Cause Analysis of Traffic Accidents in the Us Based on Exploratory Data Analysis

Yiding Zhao

The North China University of Technology, Beijing, China

m18911660733@163.com/2053770@brunel.ac.uk

Keywords: Traffic accidents, Us-accidents, Exploratory data analysis

Abstract: Nowadays, traffic accidents are frequent, to explore the causes of accidents, the method of data mining and building machine learning models for analysis proposed by the researcher before is feasible and efficient. But this article used the 2016-2020 US-Accident dataset provided by Kaggle and uses the data analysis method of Exploratory Data Analysis to roughly analyze several causes of accidents. First, the outbreak of COVID-19 led to a rapid increase in the accident rate; second, the morning and evening rush hours during weekdays were also the main cause of accidents; finally, the accident rate was higher in the second half of the year than in the first half, and more accidents occurred in cloudy weather. Using this result can provide a more intuitive understanding of the causes of accidents, which can be helpful for more in-depth research.

1. Introduction

Road traffic is an indispensable important part of life, but traffic accidents are a major public safety issue, and their frequent occurrence not only affects the safety of life but also brings serious economic losses [1][6]. Almost all the time, someone has suffered from traffic accidents in one way or another. But what causes these accidents to happen? This article roughly analyzes Kaggle's US-accident datasets using the Exploratory Data Analysis (EDA)method. This is a countrywide car accident dataset, which covers 49 states of the US. The accident data are collected from February 2016 to December 2020, Currently, there are about 1.5 million accident records in this dataset, which are collected from the state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. [2] For the rest of this paper, firstly, the preparatory work before analysis is described in Section2. The analysis process is described in Section3. Finally, Section4 describes the results of the analysis and concludes.

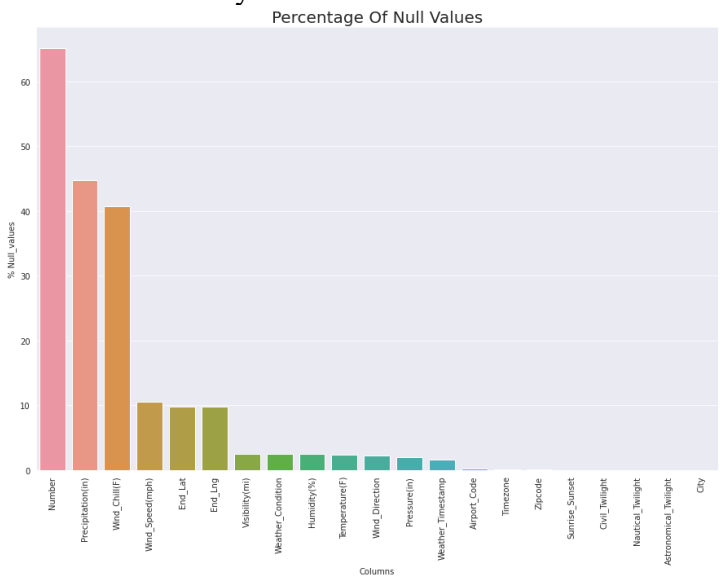

Fig.1 Percentage of Null Values

2. Previous Preparation

2.1 Data Cleaning

First and foremost, I found the dataset of the US Accident on Kaggle. But the data is very messy, which caused cannot be used directly. So secondly, we should clean this dataset, we'll check the null values and remove these columns, which have a lot of null values. Finally, we'll enter appropriate values for the required columns and perform memory optimizations to reduce data memory usage. After analysis, the proportion of empty values in each data can be visualized. Which has been drawn in Fig.1. The horizontal coordinate of Fig1 indicates each type of data in the dataset, and the vertical coordinate indicates the percentage of null values in this type of data.

The visualized chart shows that the data in the first three columns are more than 40 percent null values, so we should omit this data. There's also a big jump from the 7th to 6th column, which may cause some problems. So we should delete the top6 columns. Now, although we still have null values in our dataset, these null values do not have much impact on the next analysis and can be ignored. So the results of our processing are ready to be fed into the EDA model. In the end, to make the data more complete, we can use the '.fillna()' function. It can help us to rational fill the remaining columns with appropriate values.

2.2 Memory Optimization

Converting string datatype to categorical datatype columns is a useful way to help us reduce memory usage. Therefore, we should also convert the columns start_time and end_time to the DataTime datatype. After this operation, memory usage will be reduced from 780+mb (Mbyte) to 366+mb (Mbyte).

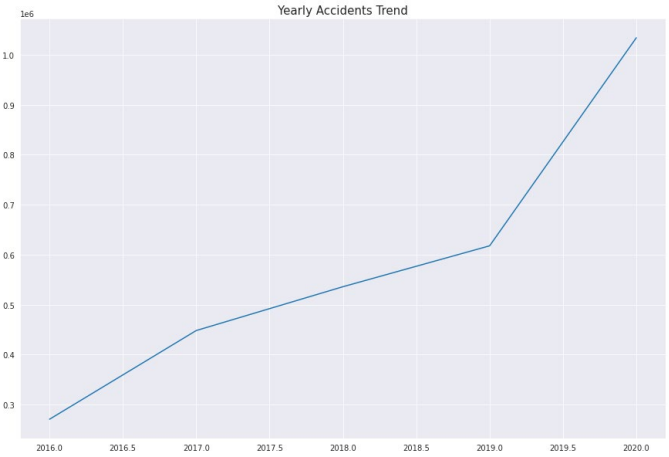

Fig.2 Number of Traffic Accidents in Us from 2016 to 2020

3. Analysis Process

3.1 Integral Analysis: Which State Has the Most Accidents?

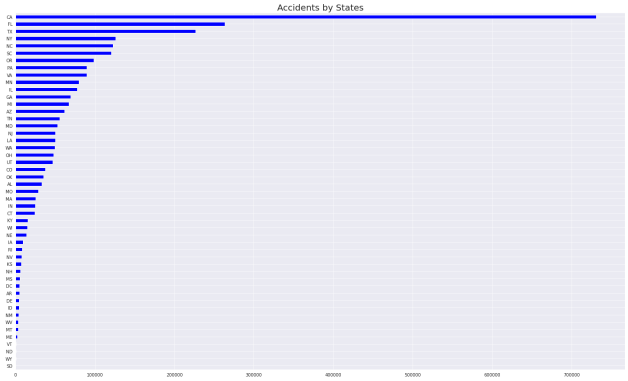

Fig.3 Yearly Accidents Trend

Although we only have accident data from 49 states, it can also fully reflection of the causes of traffic accidents in the United States. Now we can carry out a visualization of accident records, it will help us understand which states in the United States have the highest accident rate. Fig.2 is the result after visualization, its horizontal coordinate indicates the number of accidents and the vertical coordinate is the abbreviation of each state.

It can be seen in Fig.2 that states with the abbreviation CA have the highest number of traffic accidents. By looking up the abbreviations of the U.S. states, we can know that CA means California. It is the largest state in the United States, and its GDP ranking is still in the first place in the nation in 2021 [3], and its total economy is about equal to the fifth largest economy in the world, while several other states with high traffic accident rankings are also the more economically developed states in the United States. It is judged that more traffic accidents occur in states with more economic activities.

3.2 Analysis from the Perspective of Each Year

This dataset contains traffic accident data from 2016 to 2020, which we can now visualize to see how traffic accidents will look from 2016 to 2020.

As we can see in Fig3. As the years grow, accidents show an upward trend, with the highest accident growth rate from 2019 to 2020. The most current studies suggest that the coronavirus epidemic is causing Spanish people to become more aggressive in their driving styles [5]. Americans are no exception. After many years of increased safety and lower wrongful death fatalities on U.S. highways and roads, coronavirus is an outbreak. The coronavirus pandemic has made people live at home for longer periods, which led to fewer cars on the road and made speeding easier. And because many people are restricted from going out, more people are using unsafe ways to spend their lives, such as alcohol and drugs, which leads to more drunk and drugged driving, which leads to a higher risk of traffic accidents.[4]. So the outbreak of the COVID-19 in 2019 is also the main reason for the increase in the accident rate.

3.3 Analysis from the Perspective of Each Month

We are now starting to visualize the monthly accident data. Fig.4 is a visualization of the results. The horizontal coordinate represents the month and the vertical coordinate represents the percentage of accidents per year.

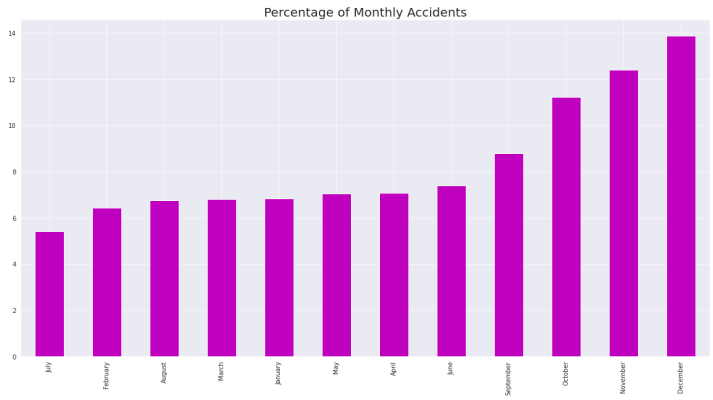

Fig.4 Percentage of Monthly Accidents

We can find from Fig4 that there are more accidents in the second half of the year than in the first half of the year. I speculate that it is most likely due to the lower temperatures and increased rain and snow in the second half of the year. So below we will focus research on the weather conditions.

1) Analysis from the Weather Conditions

As we all know, bad weather conditions can lead to serious traffic accidents. Above I speculated that the reason why more traffic accidents occur in the second half of the year than in the first half is because of the lower temperatures and increased rain and snow in the second half of the year, but Fig.5 shows that the accident rate is higher in sunny and cloudy weather conditions, so that this result negates my guess.

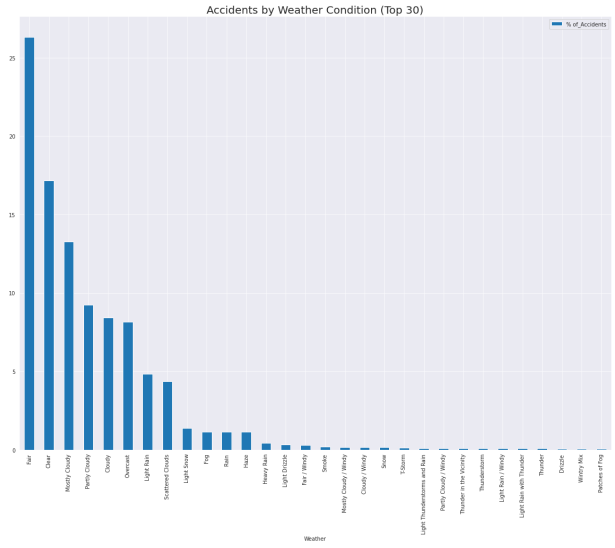

Fig.5 Accidents by Weather Conditions(Top30)

3.4 Analysis from the Perspective of Each Day

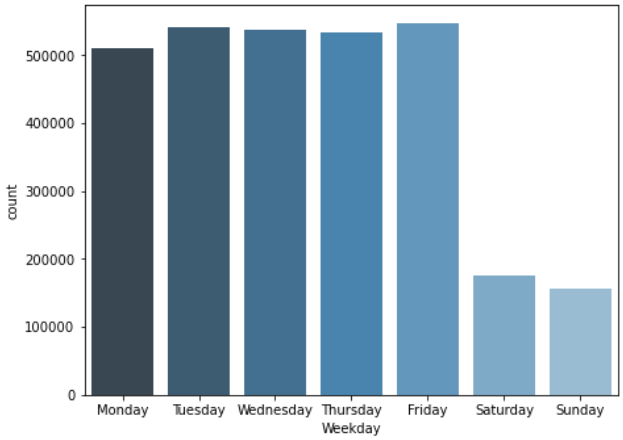

Fig.6 Number of Traffic Accidents Per Day of the Week

Now visualize the weekly accident occurrence from Monday to Sunday.

It is obvious to observe from Fig.6 that there are fewer traffic accidents on Saturday and Sunday. So I speculate that the reason for the higher accident occurrence from Monday to Friday may be related to commuting to and from work, where a higher accident rate may occur during the morning and evening peak hours, so we start the analysis below by period.

1) Analysis Below by Time

First, visualize the daily accident data for time. The visualization result is shown in Fig7. Its horizontal coordinate indicates the time and vertical coordinate indicates the number of accidents

From Fig.7, we can find that the accident rate is the highest in the morning peak (7:00-8:00) and the evening peak (16:00-17:00). This confirms our previous suspicions about the cause of the accident.

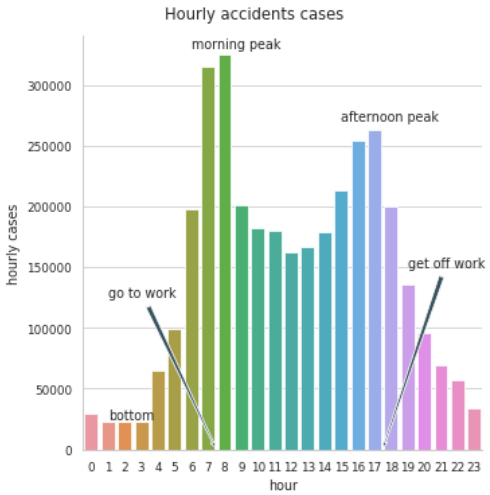

Fig.7 Number of Traffic Accidents by Time of Day

4. Result and Conclusion

1) Higher accident rates in states with stronger economies. The higher GDP indicates a stronger economy and better living conditions for the people. It is speculated that several states with higher GDP have more cars, which leads to more traffic accidents but there is no data to prove it.

2) Increases in accident rates yearly from 2016 to 2020, you can see detail in section3-B. And Surge in traffic accident growth rate linked to an outbreak of COVID-19. Therefore, it is recommended to drive with caution during the period of the COVID 19 pandemic. People should actively adjust their psychological state, reduce alcohol and drug use during the epidemic, and avoid restaurants and drugged driving. This will reduce the occurrence of traffic accidents.

3) In section3-C you can see the accident rate was higher in the second half of the year(Sept to Dec). So People should be more careful when driving in the second half of the year.

4) In section3-C-(1) you can see a higher accident rate in cloudy and sunny weather conditions. So people should be less active outside in cloudy weather, and not take it lightly when driving a car on a sunny day.

5) High accident rate on working days and significantly lower accident rate on rest days. The reason for the high accident rate on weekdays is related to the morning peak and afternoon peak. It can be seen that more details in section3-D. Therefore, people should always pay attention to traffic conditions during the morning and evening rush hours on weekdays and can change their commute to take public transportation or ride a bicycle to and from work.

References

[1] R. Tian, Z. Yang, and M. Zhang, “Method of Road Traffic Accidents Causes Analysis Based on Data Mining,” 2010 International Conference on Computational Intelligence and Software Engineering, 2010, pp. 1-4, DOI: 10.1109/CISE.2010.5677030.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] 60+ Insights Extraction-US Accident Analysis, [online] Available: https://www.kaggle.com/satyabrataroy/60-insights-extraction-us-accident-analysisK. Elissa, “Title of paper if known,” unpublished.

[3] World Population Review. GDP by State 2021 2022. Retrieved, [online] Available: https://worldpopulationreview.com/state-rankings/gdp-by- state.

[4] Traffic Accidents and Wrongful Death Cases Increase During the Covid 19, [online] Available: https://www.californiapersonalinjurylawyersblog. com/traffic-accidents-and-wrongful-death-cases-increase-during-the-covid-19-pandemi

[5] V. Corcoba, X. G. Pañeda, D. Melendi, R. García, L. Pozueco and S. Paiva, “COVID-19 and Its Effects on the Driving Style of Spanish Drivers,” in IEEE Access, vol. 9, pp. 146680-146690, 2021, doi: 10.1109/ACCESS.2021.3124064

[6] C. Parra, C. Ponce, and S. F. Rodrigo, “Evaluating the Performance of Explainable Machine Learning Models in Traffic Accidents Prediction in California,” 2020 39th International Conference of the Chilean Computer Science Society (SCCC), 2020, pp. 1-8, DOI: 10.1109/SCCC51225.2020.9281196.