Name: YAOXING CHEN Student ID: 32664443 Tutor: Jesmin Nahar, Tutorial 6.

FIT5147 Assessment 2:

The relationship between the amount of traffic violations and accidents in the United States

1. Project Introduction

1.1 Motivation:

I just got my driver's licence not long ago, but the test set by the government and the real road conditions are completely incomparable. When you are driving on the road, the weather, road conditions and your own physical condition all affect the driving safety factor.

The selection of the United States, which ranks among the top car ownership per capita, as a data source to ensure that the results of the project have some significance

In this project, in addition to exploring the weather at the time of the accident, I wanted to introduce another variable: the number of traffic violations in the area where the accident happened. Because the number and types of violations can help find the cause of traffic accidents.

1.2. Questions:

1. Among all weather conditions, which weather conditions have the most traffic accidents, will this result change due to changes in the location and time period of the accident?

2. Will the increase or decrease in the number of traffic violations significantly reflect the increase or decrease in traffic accidents, is there a certain pattern?

2.Data processing

2.1 Data wrangling

2.1.1 Data description

1. Traffic accidents in the United States between 2016 and 2022, including associated weather conditions and locations.

2. Traffic violation records that occurred between 2012 and 2017, including the place and time of occurrence, and the type of violation, most of the data from Maryland.

3. Traffic accidents in Maryland between 2015 and 2022, including associated weather conditions and locations.

2.1.2 Data size and data link:

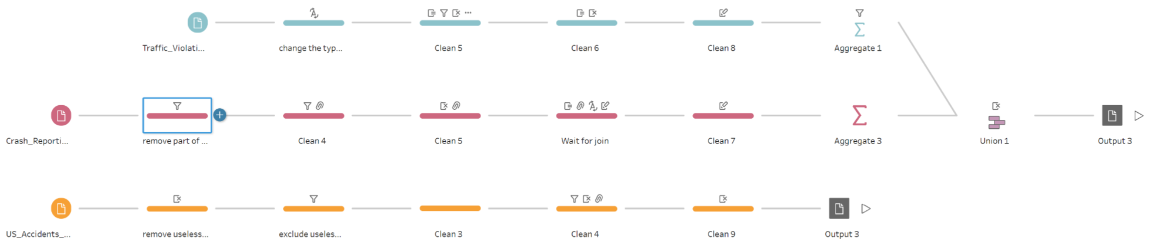
1. Tabular data: 2,845,342 rows x 47 columns (https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)

2. Tabular data: 1,087,110 rows x 35 columns (HTML) (https://data.world/jrm/trafficviolations/workspace/file?filename=Traffic_Violations.csv)

3. Tabular data: 80,000 rows x 44 columns (https://catalog.data.gov/dataset/crash-reporting-incidents-data)

TIP: After a thorough reading of the second data structure, the data on traffic violations basically comes from the state of Maryland in the United States. In order to ensure that the data analysis results are not disturbed by the gap in the amount of data. I re-collected traffic accident data in the state of Maryland – that is the third dataset.

2.1.3 Data cleaning and Checking



The picture above shows the general data cleaning process. This whole process is divided into three branches. The data sources for the first and second rows in the figure are traffic violations and traffic accidents in Maryland, respectively. The input in the third row is the traffic accident data of each state in the United States. Because the third dataset size is huge and the amount of data for each state is very different. So, it is not suitable for answering question 2, so the third dataset will be used to analyse the relationship between weather conditions, time, and location for question 1.

For first and second input of dataset:
For these two data, we first have to remove some completely unnecessary fields. For example: agency, model, year, make. These fields are production information about the vehicle and are outside the scope of this project. In addition, for the second dataset, although it contains traffic accident information in other states, the Maryland state accounts for 87% of the total. For Union with the Maryland traffic violation dataset later, only the Maryland's data will be selected for the second dataset.

When the first and second data are cleaned up, add "TypeOfEvent" ---- violation and crash to the two data sets respectively. After that, the name of the time Field of the two datasets is unified as "DateANDTime". After the two datasets are Union, we can better distinguish the source of each row of data and can use Date&Time data more conveniently.

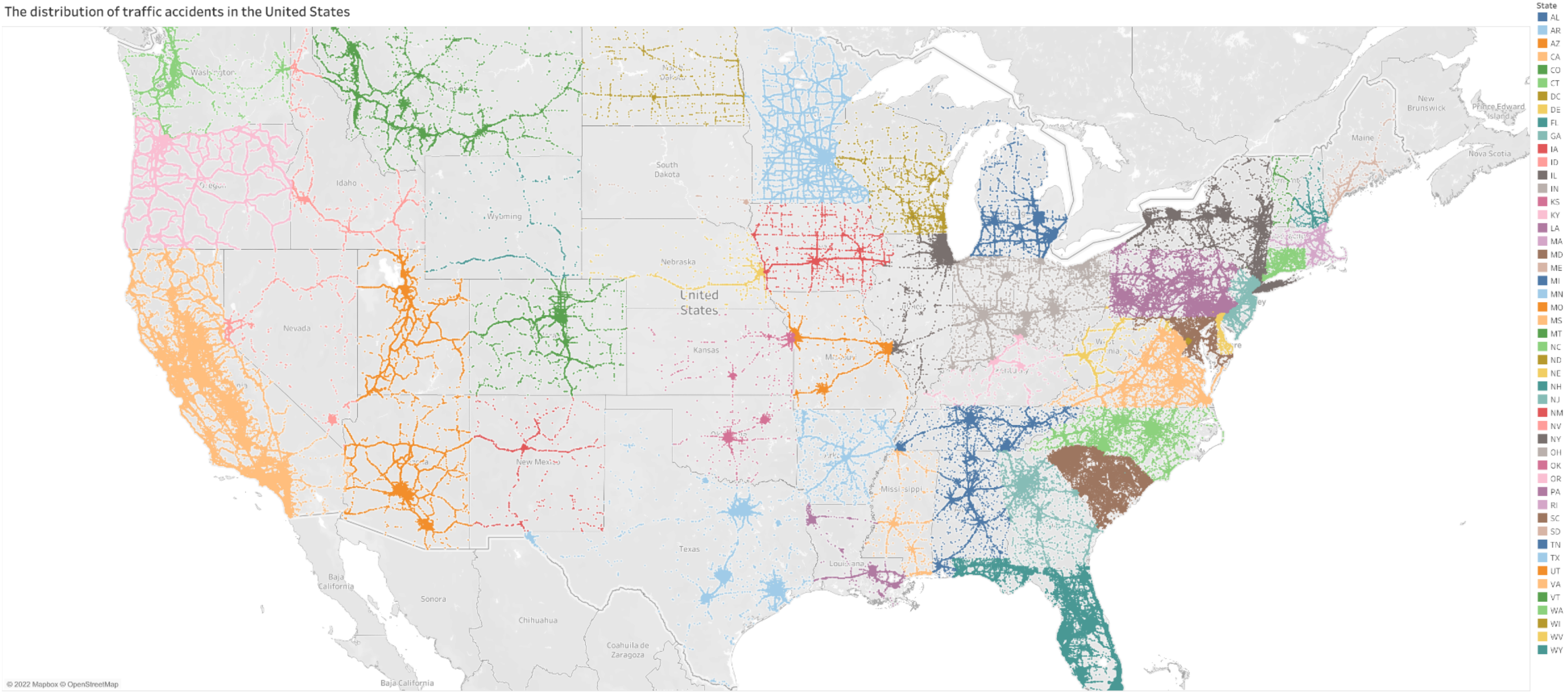
For all of input of dataset:
We need to delete some null data. The most important data types in this experiment are latitude and longitude, the time of accident and the values of various environmental conditions. If the occurrence time of the event is null, we must exclude this set of data. Fortunately, the proportion of null values in several important data selected this time basically does not exceed 3%. This dataset is very difficult to supplement, for example, most of the precipitation values are null. But we can't change all the null values to 0. Because the null value data contains a lot of data whose weather condition is Rain. In order not to affect other types of data, the main step of this data cleaning is exception.

3. Data exploration

3.1 Question 1

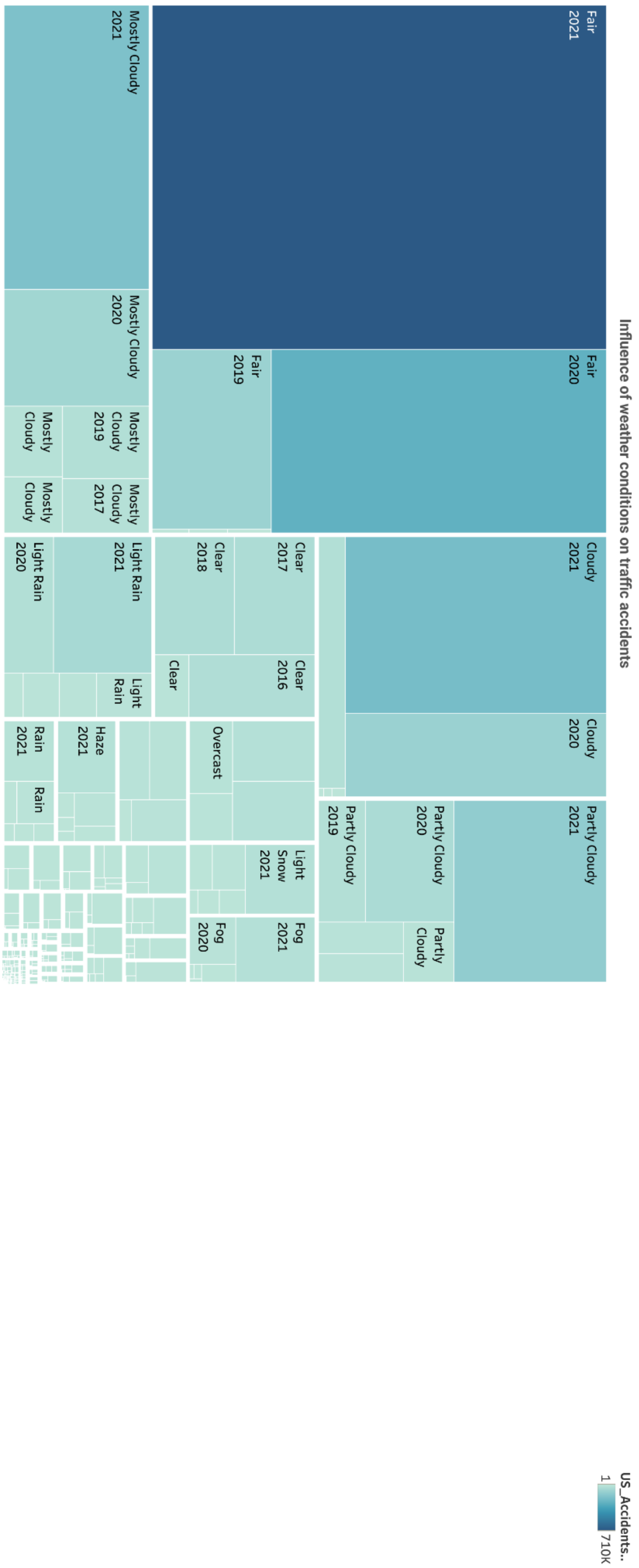
The image below is visualisations combined by the amount and geographic locations of traffic crashes across the US, and I've colored the states where the crashes occurred. We can see that without considering the time factor, most of the accidents are concentrated on the east and west coasts of the United States - California, New York and other more developed states. We can clearly see that the frequency of accidents basically decreases gradually from the inside to the outside of the city centre.

The distribution of traffic accidents in the United States



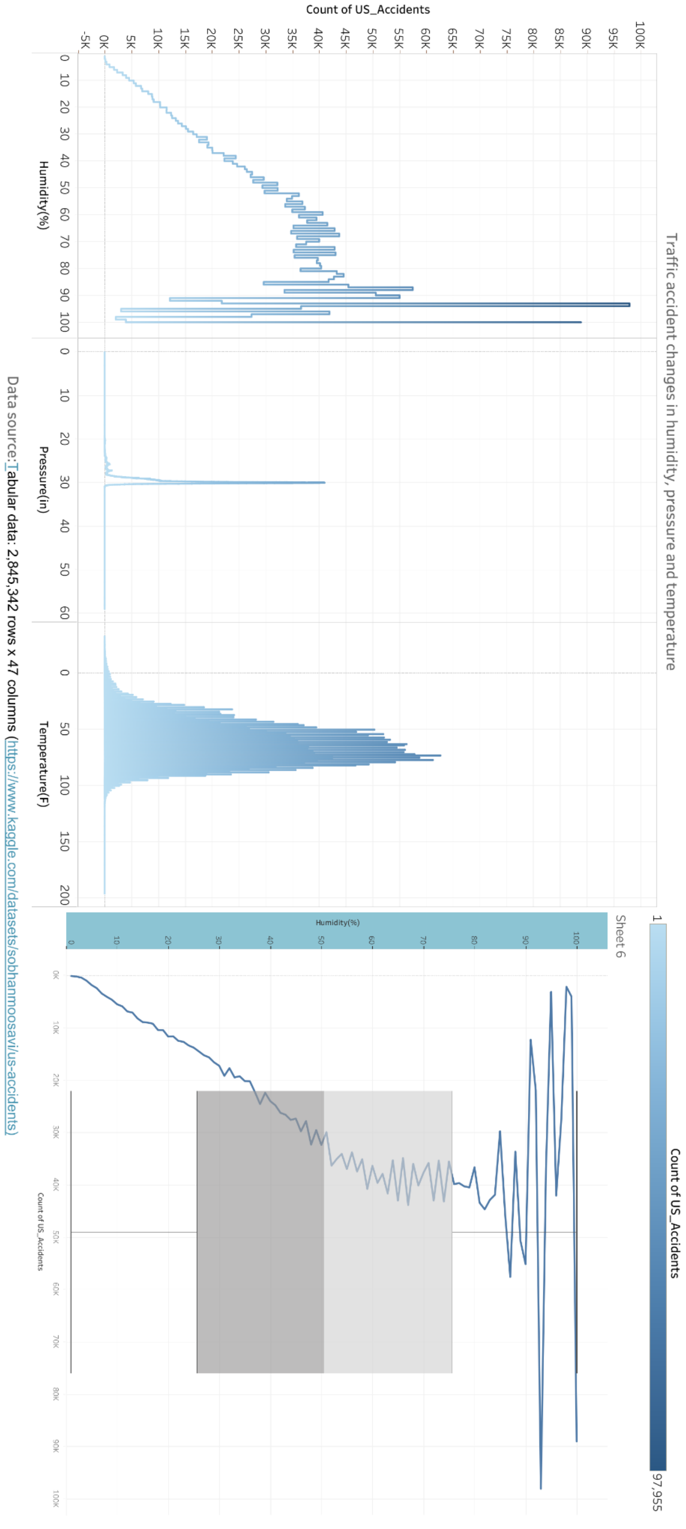
Data sourse: 2,845,342 rows x 47 columns (https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)

In the second image, we can see that, in the yearly perspective, Fair weather conditions have the most traffic accidents. But here's the surprise: The second-ranked weather situation is not the windy, rainy or foggy weather that everyone generally thinks. It's Mostly cloudy. It's rain-related weather starting at number three.



Data sourse(**The following two data need to be processed**):
Tabular data: 2,845,342 rows x 47 columns (https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)
Tabular data: 80,000 rows x 44 columns (https://catalog.data.gov/dataset/crash-reporting-incidents-data)..

According to the graph of the number of traffic accidents, humidity, temperature and pressure, we can know that the distribution of pressure is too concentrated and will not have much impact on the number of traffic accidents. It is humidity and temperature that have a significant impact on traffic accidents. After verification by box-and-whisker plots, outliner is not included in the dataset of humidity. Therefore, as the humidity increases, the number of traffic accidents will gradually increase. The humidity will fluctuate wildly between 80% and 100% --- heavy rain. The temperature and the number of traffic accidents are normally distributed.

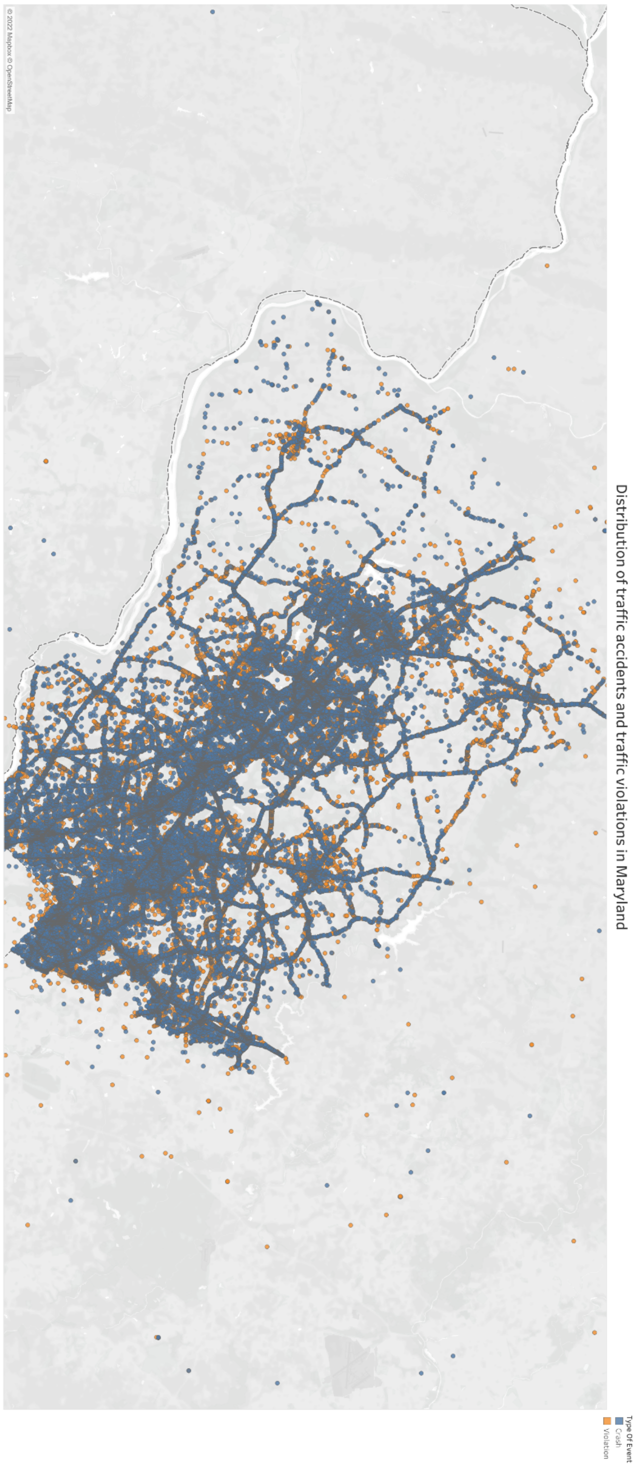


Traffic accident changes in humidity, pressure and temperature

Data source:Tabular data: 2,845,342 rows x 47 columns (https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)

3.2 Question 2

It can be accurately known from the following map: The probability of traffic accidents in places with a large number of violations is also higher. But it is different from traffic violations: the locations of traffic accidents are concentrated on the main roads of the city, and in the areas with the highest concentration of traffic accidents, the number of traffic violations has actually declined. Personal speculation: may be related to the decrease in average speed in the city.



Distribution of traffic accidents and traffic violations in Maryland

Data sourse(The following two data need to be processed):

Tabular data: 2,845,342 rows x 47 columns (https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)..

4. Data conclusion

4.1For question 1:

We can see that without considering the time factor, most of the accidents are concentrated on the east and west coasts of the United States - California, New York and other more developed states.Traffic accidents are most common in cloudy weather. Humidity, temperature can significantly affect the number of traffic accidents.

4.2 For question 2:

Traffic accidents and traffic violations are basically positively correlated within the city area. But traffic violations fell in places with the highest number of crashes.

5. Reflection

When collecting data, we should try to understand the structure of the data. This saves time when we need to supplement another dataset. In addition, when deleting useless data, it should be judged whether it will have a greater impact on the dataset itself. After the two datasets are unioned, any null value may appear. At this time, it is necessary to carefully judge whether the deletion of the null value will cause the data to be true.

6. References:

1. Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A countrywide traffic accident dataset. arXiv preprint arXiv:1906.05409.

2. Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019, November). Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 33-42).

3. Erdogan, S., Yilmaz, I., Baybura, T., & Gullu, M. (2008). Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. Accident Analysis & Prevention, 40(1), 174-181.