USA Car Accidents Severity Analysis Report

By
Utkarsha Negi

Contents

Introduction

Road accidents have become very common these days. Nearly 1.25 million people die in road crashes each year, on average, 3,287 deaths a day. Moreover, 20–50 million people are injured or disabled annually. Road traffic crashes rank as the 9th leading cause of death and accounts for 2.2% of all deaths globally. Road crashes cost USD 518 billion globally, costing individual countries from 1–2% of their annual GDP.
In the USA, over 37,000 people die in road crashes each year, and 2.35 million are injured or disabled. Road crashes cost the U.S. $230.6 billion per year or an average of $820 per person. Road crashes are the single greatest annual cause of death of healthy U.S. citizens travelling abroad.

Car accidents are unfortunately very common in the United States and the majority of these road crashes are caused by human error. While some are relatively minor, thousands of lives are taken every year by these horrible car crashes. Because your life can be at risk if you drive in an unsafe manner, it is so important to drive carefully and follow all traffic laws.

However, just because you are careful does not mean that you can assure that all other drivers on the road will do the same thing. If you are in a car crash, it may not be your fault, and you should not be held responsible for the damages caused by the ignorance or mistakes of other drivers. In such cases, you should consider protecting yourself by filing a car accident claim. There are so many damages, pains and frustrations that may arise as a result of a car accident, and it's best to guard your life above all.

Problem Statement

In an effort to reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

The model will help to analyse the answers for below questions.

Government officials and general public are lacking systems which can show
1. What is the accident-prone area in each state?
2. What day and time are safe to travel?
3. What are the factors responsible for accidents?
4. What is the severity of these accidents?
5. How many deaths happening in accidents?
6. What solution can be implemented to reduce accidents by each state?
7. How can this accident be minimised ?
8. How can the State Government improve accident-prone infrastructure?

Looking at the severity of road accidents, I decided to analyse the accidents' data to discover something useful. And here I am, sharing my results in the upcoming pages .

Data Acquisition and Cleaning

The dataset is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data 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 3.5 million accident records in this dataset.
The dataset contains 3.5 million rows and 49 columns(Quite a large dataset). A point to be noted is that even though the dataset contains data for only four years, there are 3.5 million accidents alread

In it's original form, this data is not fit for analysis. For one, there are many columns that we will not use for this model. Also, most of the features are of type object, when they should be numerical type.
We must use label encoding to covert the features to our desired data type.

Feature Description

As discussed earlier, the dataset contains 49 features, and their description is available on below.

S.No.	Attribute	Description
1	ID	This is a unique identifier of the accident record.
2	Source	Indicates source of the accident report (i.e. the API which reported the accident.).
3	TMC	A traffic accident may have a Traffic Message Channel (TMC) code which provides more detaile
4	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., lon
5	Start_Time	Shows start time of the accident in local time zone.
6	End_Time	Shows end time of the accident in local time zone.
7	Start_Lat	Shows latitude in GPS coordinate of the start point.
8	Start_Lng	Shows longitude in GPS coordinate of the start point.
9	End_Lat	Shows latitude in GPS coordinate of the end point.
10	End_Lng	Shows longitude in GPS coordinate of the end point.
11	Distance(mi)	The length of the road extent affected by the accident.
12	Description	Shows natural language description of the accident.
13	Number	Shows the street number in address field.
14	Street	Shows the street name in address field.
15	Side	Shows the relative side of the street (Right/Left) in address field.
16	City	Shows the city in address field.
17	County	Shows the county in address field.
18	State	Shows the state in address field.
19	Zipcode	Shows the zipcode in address field.
20	Country	Shows the country in address field.
21	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).
22	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.
23	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).
24	Temperature(F)	Shows the temperature (in Fahrenheit).
25	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).
26	Humidity(%)	Shows the humidity (in percentage).
27	Pressure(in)	Shows the air pressure (in inches).
28	Visibility(mi)	Shows visibility (in miles).
29	Wind_Direction	Shows wind direction.
30	Wind_Speed(mph)	Shows wind speed (in miles per hour).
31	Precipitation(in)	Shows precipitation amount in inches, if there is any.
32	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
33	Amenity	A POI annotation which indicates presence of amenity in a nearby location.
34	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.
35	Crossing	A POI annotation which indicates presence of crossing in a nearby location.
36	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.
37	Junction	A POI annotation which indicates presence of junction in a nearby location.
38	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.
39	Railway	A POI annotation which indicates presence of railway in a nearby location.
40	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.
41	Station	A POI annotation which indicates presence of station in a nearby location.
42	Stop	A POI annotation which indicates presence of stop in a nearby location.
43	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.
44	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.
45	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.
46	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.
47	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.
48	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.
49	Astronomical_Twili	Shows the period of day (i.e. day or night) based on astronomical twilight.

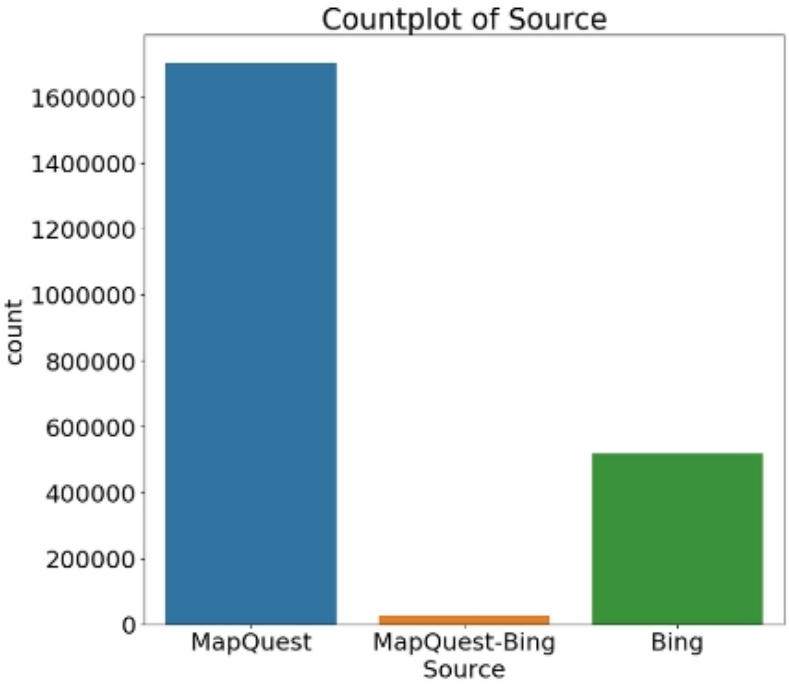
Methodology

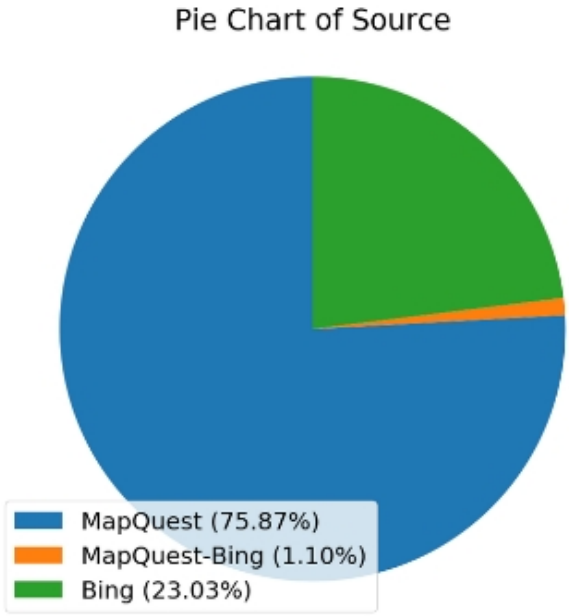
a. Exploratory Data Analysis

Before getting into the analysis part, let's look at the null values present in the dataset.
The figure below shows only those features which have null values.

TMC	516762
End_Lat	1727177
End_Lng	1727177
Description	1
Number	1458402
City	68
Zipcode	646
Timezone	2141
Airport_Code	23664
Weather_Timestamp	47170
Temperature(F)	62265
Wind_Chill(F)	1852370
Humidity(%)	64467
Pressure(in)	57280
Visibility(mi)	71360
Wind_Direction	47190
Wind_Speed(mph)	442954
Precipitation(in)	1979466
Weather_Condition	72004
Sunrise_Sunset	78
Civil_Twilight	78
Nautical_Twilight	78
Astronomical_Twilight	78

I will start by eliminating the unnecessary features first.
The feature Country contains only one entry — USA, which is quite apparent since we are dealing with the USA's dataset. Hence, I will be deleting this feature.
The feature Turning_Loop also contains one value — False. This means that there was no turning loop in the vicinity of any of the accidents. As this feature includes only one value, I'll be dropping this as well.
Let's look at the Source feature. It represents the API that reported the accident.





There are only three API sources that reported the accidents. It can be observed that most of the accidents(around 1,700,000) were reported by MapQuest, followed by Bing.

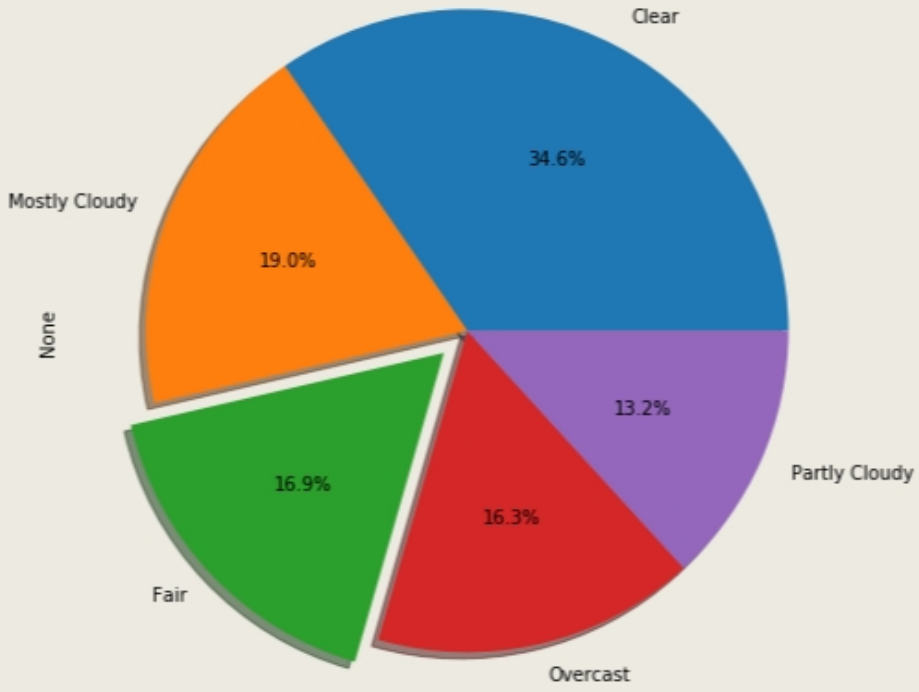
Methodology

b. Data Understanding

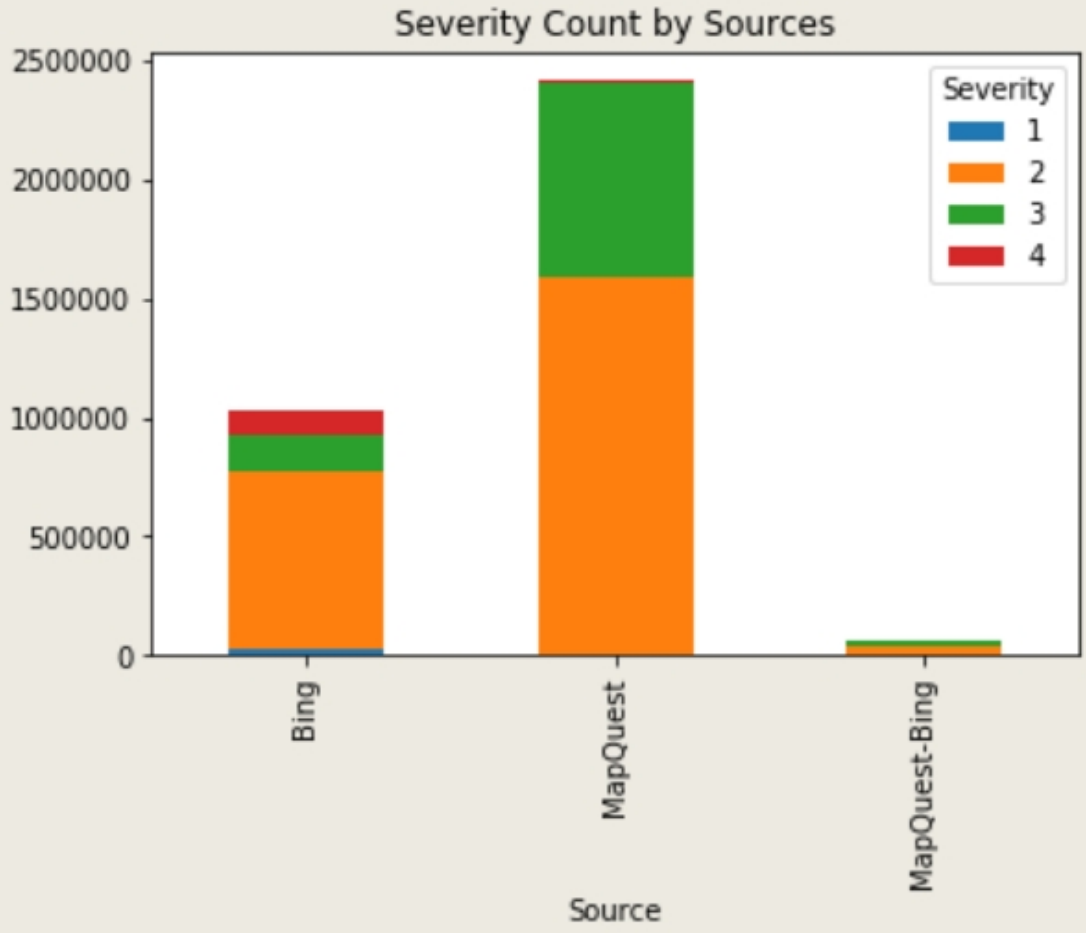
Since , we have a lot of Features , so we will focus on the Time , Weather , Bump and top 10 states where the accident occurs .see the impact of weather in the road accidents .

Top 5 weather condition with the accident percentage .

Surprisingly , mostly accidents occurred during a clear weather .



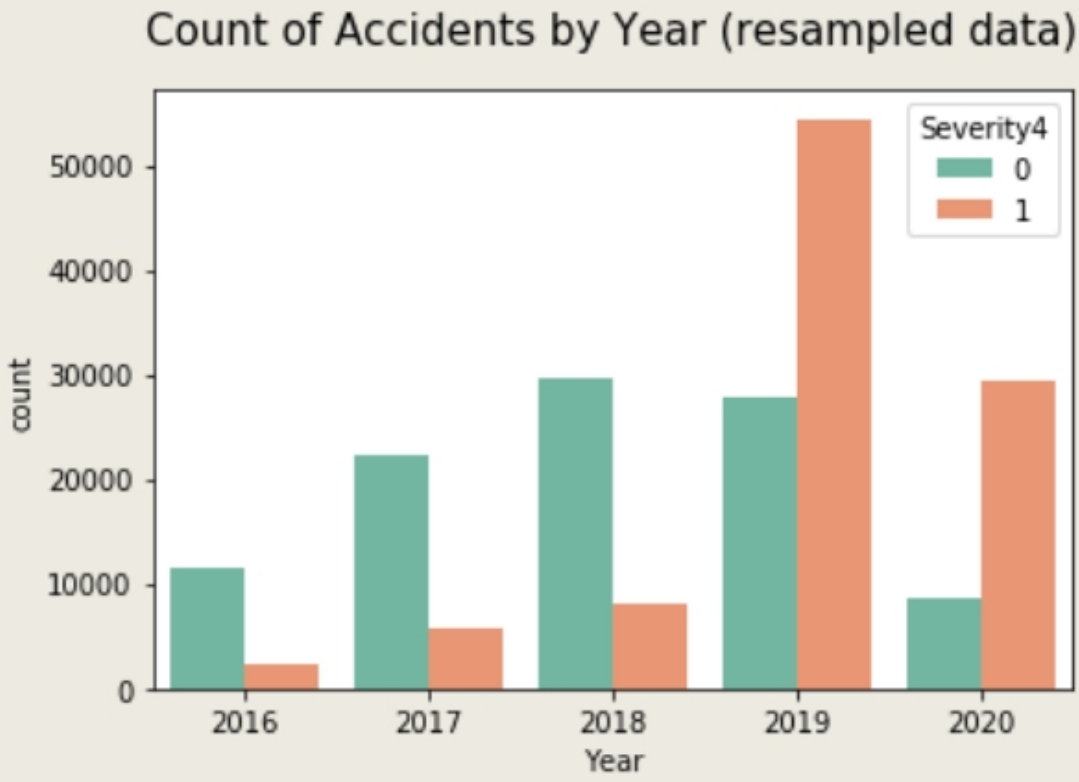
The most exciting feature is the **Severity**. It represents the severity of an accident.
Our predictor or target variable will be SEVERITYCODE because it is used measure the severity of an accident from 0 to 5 within the dataset. Attributes used to weigh the severity of an accident are WEATHER, ADDRESS , TIME, BUMP.



The plot depicts that mostly the accidents had severity equal to 2(average) followed by 3(above average), which is unfortunate. There are hardly any accidents with very low severity(0 and 1).

Let's look into Start_Time and End_Time features
The Start_Time and End_Time features depict the start and end time of an accident. To gain a better understanding, I have computed the duration of each accident.
It is interesting to see that the duration of accidents varies from minutes to years.

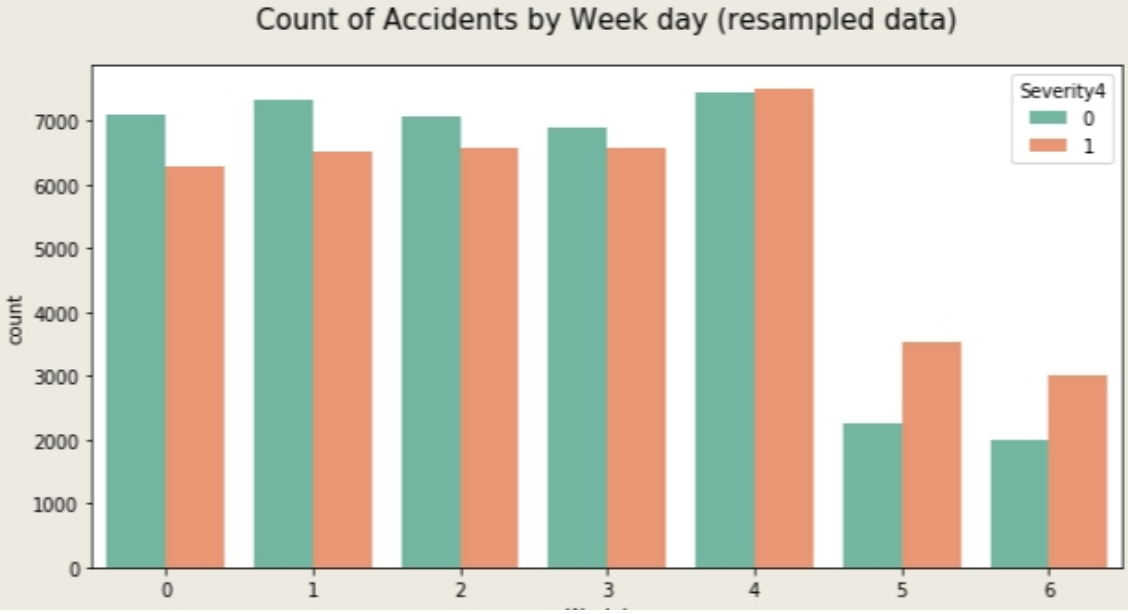
Let's see the variation from years to minutes one by one .



The plot depicts the severity 4 accidents in the form of boolean value. Since , we have to focus on the accident severity , so we will analyse with the severity 4.
The number of accidents with severity level 4 in 2019 is more than 5 times the number in 2018 while the number of other levels accidents is less. Let's back to raw data to have a look.

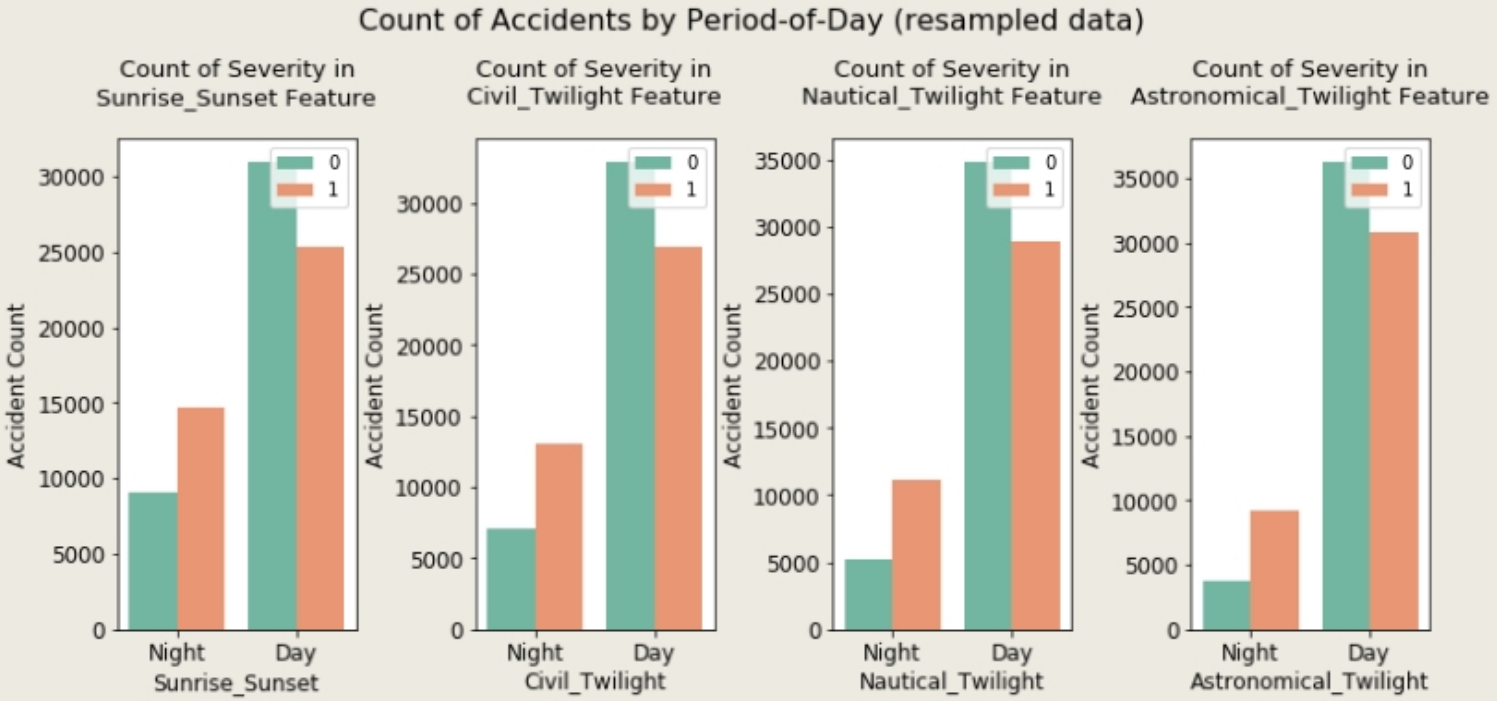
It's quite interesting that the count of other levels accidents is mostly consistent from March to December, whereas the number of level 4 accidents rapidly increased from March to May and remained stable until September then increased again from October

The number of accidents was much less on weekends while the proportion of level 4 accidents was higher.



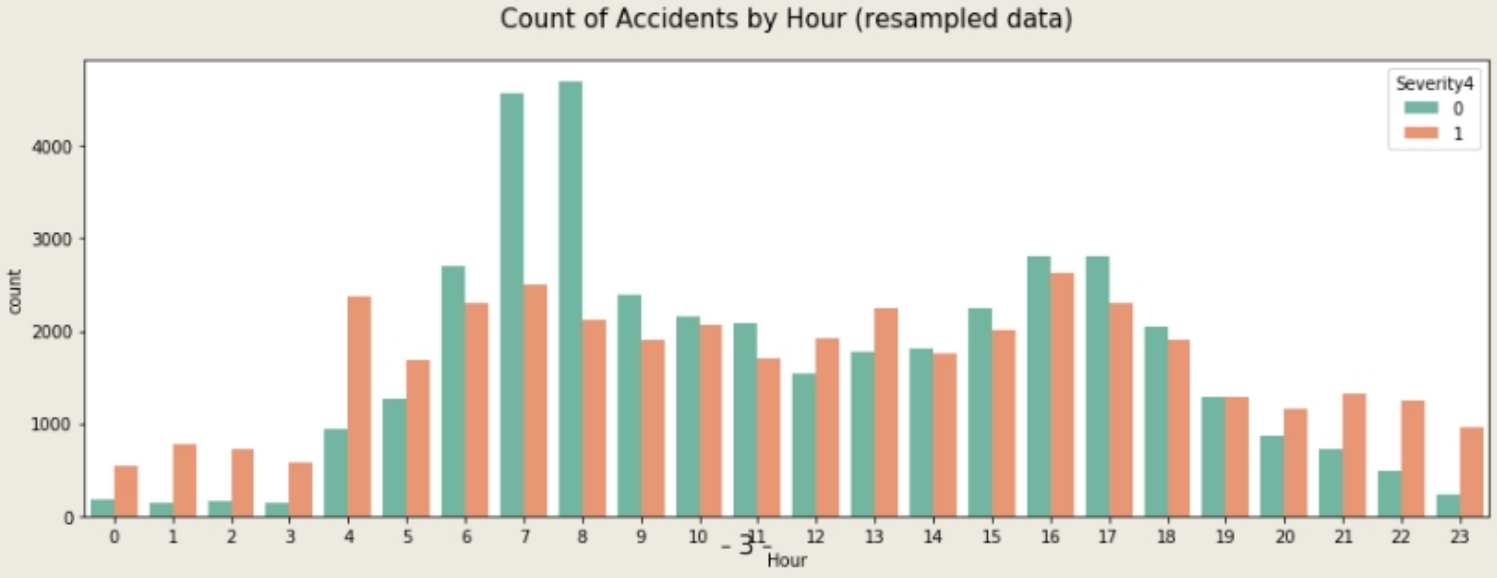
Period-of-Day

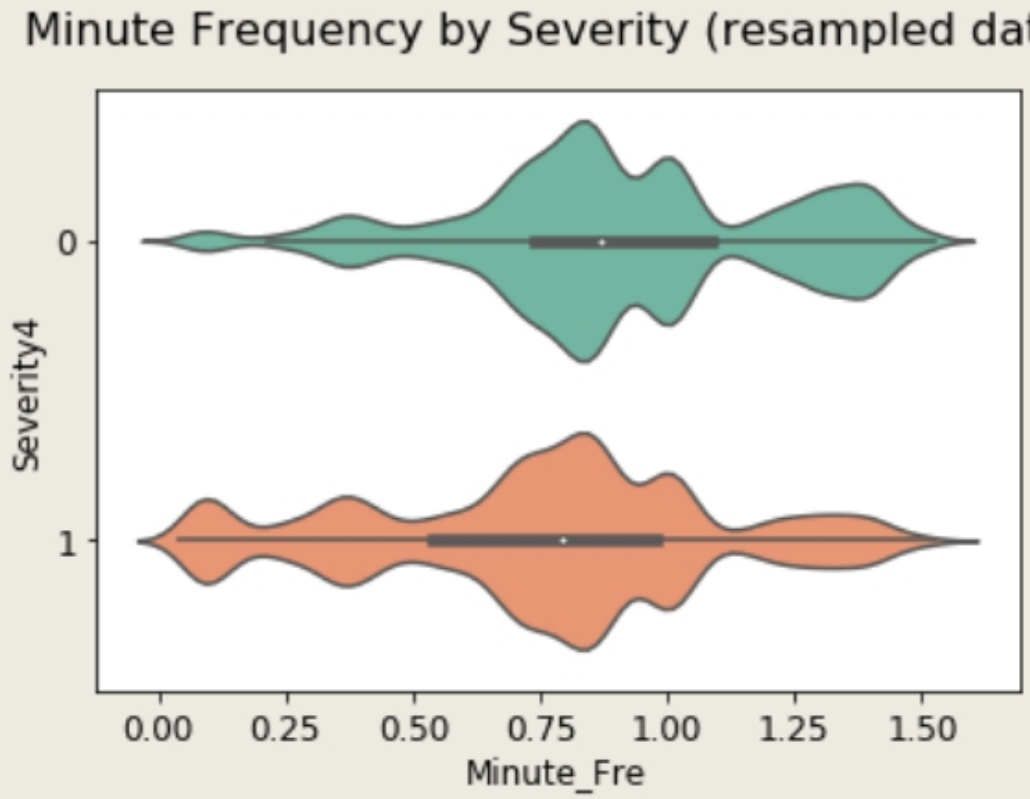
Accidents were less during the night but were more likely to be serious.



Hour

Most accidents happened during the daytime, especially AM peak and PM peak. When it comes to night, accidents were far less but more likely to be serious.

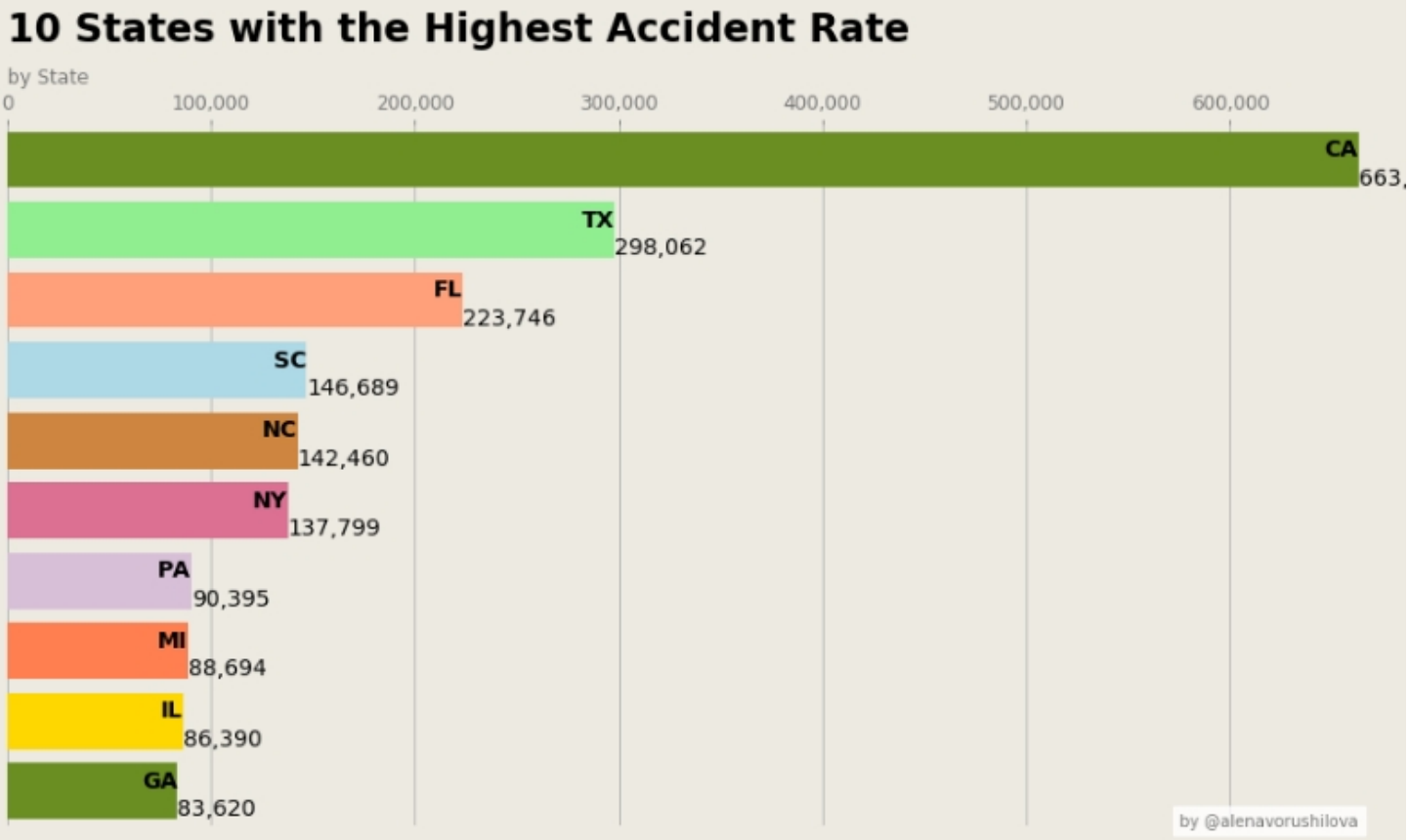




The violin plot shows that the overall minute frequency of accidents with severity level 4 is less than other levels. In other words, an accident is more likely to be a serious one when accidents happen less frequently.

Analysis on the basis of States and it's place

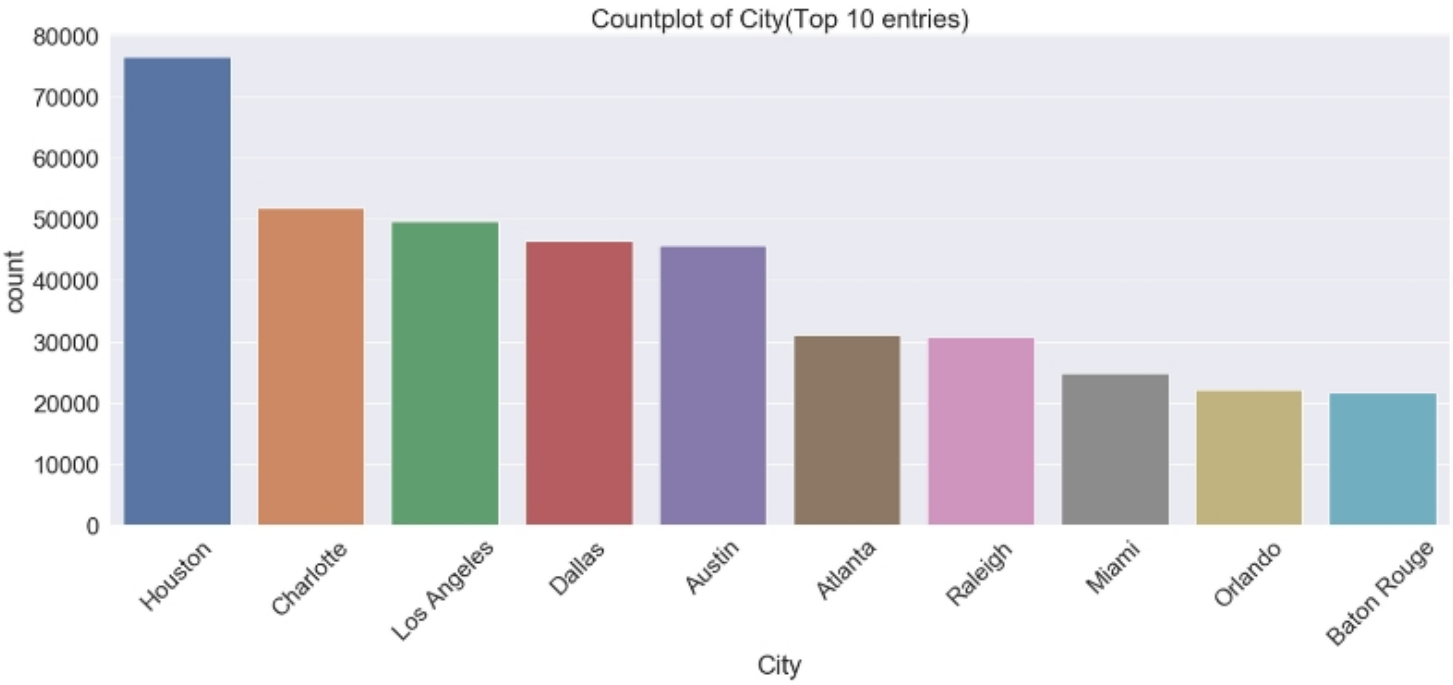
Now let’s look at the most accident-prone states in the USA.



The plot depicts that California(CA) has the most number of accidents followed by Texas(TX) and Florida(FL). It is interesting to see that the number of accidents in California is almost twice the number of accidents in Texas.

Analysis on the basis of Cities of USA

We see that most of the accidents occur in Houston, followed by Charlotte and Los Angeles.

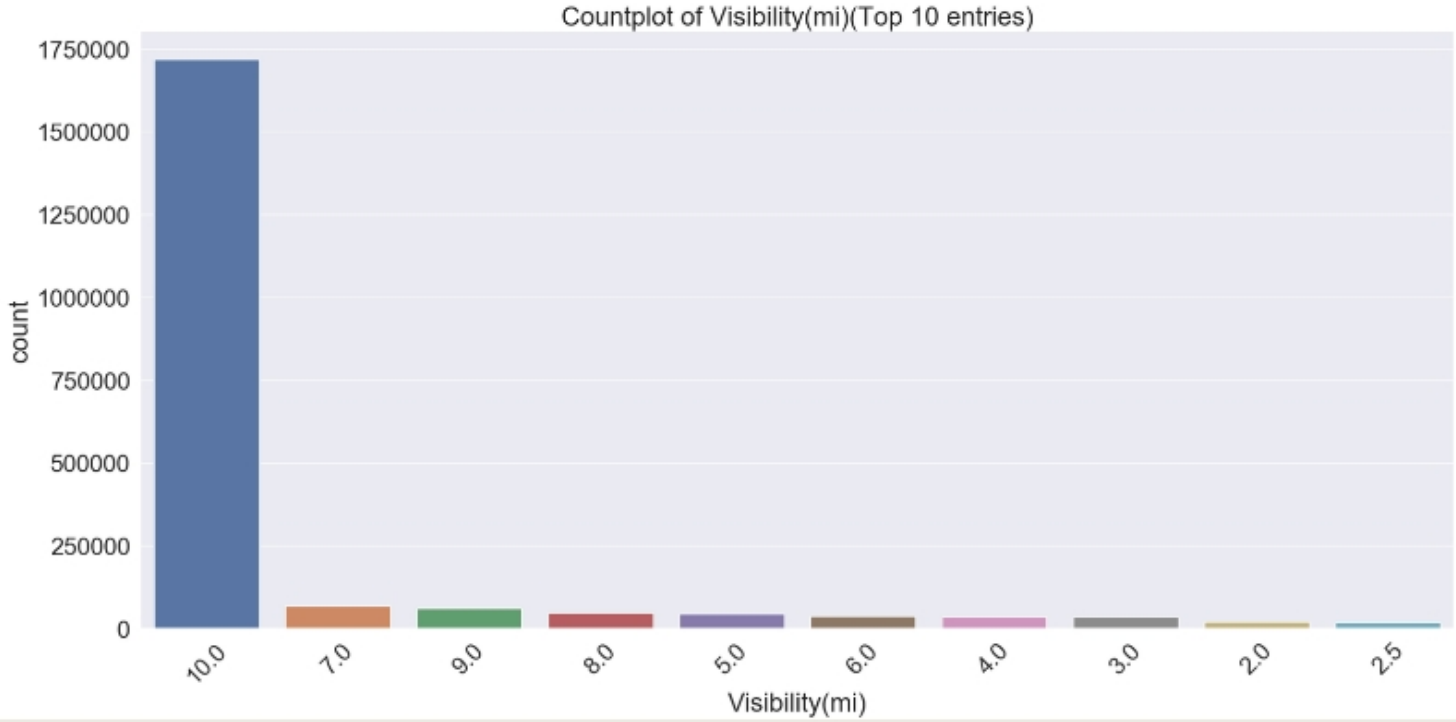


We can see that the most accident-prone city in the USA is Houston which is in Texas followed by Charlotte(North Carolina — which is number 4) and Los Angeles(California).

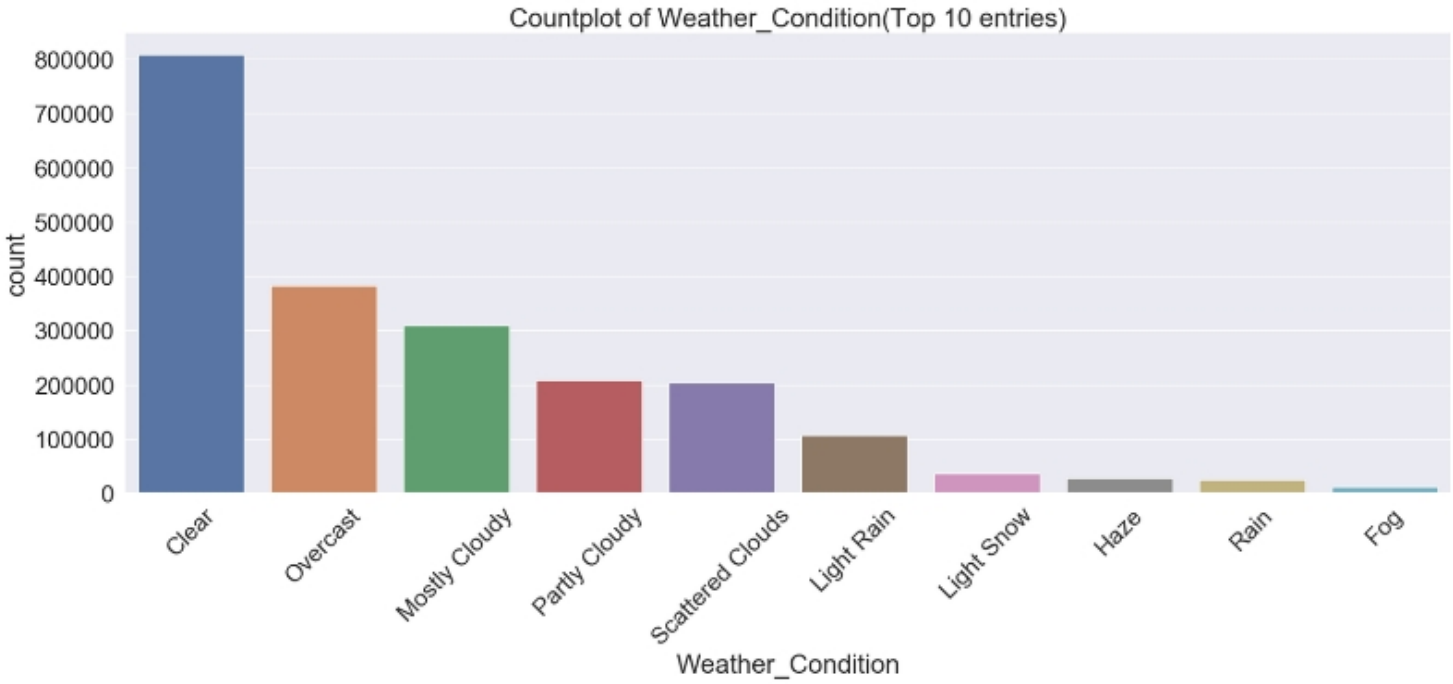
Let's go even deeper and see the visibility(mi) feature. It denotes the visibility in miles.

VISIBILITY

The below plot shows that most of the accidents occurred when visibility was high, which means that visibility is not a significant concern when it comes to accidents. This is obvious since low visibility is not the only factor.

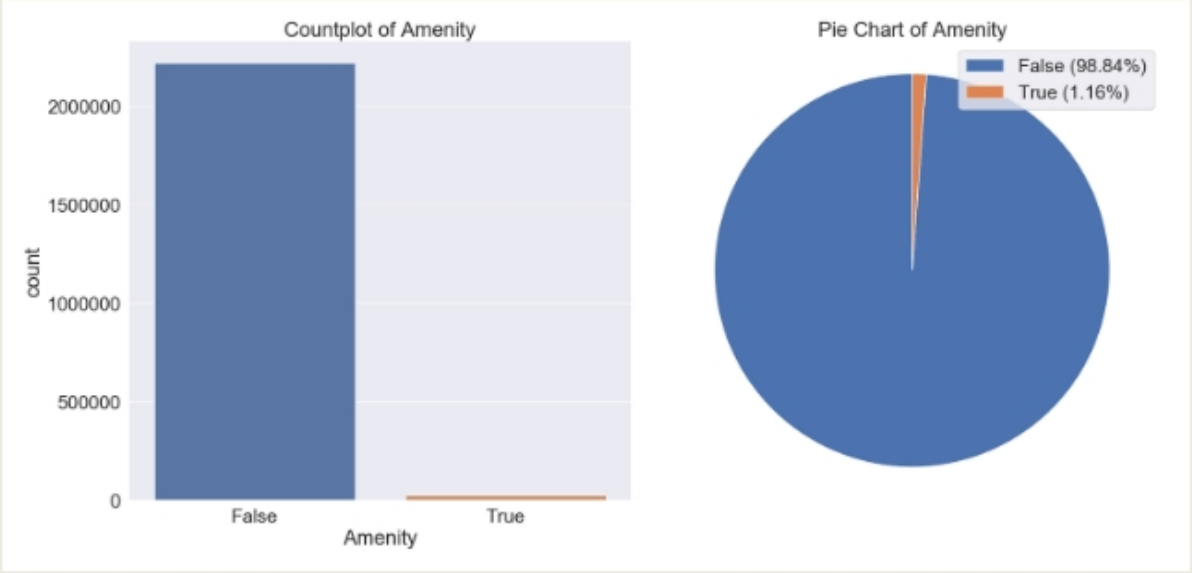


Now let's see the weather conditions during the accidents.

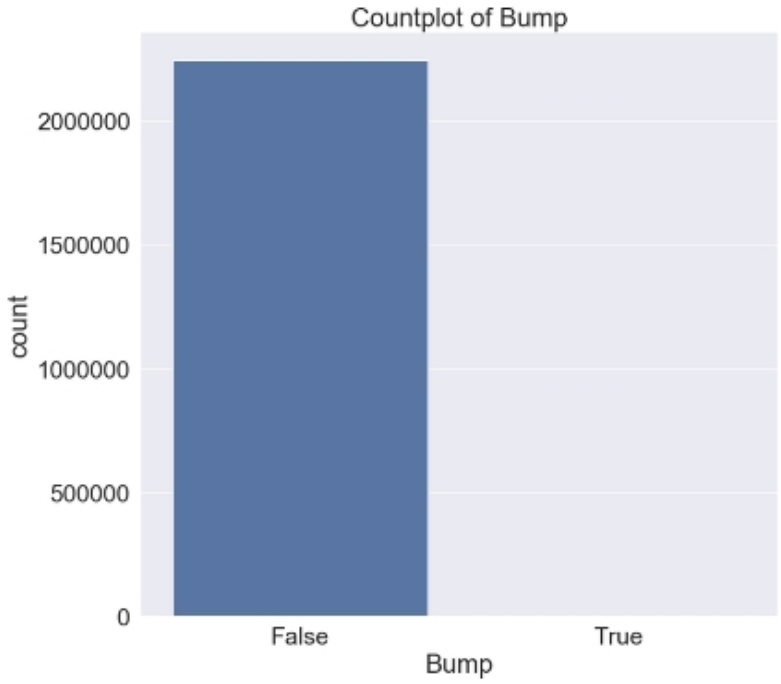


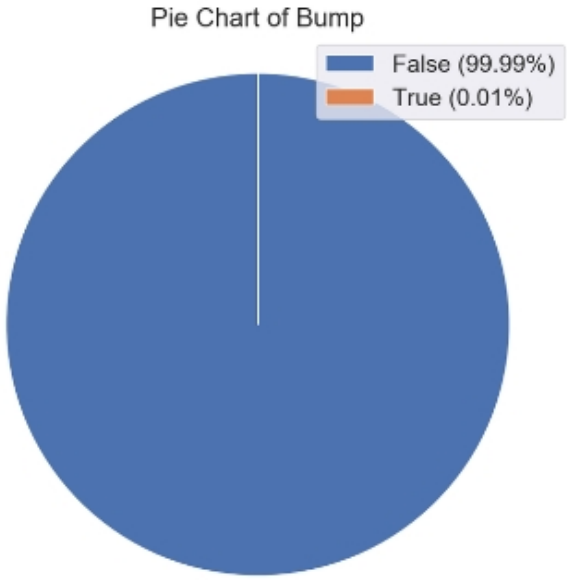
The plot depicts that the weather condition for most of the accidents was clear, followed by overcast and mostly cloudy. Overcast and mostly cloudy are reasonable factors for accidents unlike clear, which means that weather conditions also does not play an important role.

Let's look at the Amenity feature. This feature indicates the presence of amenity in a nearby location.



We see that for almost all(98.84%) accidents, there was no amenity available, which is unfortunate.
Now let's look at the bump feature. This feature indicates the presence of a speed bump or hump in a nearby location.





We see that 99.99% of the accidents were not due to a speed bump.

Our data is now ready to be fed into machine learning models.

c. Predictive Modelling

We will use the following models:

Logistic Regression
Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

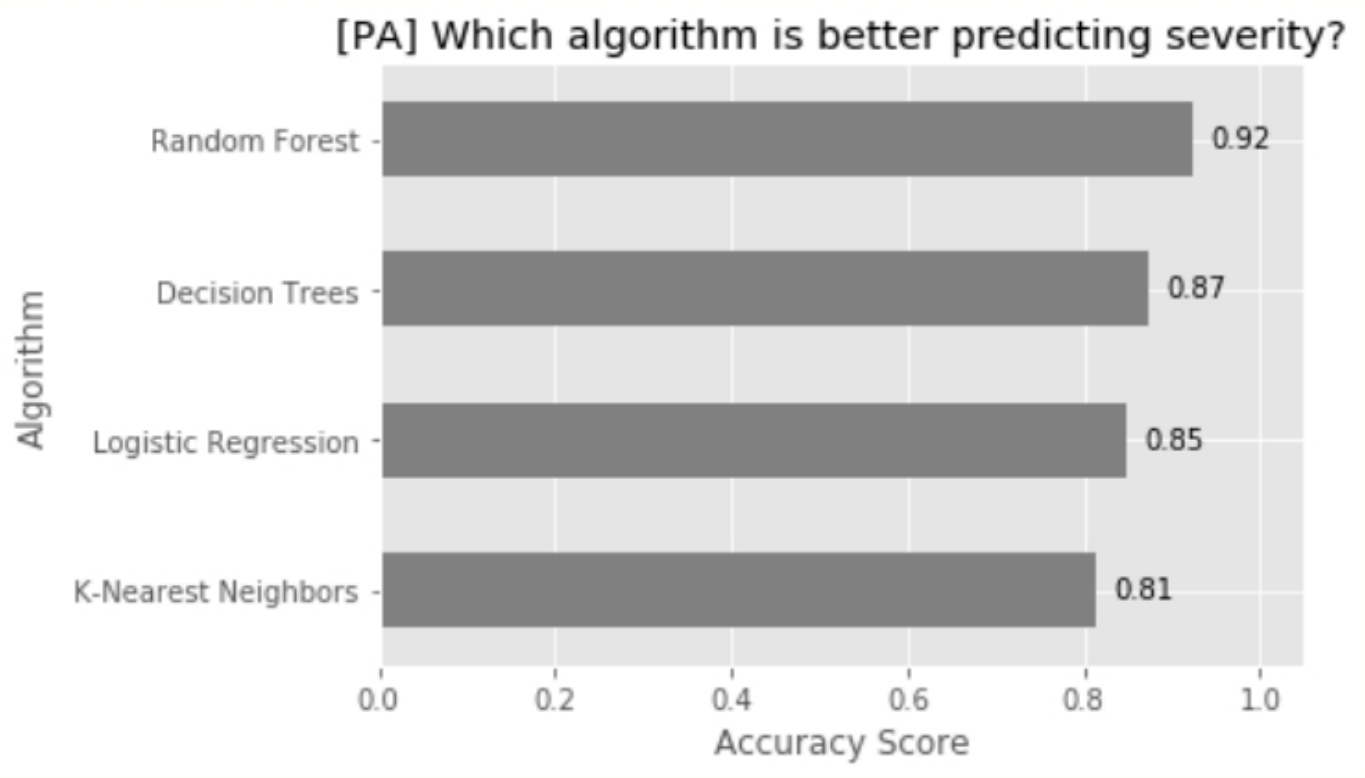
K-Nearest Neighbor (KNN)
KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

Decision Tree
A decision tree model gives us a layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.

Random Forest
Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

c. Predictive Modelling

You can go through the jupyter notebook for the predictive modelling part .
Here is the link :
https://github.com/utkarshanegi1212/Coursera_Capstone/blob/main/accident_severity_notebook.ipynb



From the above , we can see that Random Forest gives us the best accuracy .

Discussion

In the beginning of this notebook, we had categorical data that was of type 'object'. This is not a data type that we could have fed through an algorithm, so label encoding was used to created new classes that were of type int8; a numerical data type.

The USA accidents dataset, taken from Kaggle, was analyzed, and results were discussed above.

We came to a lot of exciting things like we came to know which city or state witnessed the most number of accidents in the USA, we even plotted the results on a map and also considered the severity of an accident.

Once we analyzed and cleaned the data, it was then fed through three ML models; K-Nearest Neighbour bor, Decision Tree and Logistic Regression , Random Forest . Although the Decision trees and Random Forest are ideal for this project, but the logistic regression made most sense because of its binary nature.

Conclusion

In conclusion, considering both performance and the time needed to train the model, I prefer using decision tree to make predictions. But if we care about nothing but accuracy, then I suppose random forest will be the winner.