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 Abstract

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 "The Country on Wheels"; Based on such a large 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.

2 Introduction

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. This analysis's primary method examines each attribute and finds the relationship that gets along with the U.S. car accident rate. After analyzing the data, we could figure out where and when that high resident ratio happened in the United States so the government could implement relevant strategies to control the car accident ratio. Here are some problems provided based on the data set we chose and some relevant research papers to compare later.

- Whether is the accident rate related to day or night?
- Whether the time period (day, month, year) 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 are the influence factors on the U.S. car accident rate?

3 Related Work

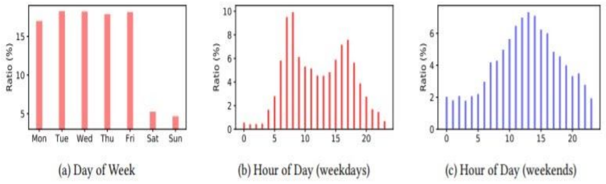

Figure 1 [1]

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.

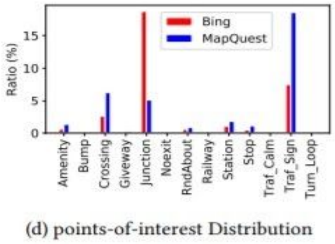

Figure 2 [2]

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

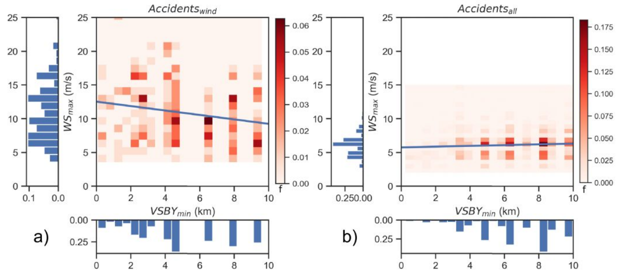

Figure 3 [3]

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.

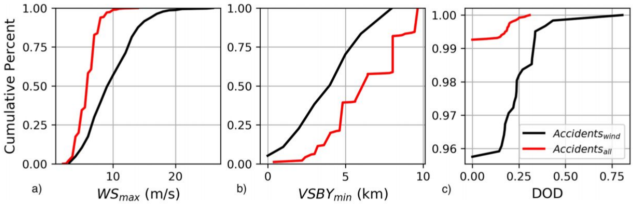

Figure 4 [4]

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 significantly impact car accidents, but DOD does not affect car accidents.

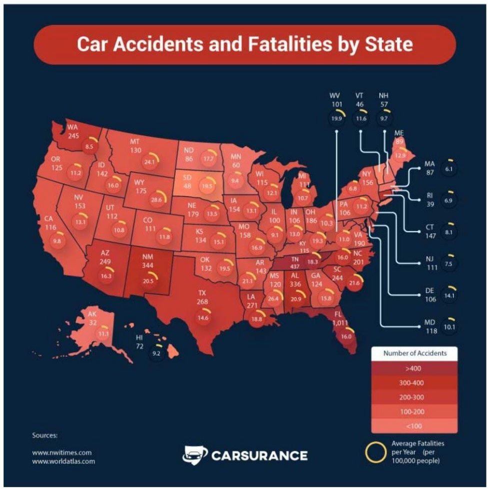

Figure 5 [5]

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.

Some of the above pictures are related work that has been completed. Our job is to generate our graphs and compare them with the above images to see any discrepancies. Based on these, we would use more data to analyze and then get more comprehensive conclusions.

4 Methodology

4.1 Dataset

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

Our dataset has about 3.5 million objects with 49 attributes, including accident 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 an 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 accidents in address record.

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

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 a speed bump or hump in a nearby location.

Crossing: the presence of crossing in a nearby location.

Give_Way: the presence of give_way in a nearby location.

Junction: the presence of a junction in a nearby location.

No_Exit: the presence of no_exit in a nearby location.

Railway: the presence of a railway in a nearby location.

Roundabout: the presence of a roundabout in a nearby location.

Station: the presence of a station in a nearby location.

Stop: the presence of a stop in a nearby location.

Traffic_Calming: the presence of traffic_calming in a nearby location.

Traffic_Signal: the presence of traffic_signal in a nearby location.

Turning_Loop: the presence of turning_loop in a nearby location.

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

4.2 Tools

- Pandas
- Numpy
- Plotly
- Jupyter Notebook
- Microsoft Excel

4.3 Main Tasks

Data cleaning must be processed more specifically to help people analyze the dataset more efficiently. 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.

4.3.1 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.

4.3.2 Graphic Analysis

Multiple visualizations are needed to implement this analysis; A histogram would be a good starting point of research 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.

(Just be a remainder, since the data set we chose does not support the whole year information at the years 2016 and 2020, so it’s unlikely to support some of the graphs below.)

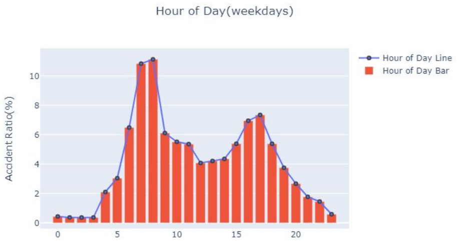

Figure 6

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

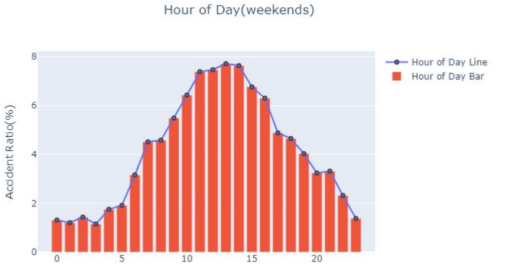

Figure 7

Figure 7: The U.S. accident rate is happening each hour of the day at the weekend.

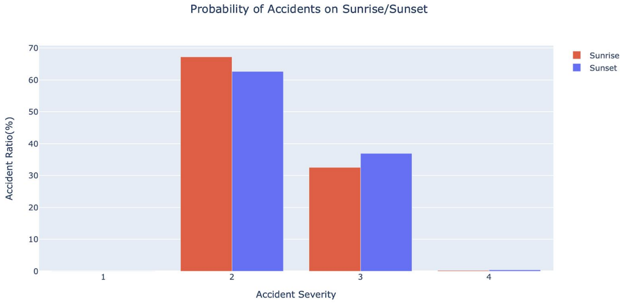

Figure 8

Figure 8: The counts of sunrise accidents and sunset accidents at different severity.

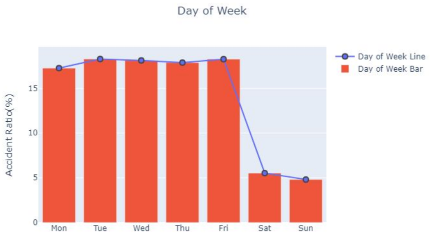

Figure 9

Figure 9: The U.S. accident rate is happening each day of the week from Monday to Sunday.

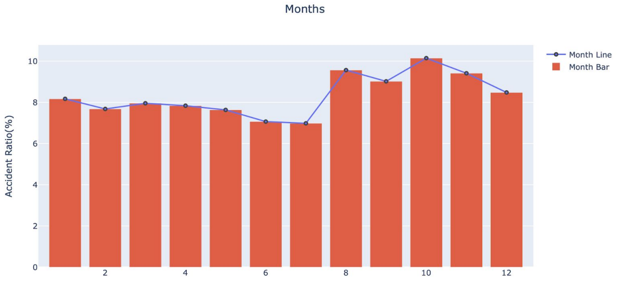

Figure 10

Figure 10: Monthly U.S. accidents that happened in 2017, 2018, and 2019.

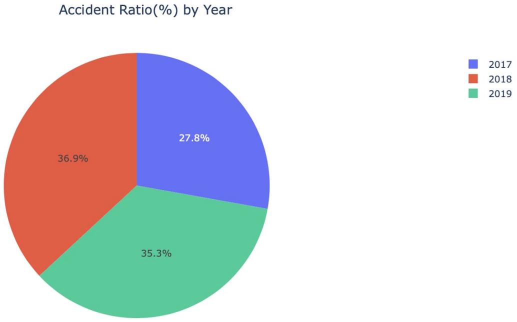

Figure 11

Figure 11: Total U.S. accident counts in 2017, 2018, and 2019. All of these years have approximately the same U.S. car accident proportion

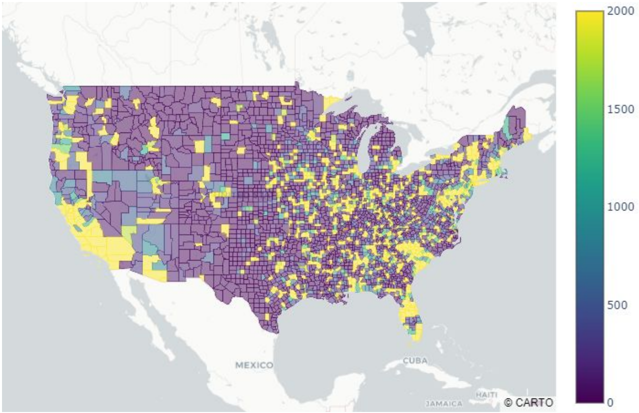

Figure 12

Figure 12: The U.S. accidents count in different counties. (Alaska, Hawaii, and Puerto Rico are not included in this map)

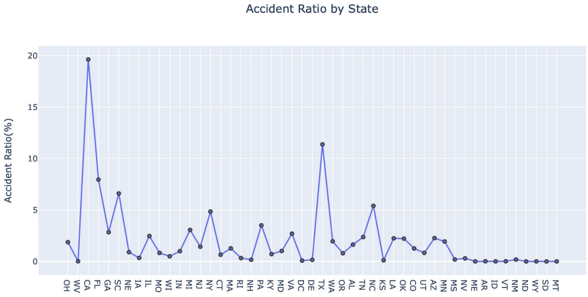

Figure 13

Figure 13: The U.S. accident rate in different states in a line chart (The state of California has the most severe car accidents ratio)

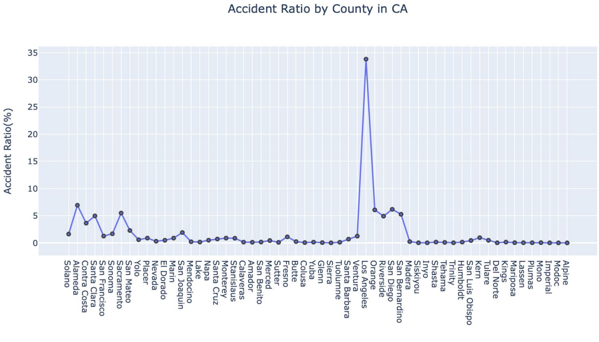

Figure 14

Figure 14: The accident rate in California as a county Level which Los Angeles has the highest accident ratio

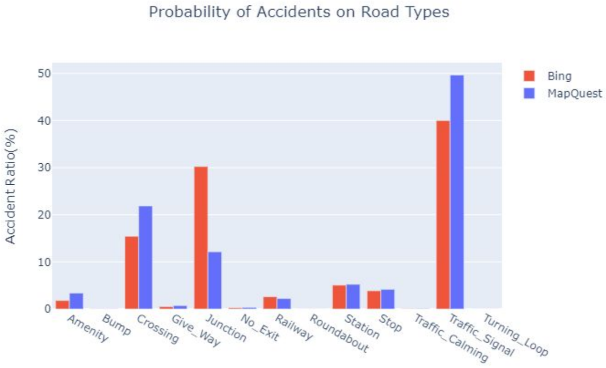
Probability of Accidents on Road Types

Figure 15

Figure 15: The probability of the U.S. accident rate on different road types collected from Bing or MapQuest.

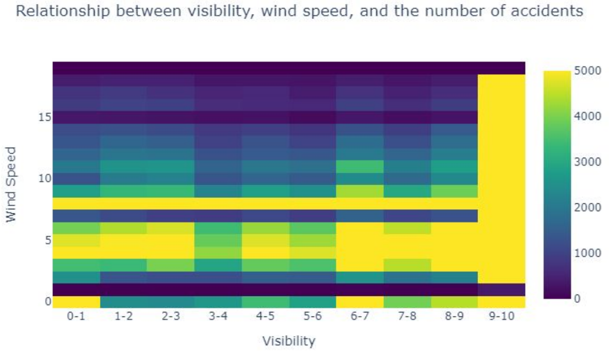
Relationship between visibility, wind speed, and the number of accidents

Figure 16

Figure 16: The accident rate happening in different wind speed and visibility conditions.

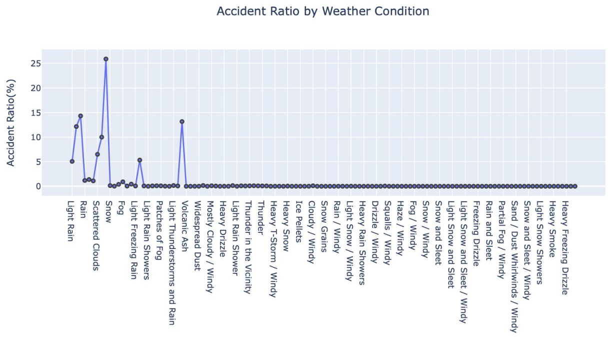
Accident Ratio by Weather Condition

Figure 17

Figure 17: The U.S. car accident rate depends on different weather conditions. Good weather conditions have a higher accident ratio.

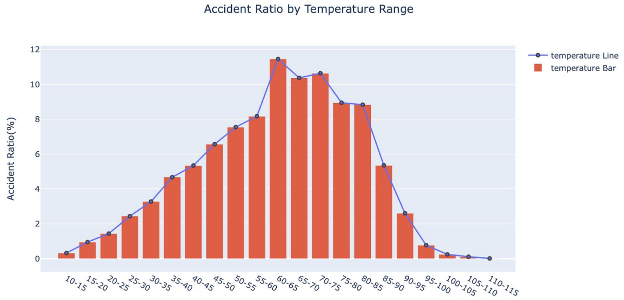
Accident Ratio by Temperature Range

Figure 18

Figure 18: The U.S. car accident rate based on temperature range, which approximately appears as a normal distribution

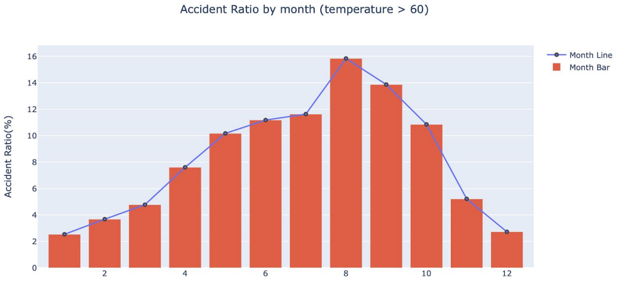
Accident Ratio by month (temperature > 60)

Figure 19

Figure 19: The U.S. car accident rate distributed only the temperature is above 60℉

4.4 Analytical Thinking

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

4.4.1 Q&A

Q: Whether the accident rate is related to day or night?
A: The accident rate is related to day or night. The reason is that based on Figure 8, the accident ratio becomes higher during the night when the accident severity level goes from two to three. When the time period transfers from day to night, the visibility will definitely be affected, and sometimes it would be the main reason for an accident.

Q: Whether the time period would affect the accident rate?
A: The time period would affect the accident rate. It usually depends on weekdays or weekends. During the weekday, it is generally 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, most people would like to have a high-quality lunch or do 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 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 challenging 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 yield a higher accident rate; however, the graph does not show the expected result. Instead, the accident rate happens most during the best visibility level and wind speed level within four mph and eight mph. As a result, we figure out that the car accident occurs typically during clear visibility and low wind speed level day. To conclude, car accidents might be positively related to the high vehicle flow rate because people are more likely to go 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 Los Angeles. The reason is that California is the state with the highest car accident ratio in the United States. When we are more specific and go detailed into California, Los Angeles has the highest accident ratio because it contains about 33% of the accident ratio in California, which is one-third of the whole state's accident ratio. Los Angeles has such a high car accident ratio because LA is the biggest city in California and it's the second biggest city in the United States; based on that huge population, that probably is the main reason why LA has the highest car accident ratio. What's more, plenty of people would go to Los Angeles as visitors and that's also another reason for the high accident ratio.

Q: The relationship is between road types and car accident rates.
A: Several road types affect the accident rate; some specific roads that have a junction, Crossing, or traffic signal would be considered as the most significant 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 kind 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 has too many factors that need to be taken care of; that's why these road types have relatively high accident ratios.

Q: What are the influence factors on the U.S. car accident rate?
A: We used to think that wind speed, visibility, the temperature 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, the level of wind speed, visibility does not affect the accident ratio very much; temperature is kind of the only environmental factor that would affect the accident ratio. When the weather is above 60℉, the accident ratio in the United States will definitely have a remarkable increase. Based on these facts, we could not continue to keep our initial assumption, which environmental factors would dominate the U.S. car accident ratio. Thus, we came out with a full belief that the vehicle flow rate would be considered the majority 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 stated in the graphical analysis above, we could realize that during the rush hour of the weekday or midday of the weekend have a higher accident rate compared to other time periods; these time periods would typically be considered as high vehicle flow rate and that can be one of the clear evidence to support this idea.

4.4.2 Information Gained from Graph

As we can see from Figure 6, the accident ratio would be much higher on the weekday compared to the weekend. For Figure 7, 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 9, the accident would more likely happen at midday during the weekend, and not too many accident rates in the morning and evening. The accident rate peak would be at 1 p.m.; after analyzing Figure 8, we figured out that the accident ratio of daytime is higher than the night when the accident severity level is two. However, when the accident severity goes to three, the night's accident ratio is higher than the daytime. In other words, we could say that the factor of sunrise and sunset would probably be the influence factor of the U.S. car accidents ratio as we combined the information between Figure 10 and Figure 18. Figure 10 shows that the accident ratio has a significant increase starting in July and does not decrease from October; as we see in Figure 18, the high accident ratio is centralized between 60℉ and 85℉; this region is usually considered during the summer months. After we combined this information, we could conclude that the car accident usually happened during the summer. The high temperature would probably cause the car driver to be distracted. 60℉ would not be considered a high temperature, but when we let the vehicle be exposed to the sun for hours, the temperature inside the car would not just be 60℉; it would usually be higher 8 to 10℉.

Figure 11 basically talks about the U.S. car accident ratio between 2017 and 2019. As we can see from the graph, even though the increasing rate of accidents is not that distinct, the car accident ratio still increases year by year. It is probably because vehicle ownership is being increased year by year as well. As stated in Figure 12, we figured out that more developed areas in the U.S. region would be more likely to have higher accident rates. To be more specific, coastal regions like California, New York, and Texas would have more severe accidents than 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 the coastal areas, there are almost yellow parts everywhere on that coastal regions. When we transferred the map chart to the line chart in Figure 13, each state's difference would be more apparent, so we find out that California has the highest accident ratio in the United States. And then, we figure out that Los Angeles has the highest accident ratio at the county-level when we go deeper into the county level; based on Figure 14, the accident ratio of Los Angeles is like ten times as much as other counties.

According to Figure 15, only the junction road type has the source gap between Bing and MapQuest; Bing's source is much higher than MapQuest's. For the other road types, it basically has similar results. More importantly, the road types with a traffic signal, junction, and crossing would be considered the most frequent location of accidents; the area with traffic signals has the highest accident rate.

Based on Figure 16, 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 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. Figure 17 shows an exciting part of this graph; the high accident ratio should be centralized at bad weather conditions instead of good weather conditions. However, the fact is totally in the opposite direction. The weather condition of clear has the highest accident ratio, about 25% of the total U.S. accident. The following weather condition, "mostly cloudy and fair" with significant proportions of accidents, is still considered a good weather condition. Nevertheless, some bad weather conditions like Snow and Light Freeze Rain only contains about 0.16% and 0.04% of the whole U.S. car accidents. Figure 18 is still a kind of graph based on the weather condition, which talks about the temperature. As we mentioned before, the graph states that the accident ratio would be much higher when the temperature is higher than the human comfort temperature. To double-check the correct analysis, we make a more specific graph to double-check that conclusion. In Figure 19, a type of bar graph, that shows the U.S. accident happened only when the temperature is above 60℉, and we figured out that the high accident ratio only happened during the summer season. After analyzing Figure 19, we could say that higher temperatures would definitely cause car accidents more frequently in the United States.

5 Evaluation

We will use an official traffic accident analysis from the authority to evaluate our analysis result. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights" by Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath, who present 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. Based on what we have in our analysis, the U.S. accident rate is happening each hour of the day on the weekday and weekend; also the accident rate The U.S. accident rate of a week from Monday to Sunday, they all have basically the same graphic result which shows that the car accident would be more likely to happen during the rush hour of the weekday and midday of the weekend. What's more, the weekday accident rate is much higher than the weekend accident rate. Afterward, we studied the relationship between the severity of the accident and day and night to find that there are more minor accidents during the daytime, but more fatal accidents happen at night. However, another research paper[6] found that there are more fatal accidents during the day than at night. Since our conclusions conflict with the findings of other studies, this viewpoint needs to be further studied.

The conclusions are almost the same for road types even though the graph we plot through our data set is slightly different from the graph from the research we mentioned at the beginning of the previous paragraph. Therefore, the accident rate is relatively high on the road around junctions and traffic signal lights.

"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. However, based on what we have data mined, the car accident would more likely happen during the high level of visibility and low wind speed, which is totally different from the paper we found. As the paper stated, car accidents are more likely to happen when visibility is low, and wind speed is very high. These differences appeared because that paper only analyzed terrible weather car accidents; however, the data set we chose is all U.S. car accident rates between 2016 and 2020. That's probably one of the main reasons for this situation. Besides, the transparent and fair weather conditions take up most accident rates in the United States because of this data set, shown in Figure 17. A more targeted data set would definitely lead to a different conclusion.

"29 Road Rage Statistics That Drivers Must Know (2020 Update)" by Sushant Mahta has a map-style graph that records the number of accidents in each state of the United States. Based on this idea, we use our data set to create a map-style graph as well. Since California has the highest accident ratio based on Figure 13, we also make a line chart, Figure 14, to show the accident ratio only in California as the county level offers more specific information on this map-style graph. "Car accident, drownings, violence: the hotter temperature will mean more deaths from injury" by Liz Hanna[7] shows how high temperatures affect the car accident. Specifically, the paper states that increased body temperature would cause a loss of concentration and fatigue. If these symptoms appeared in the driver's body, it would definitely cause a car accident. As compared to our analysis, based on Figures 18 and 19, we got a similar result: when the temperature is between 60℉ and 85 ℉, the accident rate has a distinct increase.

6 Discussion

After going over the whole project, it's crucial to do normalization on the data sets for data cleaning and graphical analysis, so it would be much easier to evaluate the graphical result. AIt'slso an excellent idea to come out with several questions to be answered; we would be more focused and concentrated on graphic analysis to increase working efficiency. However, when we go over the evaluation part, we did not do very well on comparing our research and online research paper because the result we got in our research paper is hard to find highly relevant research paper to make a comparison. Some papers do not have the information we want. For the future, we would like to produce a related mathematical model to predict where would have a relatively high resident ratio at the county level. As a result, the government could take advantage of that.

7 Conclusion

Based on the mining of our data, we found several patterns in US car accidents. Some graphical analysis like histogram, map chart, heat map, and line chart gave us a very detailed and precise result. In order to do data cleaning, we put the involved attribute's mean or mode value in the missing block. What's more, comparing our research paper and online research paper to find similarities and differences. Most of our analysis results are similar to the results of related works. To sum up, we have the following advice for drivers in the US.

1. Drivers should drive more carefully during weekdays' rush hour and weekends' noon.
2. Since there is a lot of traffic in metropolises, drivers should pay more attention to cars' distance to avoid unnecessary traffic accidents.
3. In densely populated cities, drivers should focus more on road conditions to reduce car accidents.

4. Traffic lights are the most critical parts in the traffic, pay close attention to the traffic lights, especially at the junctions and crossings.
5. During fair days and clear days, people tend to have more outdoor activities. Hence there will be more traffic jams, and drivers should drive more carefully.
6. In summer, high temperatures will cause drivers' inattention; people should avoid driving in such conditions.

8 Milestones

8.1 Milestone 1

Do data cleaning work

Due date: Oct. 11

8.2 Milestone 2

Finish data cleaning work and begin doing data selection work

Due date: Oct. 25

8.3 Milestone 3

Finish doing data selection work and begin doing graphic analysis

Due date: Nov. 8

8.4 Milestone 4

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

Due date: Nov. 15

8.5 Milestone 5

Finish the final version of our project report

Due date: Nov. 22

8.6 Milestone 6

Finish the slides for the final presentation and prepare to present

Due date: Nov. 29

9 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].

[6] K. Zhang, M. Hassan, "Crash Severity analysis of nighttime and daytime highway work zone crashes," PLoS ONE, 13-Aug-2019. [Online]. Available:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6692090/

[7] L. Hanna, "Car accident, drownings, violence: hotter temperature will mean more deaths from injury," The Conversation, 13-Jan-2020. [Online]. Available:

https://theconversation.com/car-accidents-drownings-violence-hotter-temperatures-will-mean-more-deaths-from-injury-129628

10 Appendix

10.1 Honor Code Pledge

On my honor, as a University of Colorado Boulder student, I have neither given nor received unauthorized assistance.

10.2 Individual Contribution

Tianchen Wang: Mainly focus on writing the research paper, including Abstract, Introduction, Q&A, Information Gained from Graph, Evaluation, and Reference.

Zitao Cheng: Information collection and designs of graphs

Yu Li: Mainly focus on data cleaning and graph generation