NYU TANDON SCHOOL OF ENGINEERING

US-Accident Traffic Analysis and Severity Predictive Model

Project Report

Team Members:

Aditya Ashtekar: ama1219@nyu.edu

Ishan Tickoo: it732@nyu.edu

Meghana Bachu Srinath: mbs674@nyu.edu

Contents

Abstract

On average there are approximately 6 million car accidents in the US. It is being predicted that road traffic injuries will become the 5th leading cause of death by 2030. Accurate prediction of accidents can mitigate a considerable amount of road accidents. Most of the analysis and prediction use a small dataset which leads to inaccuracies or more false positives. Here for analysis we have used a dataset that consists of around 2.4 million entries spread across the entire US. A prediction model using Random Forest Classifier is also built for the entire dataset to predict the severity of the accidents. The downstream use of this project is to notify the users about the severity of the accident so that they can choose an alternative path.

Introduction

According to the Association for safe International road travel, more than 38,000 people die every year in crashes on U.S. roadways. The U.S. traffic fatality rate is 12.4 deaths per 100,000 inhabitants.Further 4.4 million are critically injured enough to seek medical treatment. Road accidents for people 1-54 years old are the leading cause of death in the U.S. The economic and social effects of traffic accidents are costing $871 billion for US people. Road accidents with direct care expenses caused the USA more than $380 million. Several factors are responsible for accidents.

- Poor road infrastructure and management
- Non-road worthy vehicles
- Unenforced or non-existent traffic laws
- Unsafe road user behaviors and
- Inadequate post-crash care.

Road collisions can be anticipated and avoided through knowing each of these causes and by planning, proactive management, and evidence-based approaches. With the help of existing traffic accidents data we have made an attempt to identify the hotspots and weather conditions that were responsible for accidents. With the given analysis extra precautions can be taken by the government and travellers at these hotspots during the given months and the days when the weather conditions are similar. For example, Fall has more accidents than any other season.With road accidents comes the Traffic delays. Accidents with severity 0 - 1 experience least delays, 2- 3 considerate amount of time is wasted. Whereas accidents with severity 4 experience maximum delays. Several people miss meetings, appointments and things like that due to traffic delays. What if there was a method to notify people with the severity of the accidents caused and the delay that might occur in that route. This can help people switch to alternative routes. On running a Predictive model with real time data, we can predict the severity of the accident, which can be used in realtime to notify the people about it.

Big Data Technologies Used

Apache Spark

Spark is the heart of the processing in this architecture. The in-memory computation of spark enables fast computation for large datasets. Using a state-of-the-art DAG scheduler, a database optimizer and a physical execution system, Apache Spark achieves high performance for both batch and stream data.All the operations

from pre-processing to publishing the data would take place in Spark. As our dataset contains 2.9 million rows, we used the Dumbo cluster and Hadoop Distributed File System to process the data.

Data Visualization tools

A typical requirement in business use cases is the visualization of the recommendations suggested by the algorithm. Tools like Tableau, Spotfire can be used where reports can be tweaked to show relevant information to the end user. We have used Tableau for visualization.

Data set

The dataset is a countrywide list of traffic accidents, spanning 49 US states. As of February 2016, the data is continuously collected using several data providers, including two APIs that provide data on streaming event traffic. Such APIs transmit traffic data collected from a number of organizations, such as the U.S. and state highway services, law enforcement authorities, traffic monitors, and traffic controls within the road networks. There are currently around 3.0 million reports of incidents in this dataset.

Feature Description of the dataset:

1. **ID** - Unique identifier assigned for each accident record
2. **Source** - API source for each record. (Map-quest, Bing)
3. **TMC** - Traffic Message channel code to access more about the accident
4. **Severity** - Gives the level of impact on traffic due to the accident, value ranges from 1 - 4, where 1 means that delay of traffic due to accident is minimal. If level is 4 then there would be significant delay.
5. **Start_Time** - Start time of the accident
6. **End_Time** - End time of the accident
7. **Start_Lat** - Gps coordinates of starting latitude position of the accident
8. **Start_Lng**- Gps coordinates of starting longitude position of the accident
9. **End_Lat-** Gps coordinates of ending latitude position of the accident
10. **End_Lng**- Gps coordinates of ending longitude position of the accident
11. **Distance(mi)** - Distance affected due to the accident
12. **Description** - Brief overview of the accident
13. **Airport_Code** - Airport based weather station present near the accident location.
14. **State** - State where the accident occurred
15. **Street** - Street name where the accident occured
16. **Side** - Denotes weather the on the right side or left side of the accident
17. **County** - Displays the county.
18. **Timezone** - Based on the location Timezone is determined
19. **Country-** All the accidents in the database are from the United States.
20. **City** - Displays the city.
21. **Number** - Street number
22. **Zipcode** - Zipcode of the location
23. **Weather_Timestamp** - Time Stamp of the weather observation record

24. **Temperature(F)** - Temperature in Fahrenheit
25. **Wind_Chill(F)** - Wind Chill in Fahrenheit
26. **Humidity(%)** - Humidity in percentage
27. **Pressure(in)** - Pressure in inches
28. **Visibility(mi)** - Visibility in miles
29. **Wind_Direction** - Direction of the wind
30. **Wind_Speed(mph)** - Wind speed in miles per hour
31. **Precipitation(in)** - Precipitation in inches
32. **Weather_Condition** - Displays weather condition (Rain, snow,fog,etc)
33. **Amenity** - Presence of any amenity nearby
34. **Bump** - Presence of any bump nearby
35. **Crossing**- Presence of any crossing nearby
36. **Give_Way** - Presence of any Give_way nearby
37. **Junction**- Presence of any junction nearby
38. **No_Exit** - Presence of any no-exit nearby
39. **Railway-** Presence of any railway nearby
40. **Roundabout** - Presence of any roundabout nearby
41. **Station** - Presence of any station nearby
42. **Stop** - Presence of any stop nearby
43. **Traffic_Calming** - Presence of any Traffic - Calming nearby
44. **Traffic_Signal** - Presence of any Traffic signal nearby
45. **Turning_Loop** - Presence of any turning loop nearby
46. **Sunrise_Sunset** - Denotes day or night
47. **Civil_Twilight** - Denotes day or night based on civil twilight
48. **Astronomical_Twilight** - Denotes day or night based on astronomical twilight.
49. **Nautical_Twilight** - Denotes day or night based on nautical twilight

Technology Stack

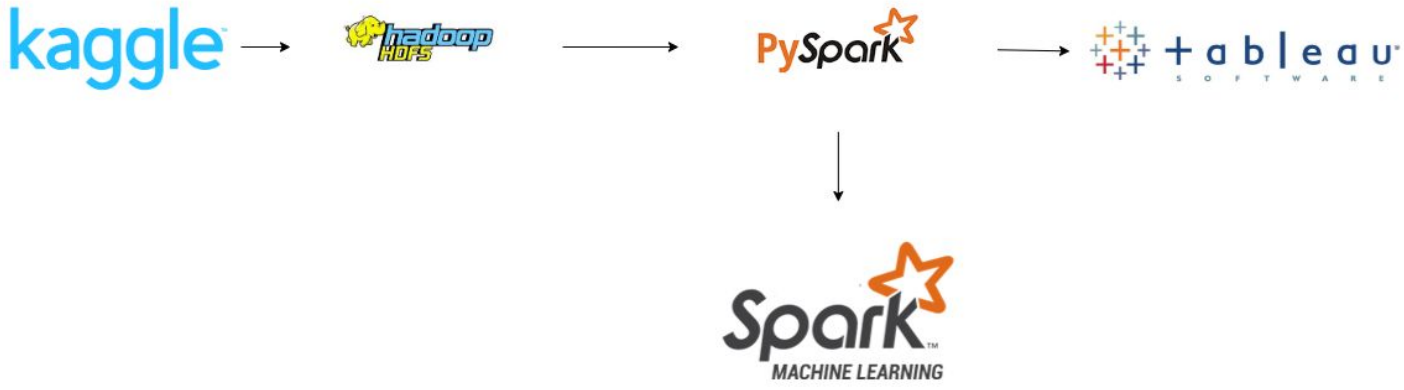
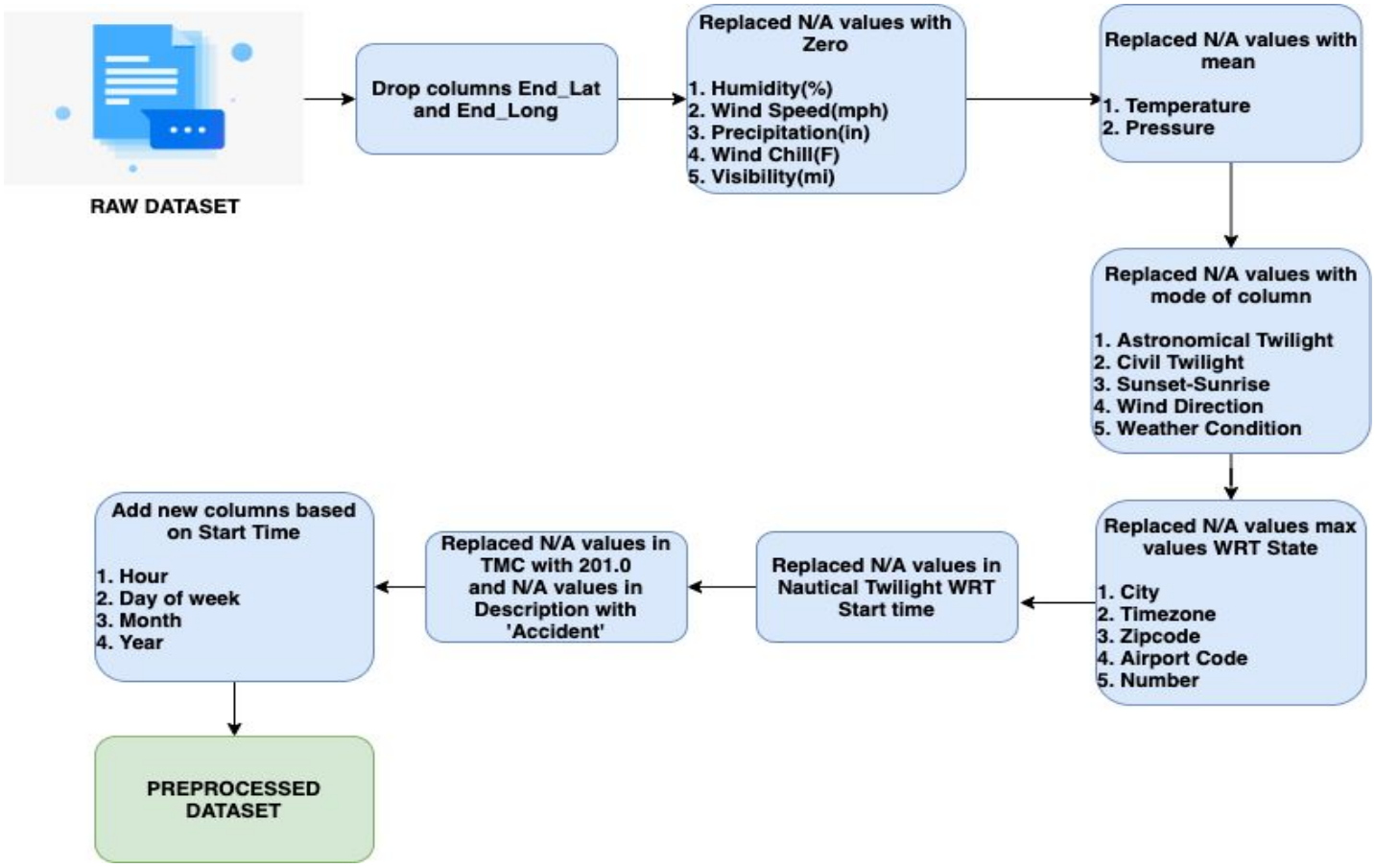


Fig 1: Technology Stack

Data Preprocessing



Data cleaning not only enhances the consistency of the data, it also improves the overall productivity in the analysis. Since data is a major advantage in many industries, unreliable data can be risky. Incorrect data will reduce the efficiency of prediction, thus reducing revenue and performance. If the enterprise has clear data so it is easy to avoid getting into these circumstances. And the way to go is to clean up the details. It removes significant errors and contradictions which are unavoidable when different data sources and used to pull in data.

We started with data cleaning by filtering the null values first. Since, there were a lot of null values present in the dataset, it would not have been feasible to delete all the rows containing null values. So for columns that contained a numeric value we replaced the null values with zero. We did this for 5 columns:'Humidity(%)', 'Precipitation(in)', 'Wind_Chill(F)', 'Wind_Speed(mph)','Visibility(mi)' because it is a really strong possibility that their values were null in some cases simple because they were zero. For example it does not always rain. Then for null values in the Temperature and Pressure columns, we replaced null values by mean of the total values in the respective columns because simply replacing the Temperature to zero would be incorrect.

For categorical columns ‘Sunrise_Sunset’, ‘Civil_Twilight’, ‘Astronomical_Twilight’, ‘Wind_Direction’ and ‘Weather_Condition’ all null values were replaced by the value that occured the maximum number of times in the accident.

For columns ‘End_Lat’ and ‘End_Lng’ there were almost all null values and in most cases their values were similar to ‘Start_Lat’ and ‘End_Lng’ so we dropped them from the dataset.

For ‘TMC’ we filled the null values with 201.0 which means Accident [3]. For categorical columns ‘City’, ‘Timezone’, ‘Zipcode’, ‘Airport_Code’, ‘Number’ we filled the na values with the values that occured the most amount of time in the State i.e. Houston is the city with the most number of accidents in Texas so all null values in the ‘City’ column with State TX were replaced by Houston. All the null values in the Description column were just replaced by the word ‘Accident’. Finally for ‘Nautical_Twilight’ we used the hour function from the ‘Start_Time’ to establish if the accident happened by night i.e. if an accident happened after 6 and before 18 then null value was replaced by day else by night.

List below shows the null values that were present before data cleaning.

ID	0
Source	0
TMC	728071
Severity	0
Start_Time	0
End_Time	0
Start_Lat	0
Start_Lng	0
End_Lat	2246264
End_Lng	2246264
Distance(mi)	0
Description	1
Number	1917605
Street	0
Side	0
City	83
County	0
State	0
Zip Code	880
Country	0
Timezone	3163
Airport_Code	5691
Weather_Timestamp	36705
Temperature(F)	56063
Wind_Chill(F)	1852623
Humidity(%)	59173
Pressure(in)	48142
Visibility(mi)	65691
Wind_Direction	45101

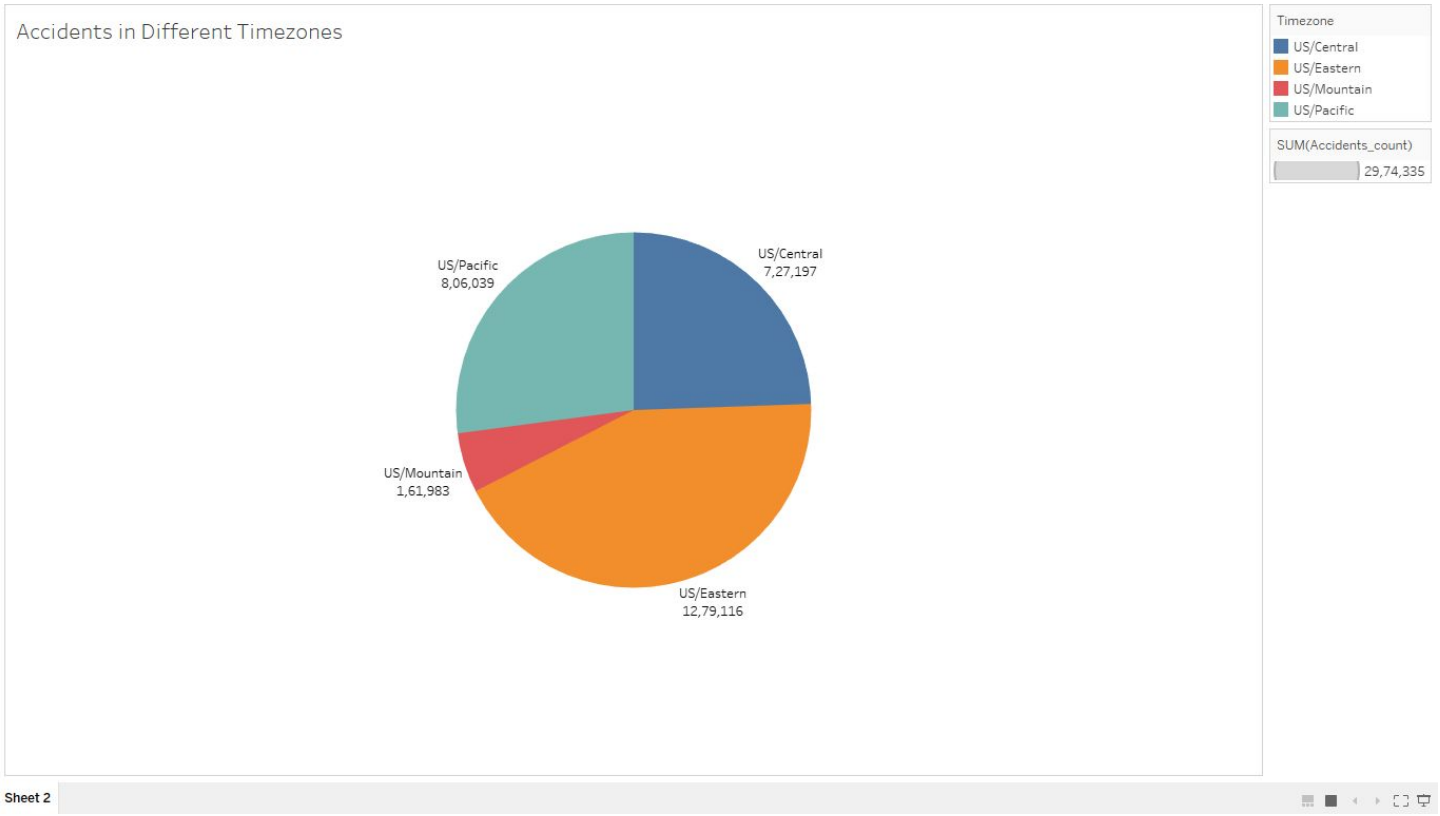
Wind_Speed(mph)	440840
Precipitation(in)	1998358
Weather_Condition	65932
Amenity	0
Bump	0
Crossing	0
Give_Way	0
Junction	0
No_Exit	0
Railway	0
Roundabout	0
Station	0
Stop	0
Traffic_Calming	0
Traffic_Signal	0
Turning_Loop	0
Sunrise_Sunset	93
Civil_Twilight	93
Nautical_Twilight	93
Astronomical_Twilight	93

Data Analysis

We performed data analysis after cleaning the data and generated some insights. The following are the results of our analysis :

TimeZone-

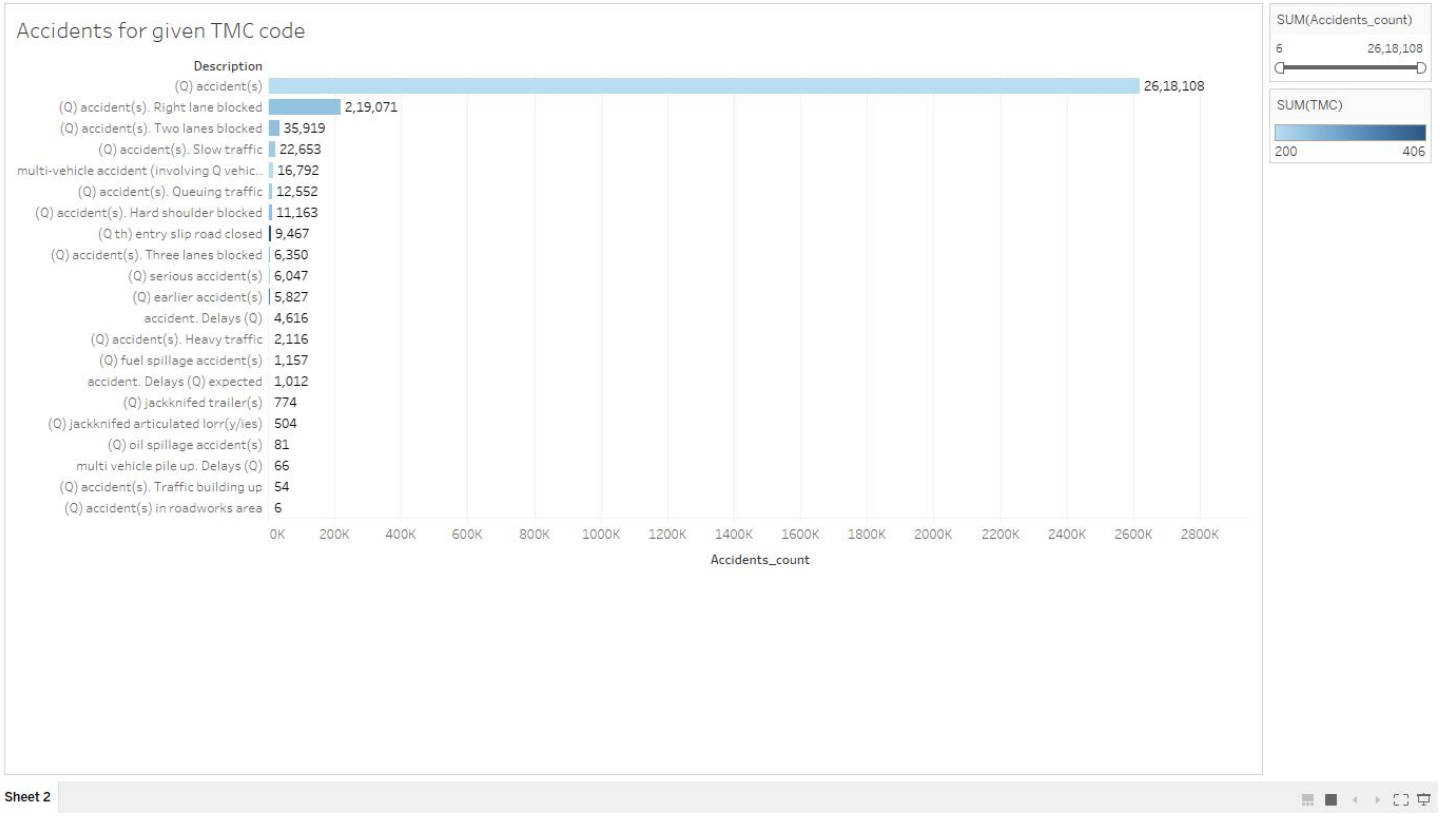
There are 4 timezones in the United States. Our data set identifies which timezone the accident occurred.



The US Eastern time zone has the maximum number of accidents.

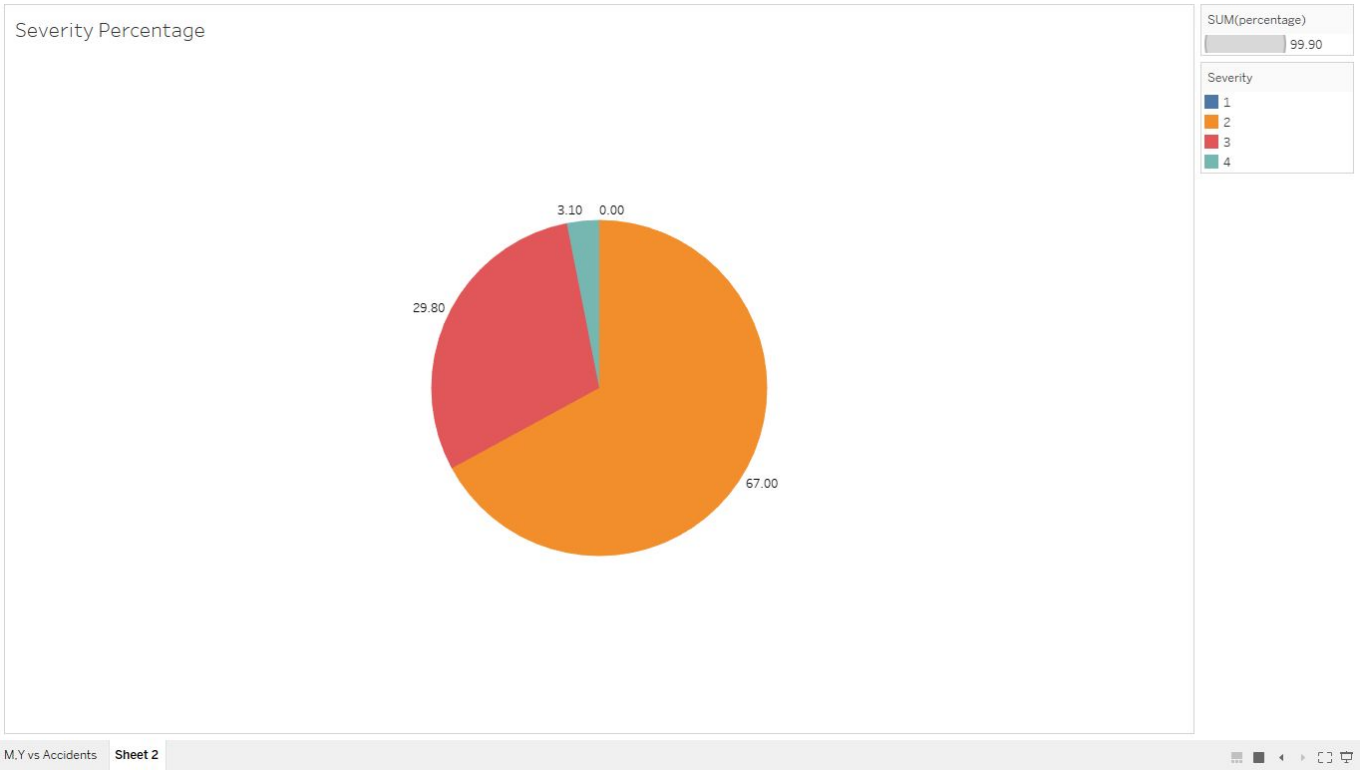
TMC Distribution-

TMC stands for Traffic Message Channel. Its aim is to provide information to mobile receivers about traffic disturbances or alerts, such as navigation devices. The details found in messages that are transmitted over FM radio, along with tv broadcasts. To pick up this inaudible signal a special radio receiver is needed. Every message contains information about an event, such as a traffic jam, and the location of that event. Both are encoded as numbers referring to lookup tables - the event code list and location code list.



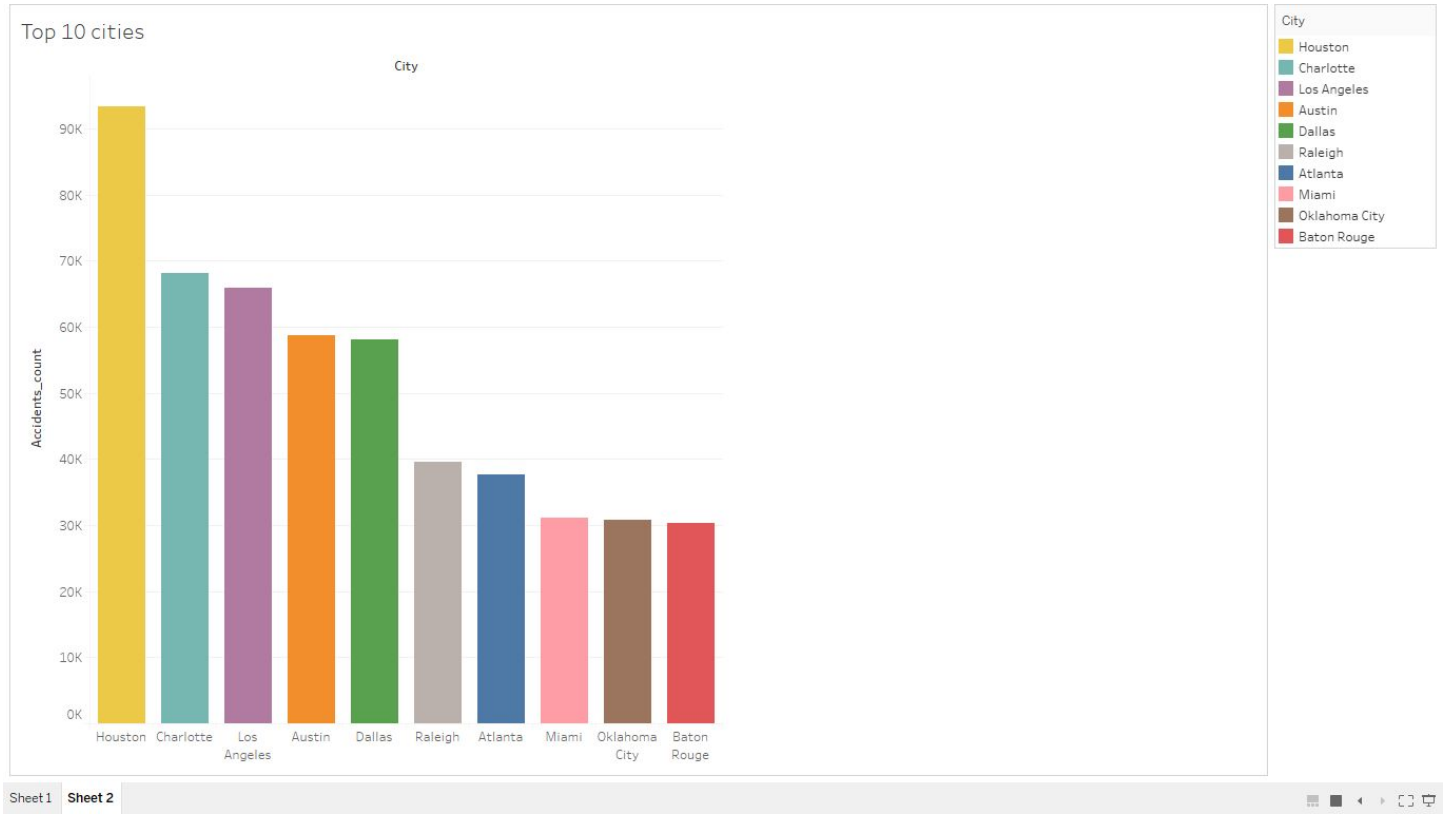
Severity Distribution:

Severity here ranges from a value of 1 to 4 where 1 being the least effect on the traffic and 4 being most time delay.



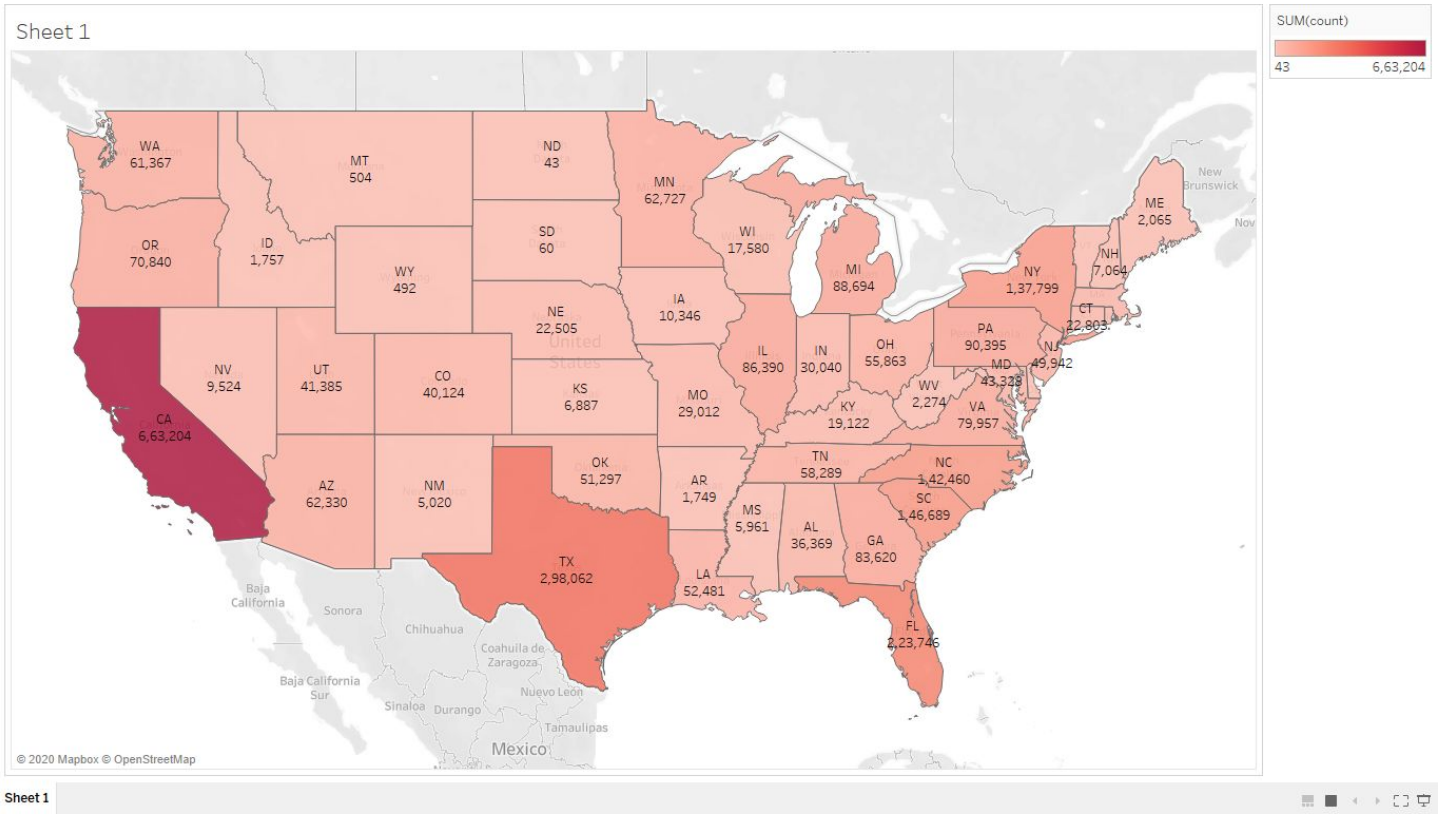
Accident distribution across the cities-

Houston,Texas has the most number of accidents. Top 10 cities along with their accident count is displayed below.



Accident distribution across the states-

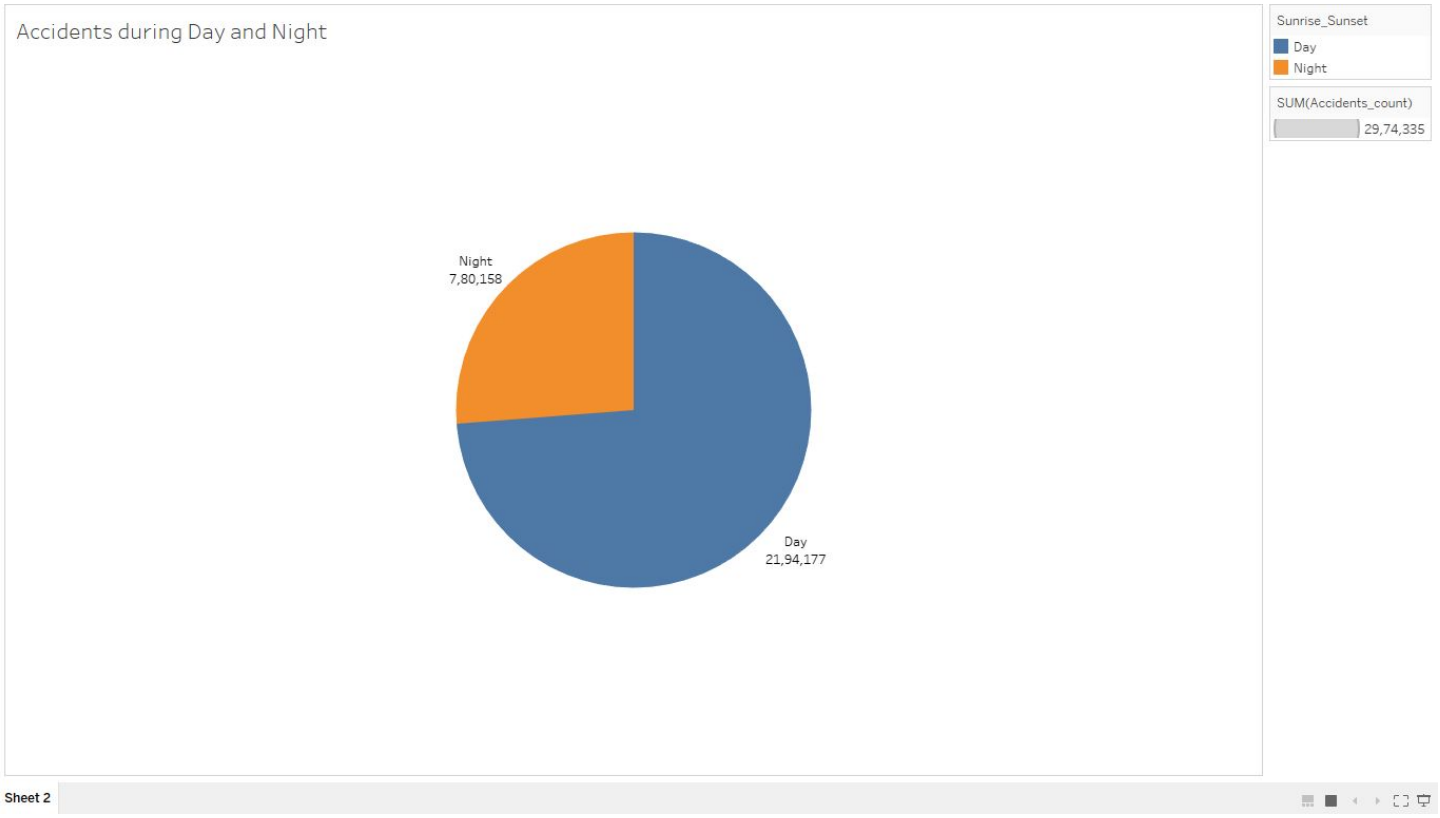
California tops the list. In three years California has encountered 663204 accidents. Top 10 states along with their accident count is displayed below.



Considering the top 15 states, its percentage distribution is displayed below.

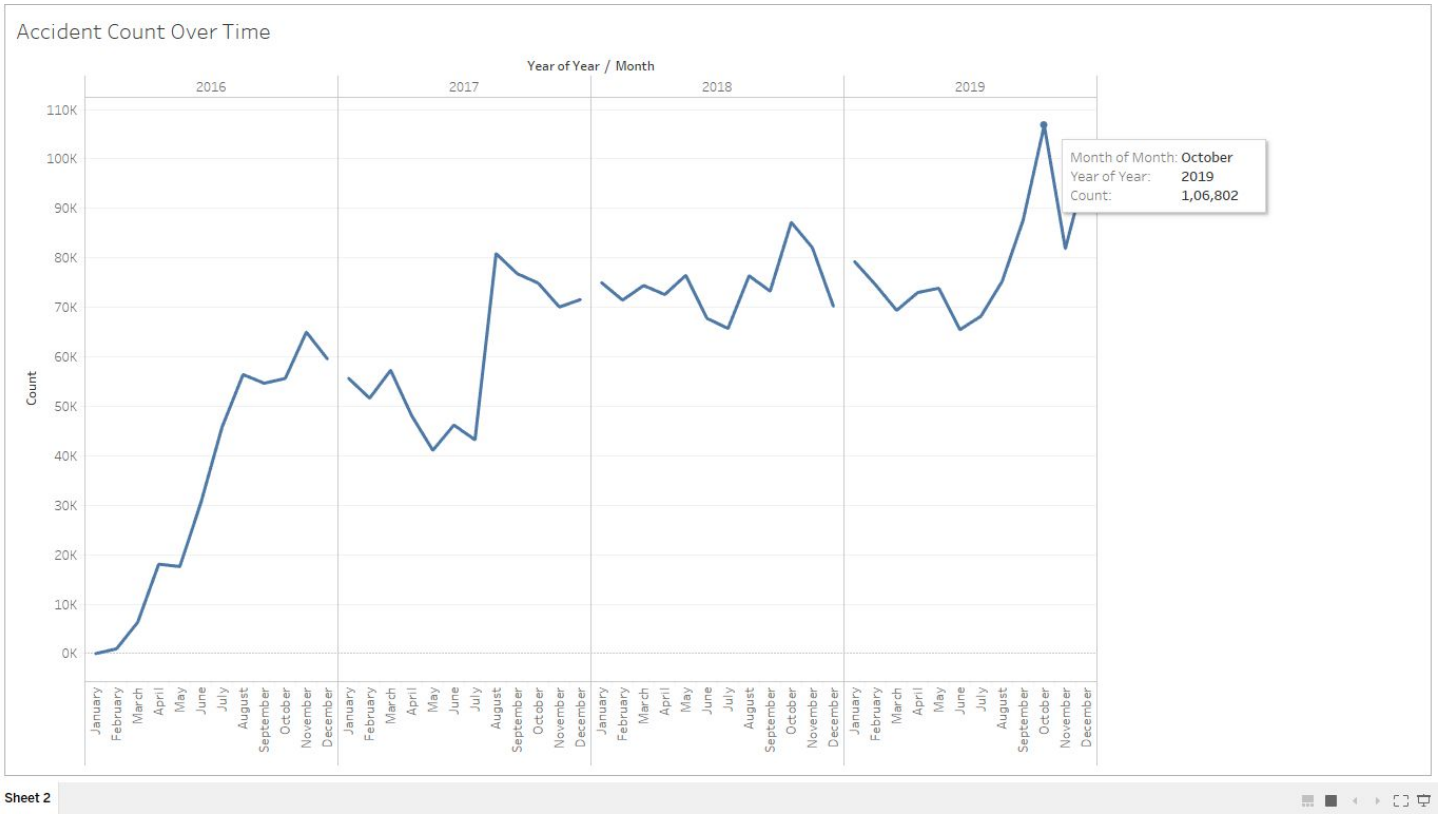
Accident distribution between day and night-

More accidents have occurred in Day compared to Night.



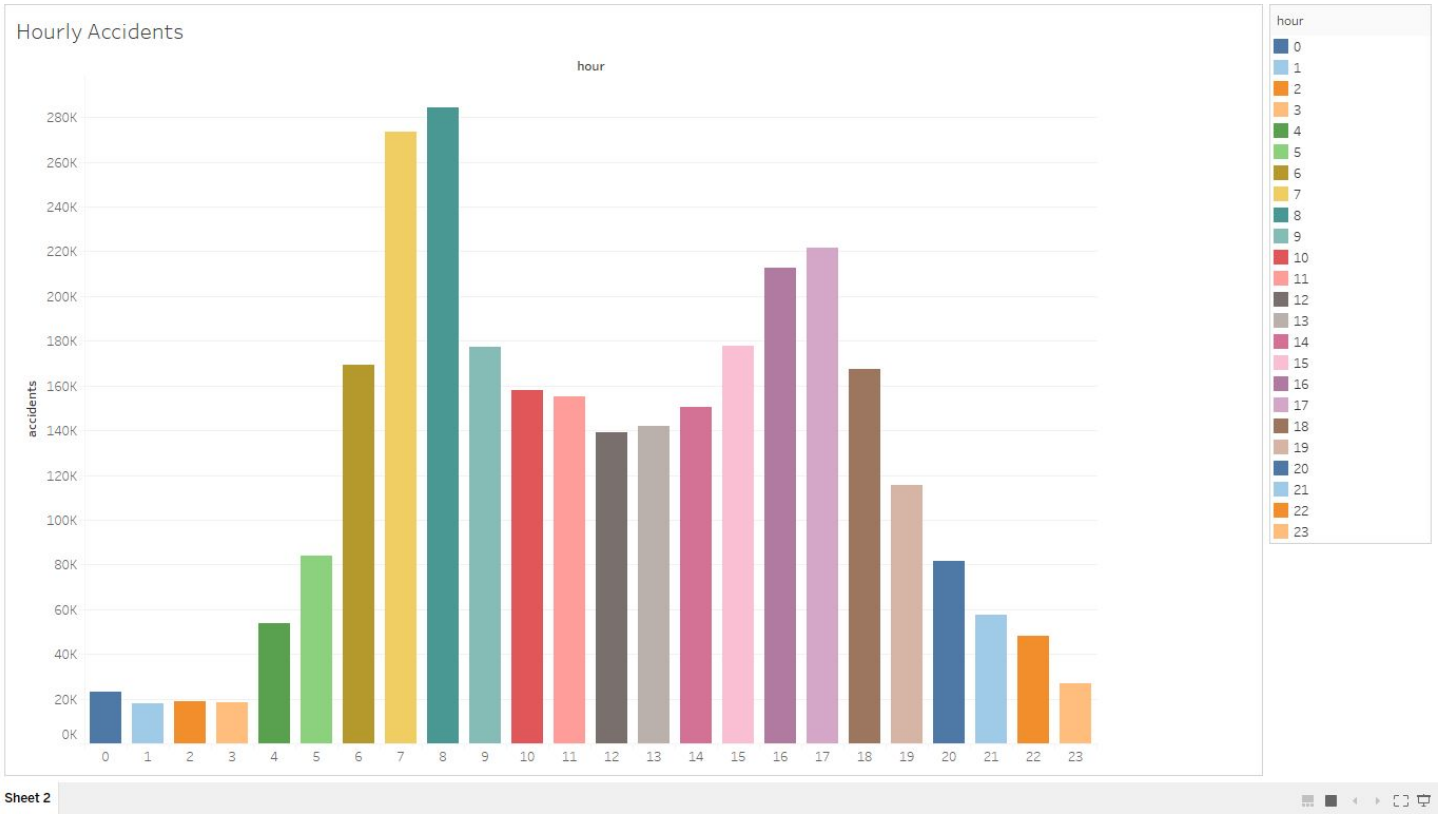
Accidents count with time -

Accidents seem to have increased over time.



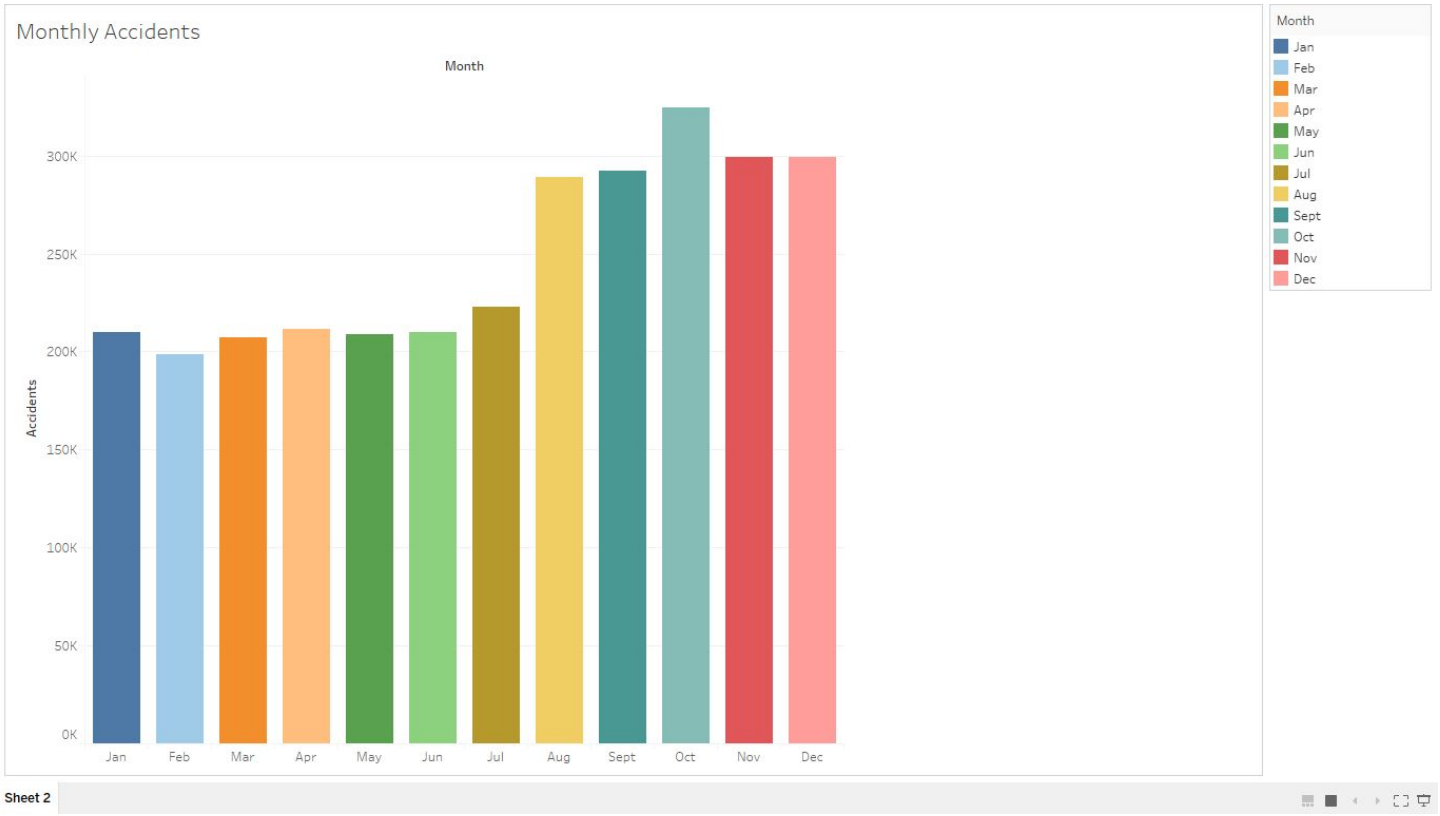
Distribution of accidents in a day on hourly basis -

Around 7AM -9AM a considerable amount of accidents have occured.

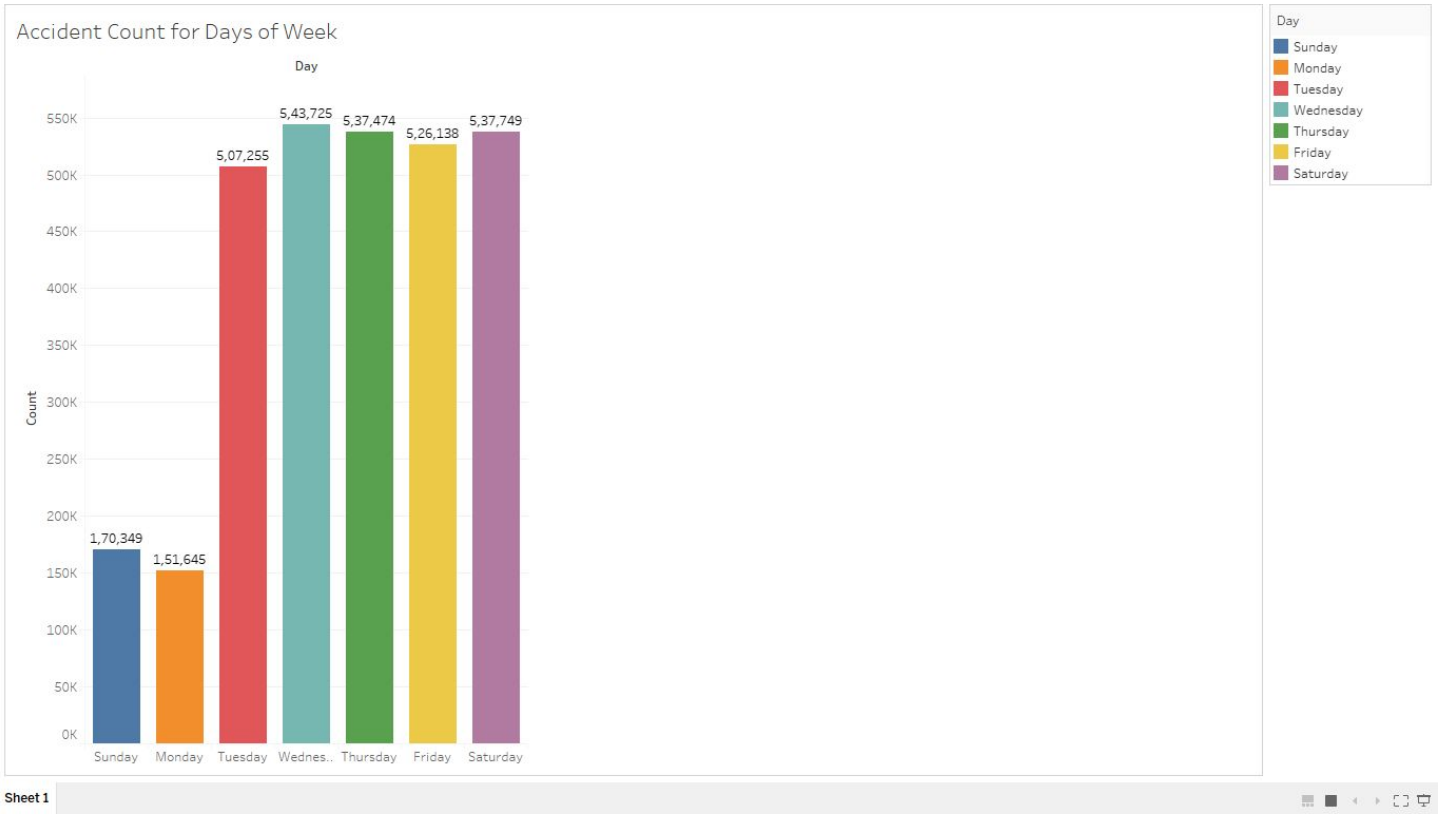


Distribution of accidents in a day on monthly basis -

Fall has more accidents compared to other seasons.

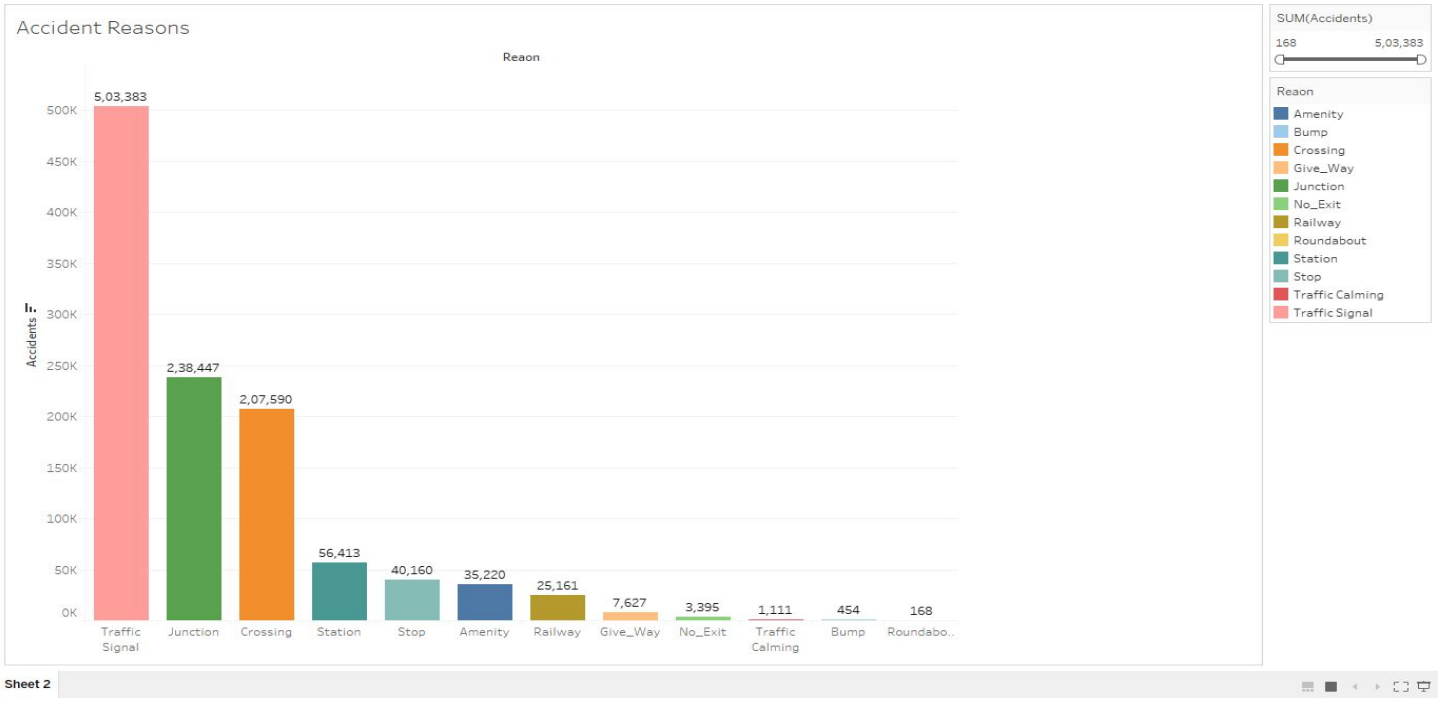


Distribution of accidents across different days of the week-



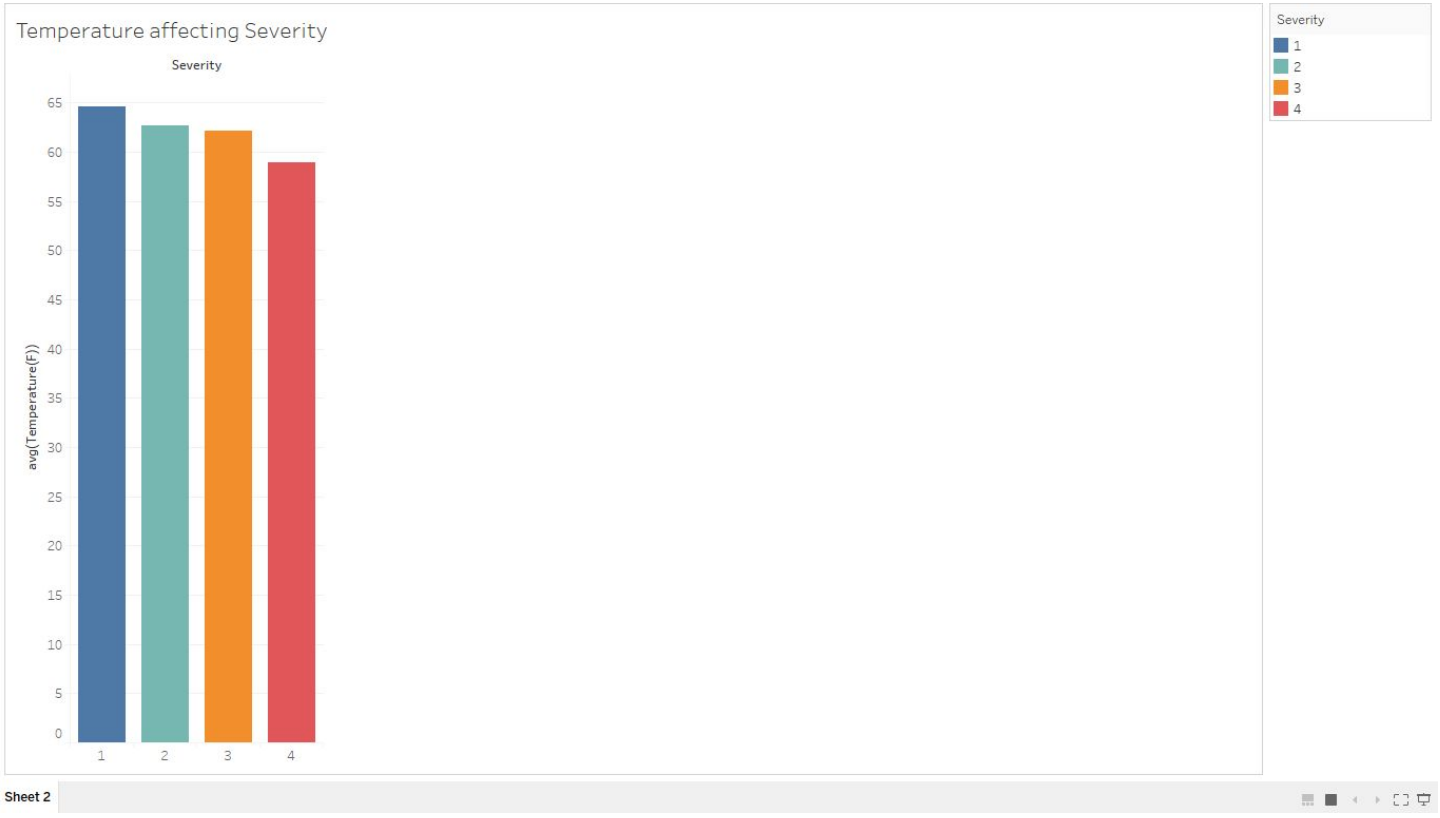
Distribution of accidents based on external factors-

Traffic signals are installed to prevent accidents. Ironically, from the below results it can be inferred that around traffic signals the number of accidents are more.



Distribution of accidents due to various weather factors-

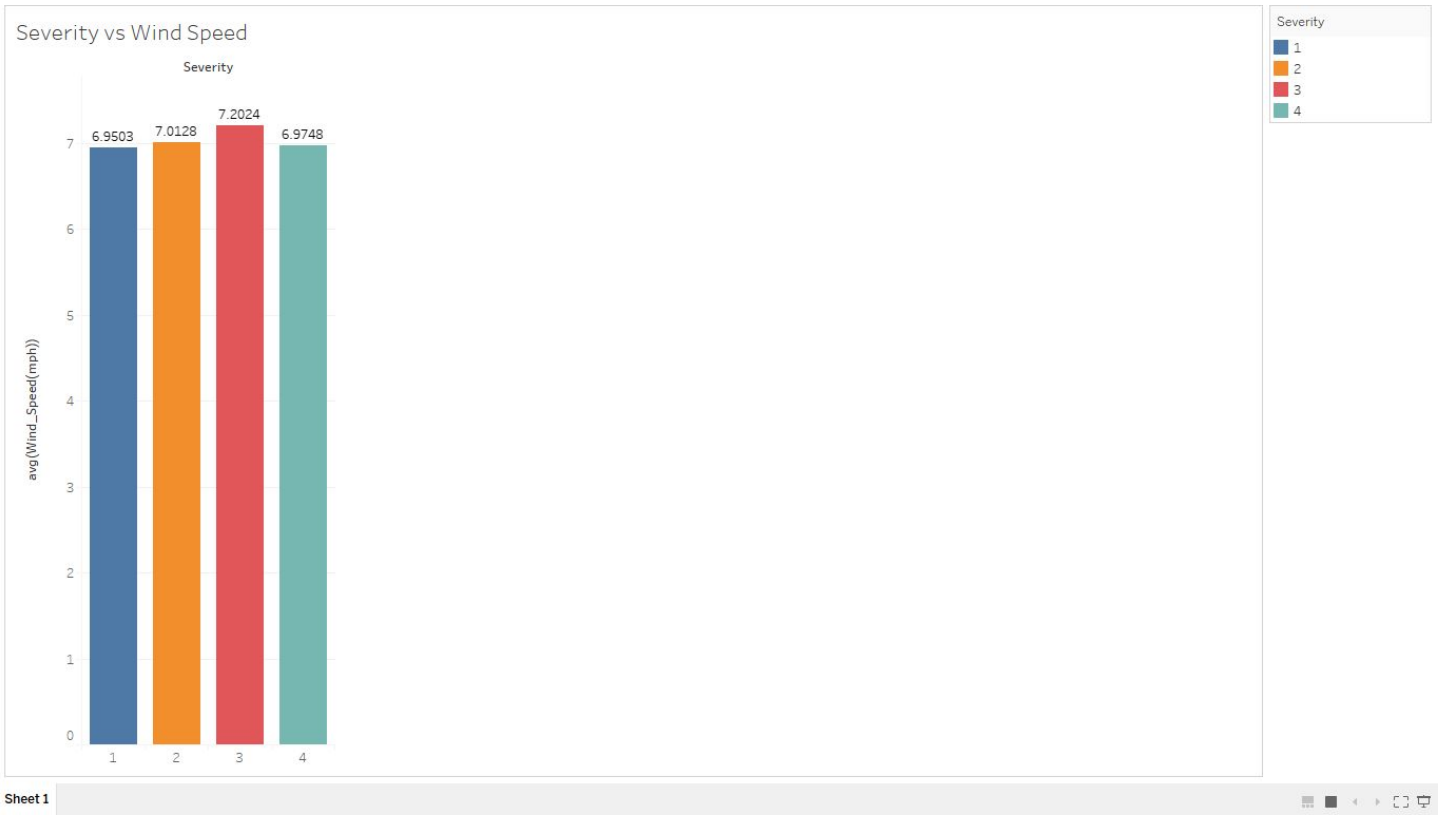
As the temperature and visibility decreases the severity of the accident increases. Whereas Wind speed, pressure has no considerable effects. As the Humidity increases severity increases too.

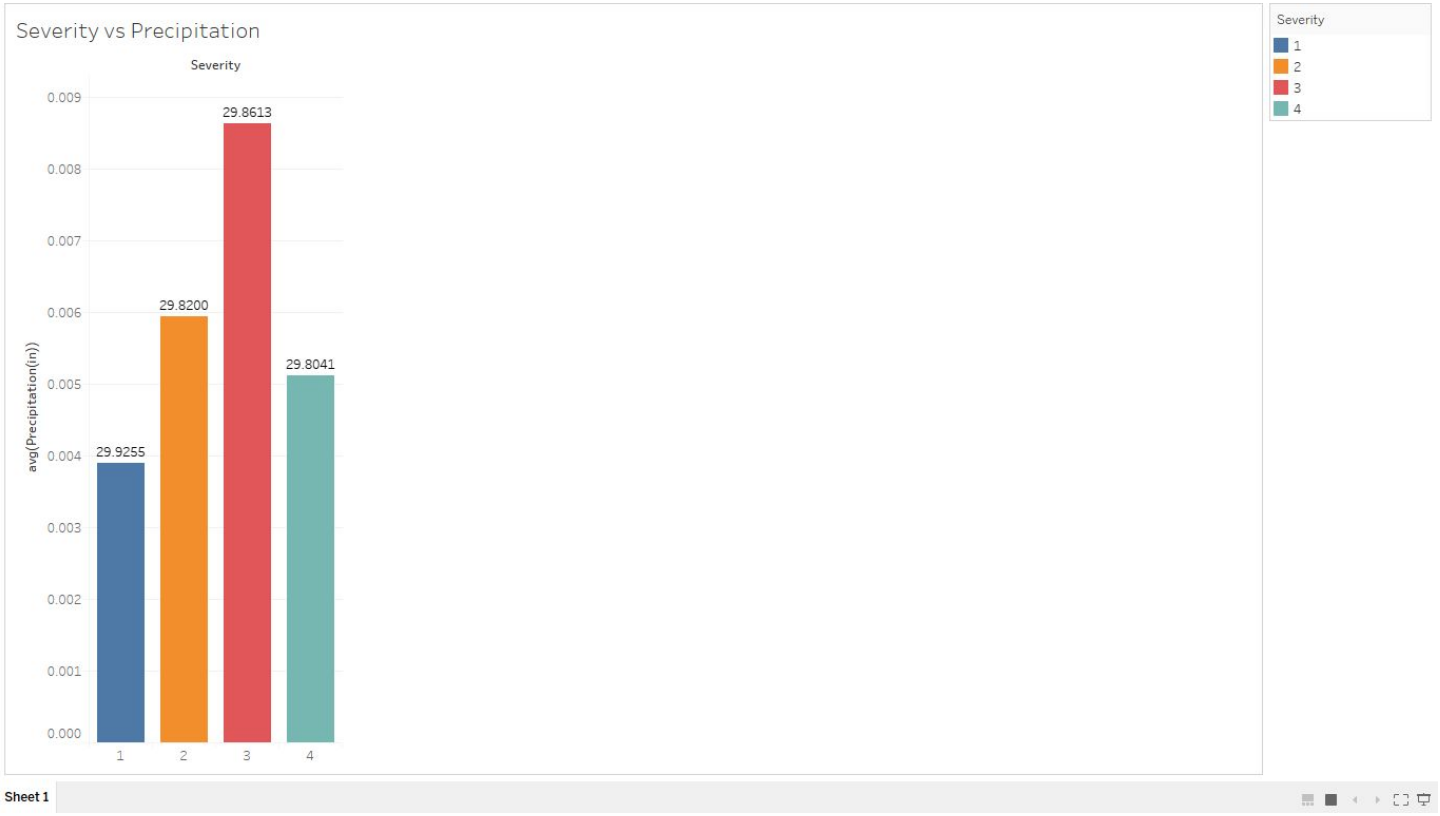



Severity vs Humidity
Severity
64.305
63.878
64.393
66.209
avg(Humidity(%))
Severity
1
2
3
4
Sheet 1


Severity vs Pressure
Severity
29.9255
29.8200
29.8613
29.8041
avg(Pressure(in))
Severity
1
2
3
4
Sheet 1


Severity vs Visibility
Severity
9.0866
8.9794
8.8908
8.8401
avg(Visibility(mi))
0
1
2
3
4
5
6
7
8
9
1
2
3
4
Severity
1
2
3
4
Sheet 1

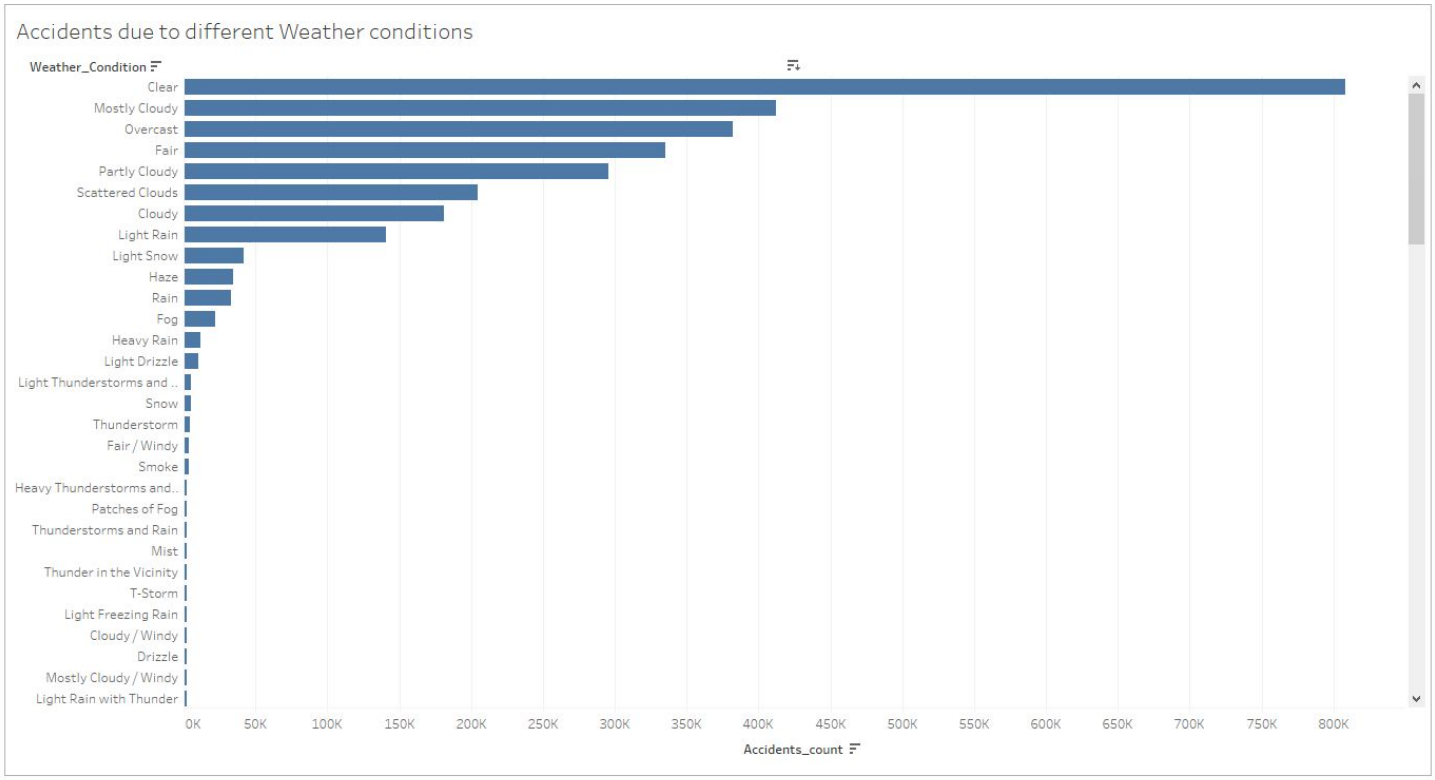

Severity vs Wind Speed
Severity
6.9503
7.0128
7.2024
6.9748
avg(Wind_Speed(mph))
0
1
2
3
4
5
6
7
1
2
3
4
Severity
1
2
3
4
Sheet 1

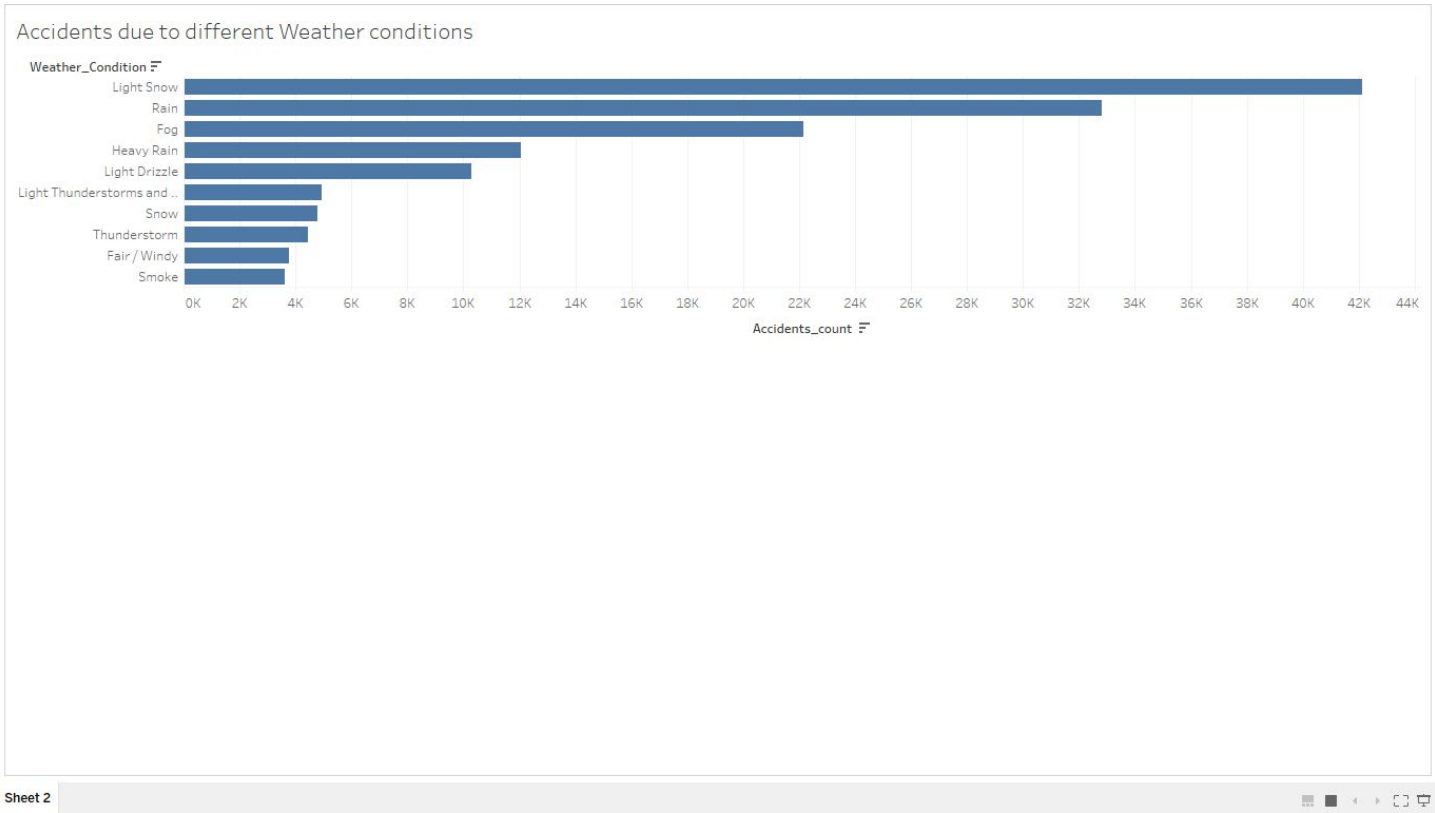

Severity vs Precipitation
Severity
avg(Precipitation(in))
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
29.9255
29.8200
29.8613
29.8041
1
2
3
4
Severity
1
2
3
4
Sheet 1

Average time of accidents for each severity -

Accidents with severity 4 approximately cause a time delay of 15 minutes on an average.

Weather conditions that have impacted accidents-





Prediction Model

The objective of the prediction model is to predict the value of 'Severity' which is used as a measure to show the delay the accident has/will cause. This process has been done manually and its definition varies from person to person who sets this value. Our model intends to do two things, firstly, to accurately predict the value of Severity so as to give the best idea of the delay caused by the accident. Secondly, to predict the value of Severity in real time as to indicate how much time the accident that just happened may cause. Severity is a measure having values 1-4 with 1 being least severe (duration of delay caused by accident is low) and 4 being most severe (duration of delay caused by accident is high).

To choose a perfect Machine Learning Classification model is always a tough job but vital. We chose to use 26 original and 5 derived attributes as these were the most vital and most likely to directly affect the duration of accident, which is denoted by Severity. Out of the total 31 attributes, 21 were categorical values. That is the reason we decided to use the Random Forest algorithm as it is efficient while dealing with categorical values.

As the dataset has approx. 3 million rows, we decided to train our model on a sample dataset. For the sample dataset, we used California state data which consisted of almost 0.6 million rows. After training the model on California state, we applied the model to test on the rest of the data set and received an accuracy of 75% in predicting the Severity.

Model Description in brief

- Features used: 31
- Trees:12
- Max depth of each tree: 16
- Train, Test data split: 70:30
- Feature Importance:
  - State: 17%
  - Side: 16%
  - Traffic Signal: 12%
  - Source: 12%
  - Distance: 8%
- **Accuracy: 75%**

Conclusion and Future Work

Traffic accidents cause a lot of chaos. It not only causes trouble only to the people who are involved in the accident, it also affects people who are struck in traffic. Most of the accident analysis is done on a smaller data set which is often not very reliable. Hence we have used a dataset which has over 2 million entries. A Prediction model to predict severity is also built.

This can be helpful if we can build an application that acts as maps and also informs the user about the possible severity and the time delay that may be caused if an accident occurs in their route.

References:

1. Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.
2. Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "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, ACM, 2019.
3. https://wiki.openstreetmap.org/wiki/TMC/Event_Code_List
4. https://www.asirt.org/safe-travel/road-safety-facts/