**Data Analysis – Traffic Accidents**

**INTRODUCTION**

As a part of the course work for Data Analysis, I have selected a dataset on US Accident, obtained from the website <https://smoosavi.org/datasets/us_accidents>.

This data set on US Accidents can be used for numerous applications such as real-time accident prediction, studying accident hotspot locations, casualty analysis and extracting cause and effect rules to predict accidents, or studying the impact of precipitation or other environmental stimuli on accident occurrence.

The purpose of this analysis is to analyze the data and find out what are the key variables that impact the severity of the traffic accidents that happened in US and ultimately predict the severity of the accidents based on given variables through data visualization and regression analysis using Python.

**DATASET**

 In this analysis I used data that is collected from February 2016 to June 2020 for the [Contiguous United States](https://en.wikipedia.org/wiki/Contiguous_United_States). In this dataset, we have about **3.5 million** traffic accidents. The dataset covers 49 states of the US. This dataset consists of 49 attributes. As part of the analysis some attributes have been excluded in data cleansing process.

**METHODOLOGY USED FOR ANALYSIS:**

## Step 1: Data cleaning:

Created a subset with variables using for this project. Creating factors of variable and removing the unknown and null values.

In this step we will check the dataset: preview data and some summary statistics; check columns for NaN (data missing); remove unnecessary columns that will not be required for analysis.

Necessary manipulations of dataset will be performed to lay the path for Data Analysis.

## Step 2: Data Visualization:

In this step the cleaned Dataset will used for various Data Visualizations which will guide our analysis to conclusions.

Data will be grouped by Severity. The codes for Severity in the dataset are as follows:

|  |  |
| --- | --- |
| Severity | Description |
| 1 | indicates the least impact on traffic  (i.e., short delay as a result of the accident) |
| 2 | indicates the moderate impact on traffic  (i.e., short delay as a result of the accident) |
| 3 | Indicates a moderate impact on traffic (i.e., moderate delay as a result of the accident). |
| 4 | Indicates a significant impact on traffic (i.e., long delay as a result of the accident). |

A number of Data Visualizations are performed:

* Plot data to visualize year over year traffic accident trend
* Plot the data by **Severity**and by**Year** to show the trend of accidents by year and by accident severity
* Plot the data by weather related variables have impact on the severity of the accidents
  + Temperature Vs Severity
  + **Humidity** Vs **Severity**
  + Wind Chill Vs Severity
  + By time of Sunrise & Sunset
  + By Months of Year
  + By the length of the road extent affected by the accident.
  + Using the heat map to know for each month, which day of the month are more likely to have more accidents.

## Step 3: Regression analysis:

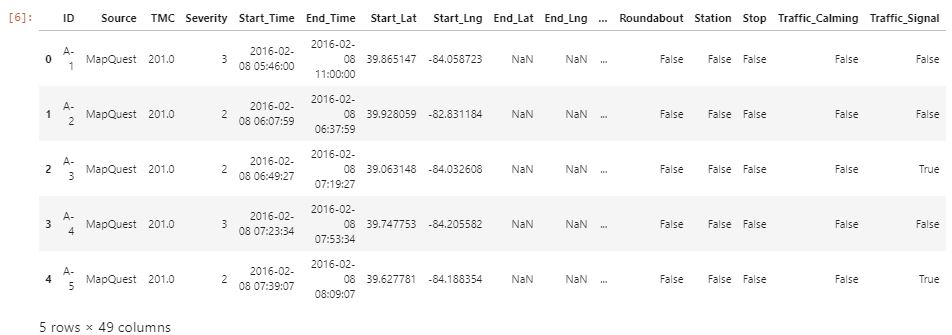
Selecting best subset of variables to perform regression analysis by using regularized Logistic regression method since the dependent variable is Severity, which is a categorical variable.

The independent variables are selected based on the previous analysis. Split data into two parts train and test. Finally fit the Logistic regression model using train data set and test this model with test dataset and finally conclusion. Finally use Confusion Matrix to see how many of the prediction are correct and how many are incorrect.

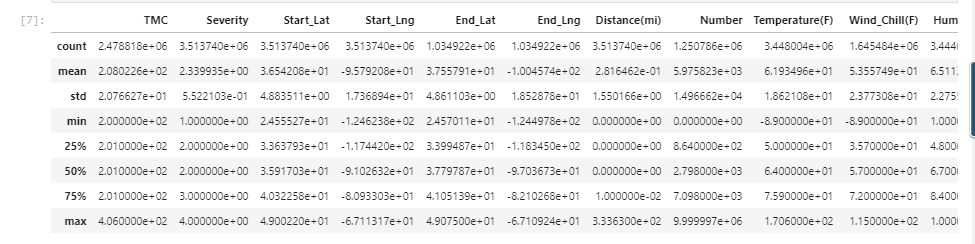
DATA ANALYSIS

**STEP 1: DATA CLEANSING**

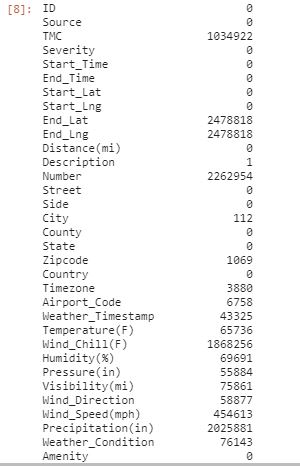
The process starts with loading datasets. Following which data is previewed as shown in the figure below.

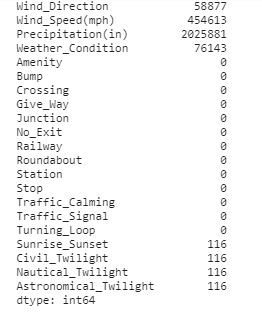


Next, we run the summary statistics to better understand the dataset.

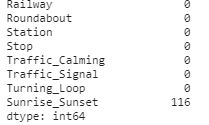
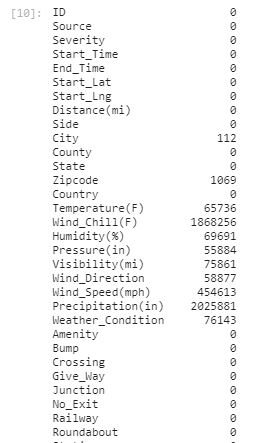


We then search the completeness of data and for any missing fields. The below figure illustrates there are considerable missing information which needs to be excluded from the analysis.





The data is manipulated to exclude the missing information and we get a more refined data set which is shown below

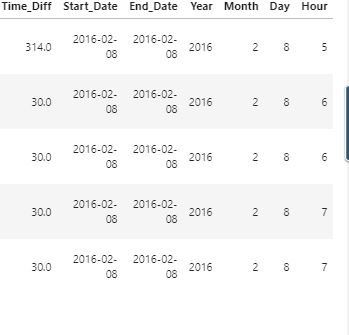
 

Some additional columns are added for preparing the dataset for further analysis.

* + Time\_Diff: Time difference between start time and end time of the accident
  + Year: Year of start time
  + Month: Month of start time
  + Day: Day of start time
  + Hour: Hour of start time

The new dataset after the data cleansing process is shown below:





**STEP 2: DATA VISUALISATION**

With cleansed dataset obtained it is time for data analysis and visualization.

The purpose of this analysis:

* + To analyze the data and find out what are the key variables that impact the severity of the traffic accidents that happened in US
  + To predict the severity of the accidents based on given variables through data visualization and regression analysis using Python.

In this step the cleaned Dataset will used for various Data Visualizations which will guide our analysis to conclusions.

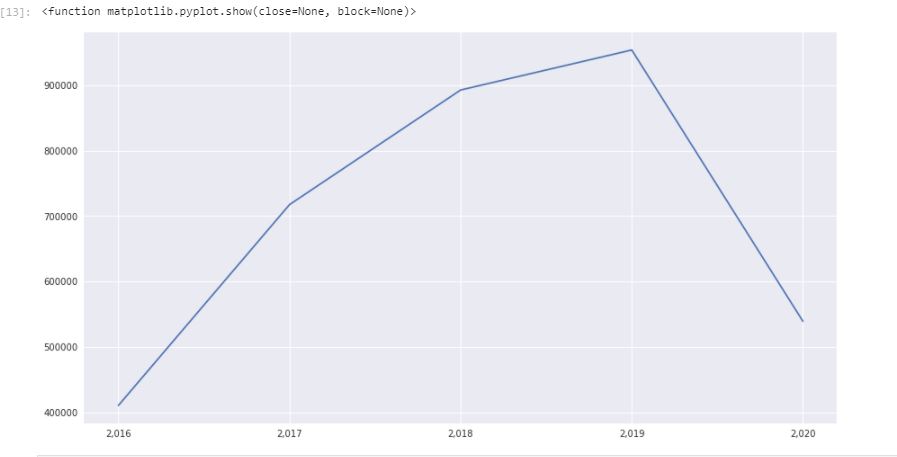
Data will be grouped by Severity. The codes for Severity in the dataset are as follows:

|  |  |
| --- | --- |
| Severity | Description |
| 1 | indicates the least impact on traffic  (i.e., short delay as a result of the accident) |
| 2 | indicates the moderate impact on traffic  (i.e., short delay as a result of the accident) |
| 3 | Indicates a moderate impact on traffic (i.e., moderate delay as a result of the accident). |
| 4 | Indicates a significant impact on traffic (i.e., long delay as a result of the accident). |

A number of Data Visualizations are performed:

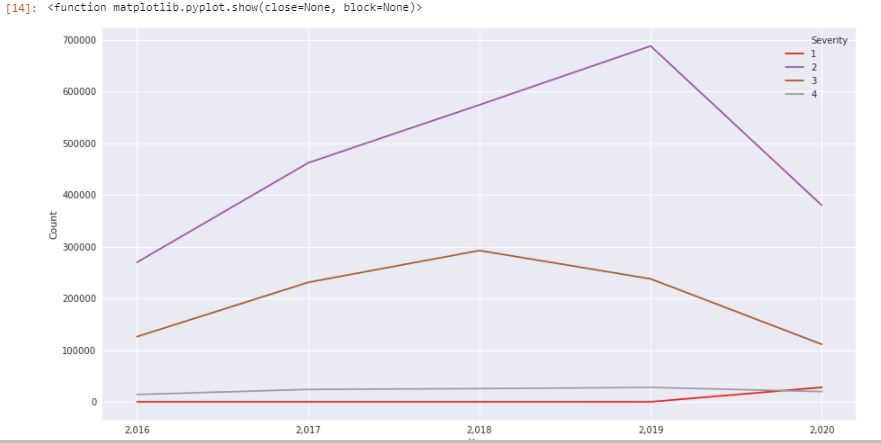
* Plot data to visualize year over year traffic accident trend.
* Plot the data by Severityand byYear to show the trend of accidents by year and by accident severity
* Plot the data by weather related variables have impact on the severity of the accidents
  + Temperature Vs Severity
  + Humidity Vs Severity
  + Wind Chill Vs Severity
  + By time of Sunrise & Sunset
  + By Months of Year
  + By the length of the road extent affected by the accident.
  + Using the heat map to know for each month, which day of the month are more likely to have more accidents.

***Plot data to visualize year over year traffic accident trend.***



From the plot it can be learned that the severity increases over the year from 2016 to 2019 and then shows a sudden decrease in 2020.

***Plot the data by Severity and by Year to show the trend of accidents by year and by accident severity***



From the plot, we can see that:

* + Severity 1 is relatively flat till 2019 and then shows an increase
  + Severity 4 is relatively flat year over year.
  + Severity 2 shows accidents are increasing year over year in a rapid speed and decrease in 2020 compared to the previous years. .
  + Severity 3 accidents have seen a decrease in 2019 and 2020 compared to 2018.

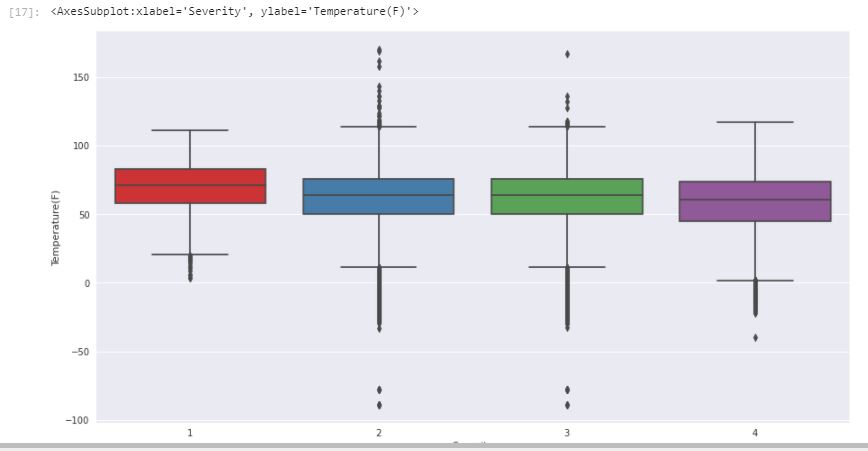
The following bar plots are also showing the similar information then observed previously





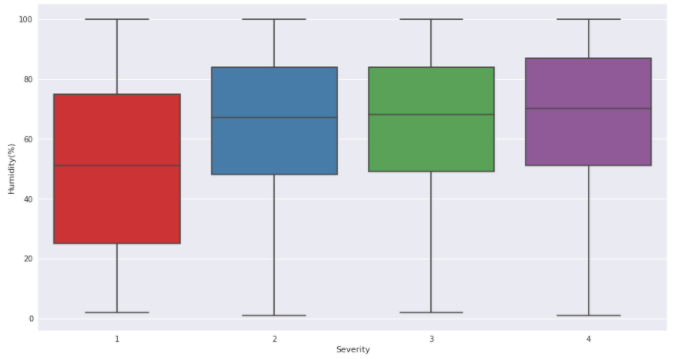
**Plot the data by weather related variables have impact on the severity of the accidents**

1. **Temperature v/s Severity**

****

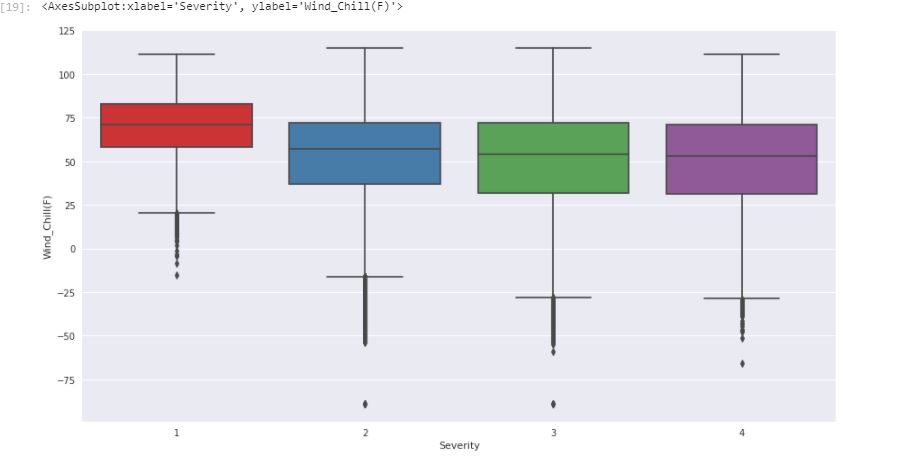
The above box plot shows that the median temperature remains approximately same for Severity 2, 3 & 4, but the median temperature is high for Severity 1 and lower for Severity 4. This indicates that the Severity increases with decrease in temperature.

1. **Humidity v/s Severity**



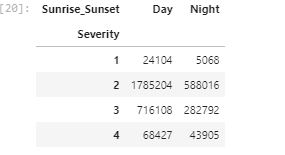
The above Box Chart indicates an impact of humidity on Severity. The higher the humidity, the higher the Severity.

1. **Wind chill v/s Severity**



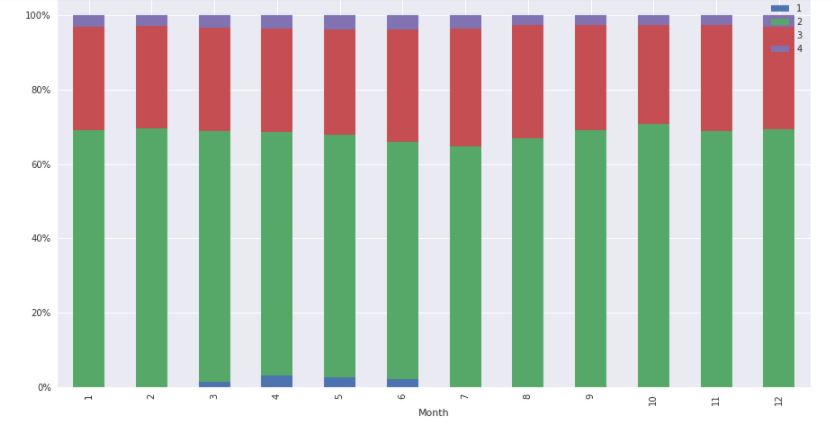
The above Box chart also indicates the impact of humidity on Severity. The lower the Wind Chill, the higher the Severity

1. **By time of Sunrise & Sunset**

****

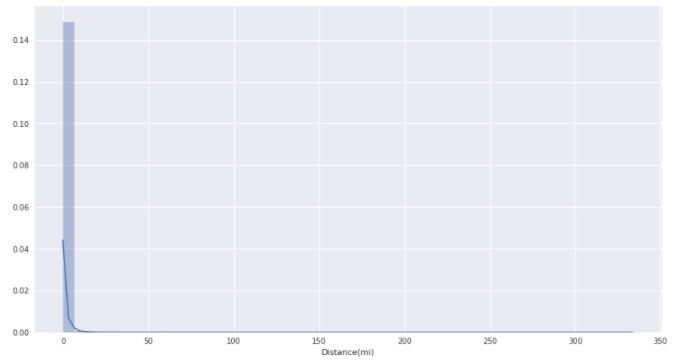
The above table summarizes the Severity for accidents during Day & Night. It appears that majority of the accidents with Severity occur during day.

1. **By Months of Year**

****

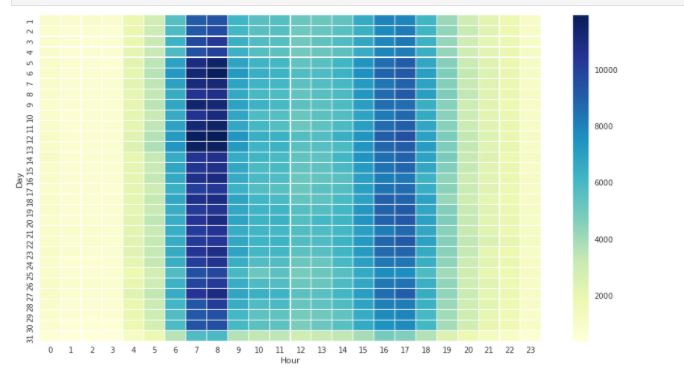
The above chart indicates that the accidents are moreover uniform during the year but the High Severity is experienced between March & June.

1. **By the length of the road extent affected by the accident.**

****

The above plot indicates that more severe accidents will affect longer distances and last longer time.

1. **Using the heat map to know for each month, which day of the month are more likely to have more accidents.**



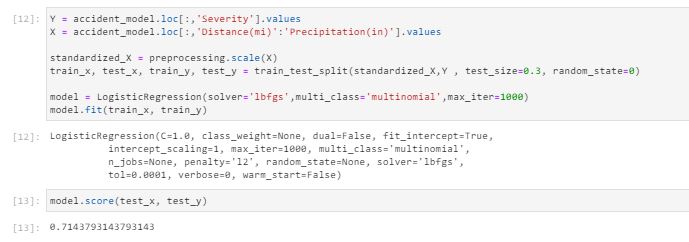
The Heat map indicates that most accidents happened between 7 and 8 hours of the day, which is the morning rush hour. Morning rush hour have much more accidents then the afternoon rush hour, which is 4 to 6 in the afternoon

**STEP 3: REGRESSION ANALYSIS:**

The independent variables are selected based on the previous analysis. Split data into two parts train and test. Finally fit the Logistic regression model using train data set and test this model with test dataset and finally conclusion. Finally use Confusion Matrix to see how many of the prediction are correct and how many are incorrect.

The following independent variables are selected based on the previous analysis:

* + Distance(mi)
  + Time\_Diff
  + Temperature(F)
  + Wind\_Chill(F)
  + Humidity(%)
  + Pressure(in)
  + Visibility(mi)
  + Wind\_Speed(mph)
  + Precipitation(in)



We use score() to evaluate the model, the model gets 0.72, which is decent for the first try.



We can see that Severity 2 and 3 are the most accidents. The model did a good job in identifying Severity 2 but not Severity 3.

Conclusion:

A significant amount of Data Analysis has been performed on various attributes of the Dataset and we have gain prime information on the traffic accident that have occurred since 2015 till date. It is now evident that there is a significant relationship between weather conditions and Severity of accidents. Also there exists a link between the period of the months and accident severity. There also exits a direct impact of visibility (Daylight) on the Severity of accidents. The heat map showed a relationship between rush hours on Severity of accidents. All these information must be used to better plan the developments in infrastructure such as roads and highways thereby reduce road accidents.