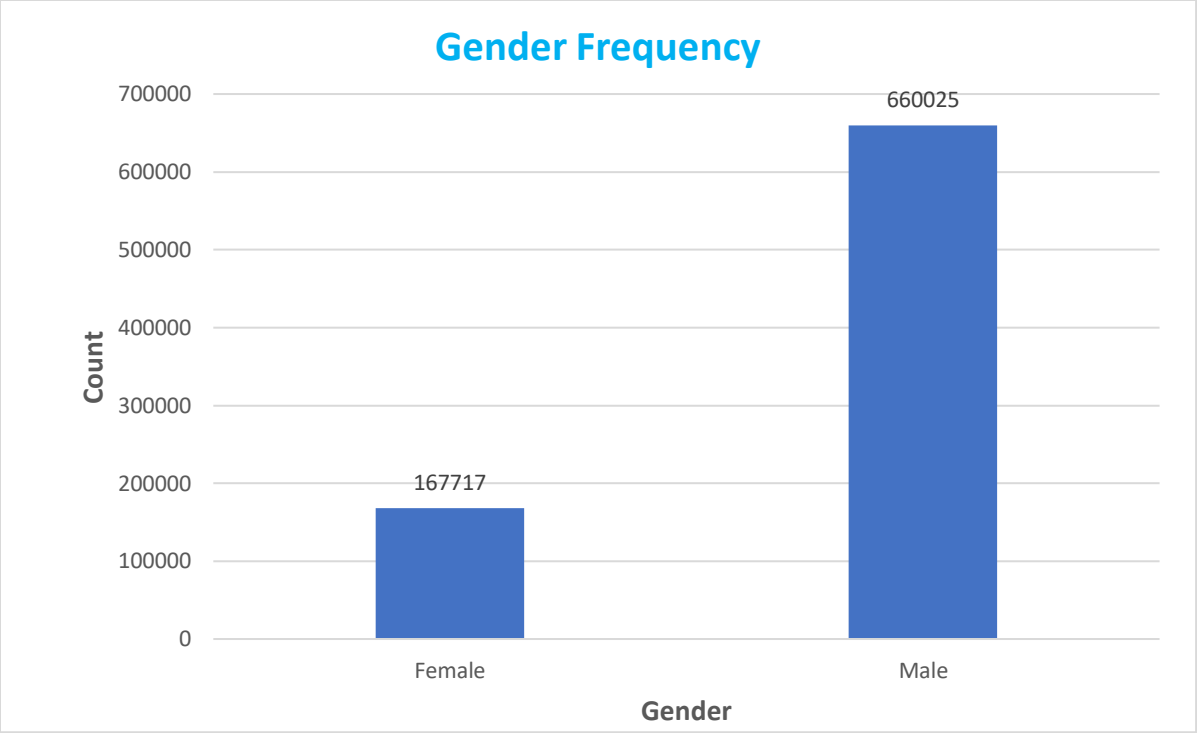
Other frequency and chi square results also found interesting patterns. The vast majority of accidents (~83%) happen in speed limits around 30mph. Starting at 40 mph, the severe and fatal accidents are higher than expected. At 50-70mph, fatalities are greatly increased, with about 840% above expected resulting in fatalities at 70mph. Looking at road conditions, about 80% of accidents happen in dry conditions, and about 20% happen in wet conditions. In dry conditions, all three severities occur at about the expected number, while for wet conditions, higher than expected result in fatalities (~108% above expected). The light attribute had especially significant results. Under conditions of darkness with no lights, severe accidents were higher than expected, while fatal accidents were much higher than expected (~411% above expected). Under conditions of darkness with lights lit and daylight, accidents of all three severities were at or slightly below expected.

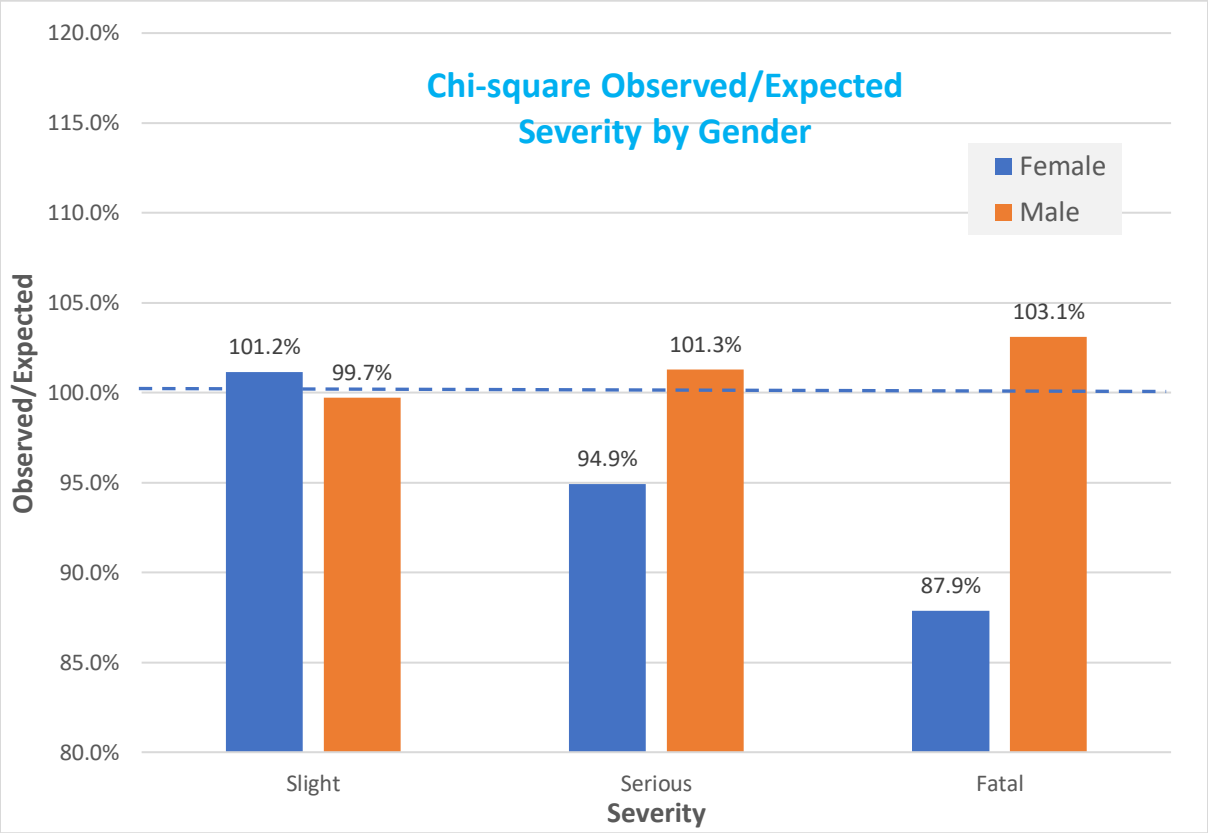
Milestones To Do

There are still significant milestones to complete. Further statistical analysis is necessary, including finding the mean, median and mode of accident age. It also may be interesting to run a linear regression, perhaps on severity by speed limit. Several frequency graphs may be useful, and I still need to run a chi squared on time vs. severity. In the next few weeks I plan to complete the final paper then move onto recording the presentation video. Results will need to be compared to results of previous studies. I would like to finish early so I have more time to devote to learning web development frameworks on my own for my intended job path, such as ReactJS.

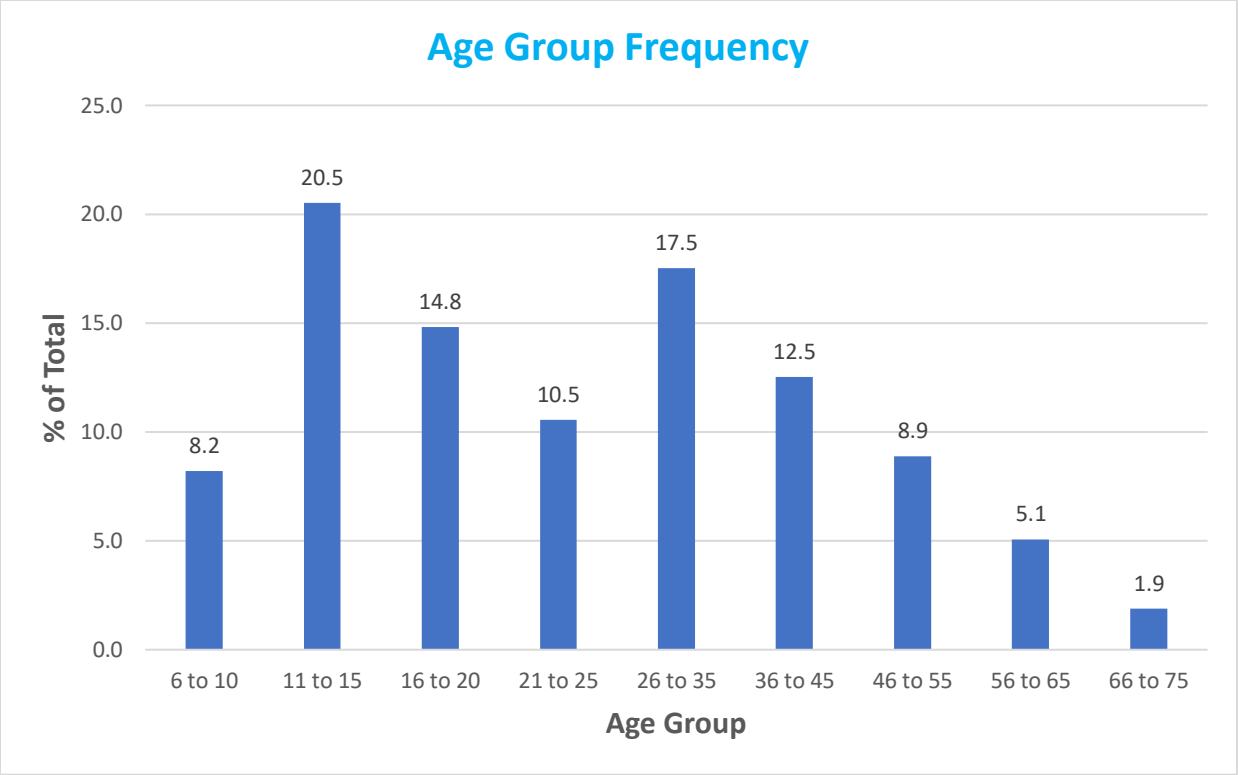
Results (Graphs)

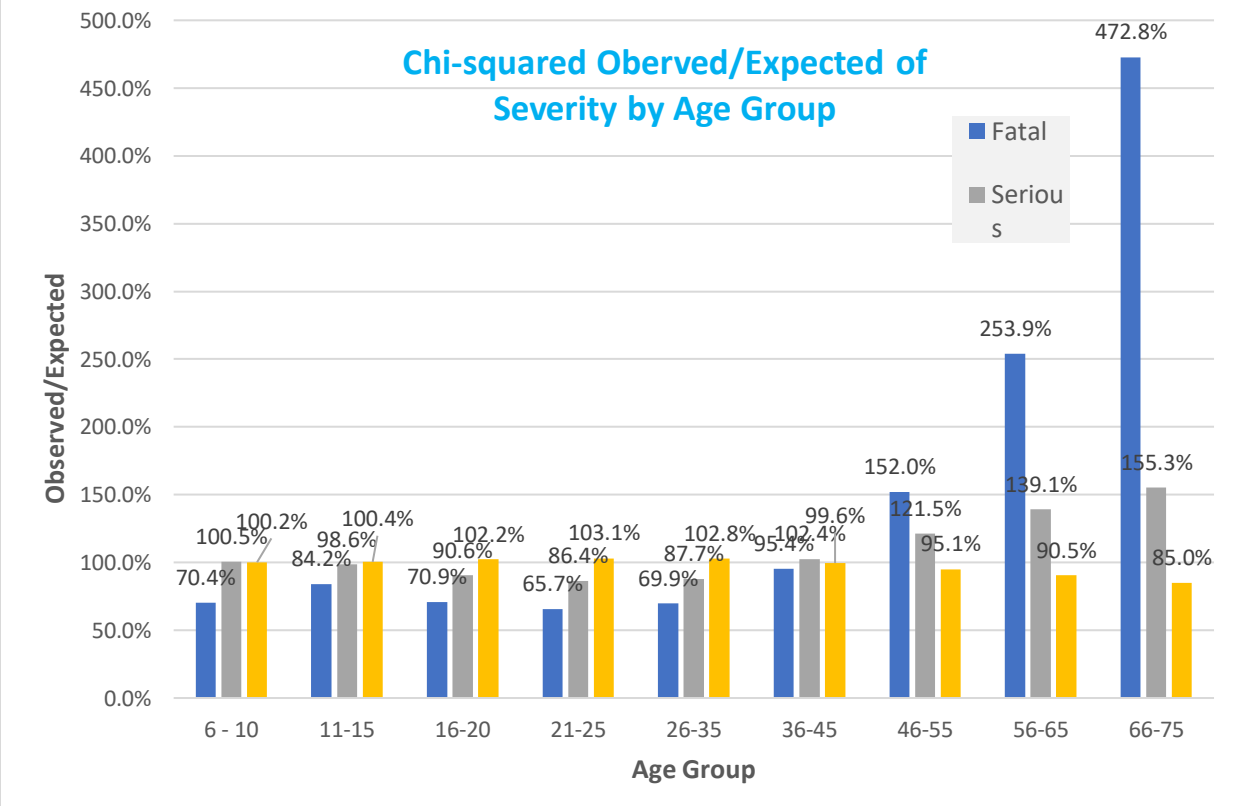




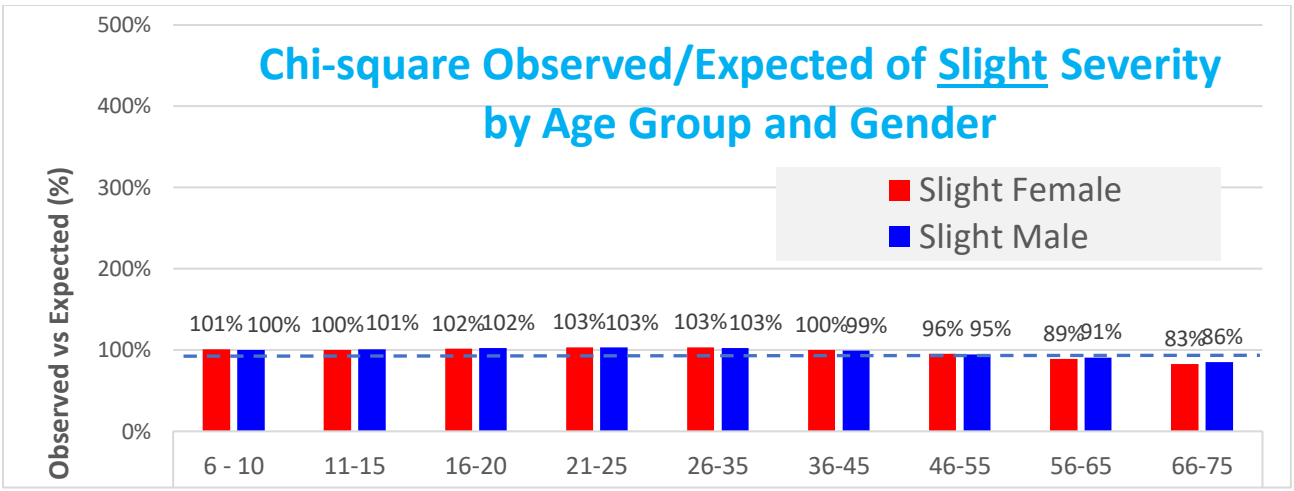


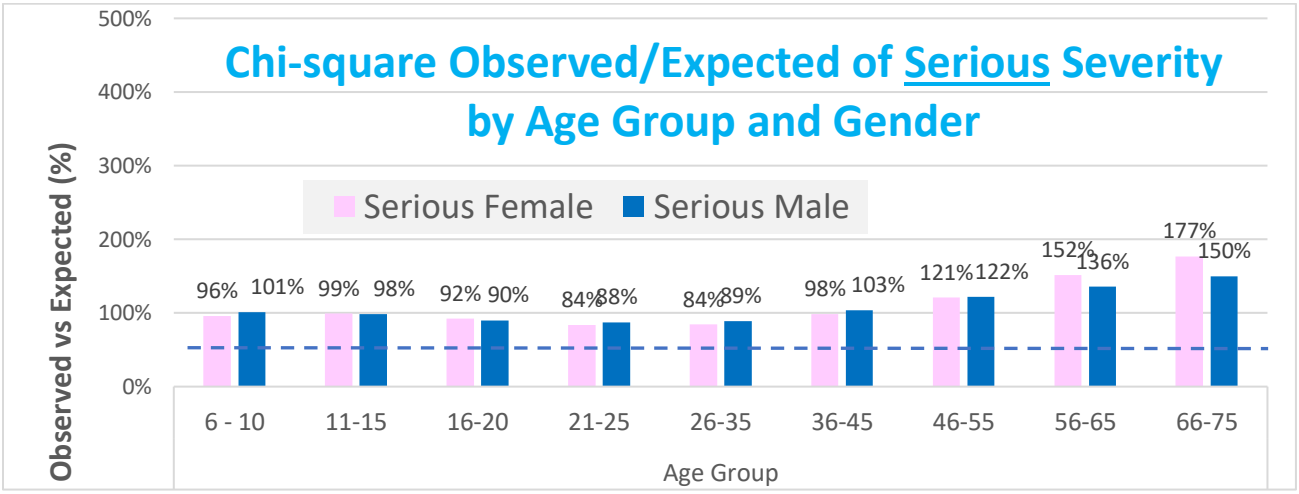
There was a significant relationship between severity and gender (DOF= 2, N=827742, x^2=139.613, p=0.000).

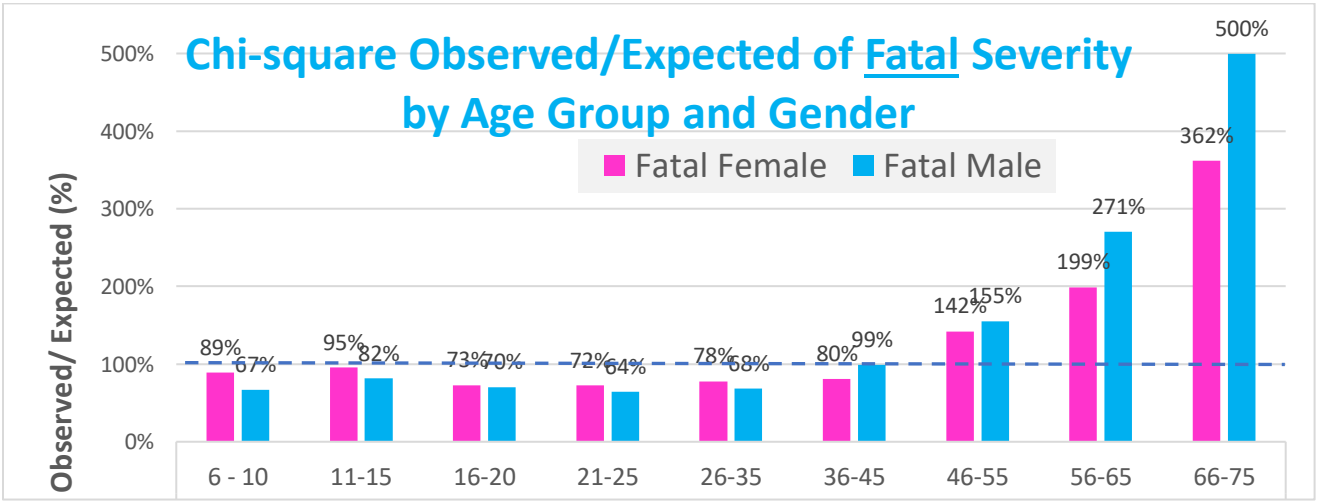




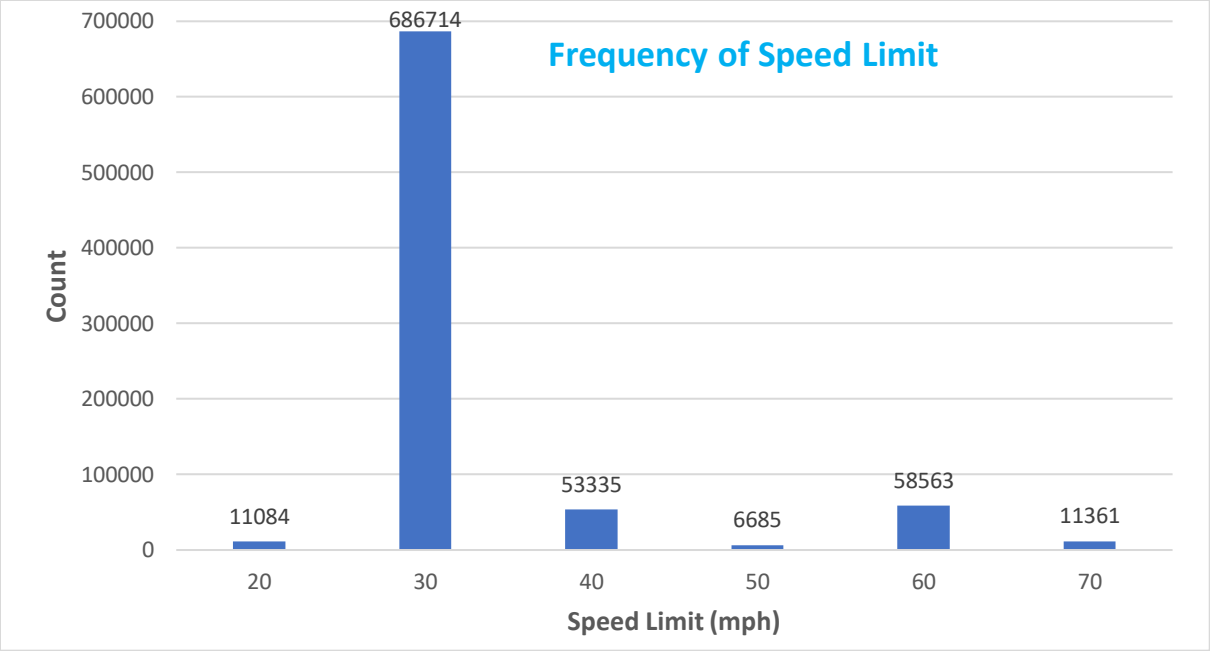
There was a significant relationship between severity and age (DOF= 16, N=827742, x^2=7370.699, p=0.000).

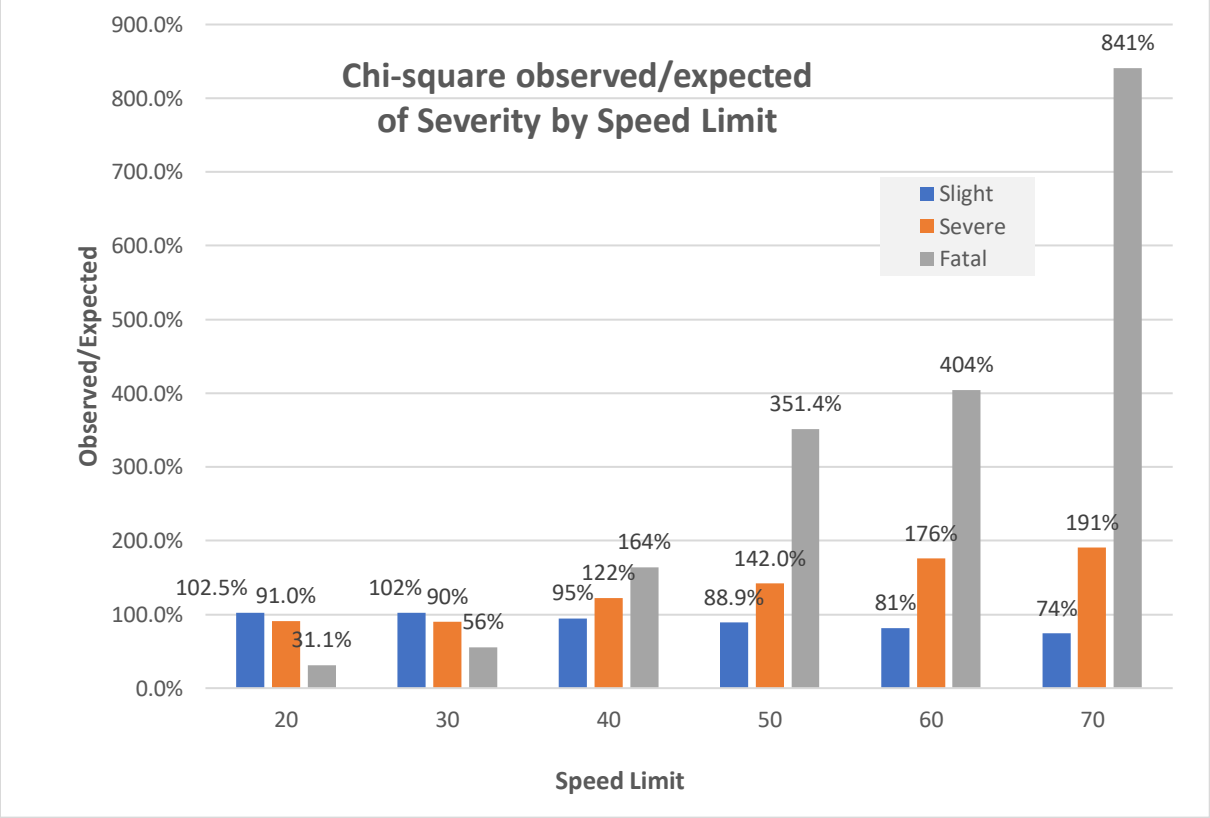




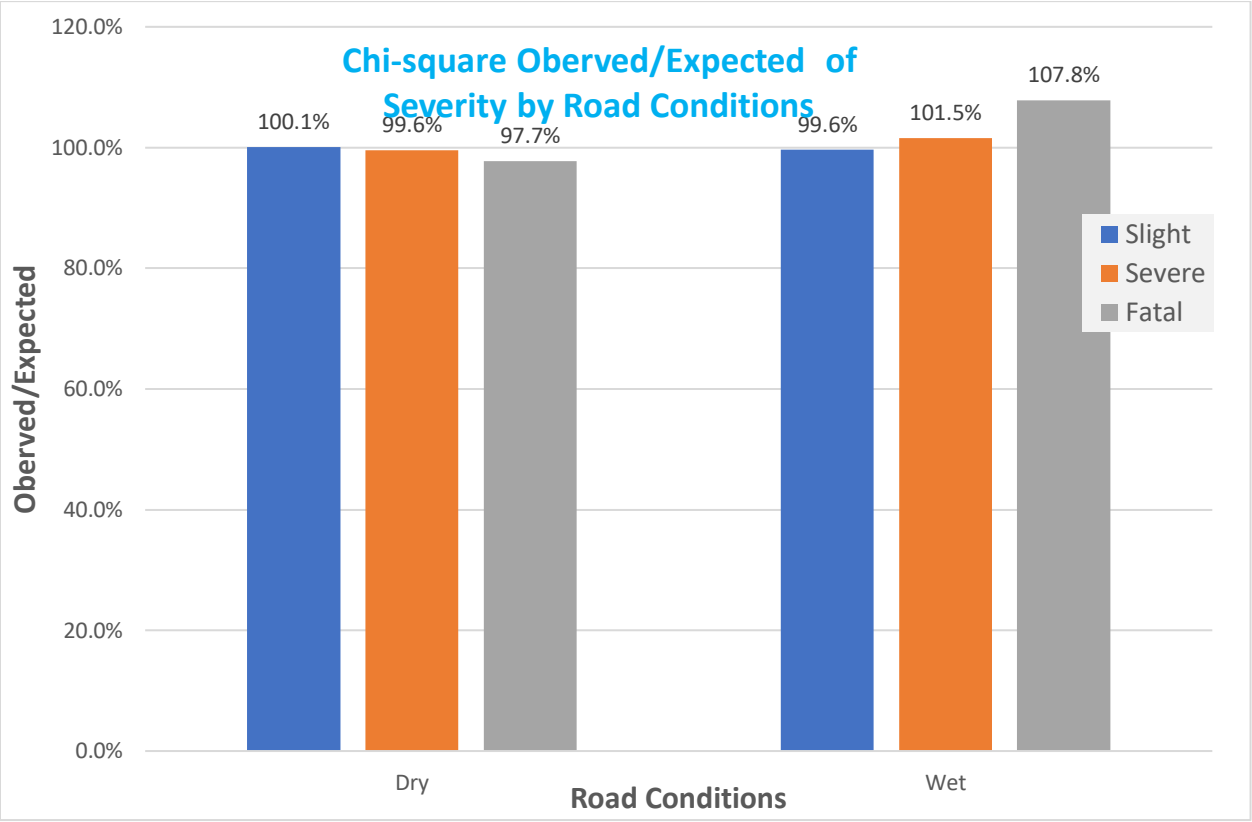


There was a significant relationship between severity and age and gender (DOF= 18, N=827742, x^2=1363.373, p=0.000).

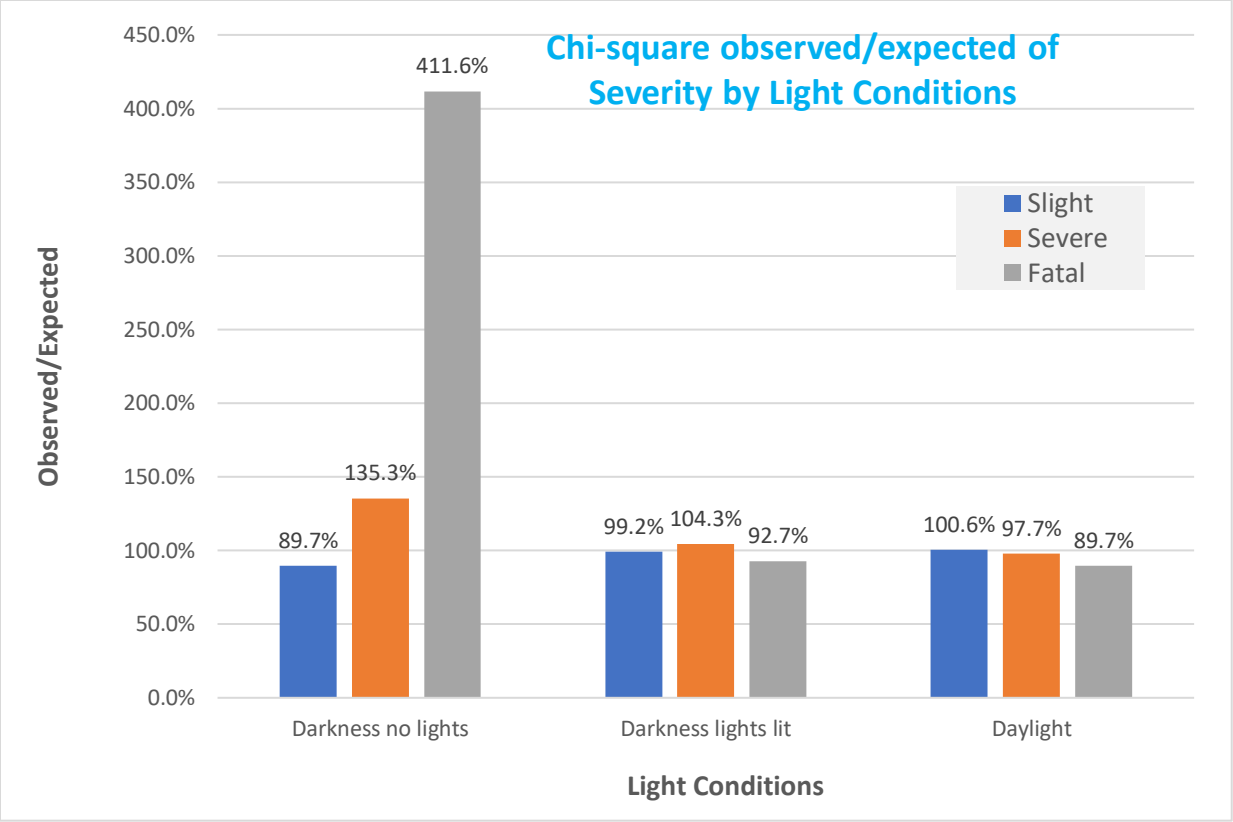




There was a significant relationship between severity and speed limit (DOF= 10, N=827742, x^2=23037.276, p=0.000).



There was a significant relationship between severity and road conditions (DOF= 2, N=818100, x^2=24.452, p=0.000).



There was a significant relationship between severity and light conditions (DOF= 6, N=827742, x^2=2923.31, p=0.000).

Sources:

1. Bicycle accidents in Ohio: What you need to know: 2022. *https://www.plevinandgallucci.com/ohio-bicyclist-deaths/#:~:text=In%20most%20years%2C%20about%2020,traffic%20fatalities%2C%20including%20pedestrian%20deaths*. Accessed: 2023-02-27.
2. Single-bicycle crashes: An in-depth analysis of self-reported crashes and estimation of attributable hospital cost: 2021. *https://www.sciencedirect.com/science/article/pii/S0001457521003845*. Accessed: 2023-02-27.
3. A cross-sectional study of characteristics of bicyclist upper and lower extremity injuries in bicycle-vehicle crashes in Ohio, United States, 2013-2017: 2021. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7923836/*. Accessed: 2023-02-27.
4. A study of bicycle and passenger car collisions based on insurance claims data: 2012. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503407/*. Accessed: 2023-02-27.
5. Bicycle accidents in Great Britain (1979 to 2018): 2021. *https://www.kaggle.com/datasets/johnharshith/bicycle-accidents-in-great-britain-1979-to-2018*. Accessed: 2023-02-27.
6. REPORTED ROAD CASUALTIES IN GREAT BRITAIN: PEDAL CYCLE FACTSHEET, 2020. *GOV.UK. https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-pedal-cyclist-factsheet-2020/reported-road-casualties-in-great-britain-pedal-cycle-factsheet-2020*.