Investigating Motor Vehicle Crashes in New York City

Casey Delaney, San Jose State University[1],
Manasa Bobba, San Jose State University[1], and
Sudheendra Katikar, San Jose State University[1]

[1]

Ongoing studies have anticipated that in 2030, car crashes will be the fifth driving reason for death around the world. Road accidents and its safety have been a major concern around the world and have been a primary concern for automotive manufactures. Road safety concerns sparked the creation of the NHTSA (National Highway Traffic Safety Administration) as well as the IIHS (Insurance Institute for Highway Safety) among others. Road accidents and reckless driving occur in every part of the world. Because of this, many pedestrians are affected too. This paper aims to analyze road accidents in one of the most populous cities, New York. The data we have used is especially informative because it is generated from a comprehensive police report template. It consists of information from all police reported motor vehicle collisions in NYC over the period of 2012-2021 and has over 1.8 million records.

keywords - clustering, k-means, A-Priori Algorithm, Association Analysis

1. Introduction

This paper has the purpose of determining contributing factors that cause higher instances of accidents in automobiles. In a span of 9 years, over 1.8 million accidents were documented in New York. By analyzing these traffic records, billions of dollars worth of damage may be recovered in future years. A 2014 report showed that 871 billion dollars per year is being wasted due to traffic collisions in the United States [Lowy 2014].

1.1 Domain introduction

It is computationally infeasible to comprehend the data in a large data set. For this reason, machine learning and data mining will be utilized to help in the understanding of this data set.

1.1.1 Machine learning

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a subset of Artificial intelligence. Machine learning algorithms build a mathematical model based on sample data known as training data. Machine learning algorithms are widely used in various applications such as email filtering, computer vision. It is mainly focused on computational statistics which focuses mainly on predicting.

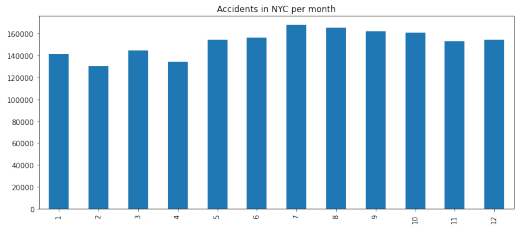

Fig. 1. Number of accidents as a function of month in NYC

1.1.2 Data mining

Data mining is a branch which involves looking for hidden, valid, and potentially useful patterns in huge data sets. It is all about discovering previously unknown relations among data. Data mining involves anomaly detection, association rule mining, clustering, classification, regression.

1.2 Analysis Method

Three methods of analyzing the data will be used. Each method has a specific purpose and will show features in the data that other methods are unable to reproduce. Association analysis with the A-Priori algorithm will be used to find a common relationship in data item sets while the K-Means clustering algorithm will be used to find a visual relation between large groups of data. The data is not limited to these methods and other methods such as dimension reduction or regression analysis may be used to discover new features in the data.

1.3 Initial Data Impressions

Performing initial data analysis has provided some inconclusive results. Figure 1 shows the number of accidents as a function of the month it occurs in. New York can contain harsh winter conditions during the months of November through April. One might hypothesize that this would cause the number of accidents per month to increase. The data shows the opposite with the number of accidents per month peaking in June at 160,000. This could be because more people choose to stay indoors during the winter months.

Figure 2 shows the number of accidents per year. From the years 2012 through 2018, there has been a steady increase in accidents. This can be due to a variety of factors such as increasing population putting more vehicles on the road. In 2019, there is a small decline in the number of accidents and

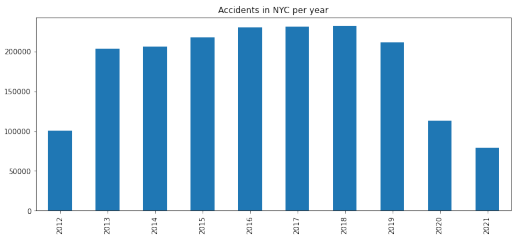

Fig. 2. Number of Cyclist Injuries as a function of year in NYC

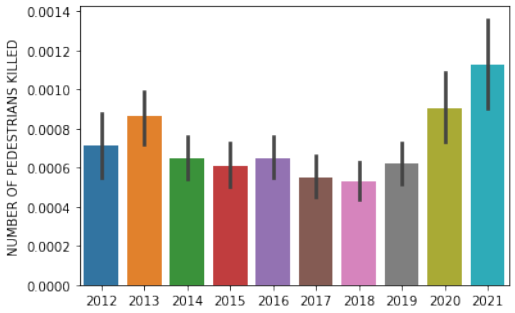

Fig. 3. Number of accidents as a function of year in NYC

2020 shows a much sharper decline. This can be a direct result from the Covid-19 pandemic creating a work-at-home situation for many citizens. The data for 2021 is incomplete since the year is not over at the time of this writing (data downloaded in October, 2021) and will therefore not be considered for this impression.

Interesting statistics begin to appear when the number of injuries are viewed. Figure 3 shows the number of cyclist injuries as a function of year. In the past 3 years, there has been a dramatic increase of cyclist injuries. The year 2021 is still incomplete at the time of this writing, yet there are more cyclist injuries in 2021 than in any of the previous years.

Finding relationships for the causes in these injuries has the potential to save many lives each year. Although the total number of deaths in New York City may seem small, this is only one city among many throughout the United State and the rest of the world.

2. Methods

To analyze the data set, three methods were used. These methods consist of association analysis using the A-Priori algorithm, clustering with K-means clustering and ... Depending on the algorithm used, the size of the data set may be reduced to decrease the processing time. All computations were performed in a notebook hosted by Google Colab and certain actions were unable to be completed due to Google Colab limitations. The entire data set was used in applicable instances where the computing power did not exceed the limitations.

2.1 Association Analysis using A-Priori Algorithm
Association analysis will be used to find common relationships between different itemsets. The A-Priori algorithm is used to determine correlations and co-occurrences exist. Generally the A-Priori algorithm performs slower than other methods on large data sets. The algorithm is relatively simple but is limited due to its slow performance. Increasing the speed is possible in this research by limiting the data to data deemed useful such as the primary vehicle involved in the car accident (Vehicle Type Code 1) and the number of pedestrians or passengers injured or killed. The majority of car crashes are limited to one or two vehicles rendering columns for vehicles 3, 4 and 5 almost useless. These columns, among others, were left out to improve computational efficiency.

2.2 Clustering with K-Means
The use of the K-Means algorithm will allow the data to be clustered based on the data set features. The algorithm has been run on columns of latitude and longitude and longitude data to determine if there are certain localities that are especially prone to accidents. To this point, data has been inconclusive. Results were also analyzed to determine if there are clusters around a specific crash time and this far, results have been inconclusive. This could be because the number of clusters must be increased to account for the large variation within the city. The size of the dataset makes it difficult to use the K-Means algorithm due to the limitations of Google Colab so the dataset was reduced in size.

3. Comparisons

4. Example Analysis

5. Conclusions

ACKNOWLEDGEMENTS

References

Joan Lowy, Associated Press. (2014, May 29). Traffic accidents in the U.S. cost $871 billion a year, federal study finds. PBS NewsHour. https://www.pbs.org/newshour/nation/motor-vehicle-crashes-u-s-cost-871-billion-year-federal-study-finds