Analysis of U.S. Car Accident

Tianchen Wang
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
tiwa4690@colorado.edu

Zitao Cheng
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
zich1081@colorado.edu

Yu Li
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
yuli9223@colorado.edu

1 Introduction

In today's society, cars as a vehicle have become a common way to travel for people. The United States has become one of the most holding quantities of car vehicles in the world. In other words, we can call the United States as "The Country on Wheels"; Based on such a big holding quantity, the U.S. car accident rate should not be very low. After examining the dataset of U.S. car accidents, there should be some interesting relationships between accident rate and different attributes like temperature, humidity, etc. that would affect the car accident. That is the primary method of analyzing U.S. car accidents, which in this project goes over the dataset about the U.S. car accident from 2016 to 2020, and the dataset has these attributes: temperature, precipitation, wind speed, pressure, humidity, etc. The primary method in this analysis is examining each attribute and finding the relationship that gets along with the U.S. car accident rate.

2 Specific Questions

- Whether the accident rate is related to day or night
- Whether the time period would affect the accident rate?
- What range of visibility would cause car accidents more frequently?
- Where is the most likely place to have a car accident(county-level)?
- The relationship is between road types and car accident types.
- What's the biggest influence factor on the U.S. car accident rate?

3 Related Work

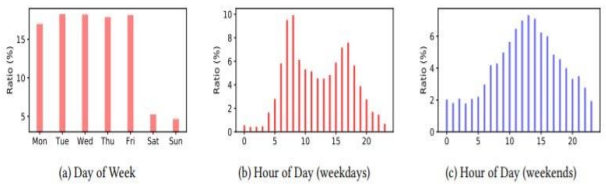
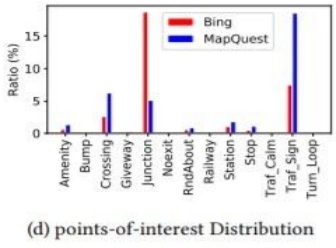


Figure 1 [1]

Figure 1: These graphs are about traffic accident analysis in terms of time. They contain a ratio of car

accidents for seven days a week and hours of workdays and weekends.



(d) points-of-interest Distribution

Figure 2 [2]

Figure 2: The graph is a histogram about the traffic accident rate based on road type. Especially, junctions and traffic signals in the nearby location have a high car accident rate.

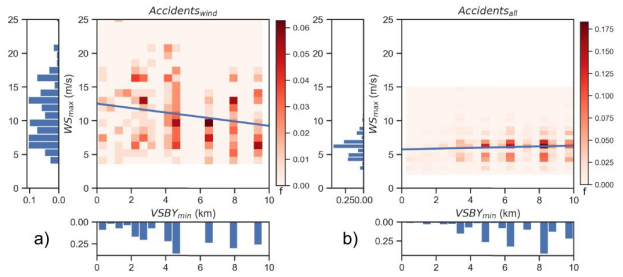


Figure 3 [3]

Figure 3: These heat maps are shown the impact of wind speed and visibility on traffic accidents, which indicates wind speed has more impact on traffic accidents than visibility.

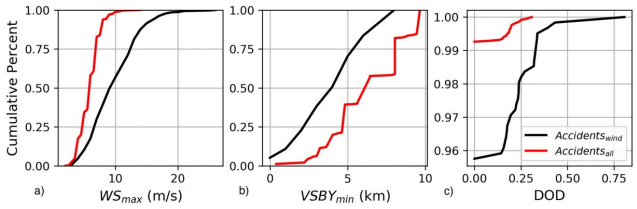


Figure 4 [4]

Figure 4: The line chart above shows the relationship between the cumulative percent of all accidents and three wind and visibility-related attributes. It indicates that Wind speed and visibility have a significant impact on car accidents, but DOD does not impact car accidents.

Figure 5 [5]

Figure 5: The geographical map reveals the number of accidents through the intensity of states' color. Rhode Island and Florida have the lowest and highest car accident rate, separately.

4 Data set

Url:
https://www.kaggle.com/sobhanmoosavi/us-accidents

Our dataset has about 3.5 million objects with 49 attributes including accident's severity, location, weather condition, road types, time, and so on.

Next, there are the attributes we used in data mining:

Source: the source of the accident report

Severity: the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).

Start_time: the start time of the accident in the local time zone.

End_time: the end time of the accident in the local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.

Start_lat: the latitude in GPS coordinate of the start point.

Start_Lng: the longitude in GPS coordinate of the start point.

County: the county of accident in address record.

Temperature(F): the temperature at the time of the accident (in Fahrenheit).

Humidity(%): the humidity at the time of the accident (in percentage).

Visibility: the visibility at the time of the accident (in miles).

Wind_Speed(mph): the wind speed at the time of the accident (in miles per hour).

Weather_Condition: the weather condition at the time of the accident (rain, snow, thunderstorm, fog, etc.)

Amenity: the presence of amenity in a nearby location.

Bump: the presence of speed bump or hump in a nearby location.

Crossing: presence of crossing in a nearby location.

Give_Way: presence of give_way in a nearby location.

Junction: presence of junction in a nearby location.

No_Exit: presence of no_exit in a nearby location.

Railway: presence of railway in a nearby location.

Roundabout: presence of roundabout in a nearby location.

Station: presence of station in a nearby location.

Stop: presence of stop in a nearby location.

Traffic_Calming: presence of traffic_calming in a nearby location.

Traffic_Signal: presence of traffic_signal in a nearby location.

Turning_Loop: presence of turning_loop in a nearby location.

Sunrise_Sunset: the period of day (i.e. day or night) based on sunrise/sunset.

5 Main Work

5.1 Data Cleaning

In order to analyze the data set more easily and increase analysis efficiency, data cleaning has to be processed more specifically. Some attributes that should be used usually have missing blocks; in that case, we decide to fill the involved attribute's mean or mode value based on the attribute's type is appropriate for data cleaning.

5.2 Data Selection

The accident dataset collected data from February 2016 to June 2020 while our project is working in 2020. Therefore, our dataset is relatively accurate.

Based on the dataset we found, there are 49 columns totally, but not all of the columns we would use to analyze the U.S. car accident; In this situation, we have to choose the useful and valuable attributes to do data mining. Some attributes like start-time, end-time, end-latitude, end-longitude, distance, description, etc. These attributes would be considered as a non-valuable attribute in the process of data selection.

Besides, we delete some irrelevant attributes such as “End_Lat”, “End_Lng”, “Number”, and “Precipitation(in). The first two columns of latitude and longitude data at the end of the accident are empty, so we delete them. The third attribute, street number, is not helpful to us, so we also decided to delete it, equivalent to data compression. We want to use the last data precipitation for analysis, but its data is seriously insufficient, and we have to discard it.

5.3 Graphic Analysis

Multiple visualizations are needed to implement into this analysis; A histogram would be a good starting point of analysis because it is similar to figure 1 and figure 2. Also, A frequency distribution graph compared with three elements would be an excellent way to get the desired result since it would be similar to figure 3. What's more, we try to find a distribution of U.S. maps, which presents the U.S. accident rate in each state, and even better, we could show the county unit's accident rate.

Figure 6(1)

Figure 6(1): The U.S. accident rate is happening each day of a week from Monday to Sunday.

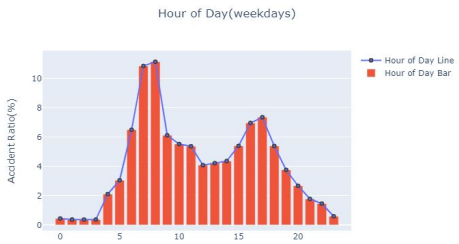


Figure 6(2)

Figure 6(2): The U.S. accident rate is happening each hour of the day on the weekday.

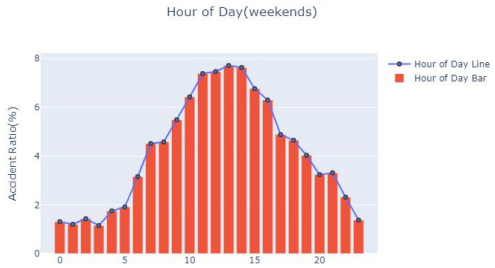


Figure 6(3)

Figure 6(3): The U.S. accident rate is happening each hour of the day at the weekend.

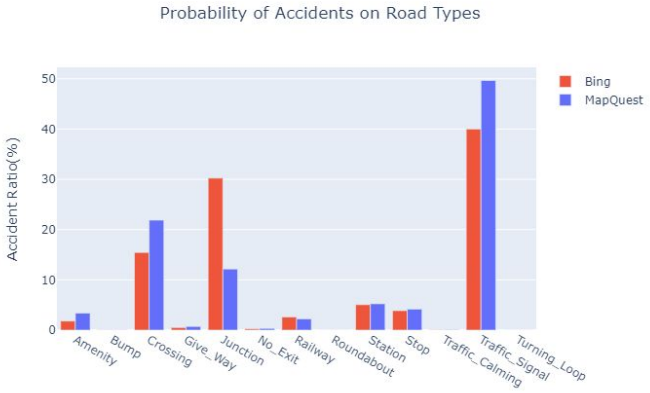


Figure 7

Figure 7: The probability of U.S. accident rate on different road types collected from Bing or MapQuest.

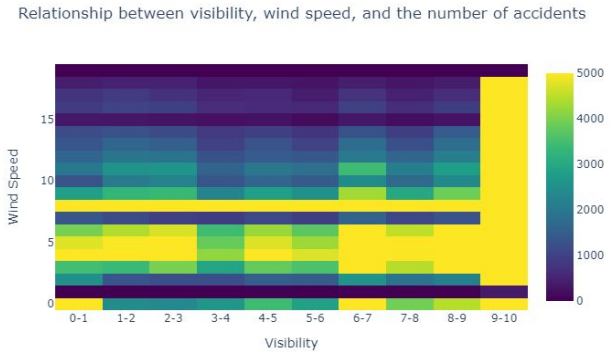


Figure 8

Figure 8: The accident rate happening in different wind speed and visibility conditions.

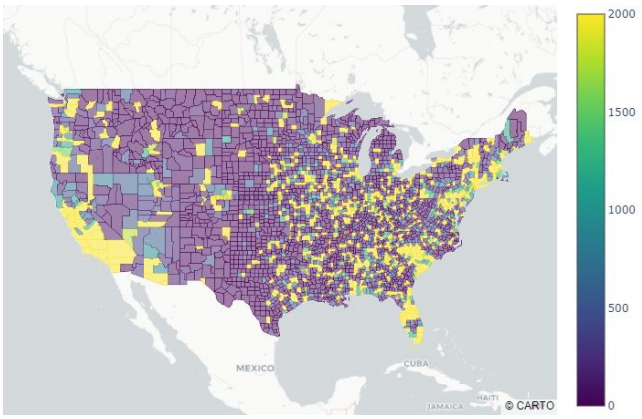


Figure 9

Figure 9: The counts of car accidents in different counties. (Alaska, Hawaii, and Puerto Rico are not include in this map)

5.4 Analyzing the data

After we finish doing the visualization, the next step is doing some calculations to find a pattern and relationship between two selected attributes.

5.4.1 Q&A

Q: Whether the accident rate is related to day or night?
A: The accident rate definitely is related to day or night. The reason is that based on both Figure 6(2) and Figure 6(3), the accident ratio does not distribute at night; instead, most car accidents appeared during the day. Because, not too many cars would be driven on the street during the night, and it will definitely decrease the accident ratio.

Q: Whether the time period would affect the accident rate?
A: The time period would definitely affect the accident rate. It usually depends on weekdays or weekends. During the weekday, it is usually during the rush hour; people need to go out for work, and it would increase the vehicle flow rate and cause the accident ratio to be much higher than the weekend. During the weekend, it is usually during the mid-day; people don't need to work during the weekend; instead, the majority of people would like to have a high quality of lunch or take some outdoor activities. As a result, it would cause the vehicle flow rate to be increased; relatively, the accident rate increased as well. So, the time period would definitely affect the accident rate.

Q: What range of visibility and wind speed would cause car accidents more frequently?
A: According to the graph, it's difficult to figure out what level of visibility would cause car accidents more frequently. Ideally, it should show that higher wind speed and low visibility would cause a higher accident rate; however, the graph does not show the expected result. Instead, the accident rate happens at most during the best visibility level and wind speed level within 4 mph and 8 mph. As a result, we figure out that the car accident normally happens during clear visibility and low wind speed level day. To conclude, the car accident might be highly related to the high vehicle flow rate, because people are more likely going out during weather conditions.

Q: Where is the most likely place to have a car accident(county-level)?
A: The most likely place to have a car accident should be some highly developed areas in the United States, for example, California, New York, Chicago, etc. More developed regions should have bigger vehicle flow rates, and because of that, the accident rate must be higher than in other regions. A highly developed region needs that amount of vehicle flow rates to develop the economy.

Q: The relationship is between road types and car accident rates.
A: Several road types definitely affect the accident rate; some specific roads that have a junction, Crossing, or

traffic signal would be considered as the biggest influence factor in a car accident. These types of roads usually have a property of mixed and high flow rates. It definitely would cause a car accident to happen. Also, a type of road that has traffic signals has the highest accident ratio. The reason is that too many people would like to run a red light since most of them don't want to wait for the next green light to come and want to save some time. Besides, the road type of crossing and junction, it has too many factors that need to be taken care of; that's why these road types have relatively high accident ratios.

Q: What's the biggest influence factor on the U.S. car accident rate?
A: We used to think that wind speed, visibility would definitely be the main reason for a car accident; when the driver view is unclear and weather condition is bad, people used to think a car accident would happen, however as we discovered on the graph, some environmental factor like the level of wind speed, visibility do not affect the accident ratio very much. We think that only the vehicle flow rate would be considered as the biggest influence factor on the U.S. car accident rate. As we all know, The United States is also known as "The Country on Wheels". Too many vehicles would cause plenty of uncertainties in transportation. As we can conclude above, during the rush hour of the weekday or midday of the weekend have a higher accident rate compared to other time periods. That would be one of the clear evidence to support this idea.

5.4.2 Information Gained from Graph

As we can see from figure 6(1), the accident ratio would be much higher on the weekday compared to the weekend. For the figure 6(2), the accident ratio would be much higher during the morning and evening rush hours from 7 to 8 a.m. and from 4 to 5 p.m. It's the regular business hours that people go out for work. Based on figure 6(3), the accident would more likely happen at midday during the weekend and not too many accident rates in the morning and evening. The peak of the accident rate would be at 1 p.m.

According to Figure 7, only the junction road type has the source gap between Bing and MapQuest; the source of Bing is much higher than MapQuest's. For the other road types, it basically has similar results. What's more, the road type that has a traffic signal, junction, and crossing would be considered as the most frequent location of accidents; the area that has traffic signals has the highest accident rate.

Based on Figure 8, we realized that when visibility is between 9 and 10, the accident rate happens very often no matter the level of winter speed; also, when the level of wind speed is 8, the accident rate happens very often no matter the level of visibility. However, we do clearly know that when visibility is between 0 and 1, the accident rate happens a lot. To conclude, when the visibility is very high and the level of wind speed is very low, the probability of accident rate would become very high.

As stated in Figure 9, we figured out that more developed areas in the U.S. region would be more likely to have higher accident rates. Be more specific, coastal regions like California state, New York State and Chicago would have higher accidents compared to Ulta, New Mexico, and Montana. The mid-region of the United States would be relatively considered as a low accident rate region since we barely found the yellow part in this region; compared to coastal regions, there are almost yellow parts everywhere on that coastal regions.

5.5 Evaluation

We will use an official traffic accident analysis from the authority to evaluate our analysis result. "A Countrywide Traffic Accident Dataset." by Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath presents the same

dataset we will work on, along with a wide range of insights gleaned from this dataset with respect to the spatiotemporal characteristics of accidents will be mainly used for our comparison. “Characterizing the Role of Wind and Dust in Traffic Accidents in California” by Abinash Bhattachan Gregory S. Okin Junzhe Zhang Solomon Vimal Dennis P. Lettenmaier will also be used to help us analyze our results about wind speed and humidity. Besides, we will use car accident statistics in 2020 to evaluate our analysis of county-level places.

5.6 Tools

- Pandas
- Numpy
- Ploly
- Jupyter Notebook
- Microsoft Excel

6 Milestones

6.1 Milestone 1

Do data cleaning work

Due date: Oct. 11

6.2 Milestone 2

Finish data cleaning work and begin doing data selection work

Due date: Oct. 25

6.3 Milestone 3

Finish doing data selection work and begin doing graphic analysis

Due date: Nov. 8

6.4 Milestone 4

Finish all code part of our project and begin doing the first draft of the final report

Due date: Nov. 15

6.5 Milestone 5

Finish the final version of our project report

Due date: Nov. 22

6.6 Milestone 6

Finish the slides for the final presentation and prepare to present

Due date: Nov. 29

6 REFERENCES

[1] S. Moosavi, M. Hossein, S. Parthasarathy, R. Teodorescu, and R. Ramnath, “Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights,” arXivLabs, 2019. [Online]. Available: https://arxiv.org/abs/1909.09638. [Accessed: 2020].

[2] S. Moosavi, M. Hossein, S. Parthasarathy, R. Teodorescu, and R. Ramnath, “Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights,” *arXivLabs*, 2019. [Online]. Available: https://arxiv.org/abs/1909.09638. [Accessed: 2020].

[3] A. Bhattachan, G. S. Okin, J. Zhang, S. Vimal, and D. P. Lettenmaier, “Characterizing the Role of Wind and Dust in Traffic Accidents in California,” AGU Journals, 28-Oct-2019. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019GH000212. [Accessed: 2020].

[4] A. Bhattachan, G. S. Okin, J. Zhang, S. Vimal, and D. P. Lettenmaier, “Characterizing the Role of Wind and Dust in Traffic Accidents in California,” AGU Journals, 28-Oct-2019. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019GH000212. [Accessed: 2020].

[5] S. Mehta, “29 Road Rage Statistics That Drivers Must Know (2020 Update),” *carsurance*, 26-Feb-2020. [Online]. Available: https://carsurance.net/blog/road-rage-statistics/. [Accessed: 2020].