US Accidents Data (4.2 Million Records) – Data Mining

Dataset Link: https://smoosavi.org/datasets/us_accidents

Data Description and Types

This is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about **4.2 million** accident records in this dataset.

Table 1: Data Description and Types

Column	Description	Data Type
ID	This is a unique identifier of the accident record.	object
Source	Indicates source of the accident report (i.e. the API which reported the accident.).	object
TMC	A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.	float64
Severity	Shows 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).	int64
Start_Time	Shows start time of the accident in local time zone.	object
End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	object
Start_Lat	Shows latitude in GPS coordinate of the start point.	float64
Start_Lng	Shows longitude in GPS coordinate of the start point.	float64
End_Lat	Shows latitude in GPS coordinate of the end point.	float64
End_Lng	Shows longitude in GPS coordinate of the end point.	float64
Distance(mi)	The length of the road extent affected by the accident.	float64
Description	Shows natural language description of the accident.	object
Number	Shows the street number in address field.	float64
Street	Shows the street name in address field.	object
Side	Shows the relative side of the street (Right/Left) in address field.	object
City	Shows the city in address field.	object
County	Shows the county in address field.	object

State	Shows the state in address field.	object
Zipcode	Shows the zipcode in address field.	object
Country	Shows the country in address field.	object
Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	object
Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	object
Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	object
Temperature(F)	Shows the temperature (in Fahrenheit).	float64
Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	float64
Humidity(%)	Shows the humidity (in percentage).	float64
Pressure(in)	Shows the air pressure (in inches).	float64
Visibility(mi)	Shows visibility (in miles).	float64
Wind_Direction	Shows wind direction.	object
Wind_Speed(mph)	Shows wind speed (in miles per hour).	float64
Precipitation(in)	Shows precipitation amount in inches, if there is any.	float64
Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	object
Amenity	A POI annotation which indicates presence of amenity in a nearby location.	bool
Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	bool
Crossing	A POI annotation which indicates presence of crossing in a nearby location.	bool
Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	bool
Junction	A POI annotation which indicates presence of junction in a nearby location.	bool
No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	bool
Railway	A POI annotation which indicates presence of railway in a nearby location.	bool
Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	bool
Station	A POI annotation which indicates presence of station in a nearby location.	bool
Stop	A POI annotation which indicates presence of stop in a nearby location.	bool
Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	bool
Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.	bool
Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	bool
Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	object
Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	object
Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	object

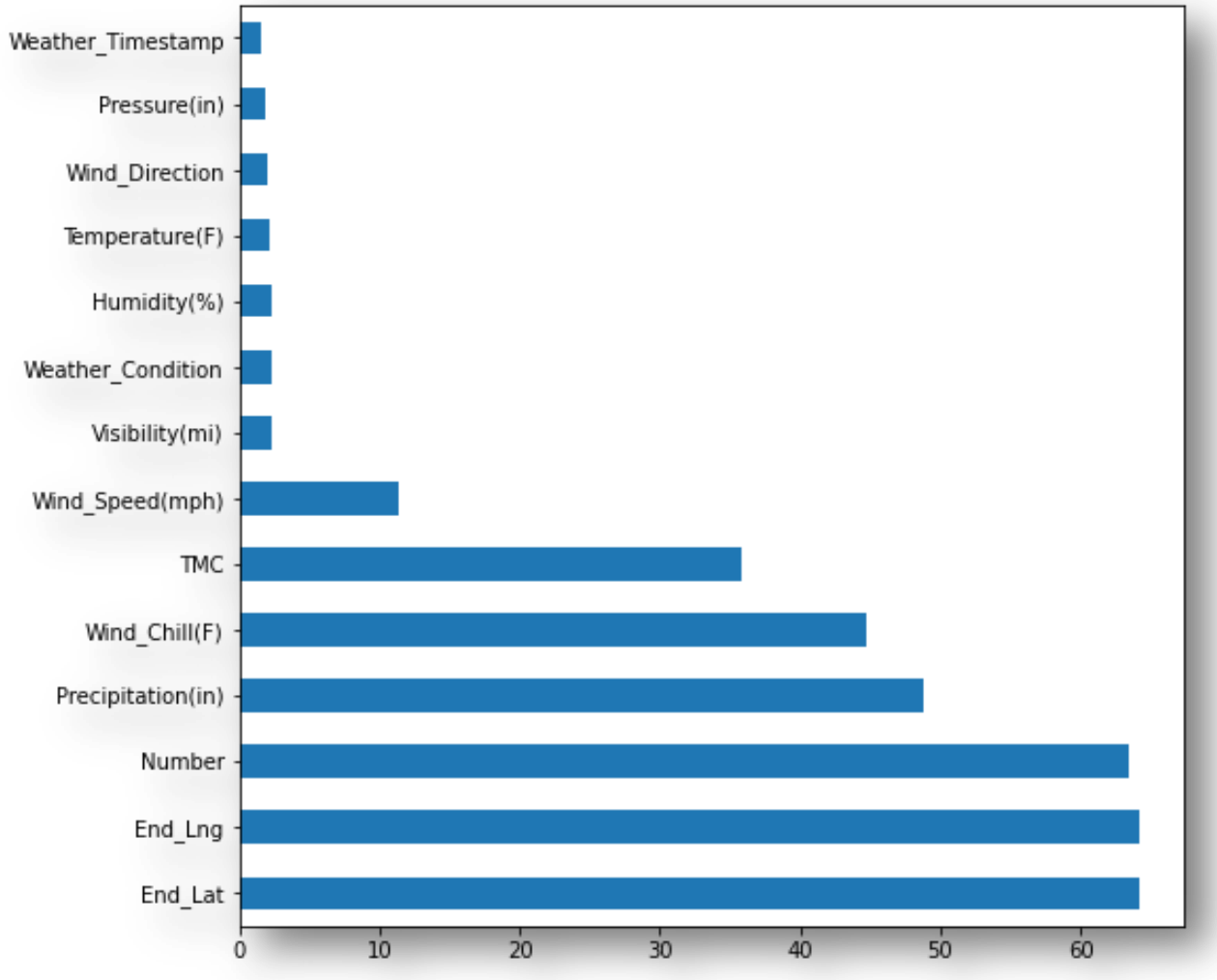
Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	object

Data Preparation and Cleaning

1. Load the file using Pandas
2. Look at some information about the data & the columns
3. Fix any missing or incorrect values

Missing Values: 13 columns have missing values, namely End_Lng,Number, Precipitation(in), Wind_Chill(F), TMC, Wind_Speed(mph), Visibility(mi), Weather_Condition, Humidity(%), Temperature(F), Wind_Direction, Pressure(in), and Weather_Timestamp.

Figure 1: Visualizing missing values



From Table 1 we see that the columns ID, Number, End_Lng, End_Lat have a lot of missing values and also they are not very relevant. So we drop these columns.

Exploratory Data Analysis & Visualization

We will concentrate our study on 5 main columns, namely City, State, Start_Time, Start_Lat, Start_Lng.

Cities

First we visualize the distribution of accidents using histogram and distribution plots which is in Figure 3 and Figure 4.

Figure 2: Code Snippet 1

```
plt.figure(figsize=(8,8))
sns.histplot(cities,log_scale=True)
plt.show()

plt.figure(figsize=(8,8))
sns.distplot(cities)
plt.show()
```

Figure 3: Histogram of accident frequency by Cities

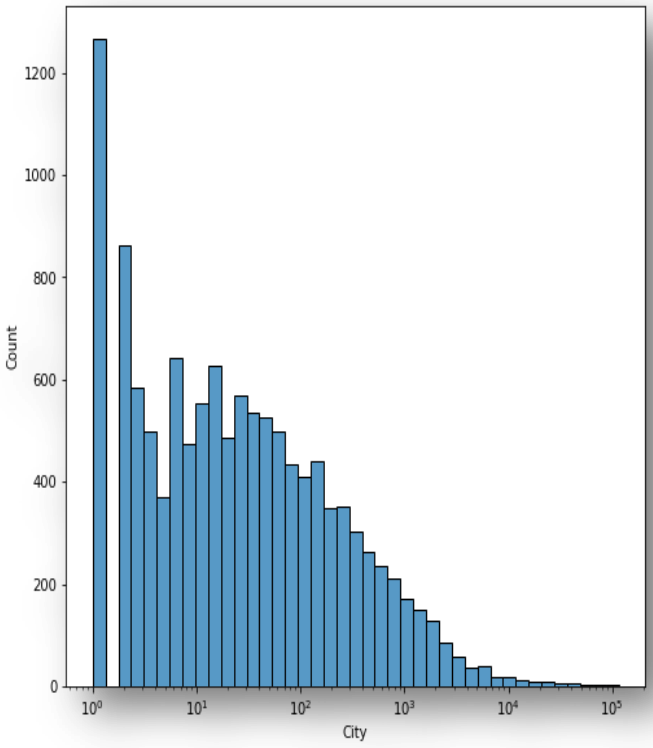
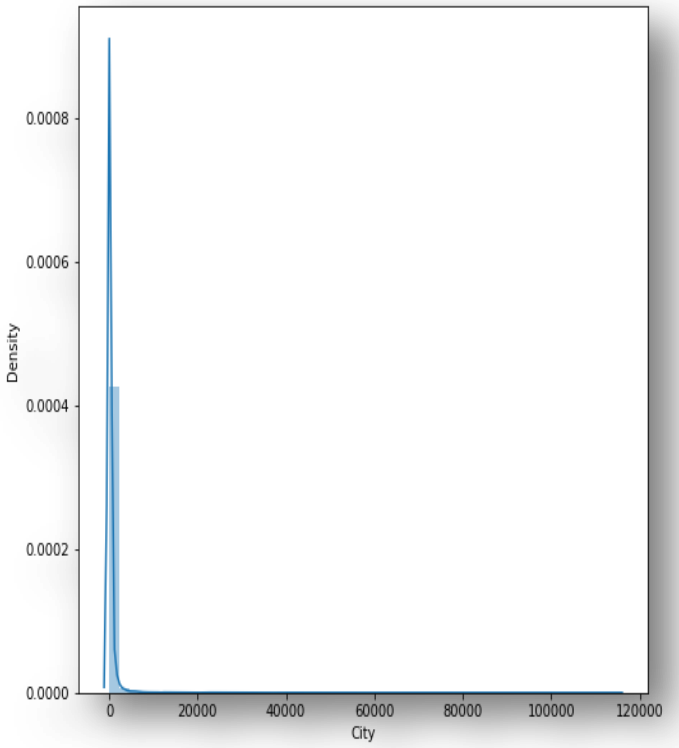
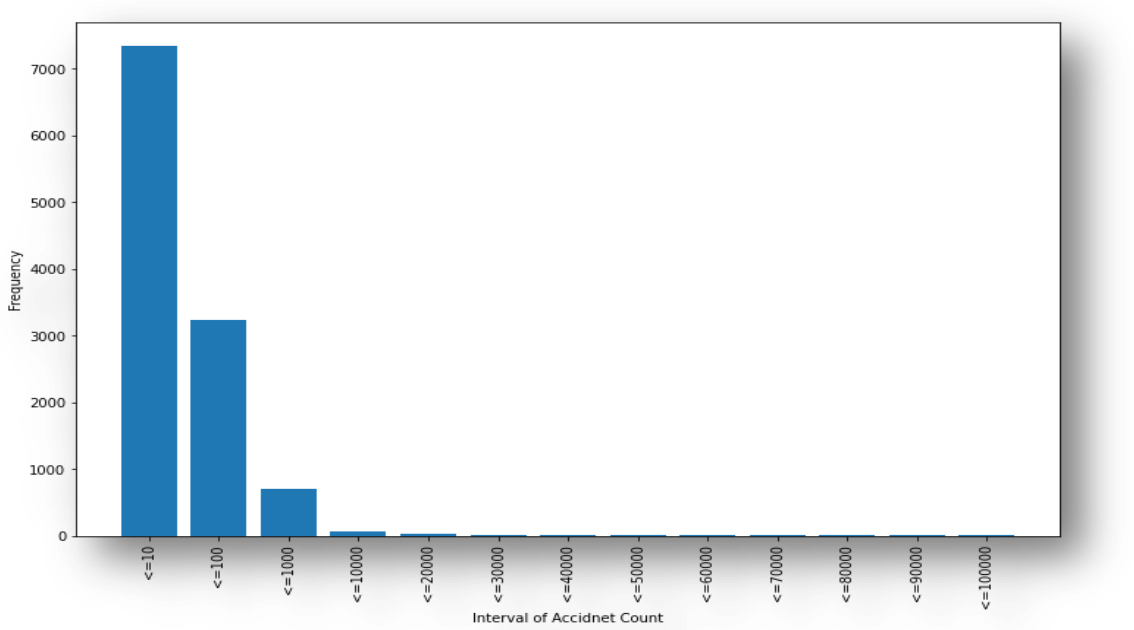


Figure 4: Distribution Plot of accident frequency by Cities



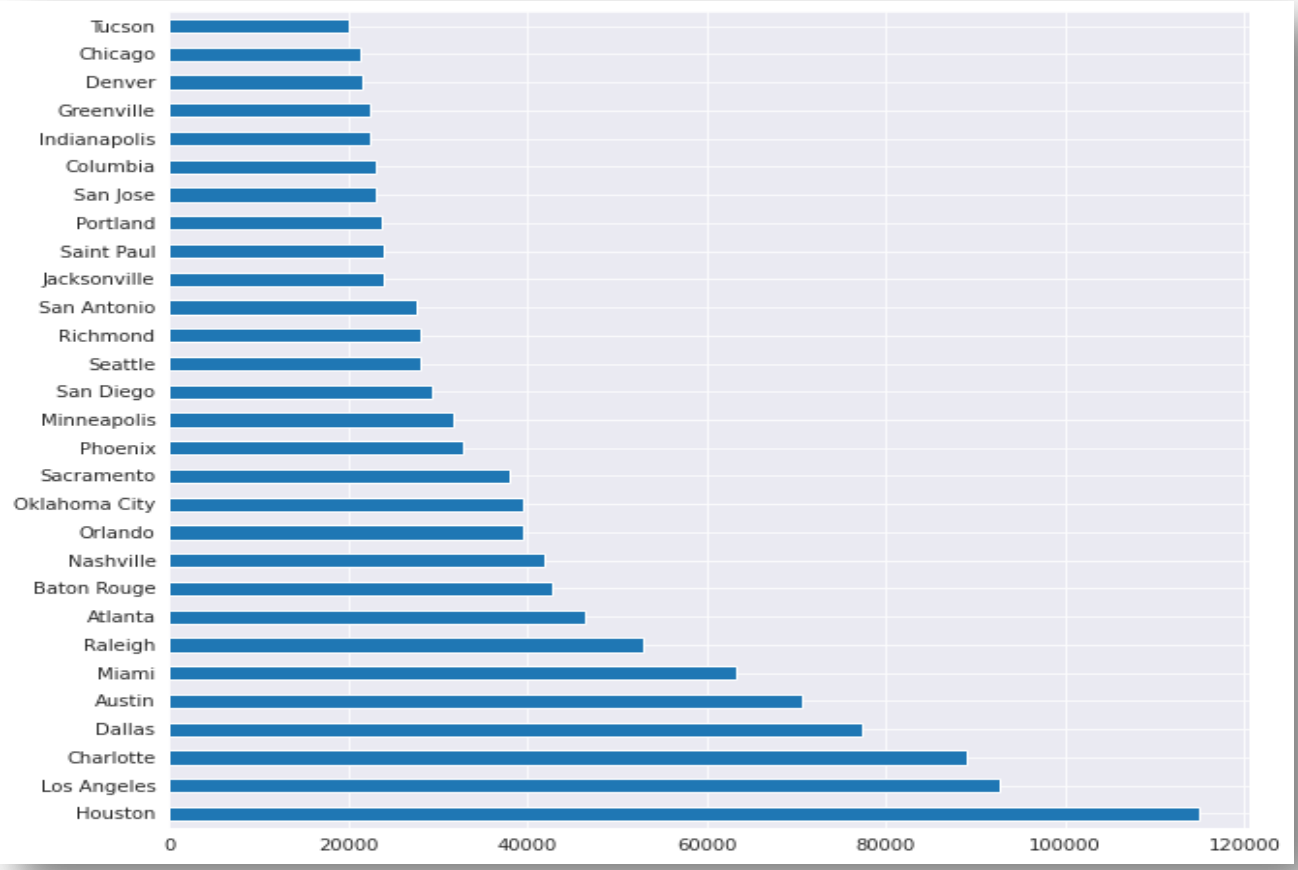
From Figure 3 and Figure 4 it is well evident that the distribution of accidents by cities is right skewed and has a long tail. It shows that most of the cities have very low accident rates and only a handful of them have high accident rates. To further investigate this study we break the frequency into intervals and visualize the data which is Figure 5. This visualization shows that 7000 cities have less than 10 accidents, around 3000+ cities have lee than 100 accidents and so on. So next we set the bar a

Figure 5: Interval breakdown of accidents



convenient number of 20,000 accidents and less and find out the % of cities having more than 20,000 accidents and the ones having less than 20,000 accidents. This breakdown further simplifies our studies and we see that about 99.76% of the cities have lee than 20,000 accidents between the date range of this data and only about 0.24% cities have more than 20,000 accidents. And the cities having the most number of accidents are visualized in Figure 6.

Figure 6: Accidents having more than 20,000 accidents



States

First we see the distribution of accidents by the states using Histogram and a distribution plot which is in Figure 7 and Figure 8.

Figure 7: Histogram of accident frequency by States

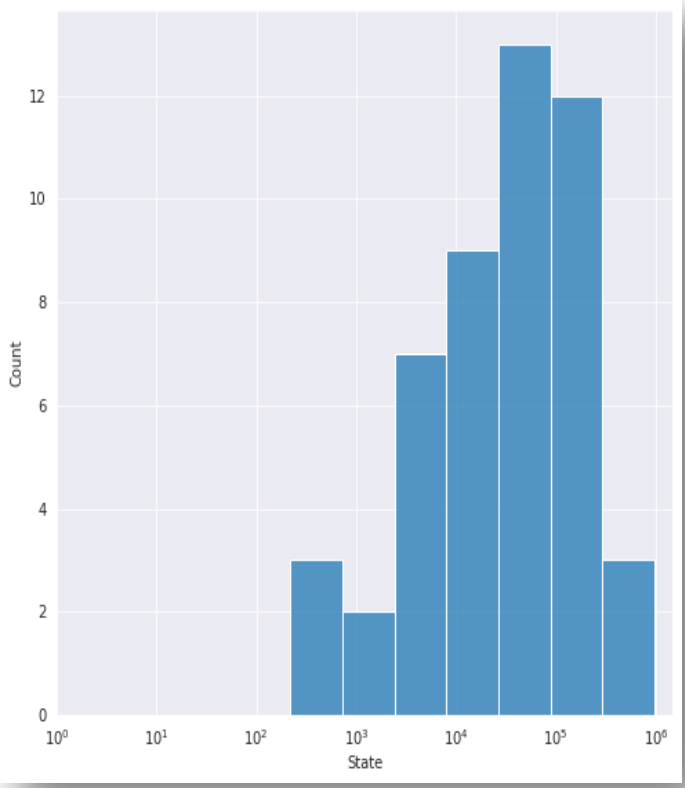
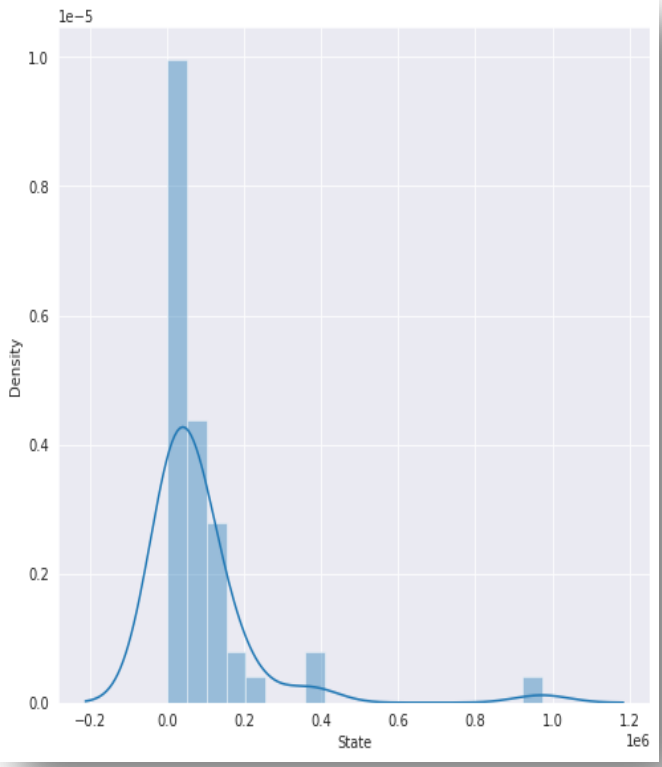
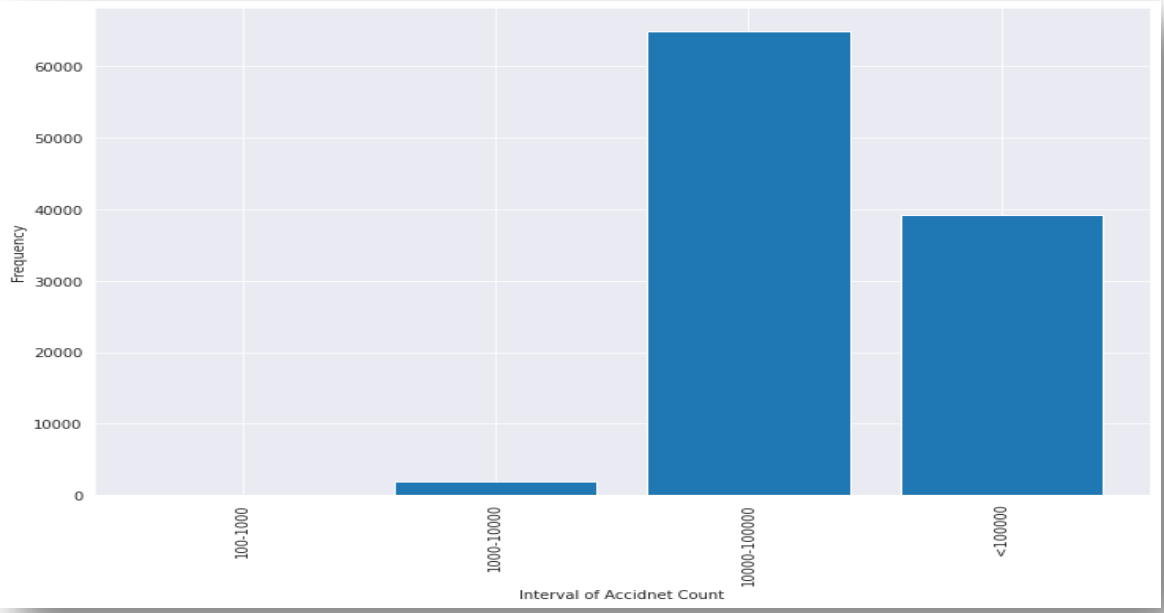


Figure 8: Distribution Plot of accident frequency by States



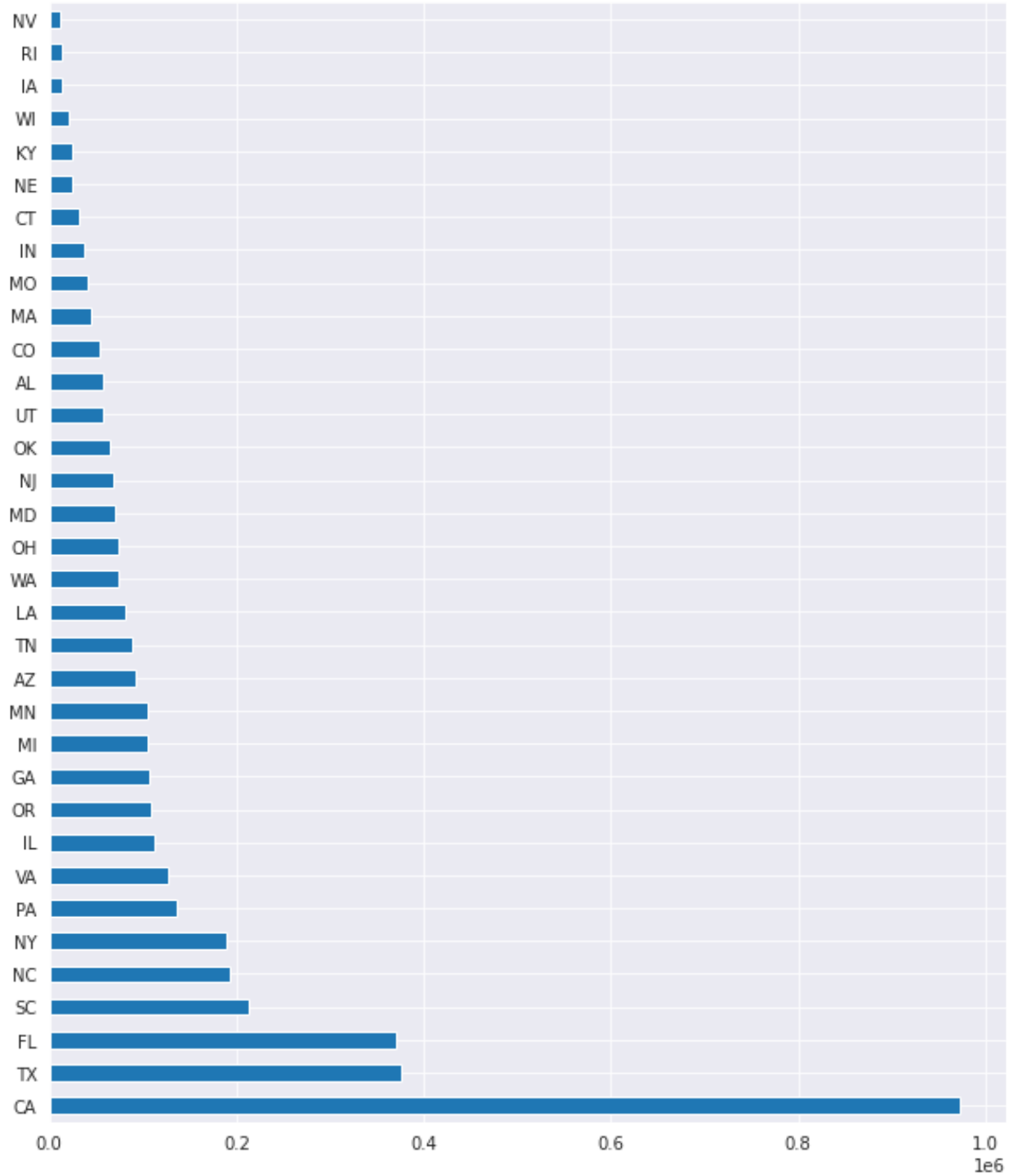
From Figure 7 and Figure 8 it is evident that the distribution has some sort of bell curve (normal distribution) or to be precise, it has a Weibull distribution. So to further look into the actual segregation of the accidents we create a Count Interval graph and visualize the accidents in Figure 9.

Figure 9: Count Interval visualization of Accidents- State wise



This visualization shows that very few states, about 26.5% have less than 10,000 accidents. 73.5% sates have more than 10,000 accidents. Furthermore, the states having more than 10,000 accidents are visualized in Figure 10.

Figure 10: States having more than 10,000 accidents



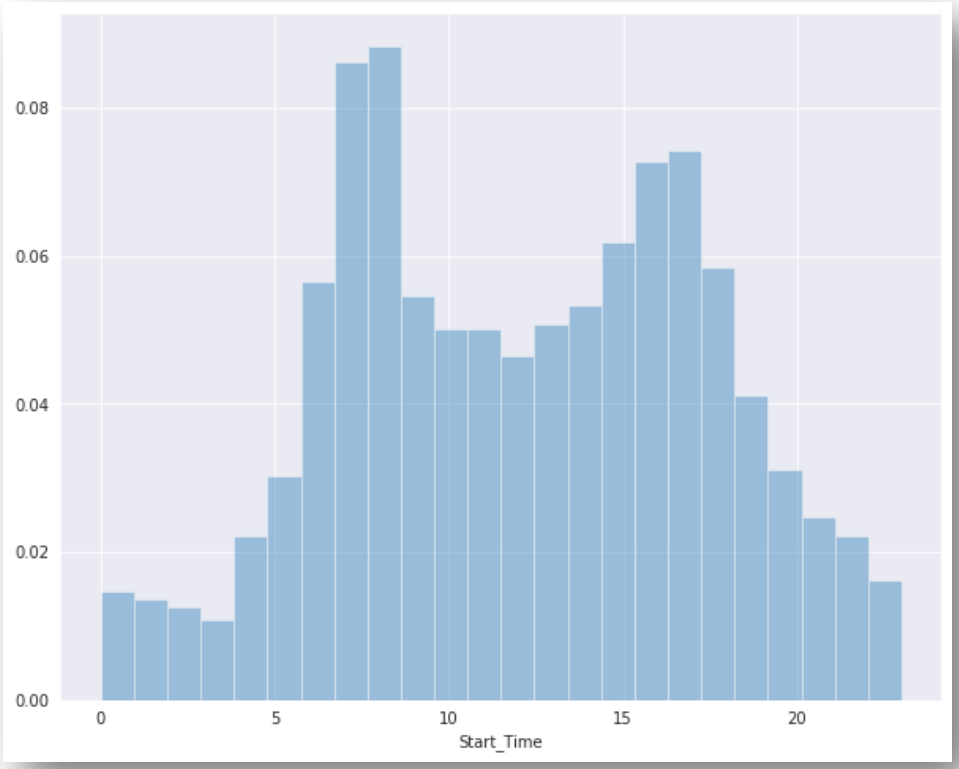
Start Time

The Start Time column contains the date and the time of the accident. This provides a scope to study the data in three ways: **Hourly, Daily** and **Monthly.** From Table 1 we see that this column is not a datetime object. So first we convert this into a datetime object go forward with our analysis.

Hourly

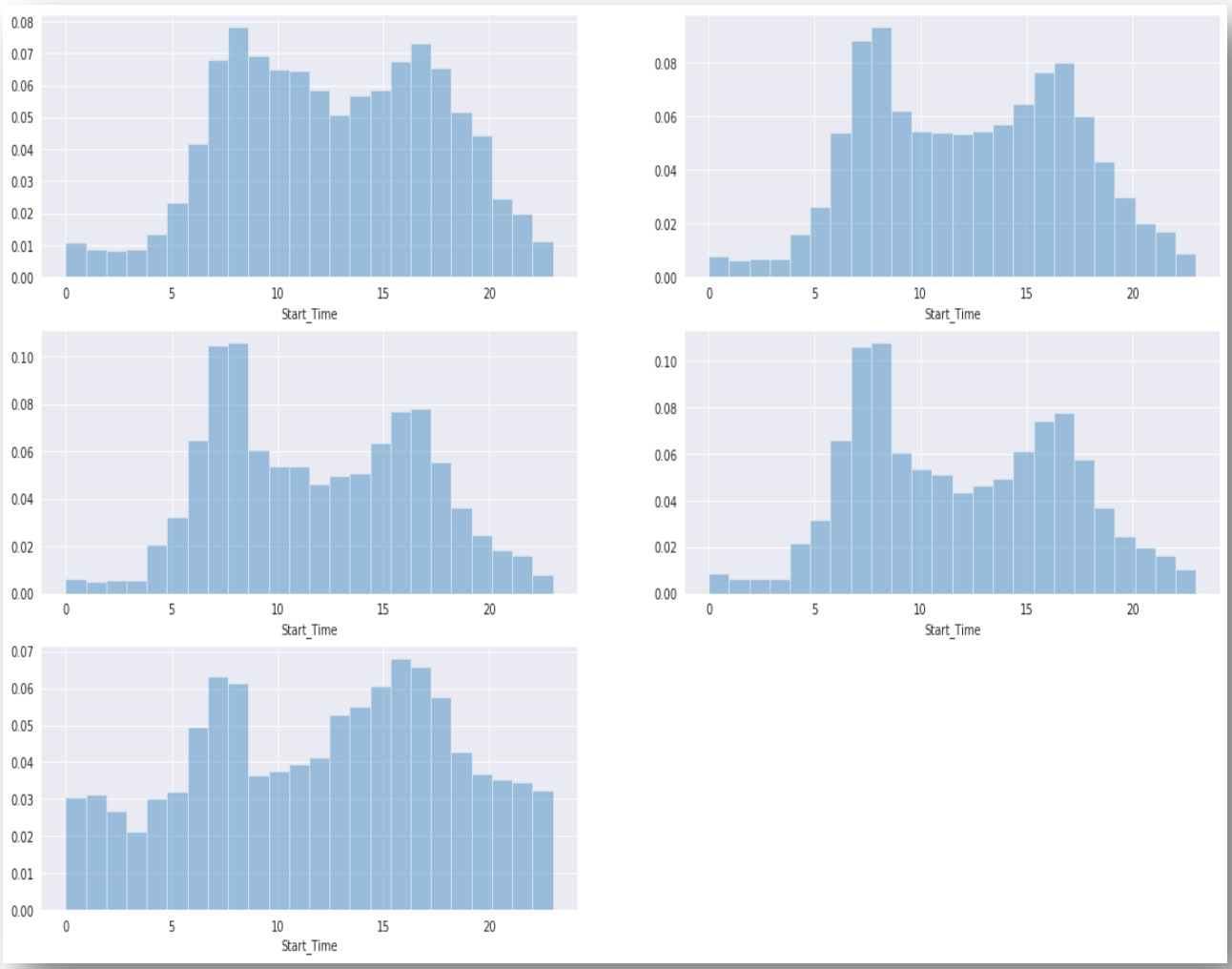
First we plot the distribution of hourly accidents which is in Figure 11 and see that a high percentage of accidents occur between 6 am and 10 am. This can be because people are mostly in a hurry during this time due to office timing. And again there is a spike between 3 pm to 6 pm. This is the average trend for

Figure 11: Hourly Distribution of accidents.



the years 2016-2020. Next to check whether this trend is the same for all the years, we breakdown the data year-wise and visualize the distribution which is in Figure 12. This breakdown shows that the average trend is similar for the years 2018 and 2019 and other years have a different distribution. For instance, the year 2020 has a spike between 6-9 am and again between 2-7 pm; year 2016 has a general spike throughout the day and 2017 has spikes between 7-9 am and 4-6 pm.

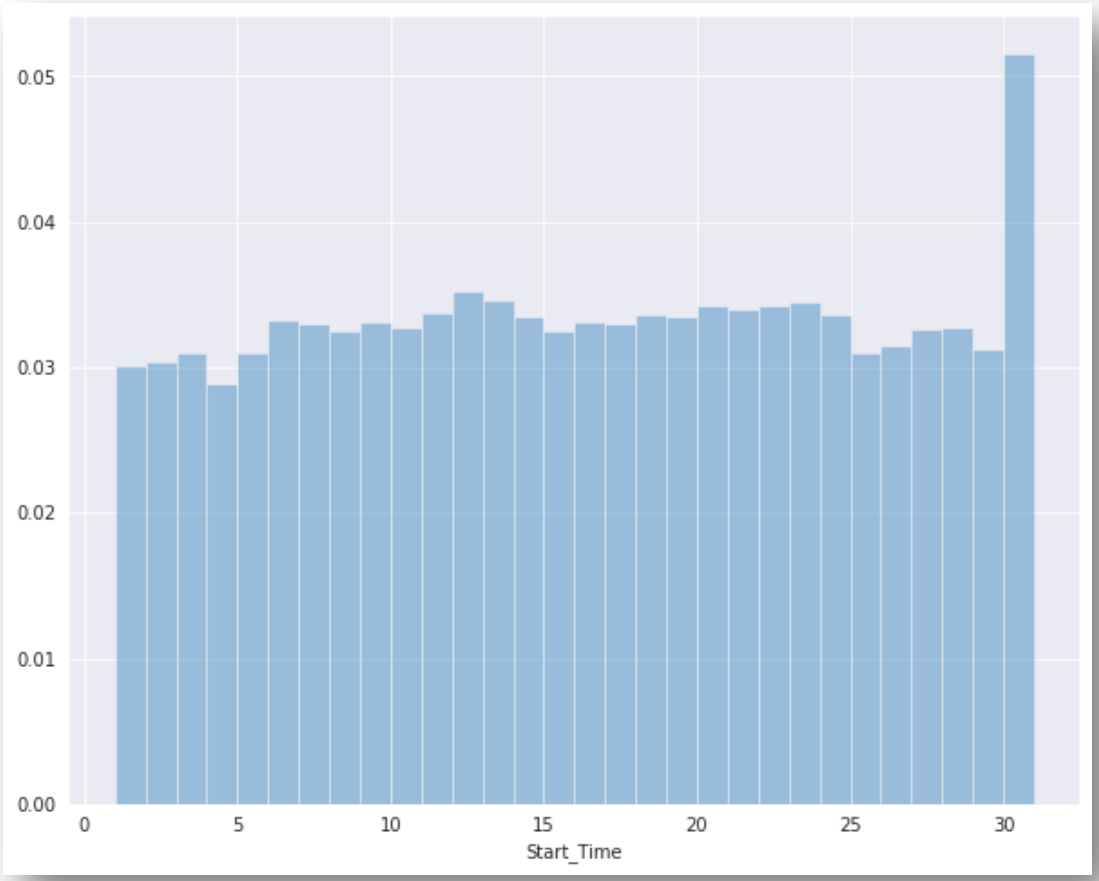
Graph 12: Hourly Analysis of accidents – Year wise



Daily

The distribution of hourly accidents is plotted in Figure 13. This graph shows that the distribution is more or less same except for the last day of the month where there is a huge spike (considering a month is composed of 30 days). This trend is a bit strange to occur so we breakdown our analysis year-wise.

Figure 13: Daily distribution of accidents



The year-wise trend plotted in Figure 14 shows that this trend is similar for every year, i.e. from 2016-2020. Though there is no particular reason for this, but this is an odd trend.

Figure 14: Daily distribution of accidents – Year wise

Monthly

The monthly distribution of accidents, visualized in Figure 15 shows that the number of accidents is higher during the winter months whereas summer months have low accident rates. As there is no particular reason for such a type of trend, so we break down the data year-wise and visualize them

Figure 15: Monthly distribution of accidents

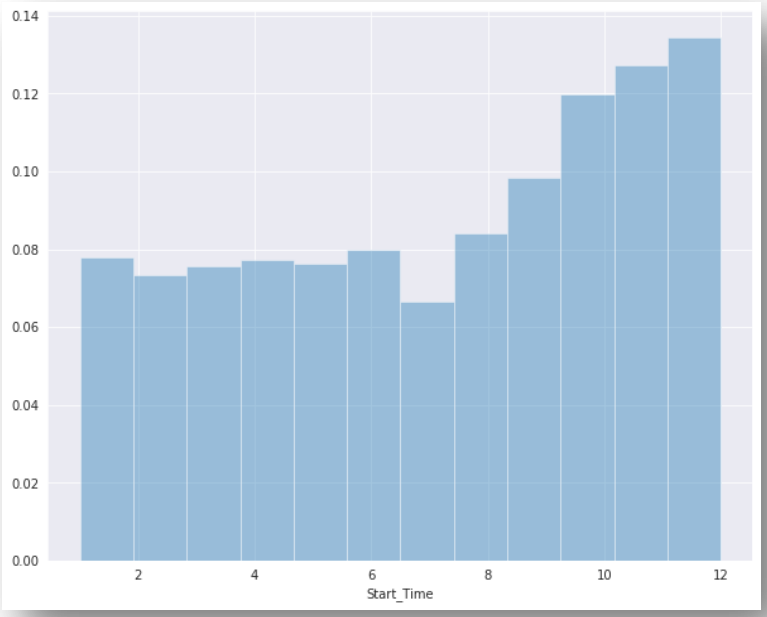
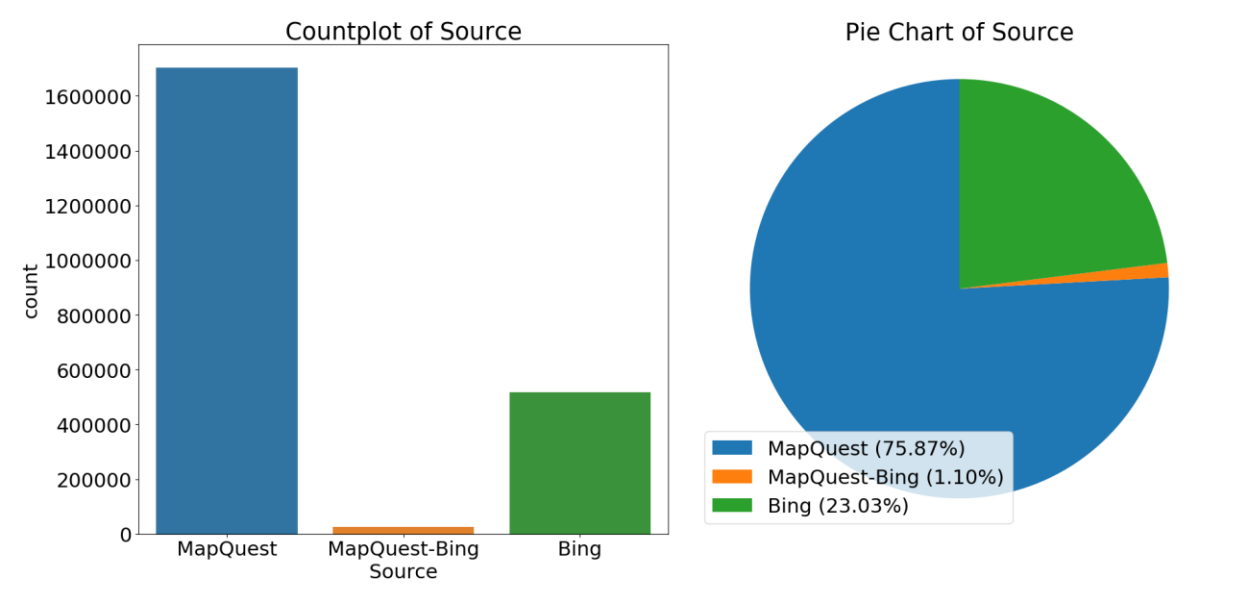


Figure 16: Monthly distribution of accidents- year wise

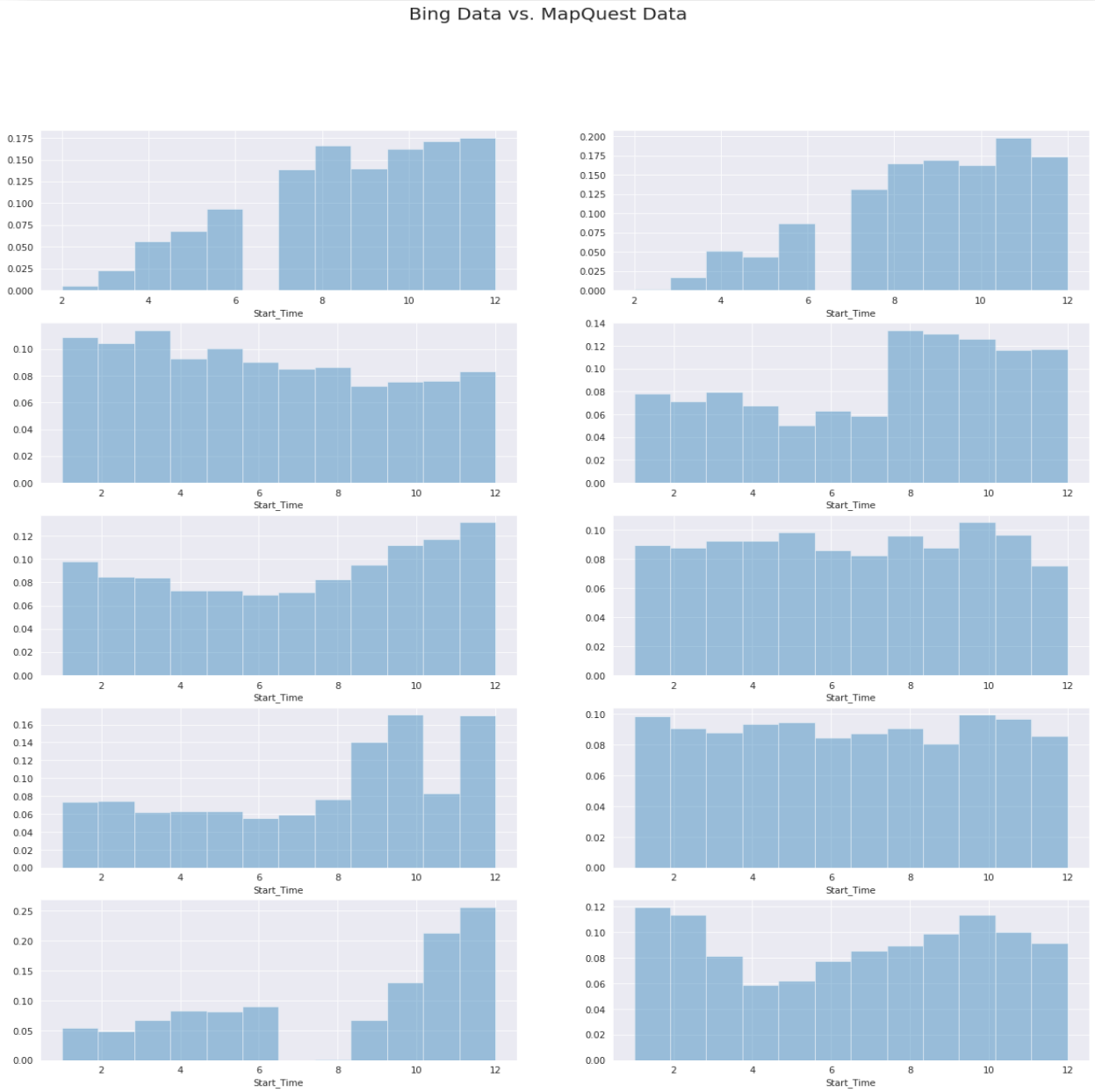
in Figure 16. This breakdown shows that some of the years have missing data. So to check the authenticity of the data by its source, we use the source column. There are three sources of data shown in Figure 17: Bing, MapQuest and Mixed. Bing provides 23.03% of the data, MapQuest provides 75.87%

Figure 17: Countplot and Pie Chart of the Source of Data



of the data and 1.1% of the data comes from other sources. So we plot the data of Bing vs. MapQuest, year-wise in Figure 18. This visualization shows that the year 2016 has a lot of missing data from both the sources.

Figure 18: Bing vs. MapQuest Data – Year wise

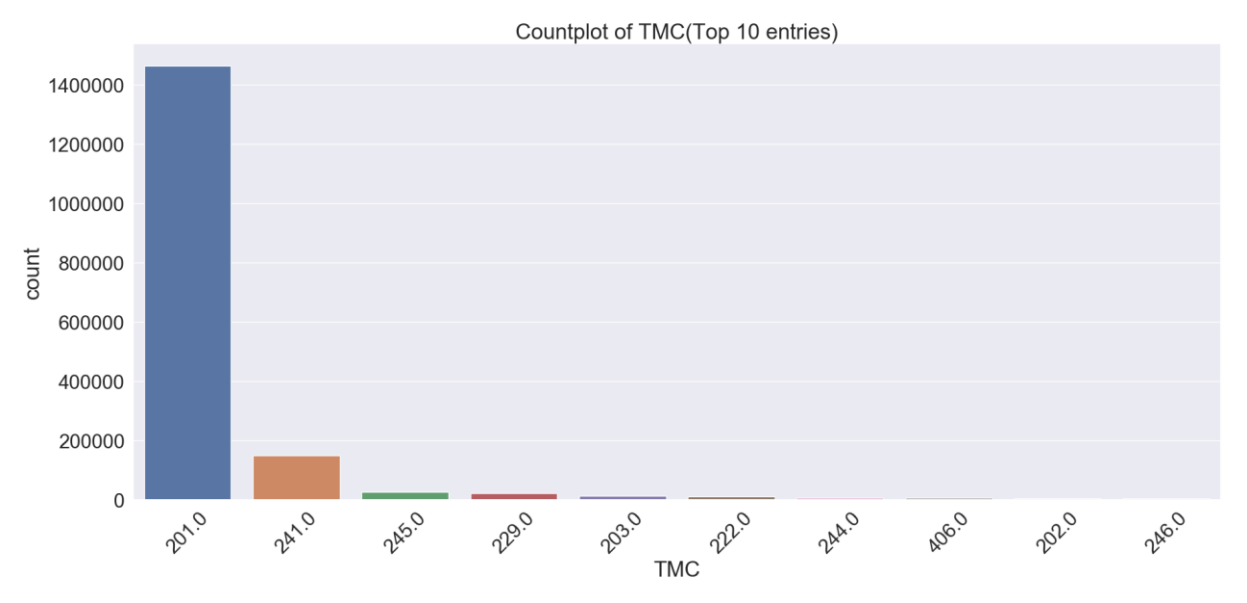


Bing has some missing data for 2020 as well. And the trends are different for different sources, also for each year.

TMC Traffic Message Channel

The TMC column (Traffic Message Channel) contains the codes used while the accidents were being reported. TMC is a technology that allows drivers to receive traffic and travel information. It is digitally encoded into RDS Type 8A groups using the Warning C or TPEG protocol and transmitted over traditional FM radio broadcasts. It can also be received via satellite radio or digital audio broadcasting. TMC enables the silent transmission of complex information appropriate for replication or display in the user's native language without interfering with audio broadcast services. Plotting this feature in Figure 19 shows that the channel 201has been mostly used for reporting the incident.

Figure 19: Countplot of Traffic Message Channel



Severity

This column represents severity of an accident where 0 and 1 represents low severity, 2 represents average severity and 3 and 4 represents highly severe accidents. Most of the accidents fall under 2 and 3 level of severity and a handful of them fall under category 4, the highly severe ones. The visualization is visualized in Figure 20.

Figure 20: Countplot and Pie Chart of Severity

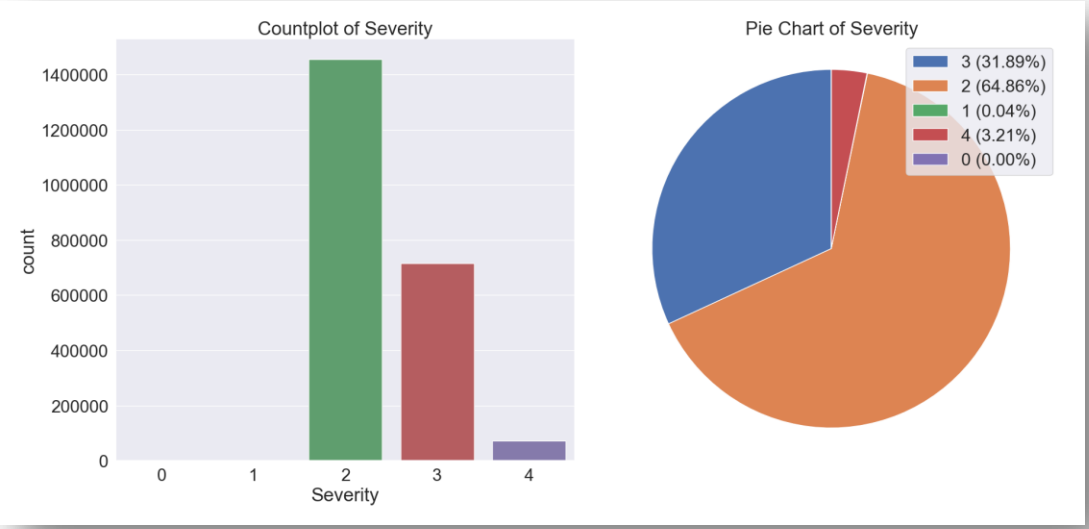


Figure 21: Severity Trend

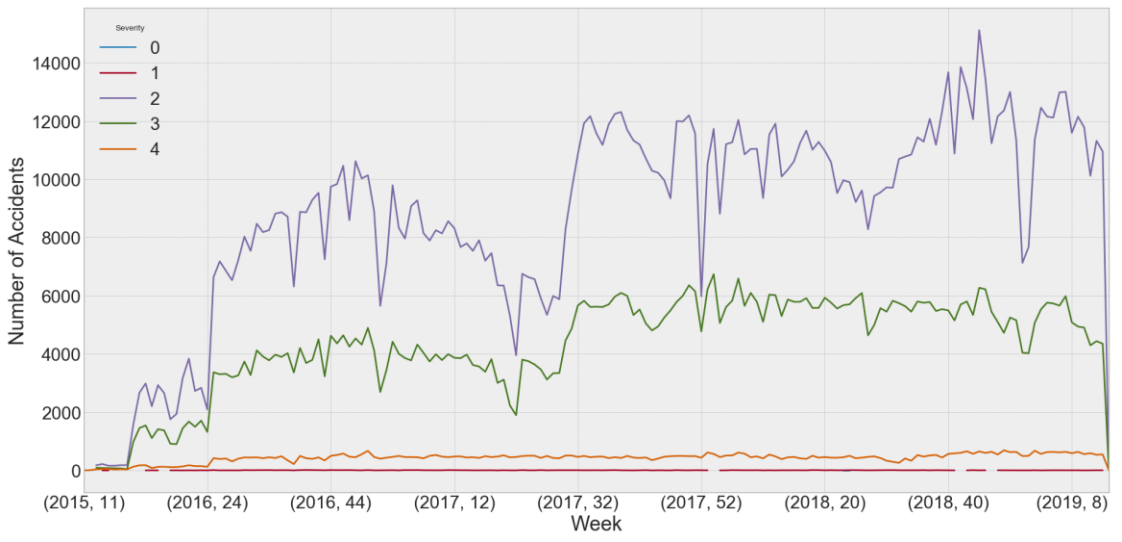
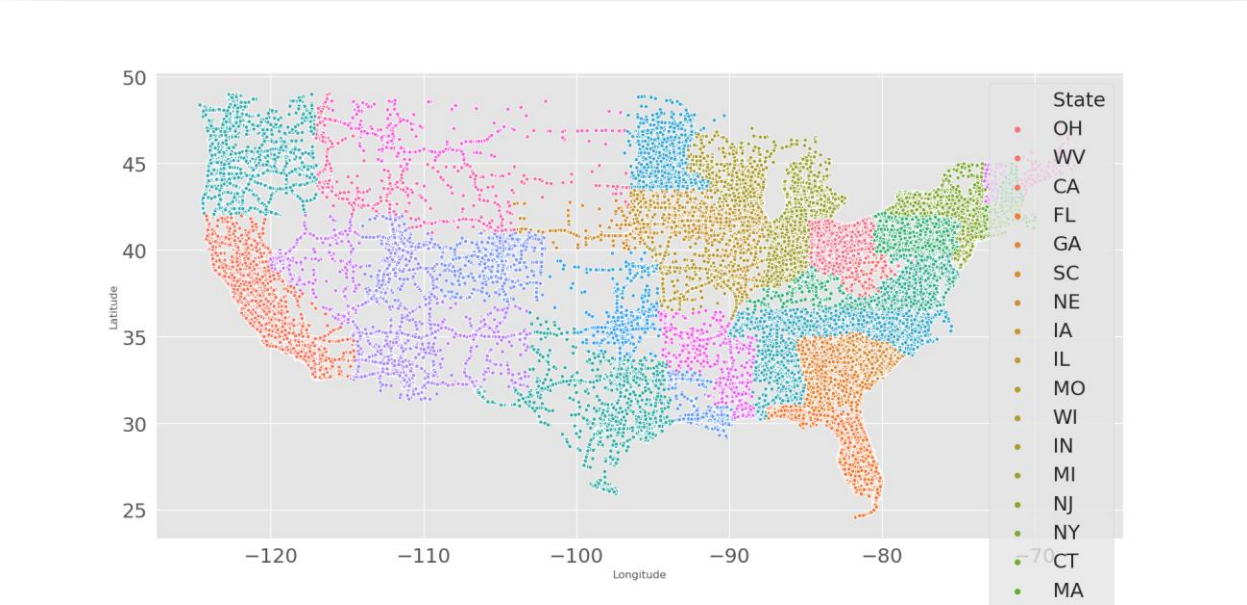


Figure 21 on the other hand shows the trend of severity of each kind. The number of injuries has risen over time with each severity level, as can be seen. This is concerning and necessitates immediate action. Despite a decline in the number of incidents in 2017, the number of accidents rose after that around week 12. Accidents with a severity of 2 are the most common and have raised the most, followed by accidents with a severity of 3 and then accidents with a severity of 4.

Start Latitude and Start Longitude

The Start Lat and Start Lng attributes are useful because they can be plotted on a map to determine the exact location of the accident. We begin by creating a scatterplot between the two which is in Figure 22.

Figure 22: Start_Lat and Start_Lng



The scatterplot is appealing, but it is concerning that it covers nearly every corner of the United States, implying that the events occurred in a vast number of places over the past few years. We also see that the east coast has a lot more accidents than the west coast.

Start Latitude, Start Longitude and Severity

To get a better idea, we plot the accident site on the USA map using the coordinates given in the dataset for each severity.

Figure 23: Start_Lat and Start_Lng for severity=0

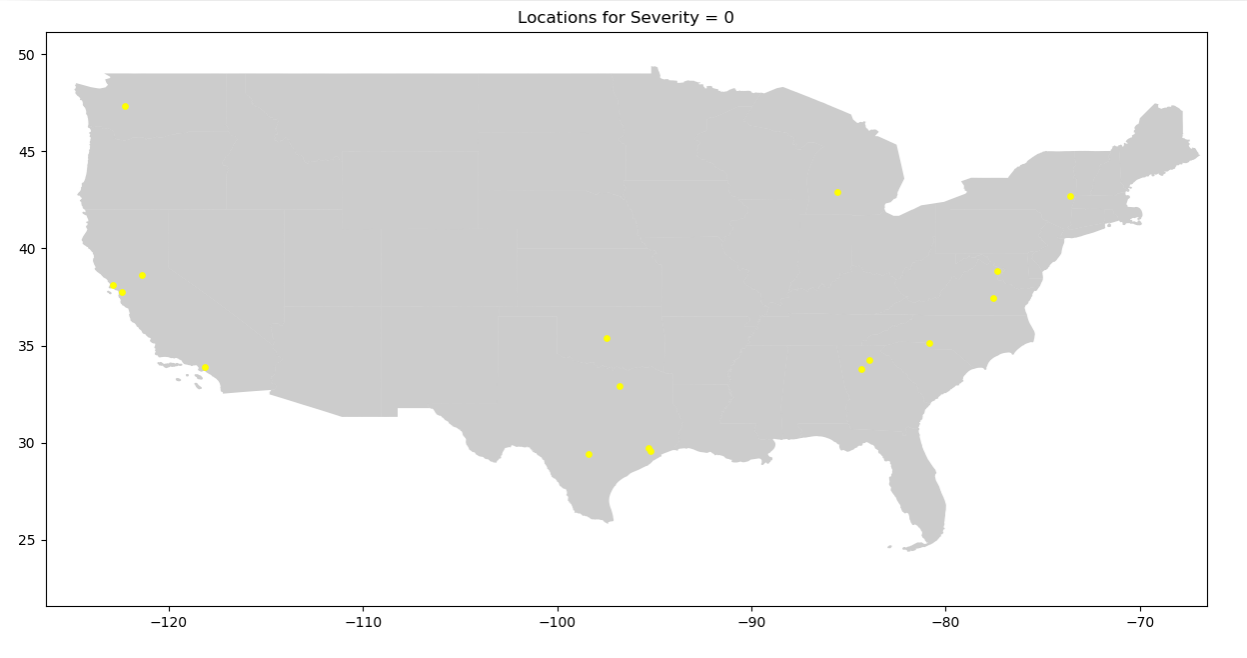


Figure 24: Start_Lat and Start_Lng for severity=1

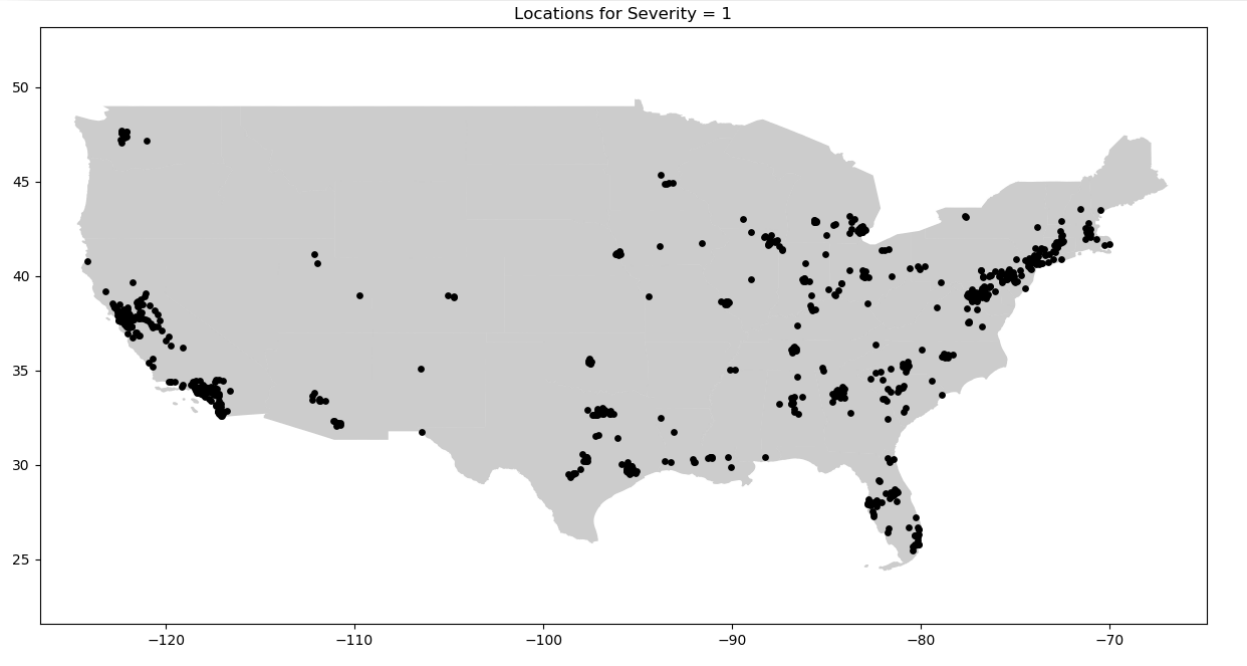


Figure 25: Start_Lat and Start_Lng for severity=2

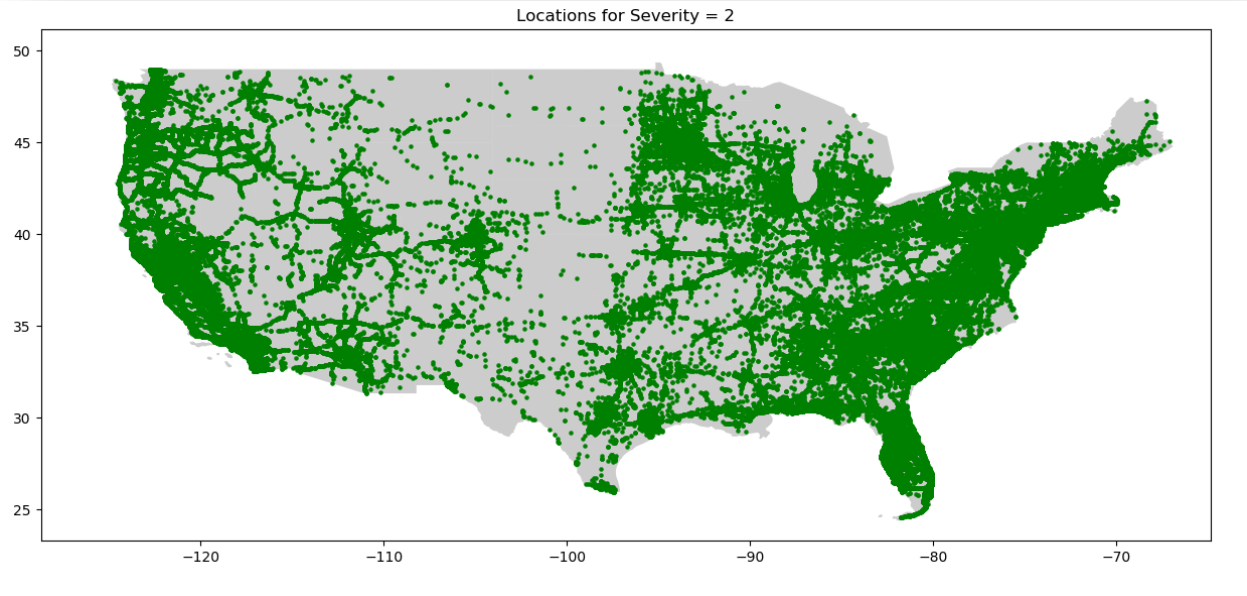


Figure 26: Start_Lat and Start_Lng for severity=3

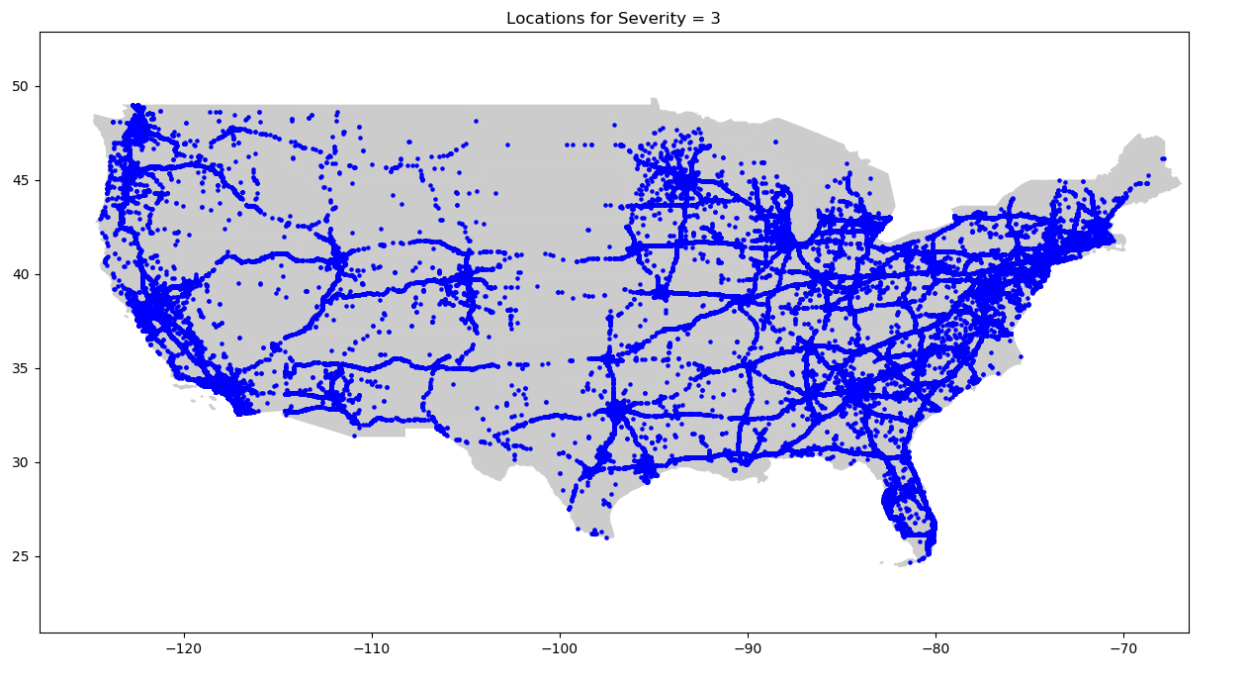
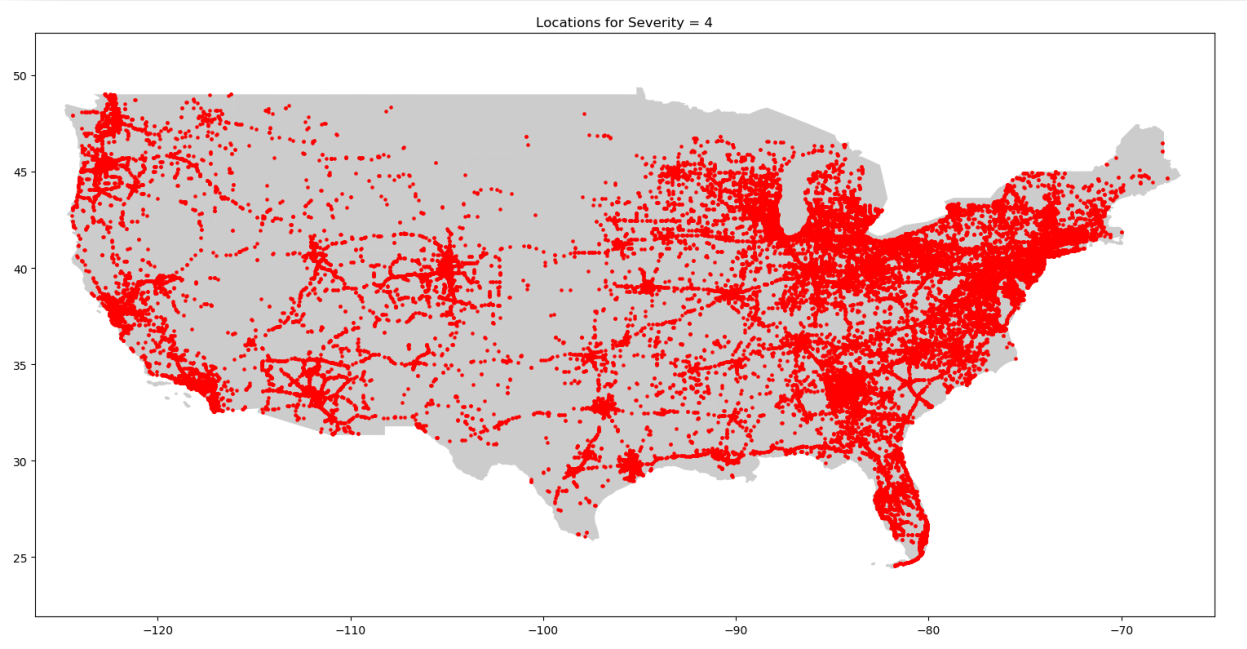


Figure 27: Start_Lat and Start_Lng for severity=4



We can deduce from the above plots that the majority of incidents happened in the Eastern and Western parts of the United States, which explains why the majority of accidents occurred in timezones such as Eastern Standard Time and Pacific Standard Time.