Nicolas Mavromatis (nima6629@colorado.edu)

CSPB 4502

Bike Accidents in Great Britain-Data Mining Project

Part 4 Project Final Report

Abstract

A data mining analysis of bicycle accident data in Great Britain from 1970-2018.

The severity of bicycling accidents was examined as a function of road conditions, weather conditions, and attributes of the cyclists. The following questions were investigated: Is there a specific age range or gender most associated with severe biking accidents? Do certain speed limit zones correlate with high numbers of severe accidents? Are certain times of day most associated with severe accidents? If so, could time of day be explained by brightness? Finally, do certain road conditions associate with severe accidents? The results were that 80% of all accident victims were male, and males were more likely to be involved in fatal accidents than females. Bicyclists 65 years and older were far more likely to be involved in a serious or fatal accident. Most accidents occurred in locations with the speed limit around 30 mph, although in 70 mph locations, accidents were much more likely to be fatal. Fatal accidents were much more common during 12:00 AM-8:00 AM and 7:45 PM-12:00 AM, with the vast majority of accidents happened under conditions of darkness. Most accidents occurred in dry conditions, but wet conditions were more slightly more likely to result in fatalities.

Introduction and Related Work

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 the most serious or even fatal 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 was illuminating to perform a data mining analysis and better understand the main factors that correlate with bicycle accidents.

A small study (350 sample points) out of Denmark looking at single-bicycle crashes provides some interesting results. Winter road maintenance was crucial to deter accidents, and that in warm weather there was 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. 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 severe biking accidents? Do certain speed limit zones correlate with high numbers of severe accidents? Are certain times of day most associated with severe accidents? Could this be explained by brightness? Finally, do certain road conditions associate with severe accidents? The limited work done on this topic found some interesting results, but with only small sample sizes.

Data Set

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. 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].

Techniques Applied

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 (I.E., roundabout, single carriageway, etc.) turned out to be relatively useless, in that 80% of the data occurred on the same road type, so it wouldn’t yield interesting results. Thus, weather and road conditions were not 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. All analyses were performed in SPSS, thus, the time format of the time variable was changed to match the format required for SPSS by removing semi colons. Time was then 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.

The data set is mostly frequency data, so Chi-square results were most appropriate to determine if statistically significant relationships exist between various attributes and accident severity. Cluster analysis would only be appropriate for unlabeled data, and frequent set analysis would be appropriate for transactional data, so these were not applied. All these tests resulted in significant results. The results were graphed in Excel by taking the observed number divided by the expected. Frequency graphs were also produced in Excel to understand the spread of the data. Other statistical measures like mean, median, mode and linear regression were also applied where appropriate.

There are several tools that were instrumental. Initially, I considered using Python and associated libraries to explore the data, and did initial exploration in this framework. 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-square 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.

Key Results and Applications

Many significant results were found. Chi-square 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 results to understand the spread of the data.

Firstly, the overwhelming majority of accidents had 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 were male, and about 20% were female (see fig. 2). The chi-square results for gender show that male and females had about the same number of slight accidents, but that males had slightly more serious accidents than expected (~101% above expected), compared to females who had less serious accidents than expected (~95% below expected). Males had an even higher observed number of fatalities than expected (~103% above expected), while females had a far lower number of fatalities than expected (~88% below expected) (see fig. 3). The chi-square results of severity by age are also significant. The mean age was 26-35, while the median age was 21-25. The modal age was 11-15 (see figs. 4 and 5). The number of serious and slight accidents from age 6-45 were about expected or lower, while fatal results were significantly lower than expected. From age 46-75, fatalities were far higher than expected and serious accidents were also higher than expected. From age 56-75, the observed fatalities were much higher than expected (see fig. 6). Chi-square results of severity by both age and gender showed some interesting results. Females and males both had similar numbers of slight accidents, about the same as expected. Females tended to have a higher number of serious accidents than males, but both increased with age. Both males and females had a higher number of severe accidents from age 46-75, with males having a much greater number. At age 66-75, females had ~362% above expected fatal results, while males had ~500% above expected fatal results (see figs. 7,8,9). Clearly, there are significant patterns associated with age and gender and accident severity.

Other frequency and chi-square results also found interesting patterns. The vast majority of accidents (~83%) happened in speed limits around 30mph (see fig. 10). Starting at 40 mph, the severe and fatal accidents were higher than expected. At 50-70mph, fatalities were greatly increased, with about 840% above expected (see fig. 11). Examining road conditions, about 77% of accidents happened in dry conditions, and about 23% happened in wet conditions (see fig. 12). In dry conditions, all three severities occurred at about the expected number, while for wet conditions, higher than expected resulted in fatalities (~108% above expected) (see fig. 13). The chi-square results for severity by time (in deciles) was especially significant and illuminating. Significantly higher than expected fatalities occurred during 12:00 AM-8:00 AM and 7:45 PM-12:00 AM (~133% and 149% above expected, respectively). Additionally, there were - more severe accidents than expected at 12:00 AM-8:00 AM 7:45 PM-12:00 AM (~103 and 115 percent above expected, respectively) (see fig. 16). The light conditions had especially significant results and converged with the findings involving time of day. The overwhelming majority of accidents happen in conditions of daylight (80%), with less than 20% happening in darkness with lights lit or unlit (see fig. 14). 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 (see fig. 15).

The results of this data mining project confirm some findings from previous studies. Generally, other studies found that above 65 years of age, injuries and fatalities are far higher. The results of this study were similar, where above 65 years of age, fatal accidents were ~362% and ~500% above expected in females and males, respectively. One study also found that most bicyclists getting in accidents are male, and this study found that 80% are male. Another finding from a previous study was that most accidents happen between 7-8 AM and 4-5 PM. This study looked at time in deciles, so not at the granularity of single hours, but found most fatal accidents occur from 12:00 AM-8:00 AM and 7:45 PM-12:00 AM (~133% and 149% above expected respectively). These findings are consistent with 7-8 AM being the most dangerous time from the other study, but inconsistent with 4-5 PM being a dangerous time. Also, accidents were found to occur in dry conditions about 70% of the time, compared to the finding of this study that 80% of accidents happen in dry conditions.

There are a few hypotheses that can perhaps explain these results. There isn't data on how many males versus females bike overall without getting in accidents, so the results can't be normalized, but the large skew toward males indicates that more males may bike overall, and/or may engage in riskier behavior when biking. As age increases above 65 years, serious and especially fatal accidents are much more likely to occur, probably because age leads to higher susceptibility to injury and death due to factors like a reduced capacity to heal. Most accidents happen in lower speed limits (~30 mph) likely because most bikers travel in lower speed limit zones, so they are more likely to be in contact with cars. At 70 mph, fatalities are greatly increased, which makes sense as a car going this fast is much more likely to cause death than at lower speeds. Most accidents happen under dry conditions, which is perhaps due to more bikers on the road during nice conditions, rather than during more dangerous wet conditions. However, wet conditions are slightly more likely to result in fatalities. Fatalities were far more likely to occur during 12:00 AM-8:00 AM and 7:45 PM-12:00 AM, during the night and early morning. This is probably due to the same reason that fatalities were far more likely under conditions of darkness with no lights, probably due to reduced visibility of the drivers and bikers involved.

Several predictions and suggestions can be made based on these results. Bikers above 65 years should be very cautious biking, and perhaps refrain from biking in busy areas. Males should be aware they are more likely to be involved in an accident, although this may be due simply to a great majority of bikers being male. Bikers should avoid the roads during wet conditions, as they are more likely to result in fatalities. They should also avoid biking during late dark hours where visibility is reduced. Bikers armed with this knowledge may be able to avoid the potential for an accident by modifying their habits. Perhaps bicycle safety classes relaying this information could decrease the number of accidents.

The data used in this study suffered from several limitations, however. The data was collected only in Great Britain, and may be vastly different for other populations. There was no data on helmet usage to compare to other studies. There was very little data on snowy and other conditions, so analyses could only be conducted on wet vs. dry conditions. Road type was not a useful variable, as most accidents occurred on the same road type. Crucially, there is also no population data to normalize by, such as total number of male and female bikers that do not experience accidents. Future studies should collect data from more diverse sources in other regions, and find population data to normalize the results.

<u>Results (Graphs)</u>

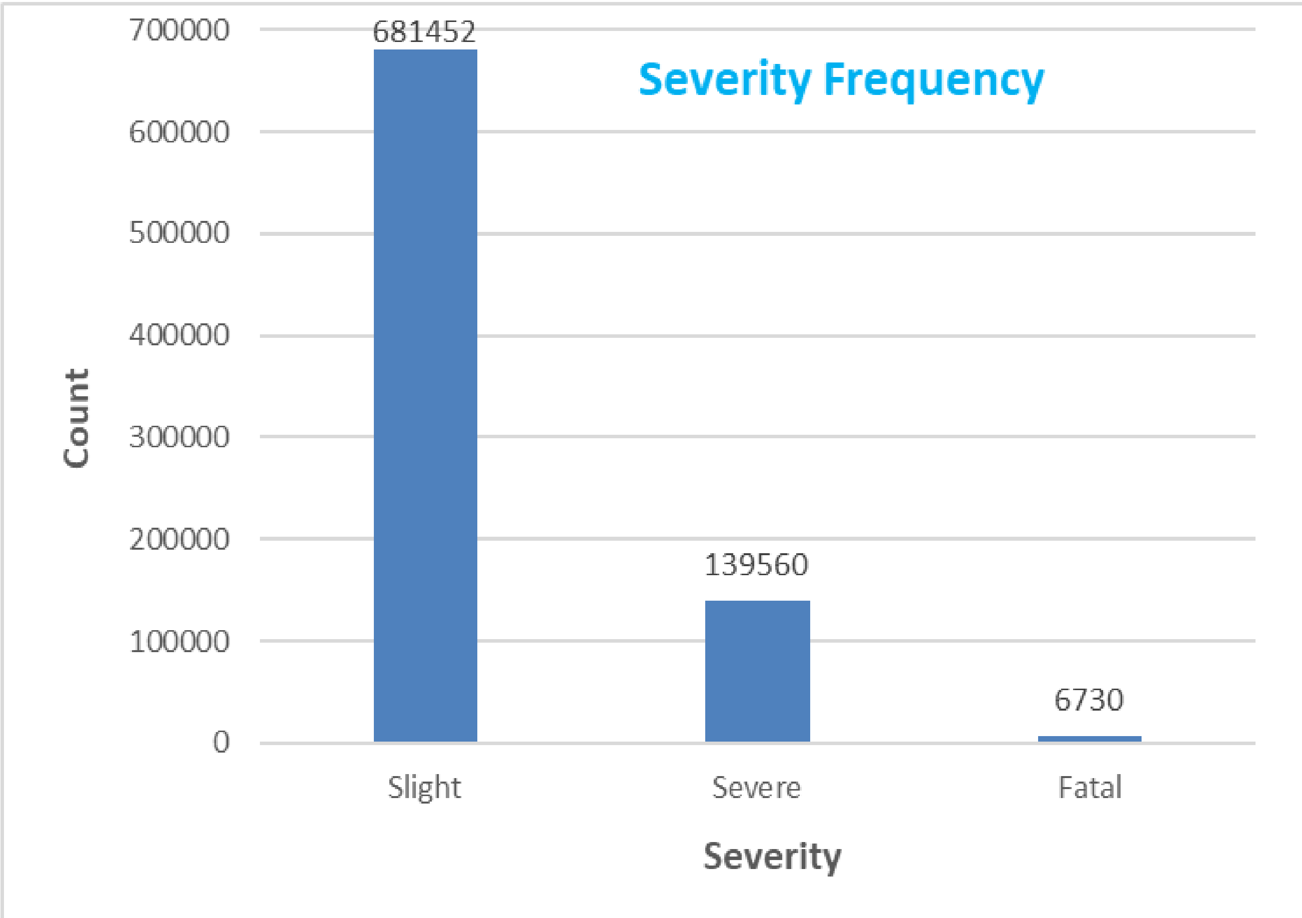
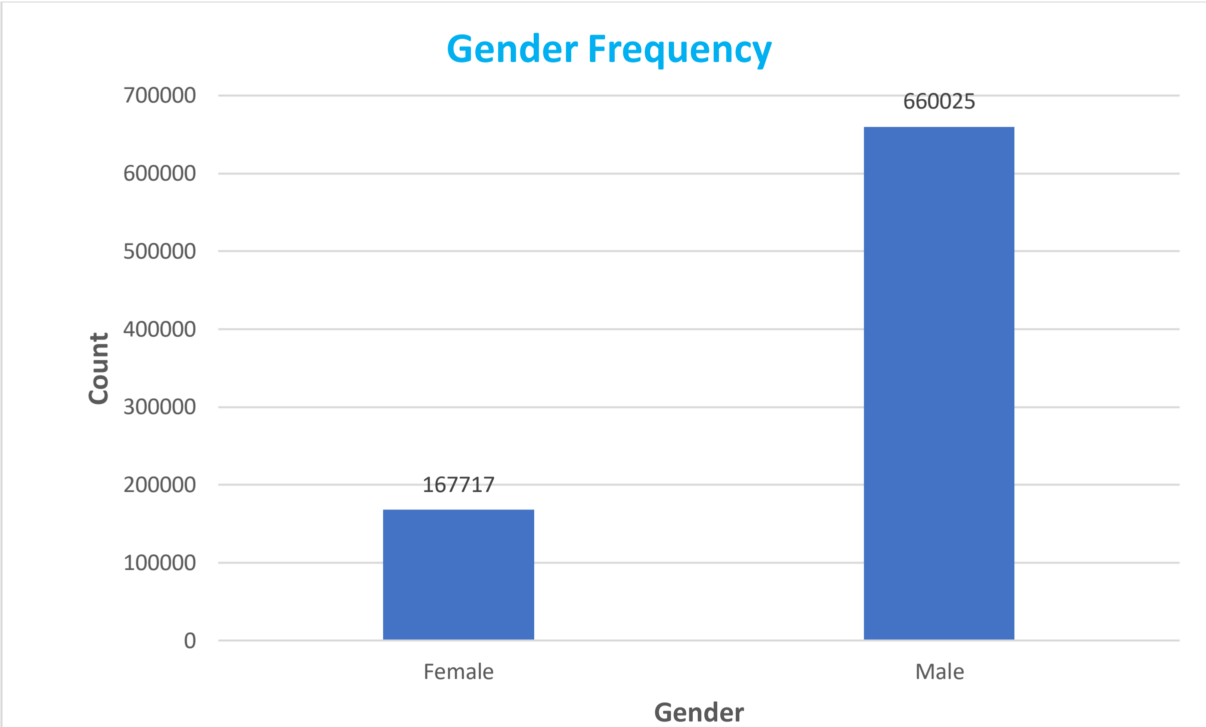
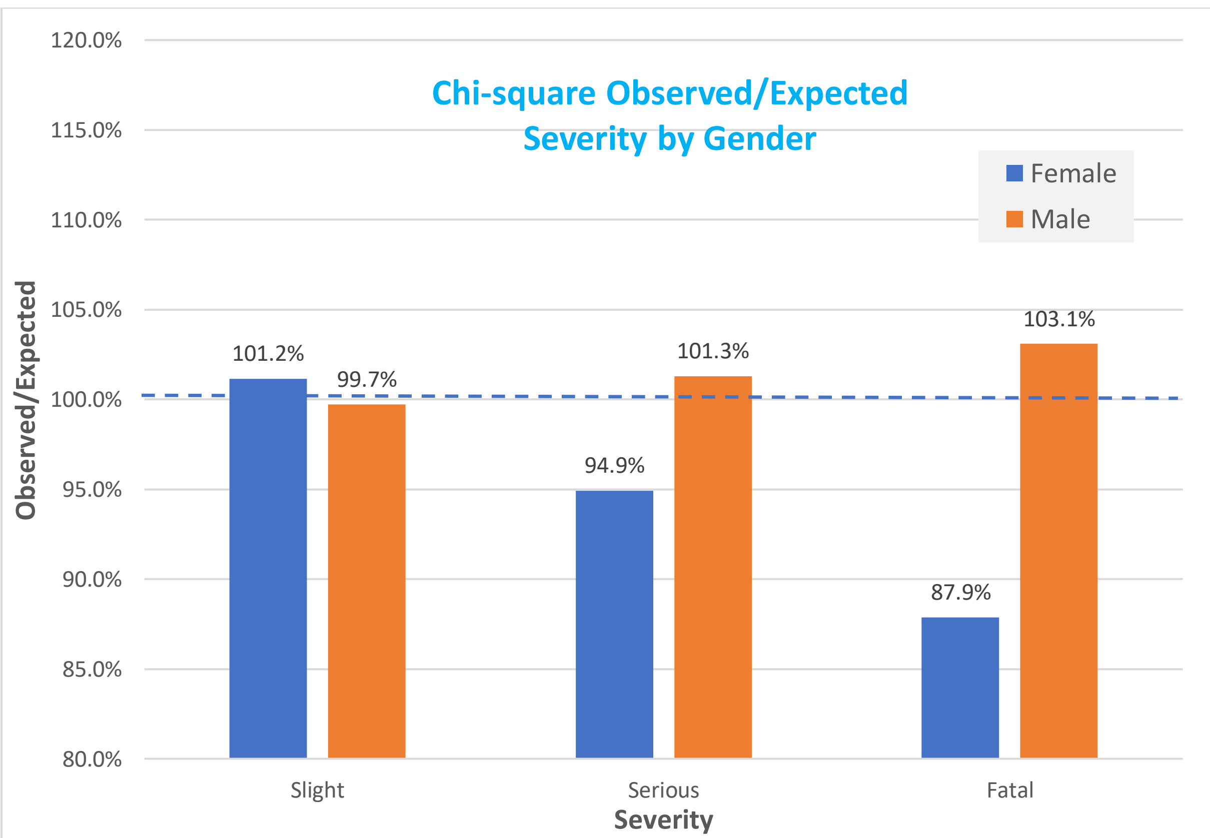
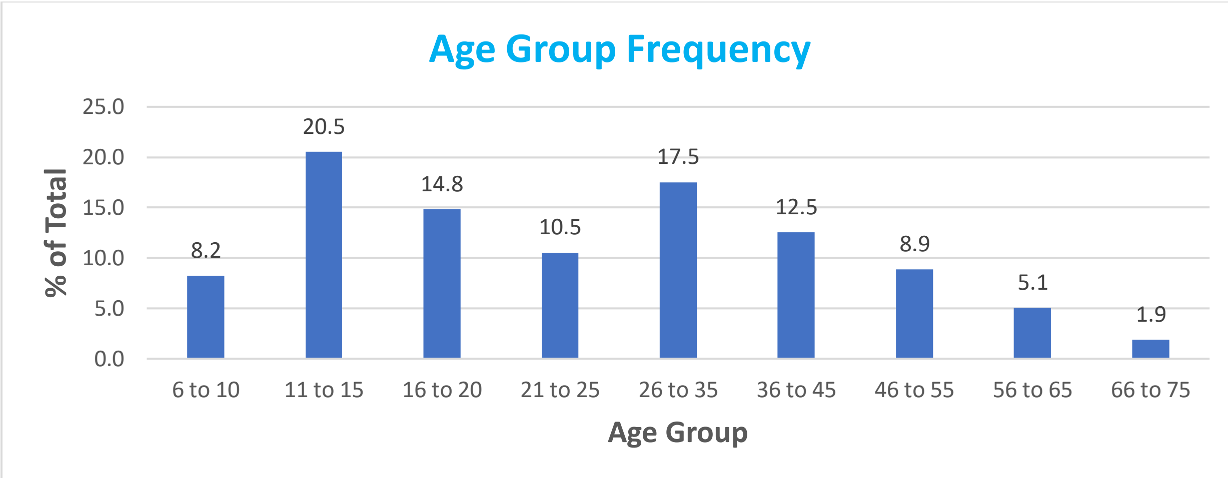


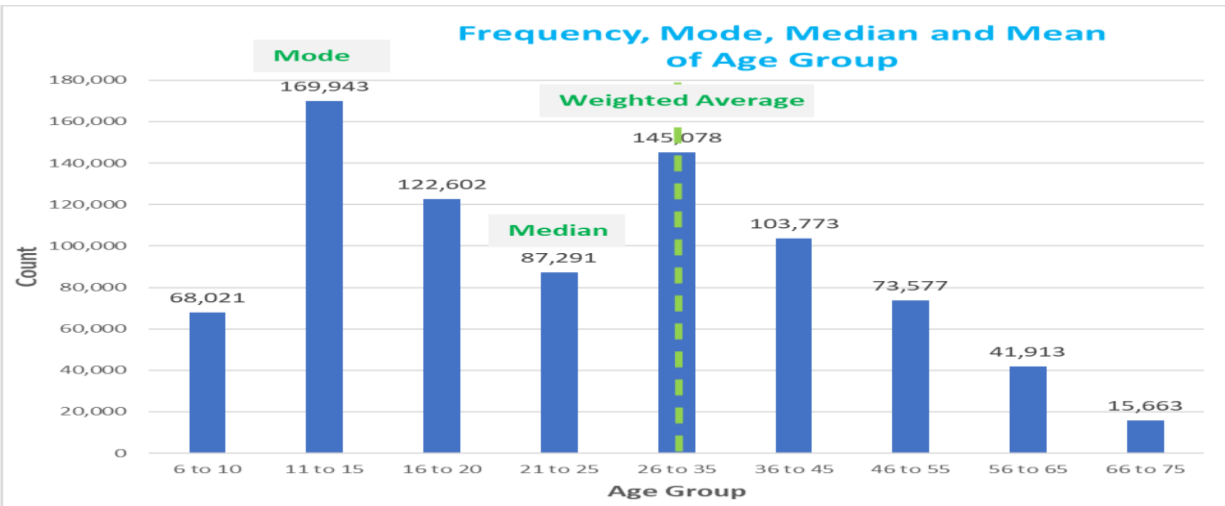
Fig. 1: Severity frequency Results

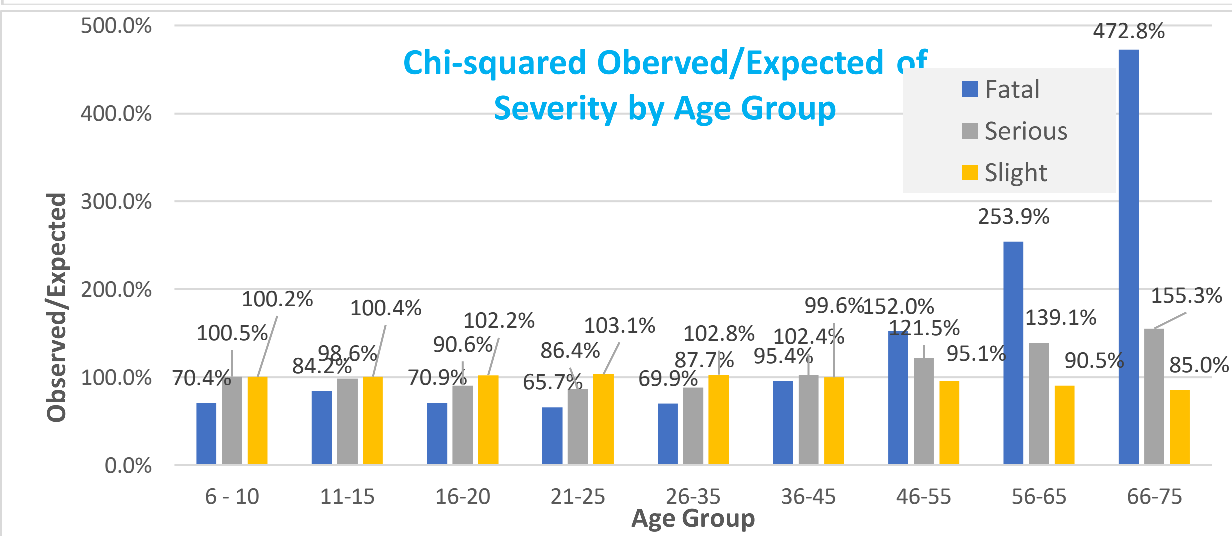




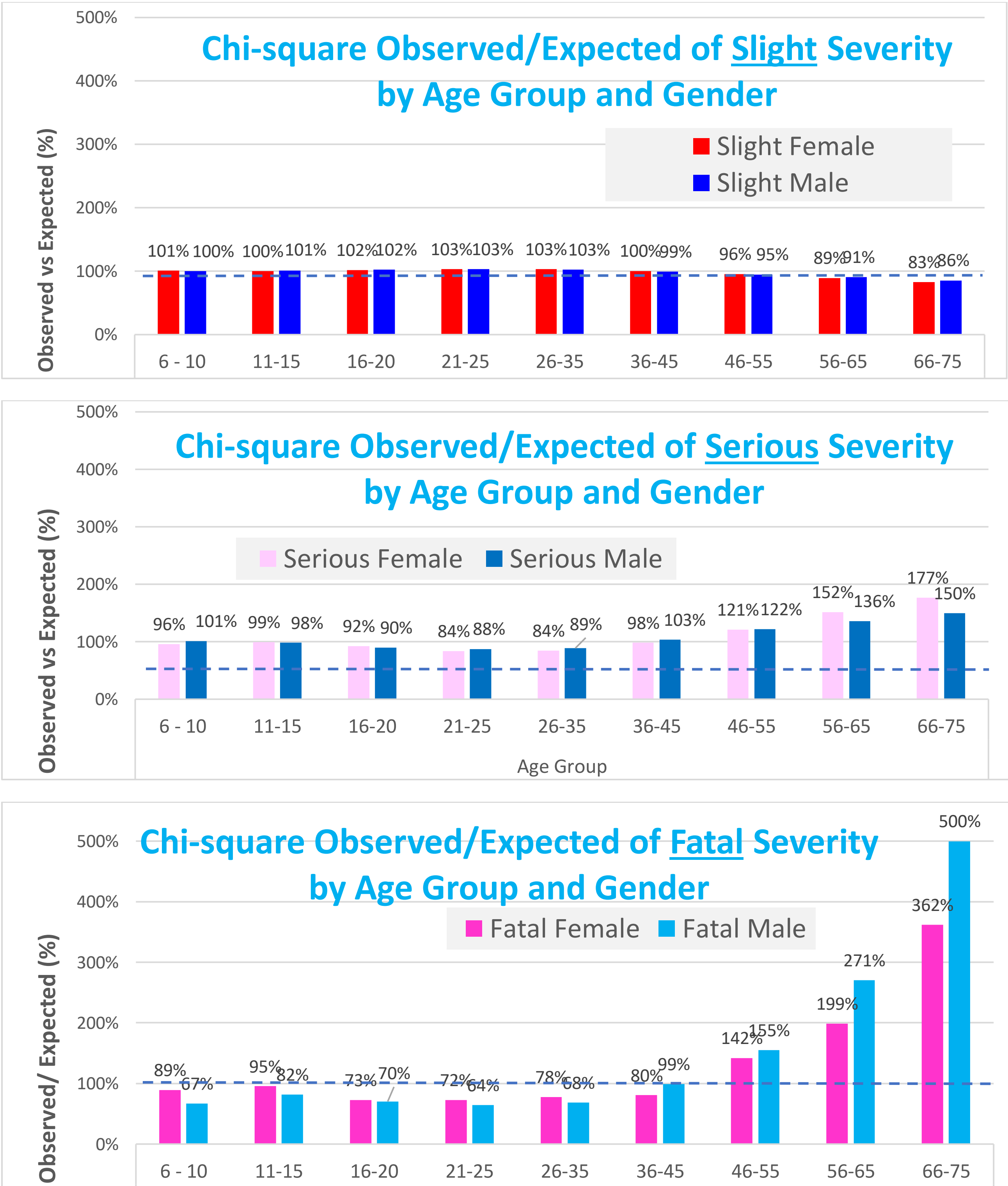
Figs. 2 and 3: Gender frequency and chi square severity by gender results. There was a significant relationship between severity and gender (DOF= 2, N=827742, x^2=139.613, p=0.000).



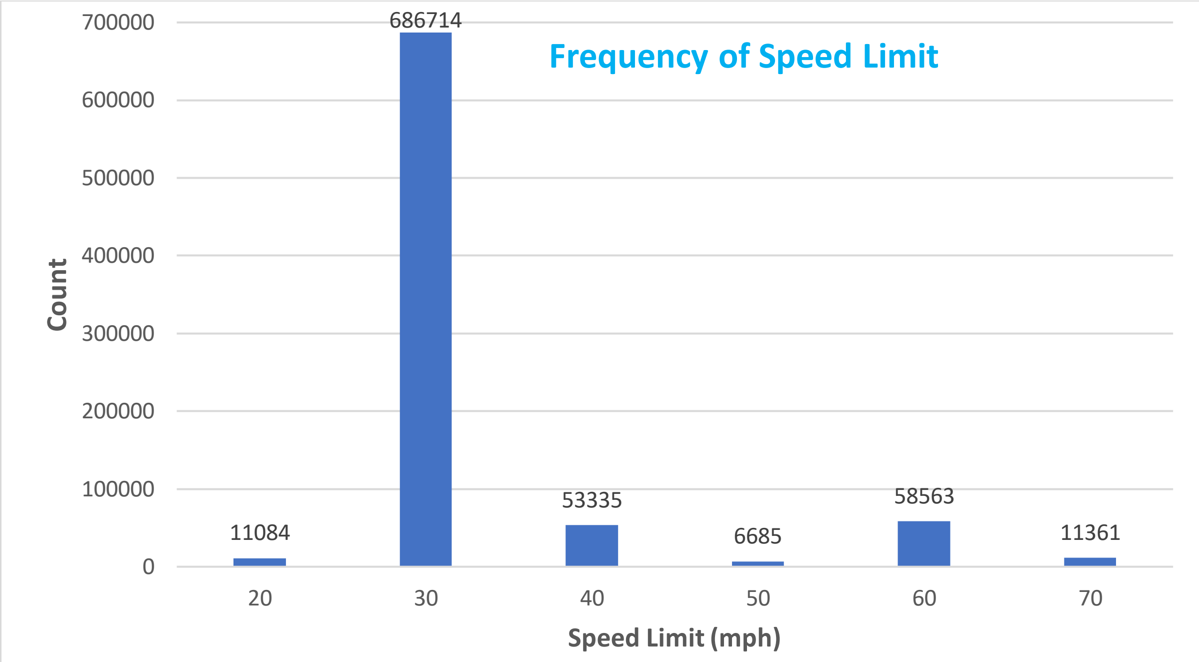


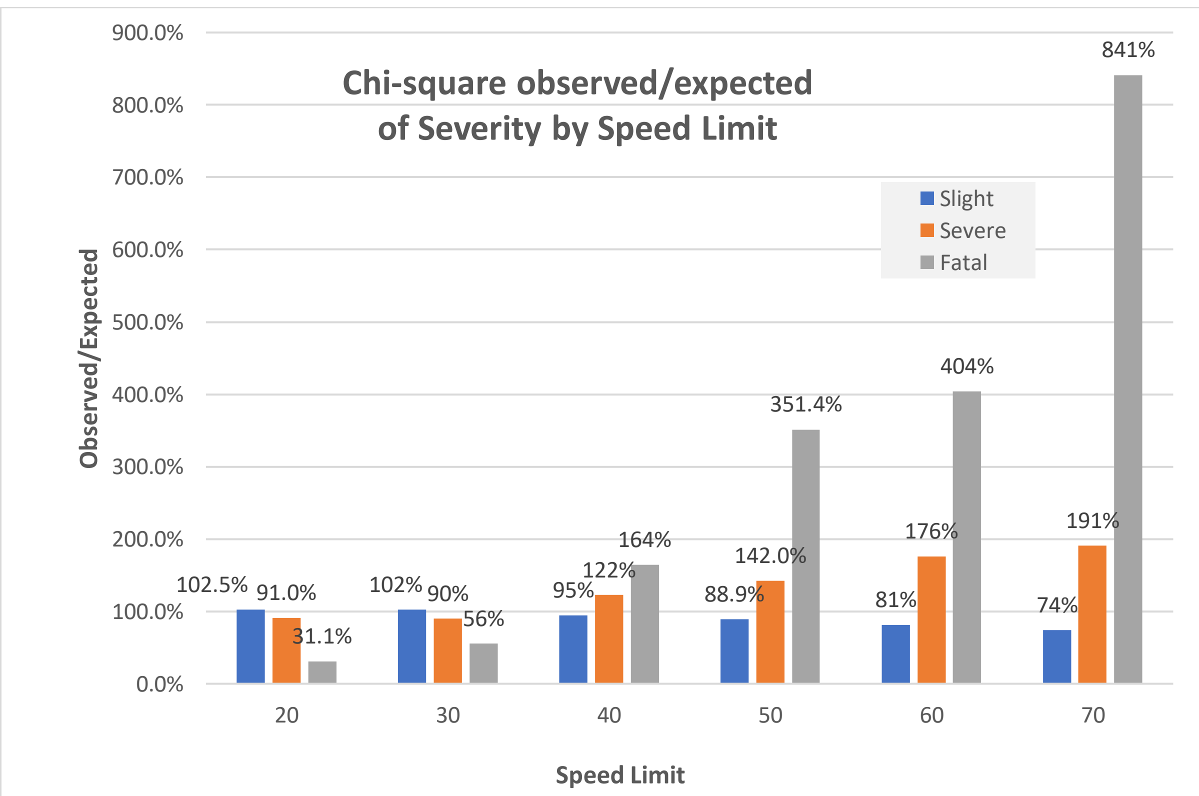


Figs. 4, 5, and 6: Age group frequency, misc. stats, and chi square severity by age results. There was a significant relationship between severity and age (DOF= 16, N=827742, x^2=7370.699, p=0.000).

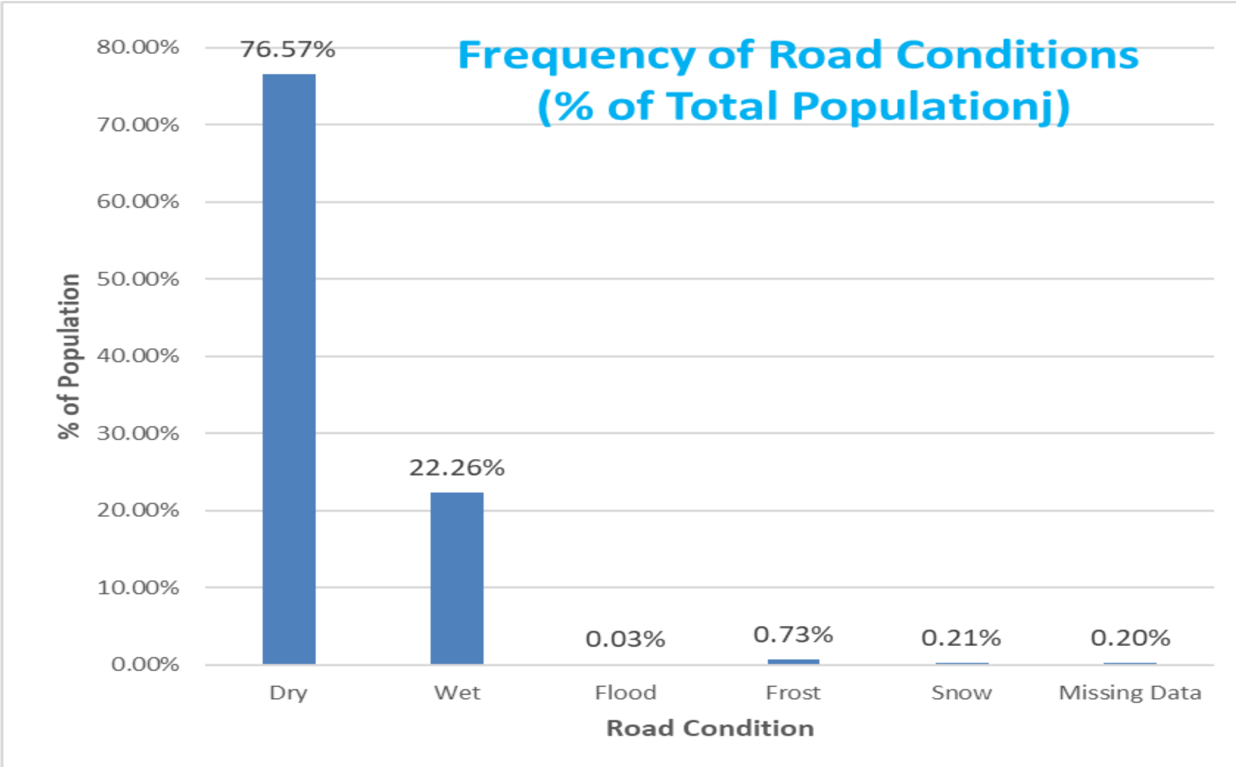


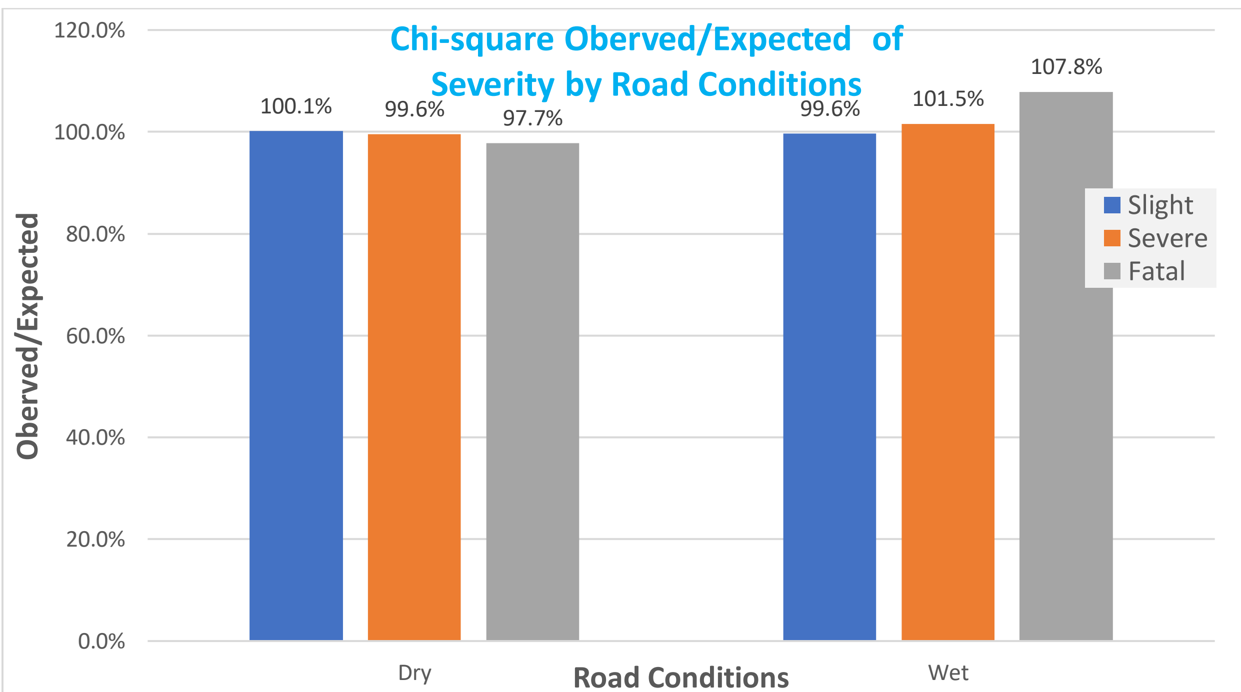
Figs. 7, 8, and 9: Chi square severity by age group and gender results. There was a significant relationship between severity and age and gender (DOF= 18, N=827742, x^2=1363.373, p=0.000).



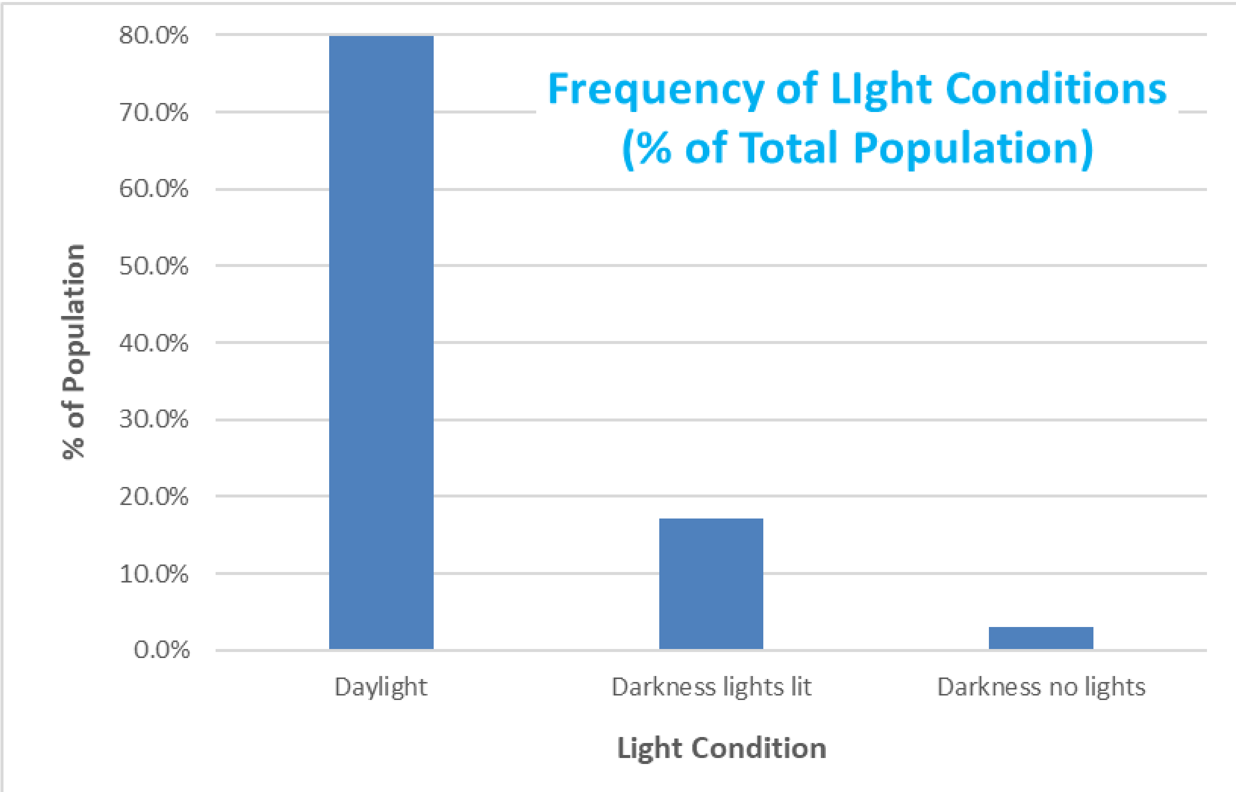


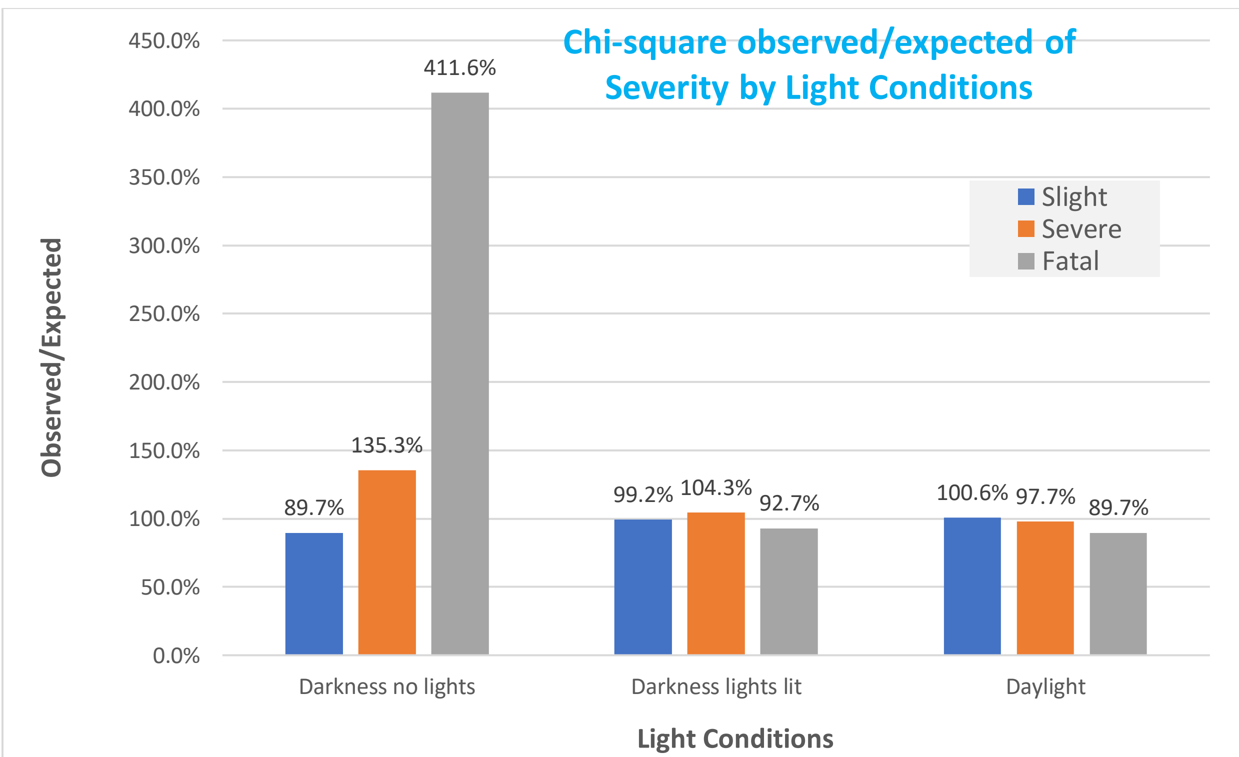
Figs. 10, and 11: Speed limit frequency and chi square results of severity by speed limit. There was a significant relationship between severity and speed limit (DOF= 10, N=827742, x^2=23037.276, p=0.000).





Figs. 12 and 13: Road condition frequency and chi square results of severity by road condition. There was a significant relationship between severity and road conditions (DOF= 2, N=818100, x^2=24.452, p=0.000).





Figs. 14 and 15: Light condition frequency and chi square results of severity by light conditions. There was a significant relationship between severity and light conditions (DOF= 6, N=827742, x^2=2923.31, p=0.000).

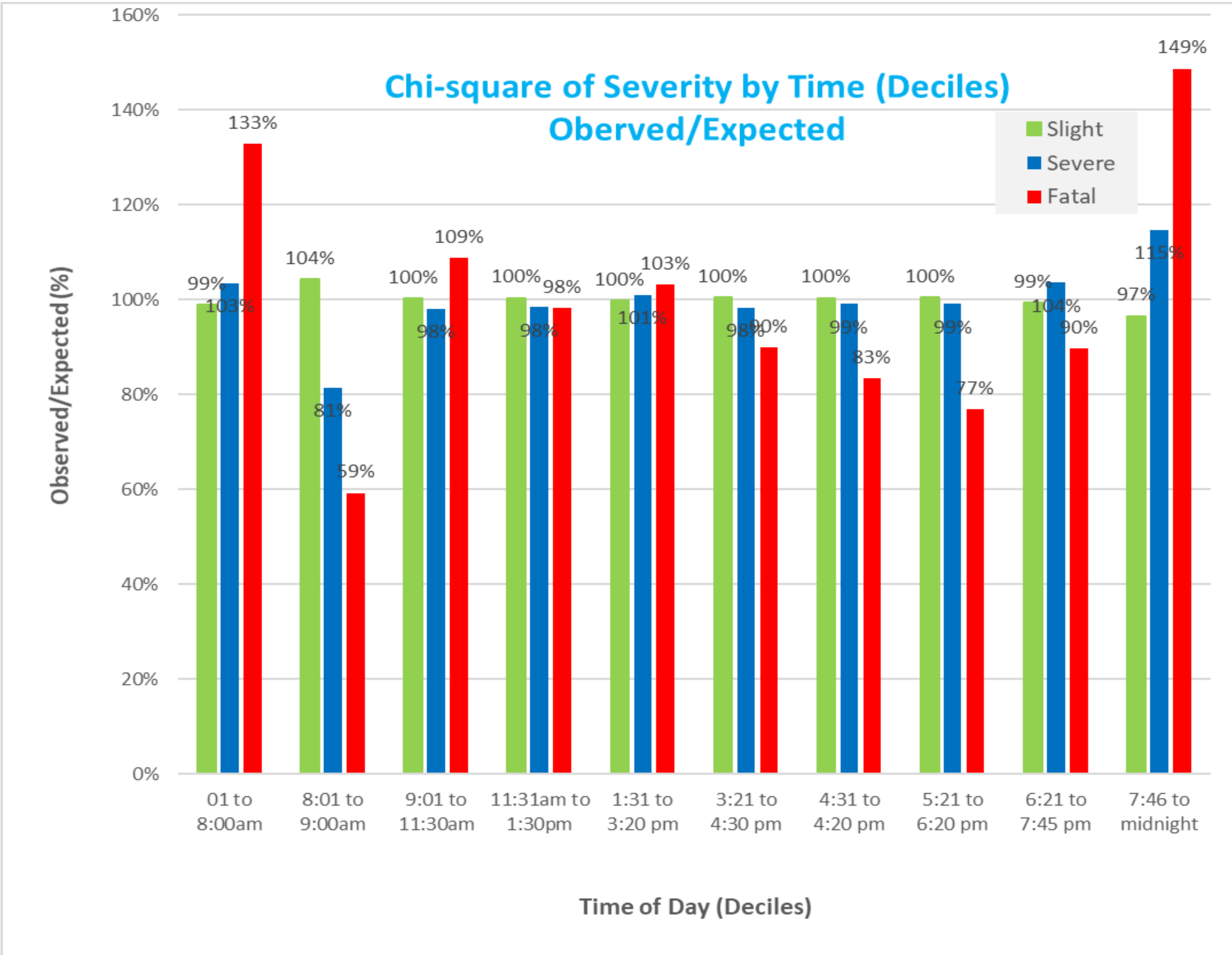


Fig. 16: Chi square results of severity by time in deciles. There was a significant relationship between severity and time (DOF= 10, N=827742, x^2=4501.45, p=0.000).

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.*