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

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.

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.

Related Work

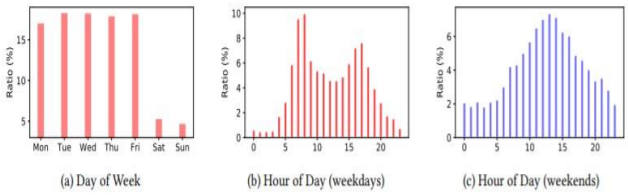

Figure 1 [1]

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.

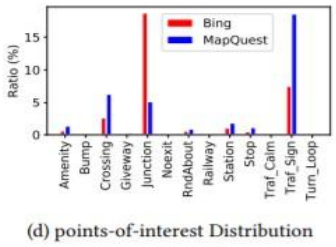

Figure 2 [2]

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

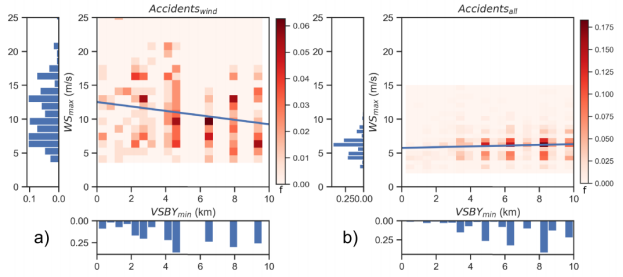

Figure 3 [3]

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.

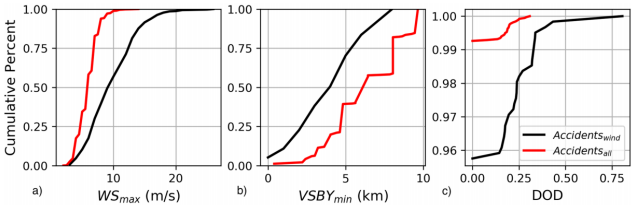

Figure 4 [4]

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.

Proposed Work

Data Cleaning and Data Selection

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 value as an appropriate way for data cleaning.

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.

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.

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. For example, suppose the two chosen attributes are the car accident rate and visibility. In that case, when we draw a line chart, it should show a positive line or negative line to reveal the relationship between two attributes, and we can get an equation corresponding to that line.

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.

Milestones

Milestone 1

Do data cleaning work

Due date: Oct. 11

Milestone 2

Finish data cleaning work and begin doing data selection work

Due date: Oct. 25

Milestone 3

Finish doing data selection work and begin doing graphic analysis

Due date: Nov. 8

Milestone 4

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

Due date: Nov. 15

Milestone 5

Finish the final version of our project report

Due date: Nov. 22

Milestone 6

Finish the slides for the final presentation and prepare to present

Due date: Nov. 29

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].