PYTHON PROJECT

DATASET: ROAD ACCIDENTS-2019

AGENDA:

* Executive Summary
* Project Motivation/Background
* Data Description
* Data Transformation/Exploratory data analysis
* Models and Analysis
* Findings and Managerial Implications
* Conclusions
* Appendix: Python codes with proper documentations
* References

Executive Summary:

* Exploring the US accidents data for the year 2019 and try to validate key objective as per the Dataset.
* United States is one of the biggest busiest countries in terms of road traffic with nearly 280 million vehicles in operation and more than 227.5 million drivers holding a valid driving license. The level of traffic is one of the reasons leading to more traffic accidents from 2016 till Dec 2002, there were some 12 million vehicles involved in crashes in the United States.
* Representing data based on source, severity, time, latitude, longitude, distance, location, weather condition etc... so that we can get the clear understanding about the US road accidents by using the following data set.
* Usage of the python programming language and importing the issues causing by road accidents.
* US-Accident’s data set can be used for numerous applications such as real-time car accident prediction, studying car accidents hotspot locations, casualty analysis and extracting cause and effect rules to predict car accidents, and studying the impact of precipitation or other environmental stimuli on accident occurrence. The dataset which we consider can also be useful to study the impact of traffic behavior and accidents.

Project Motivation/Background:

* 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.
* We have considered the dataset for road accidents for the year 2019. In the dataset, we have done processing, cleaning, preparing, exploring the data, and have done some required predictions.
* The following are the key issues we figured out with this project:

**DATA DESCRIPTION:**

* This is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is collected for the year 2019, which covers various aspects like temperature, wind, time at which accident occurred, location, time zone etc. Each record is represented with a unique ID.
* The dataset contains 50,367 records of the road accidents and 50 columns.

**Understanding Source Dataset: -**

* **Record/source API identifiers:**

ID, Source, TMC

* **Accident properties:**

Severity, Start\_Time, End\_Time, Start\_Lat, Start\_Lng, End\_Lat, End\_Lng, Distance(mi)

* **Location properties:**

Description, Number, Street, Side, City, County, State, Zip code, Country, Time zone, Airport\_Code

* **Weather Condition Properties:**

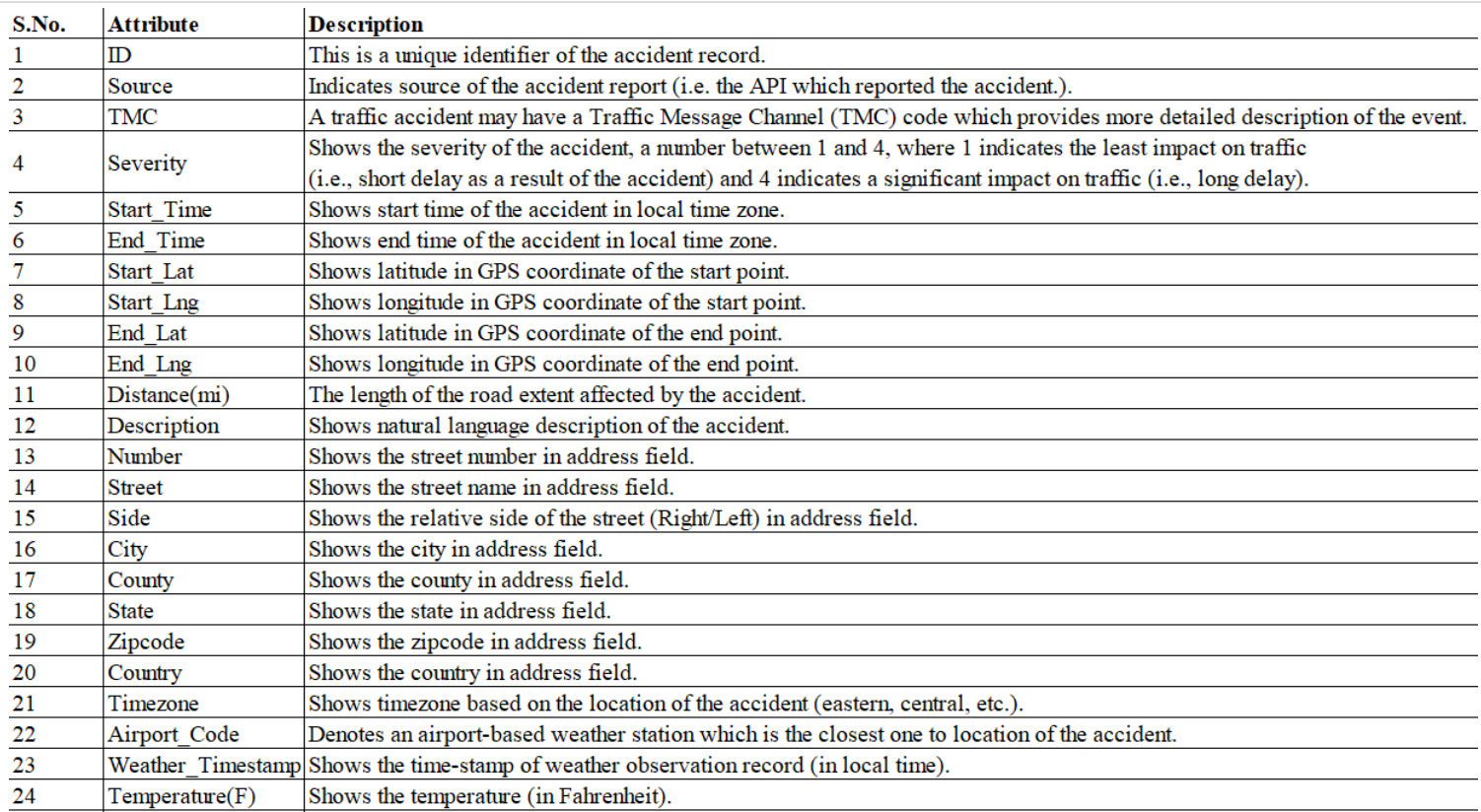
Weather\_Timestamp, Temperature(F), Wind\_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind\_Direction, Wind\_Speed(mph), Precipitation(in), Weather\_Condition, Sunrise\_Sunset, Civil\_Twilight, Nautical\_Twilight, Astronomical\_Twilight

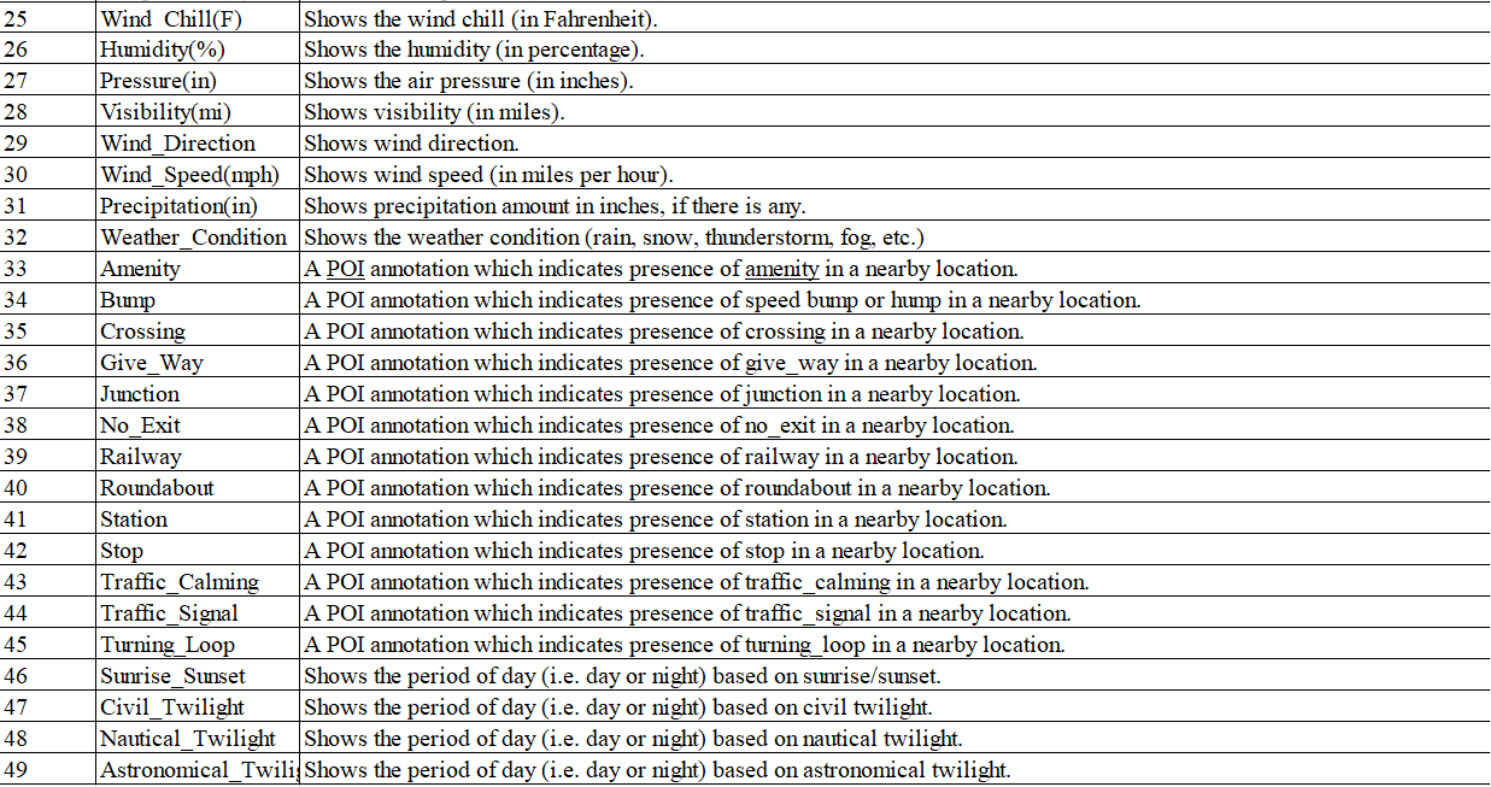
* **Nearby landmark properties:**

Amenity, Bump, Crossing, Give\_Way, Junction, No\_Exit, Railway, Roundabout, Station, Stop, Traffic\_Calming, Traffic\_Signal, Turning\_Loop,

**FEATURE DESCRIPTION OF THE DATASET:**

Below are the detailed explanations of the columns we used in the dataset:

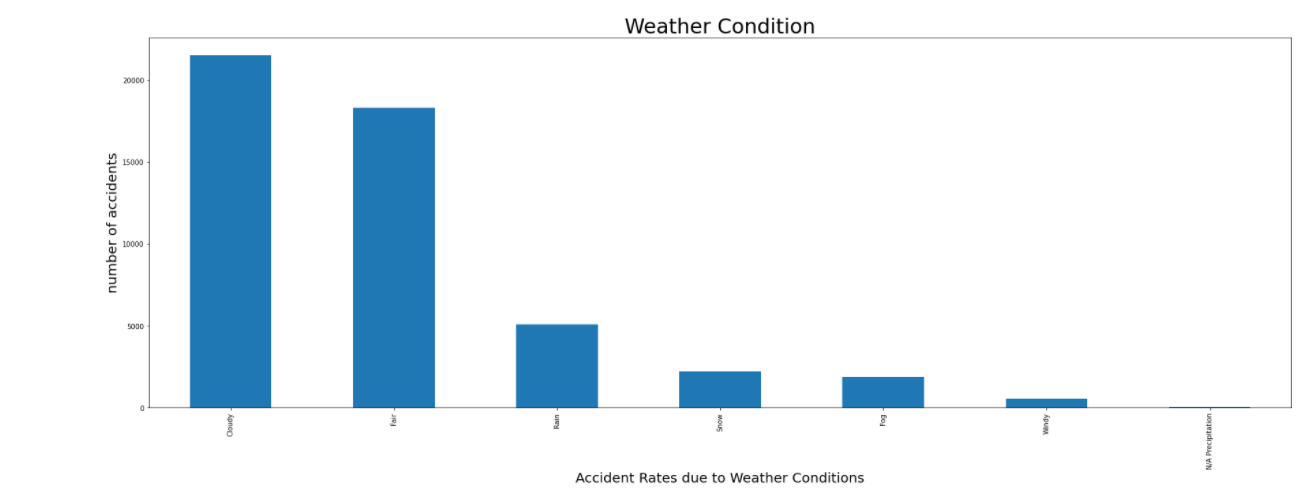
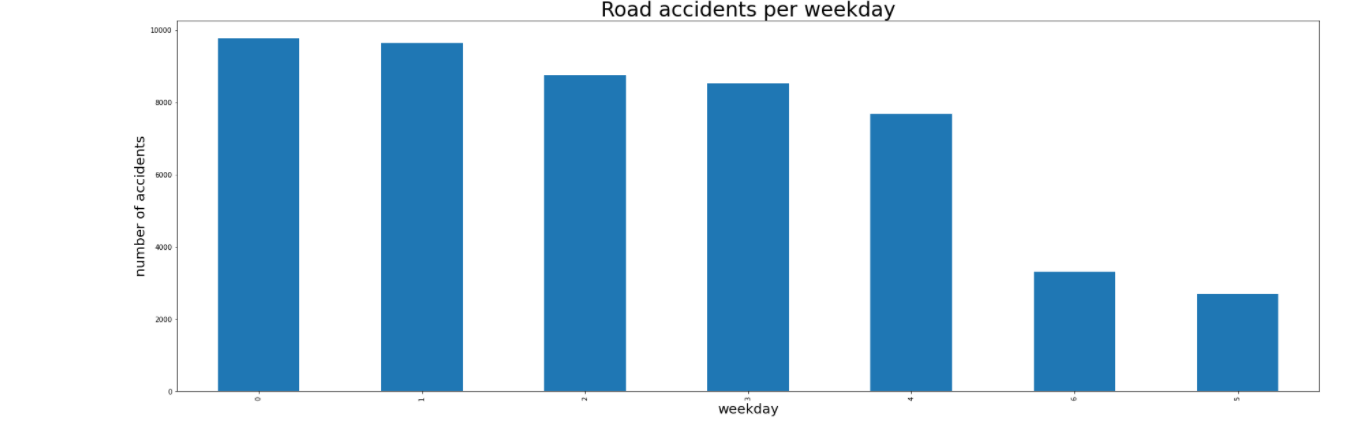


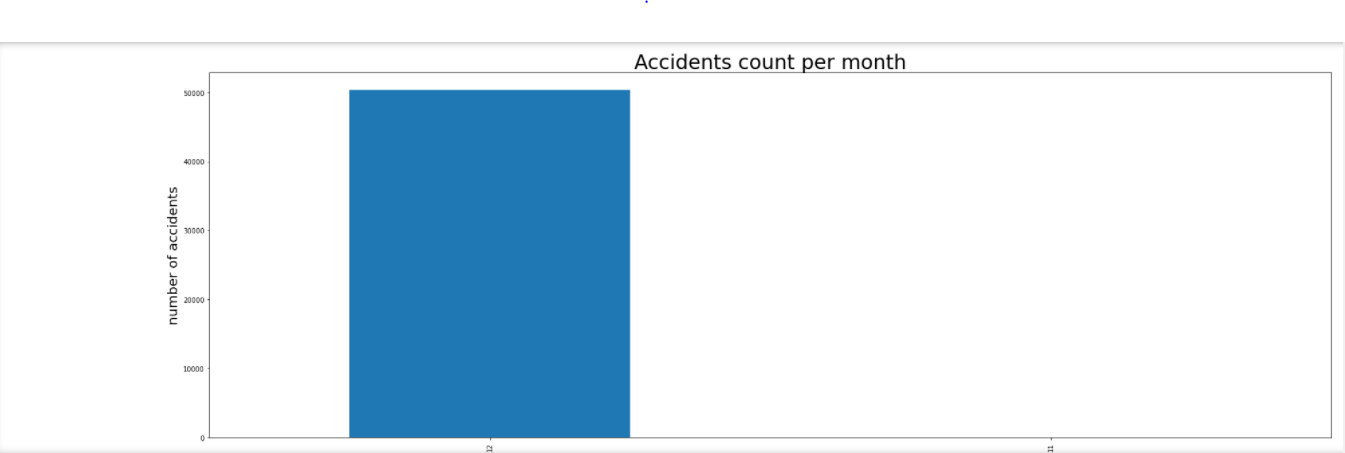


DATA TRANSFORMATION/EXPLORATORY DATA ANALYSIS:

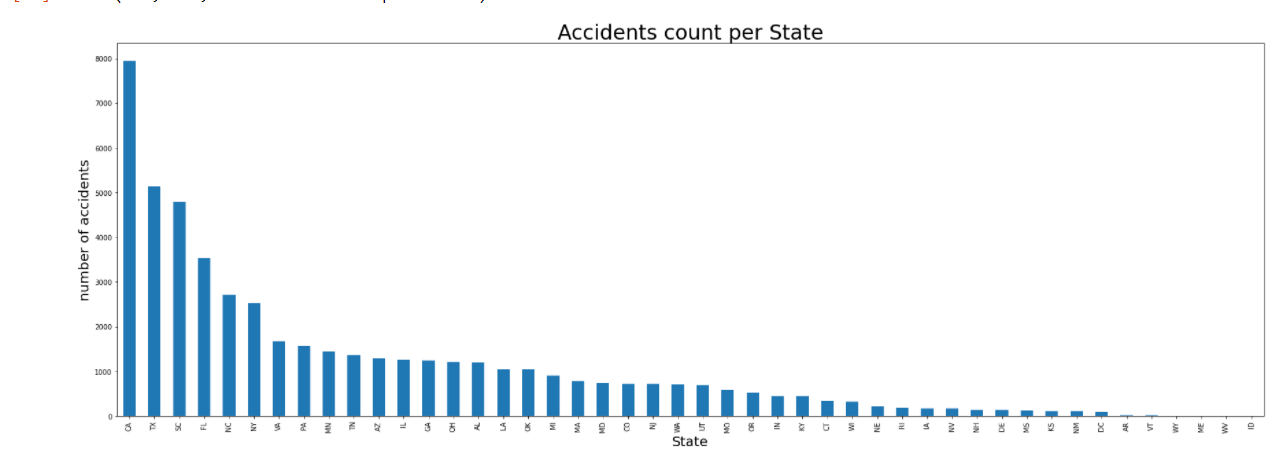
we have done the data processing, data cleaning, data exploration, data validation as per the following:

* import libraries and dataset.
* Get the basic information about the dataset: shape and info () and column.
* What are the data types?
* Cast Start Time to datetime, Extract year, month, weekday and day, Extract hour and minute.
* Source: reduce the number of sources by summarizing them and turn it into dummies.
* Time zone: split and turn them into dummy variables.
* “Start\_Lat”, “End\_Lng”, “Number”, "Severity": Replace the missing of Number with mean.
* “Wind\_speed”, “visibility”, “pressure”, “humidity”,” Wind\_direction:, “wind\_chill”, “Temperature”, “weather\_timestamp”: Replace the missing data with NaN value
* drop TMC.
* Group by weather condition and in which weather conditions does most accidents occur.
* Bar graph weather condition
* Application

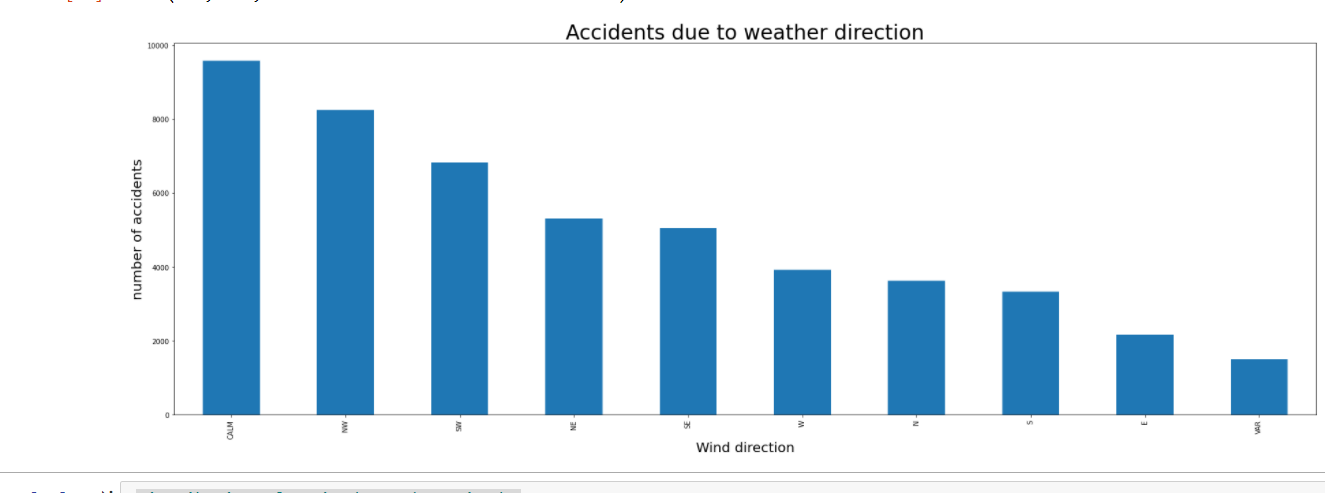
  Description automatically generated with medium confidence
* replace weather conditions in simple terms with bar graph.
* 
* Analyze Accident Number Per Day of the Week.
* Road accidents per weekday - bar plot.
* 
* Which day dose accidents occur the most?
* Accidents count per month.



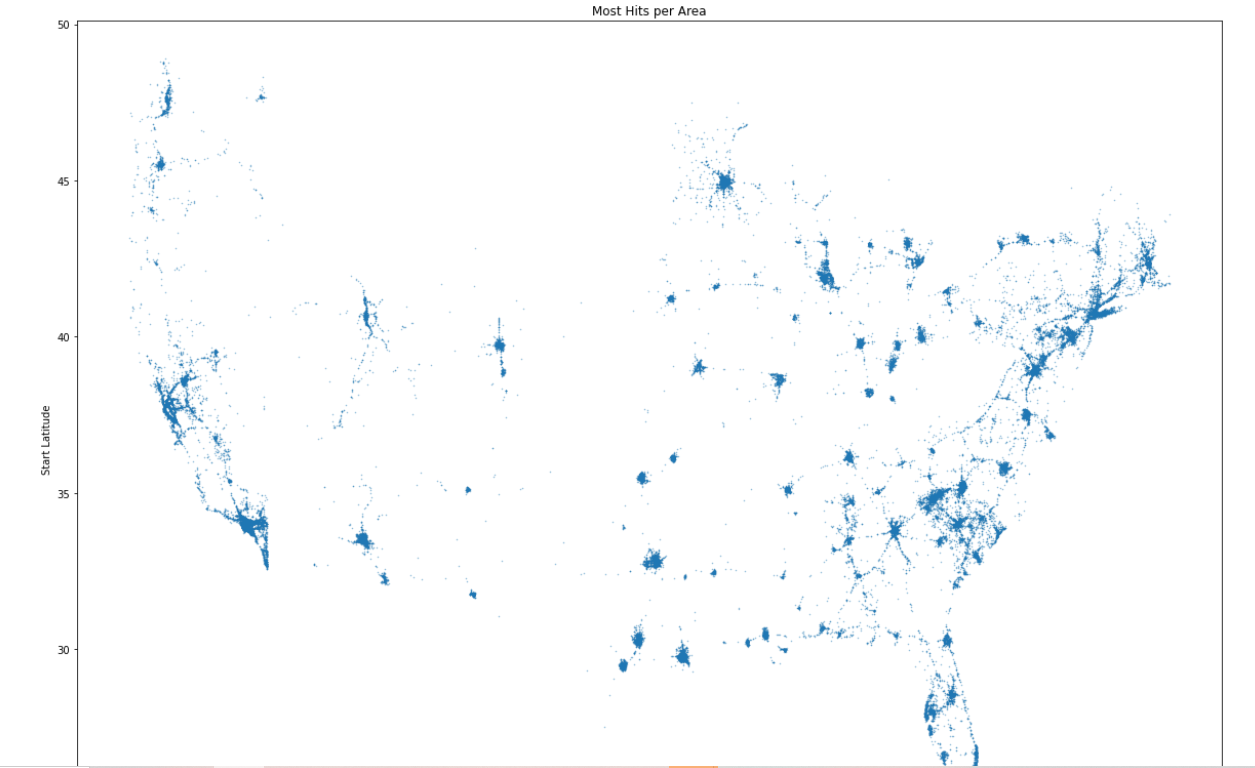
* Analyze Accidents count for particular state with a bar graph.



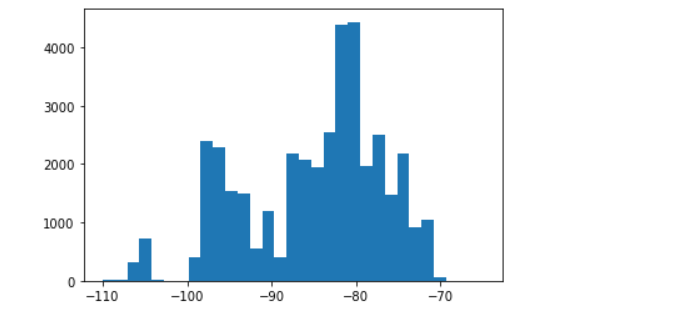
* Replacing the wind directions in simpler terms.

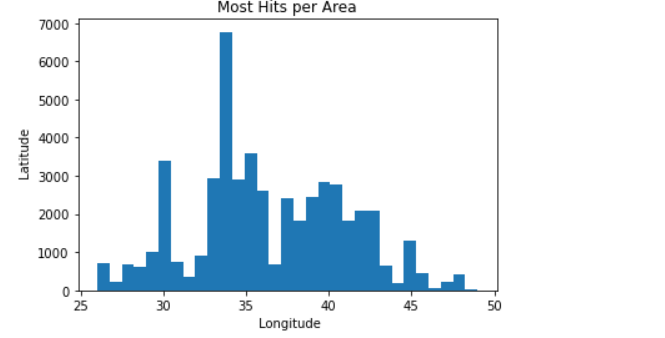


* plotting the Lat against Long could show the map of the area

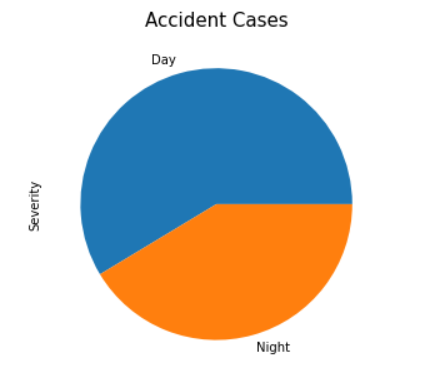


* Distribution of Latitudes and Longitudes.

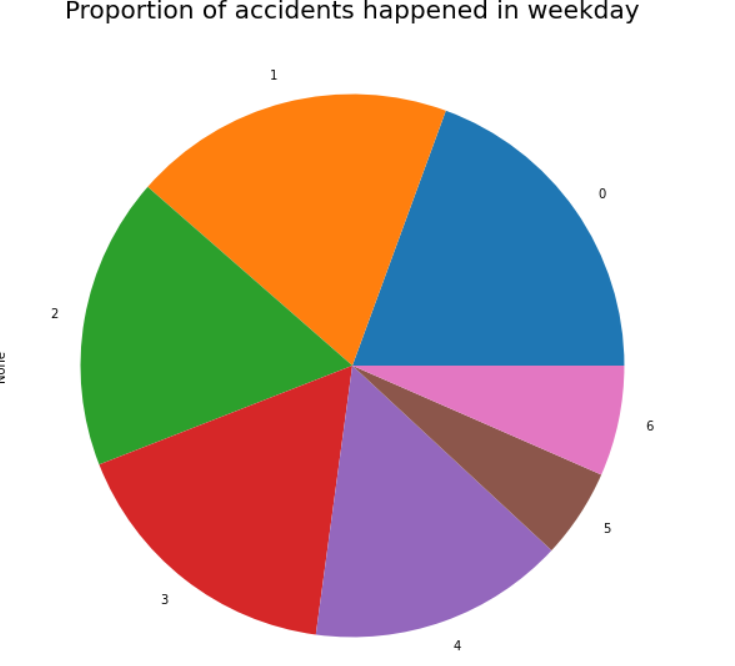




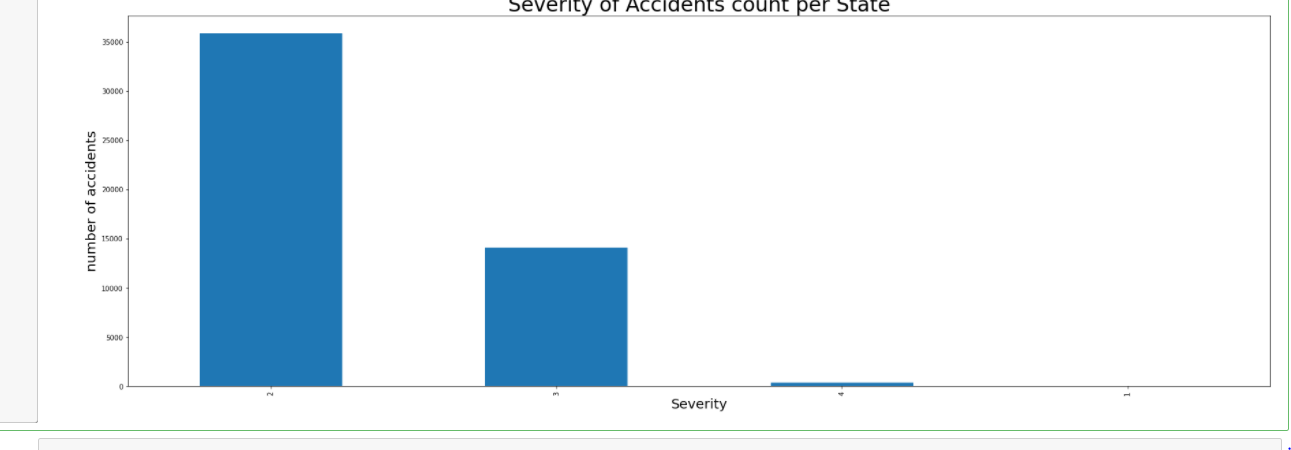
* Which State has highest accident cases is seen above.

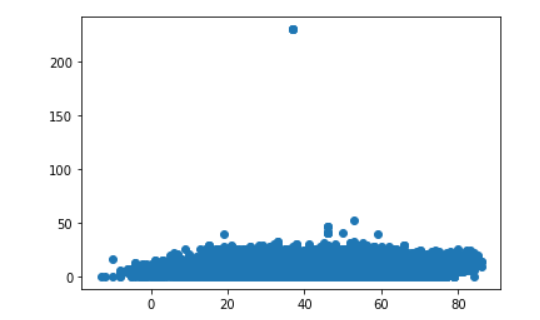


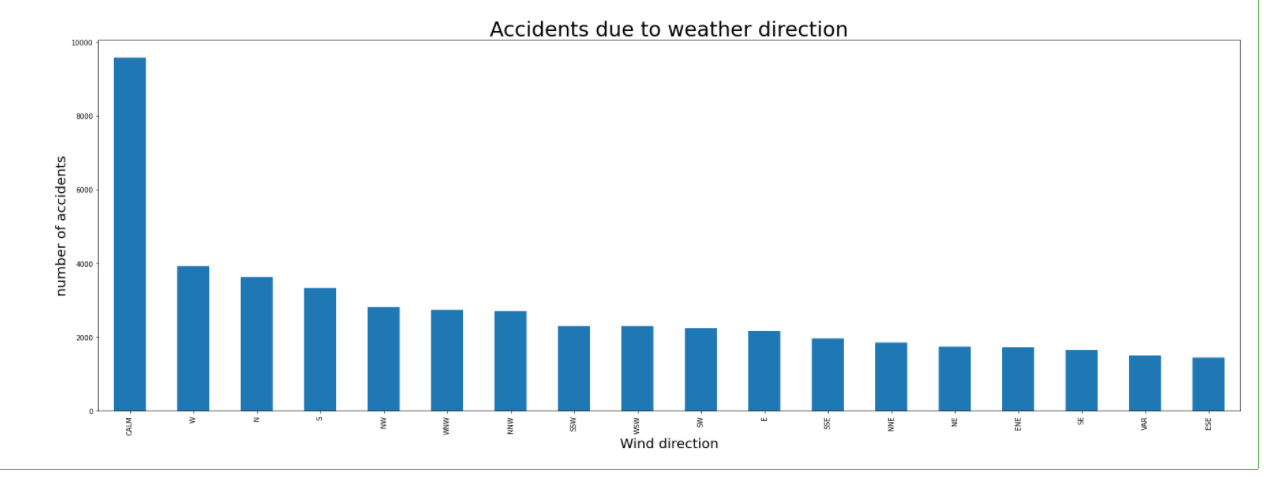
* Creating Dummies for Columns :('Amenity','Bump','Crossing','Give\_Way','Junction','No\_Exit','Railway','Roundabout','Station','Stop','Traffic\_Calming','Traffic\_Signal','Turning\_Loop','Civil\_Twilight','Nautical\_Twilight','Astronomical\_Twilight','Weather\_Condition','Wind\_Direction','Weekday')
* Accidents happened in weekday with a pie diagram.



* Factors Affecting Accident Severity
* Examine the relationship between accident severity and other accident information such as time, weather, and location.



* scatter plot: wind\_speed vs temperature. 
* Analyze Accident Number due to Weather Direction.



* Dropping Columns.
* generate the independent variable and dependent variable.
* Linear Regression.
* Decision Tree Model.
* KNN Model.
* Model Analysis.

MODELS AND ANALYSIS:

We have done the following models:

* Linear Regression Model

Accuracy score for the logistic regression model is: 0.7534965034965035.

* Decision Tree Model

Accuracy score for the Decision tree model is: 0.756993006993007.

* KNN Model

Accuracy score for KNN model is: 0.7086713286713286.

Findings and Managerial Implications:

* We must explore different algorithms as a team and share our understanding among us for better knowledge.
* We must explore about the modules in python and use the concepts of classification and Regression.
* We must perform Prediction analysis on different reasons for a road accident to occur.

Conclusions:

We have done different model analysis and figured out that, Decision Tree has better accuracy rate, and the accuracy is: 0.756993006993007.

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

Appendix:

Below is the code for python and dataset which we preferred.

We have selected few records from the dataset collected as per our convenience.

References:

https://www.kaggle.com/luisabrego/usaccidents2019