Nicolas Mavromatis (nima6629@colorado.edu)

CSPB 4502

Bike Accidents in Great Britain-Data Mining Project

Part 2 Project Proposal Paper

Bicycle accidents account for significant deaths, with about 1,000 killed annually nationwide. Alarmingly, mortalities from bicycle accidents reached a 28 year high in 2018, perhaps partially due to the increasing popularity of biking [1]. This suggests 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. 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 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 and type, 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, 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 is that most accidents occur in an urban area, half 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 road types or 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. Some 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. 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 to see if any 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. 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 will be instrumental in finding patterns and presenting results. Lastly, github is the industry standard for version control. Other tools will be assessed as necessary.

It is important to have milestones to keep progress moving forward. I already made progress cleaning the data, joining the two excel sheets and adding a season attribute. First, I will need to bin the data by time and other attributes. I plan to spend the next week or two performing statistical exploration to find interesting patterns and answer the questions posed above. I will run chi-squared tests and tabulate and graph the results of these and other statistical analyses. After a few weeks, I will add to the paper and include these findings. Afterwards, I'll finish the presentation. I plan to be mostly done with the project in about a month.

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.