**SEVERITY PREDICTION**

**IN**

**US ACCIDENTS DATASET**

**Author Name: SUDHIR SAHARAN**

**TABLE OF CONTENTS**

1. **ABSTRACT3**
2. **INTRODUCTION4**
   1. **DATA 5**
   2. **OBJECTIVES AND IMPORTANCE/ BUSINESS PROBLEM**
3. **MATERIALS AND METHODS6**
   1. **DATA PRE-PROCESSING 6** 
      1. **HANDLING MISSING VALUES 6**
      2. **DATA FORMATTING 8**
      3. **CORRELATION ANALYSIS8**
      4. **SYSTEM ARCHITECTURE 9**
      5. **SOFTWARE AND HARDWARE PLATFORM 9**
4. **RESULTS 10** 
   1. **VISUALIZATIONS USING R 10**
   2. **VISUALIZATIONS USING PYTHON 12**
   3. **VISUALIZATIONS USING TABLEAU 14**
   4. **CLASSIFICATION 17**
      1. **LOGISTIC REGRESSION 17**
      2. **DECISION TREE CLASSIFIER 18**
      3. **RANDOM FOREST CLASSIFIER 18**
   5. **NATURAL LANGUAGE PROCESSING 20**
   6. **HYPOTHESIS TESTING 21**
   7. **SQL 22**
5. **LIMITATIONS & FUTURE RESEARCH 25**
   1. **LIMITATIONS 25**
   2. **FUTURE RESEARCH 25**
6. **CONCLUSION 26**
7. **REFERENCES 27**

**1. ABSTRACT**

Many deaths and injuries are caused every year due to traffic accidents. The main reason behind the traffic accidents is the traffic congestion and delay.Reduction of traffic accidents is one of the biggest public safety challenge.Previously, have done analysis and predictions on the accident data,but those were not much effective, may be due to limited coverage, either old or private, not involving any environmental factors etc. This dataset of US Accidents is about the accidents of 49 States from February 2016 to December 2019. It has 3 million records and 49 attributes.

The state California has the highest number of accidents, in my project I performed a detailed analysis on the records of California state, used various machine learning models and identified the model that best fits for the data and predicts the accident severity.

**2.INTRODUCTION**

**DATA**:

Moosavi, Sobhan . "US Accidents (2.25 million Records) - A Country Wide Traffic Accident Dataset (2016-2019) ", <https://www.kaggle.com/sobhanmoosavi/us-accidents>, Accessed on 29th March, 2020.

There have been a lot of studies which try to predict or provide analysis on accident data as they are the major source of death injuries and traffic congestions. The studies have not been very affective due to the limited data available.  
This data is about the car accidents that took place in the 49 states from February 2016 to December 2019. This data set has several data providers and API’s which broadcast traffic data captured by various entities. This data set contains around 3 million records.

**ATTRIBUTES:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| ID | Number | This is a unique identifier of the accident record |
| Source | Factor | Indicates the API which reported the accident |
| TMC | Number | Traffic Message Chanel, Provides a detailed information about the incident. |
| Severity | Factor | Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic and 4 indicates a significant impact on traffic. Based on initial analysis this will be the classification variable for the dataset. |
| Start Time | Timestamp | Start time of the accident in local time |
| End Time | Number | End time of the accident in local time zone |
| Start\_Lat | Number | Latitude in GPS coordinate at start time |
| Start\_Lng | Number | Longitude of GPS coordinate at Start time |
| End\_Lat | Number | Latitude of GPS coordinate at end time. |
| End\_Lng | Number | Longitude of GPS coordinate at end time |
| Distance | Number | The length of the road affected. |
| Description | Number | Natural Language description of the incident. |
| Zipcode | Number | Shows the Zip code in the address field. |
| Temperature(F) | Number | Shows the temperature |
| Wind\_chill(F) | Number | Shows the wind chill |
| Visibility | Number | Shows visibility |
| Wind\_speed | Number | Shows wind speed |
| Weather\_condition | Factor | Shows the weather condition (rain, snow, thunderstorm, fog, etc.). |
| Amenity | Boolean | A Point-Of-Interest (POI) annotation which indicates presence of amenity in a nearby location. |
| Bump | Boolean | A POI annotation which indicates presence of speed bump or hump in a nearby location. |
| Crossing | Boolean | A POI annotation which indicates presence of crossing in a nearby location. |
| GiveWay | Boolean | A POI annotation which indicates presence of giveaway sign in a nearby location. |
| Junction | Boolean | A POI annotation which indicates presence of junction in a nearby location. |
| No\_exist | Boolean | A POI annotation which indicates presence of no exit sign in a nearby location. |
| Railway | Boolean | A POI annotation which indicates presence of railway in a nearby location. |

**2.1 OBJECTIVES AND IMPORTANCE / BUSINESS PROBLEM**

This data set has been developed at Cornell University to perform accident analysis and prediction with a data size of 1GB,as the dataset is very large we will use a subset of the data which contains 70,000 records.  
The questions aimed to answer through analysis are :  
1)identifying the major cause of accidents so that people can be warned based on these predictions.  
2)Identifying a point of interest, that is in which areas accidents occur the most.  
3)identifying what are the reasons for high severity accidents.  
4) Identifying during which months and days accidents are frequent.  
5) Identifying accident hotspot locations.  
6) Identifying the threshold values of temperature, wind or visibility beyond which accidents are occurring.  
7) Predicting the severity, POI of the accidents using Machine Learning technique and comparing their accuracies.

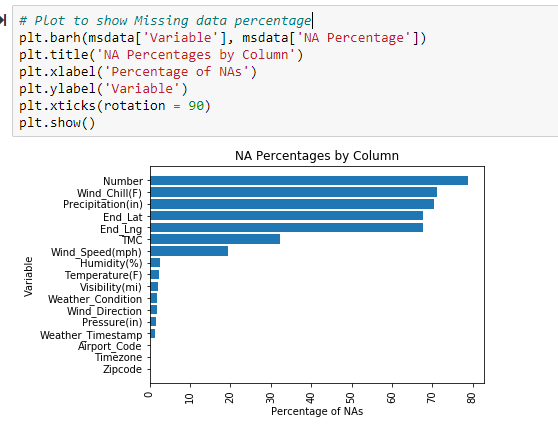
**3. MATERIALS AND METHODS**

**3.1 Data Preprocessing**

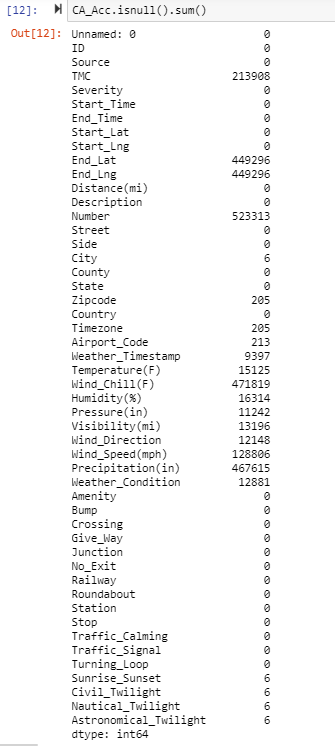
The raw data is noisy and is not in a proper form to set into the model. It has few errors and missing values.By using raw data which is inconsistent could lead to improper and not accurate visualizations and conclusions. By using data mining techniques, we convert the raw data into an understandable format that can be ready to fit the models. Performing preprocessing on raw data is a very important task to get accurate outcomes from the data.

**3.1.1 Handling Missing Values**

Handling of missing values is a major step in data pre-processing. If the data is not handled properly, we will end up with inaccurate conclusions of the data. There are many methods to handle missing data, like, dropping the records with missing data, replacing with a user-defined constant, replacing numerical data with mean or median and categorical data with mode. Dropping the records cannot be done when there is a large set of missing records, as it results in a huge data loss.[6]



**Fig1.** Percentage of Missing values in accidents dataset.



**Fig 2.** Represents total number of missing values for each attribute

From above figure, we observe that the attributes: Number, Wind\_Chill(F),Precipitation(in),End\_Lat and End\_Lng have more than 60% of data missing in them.

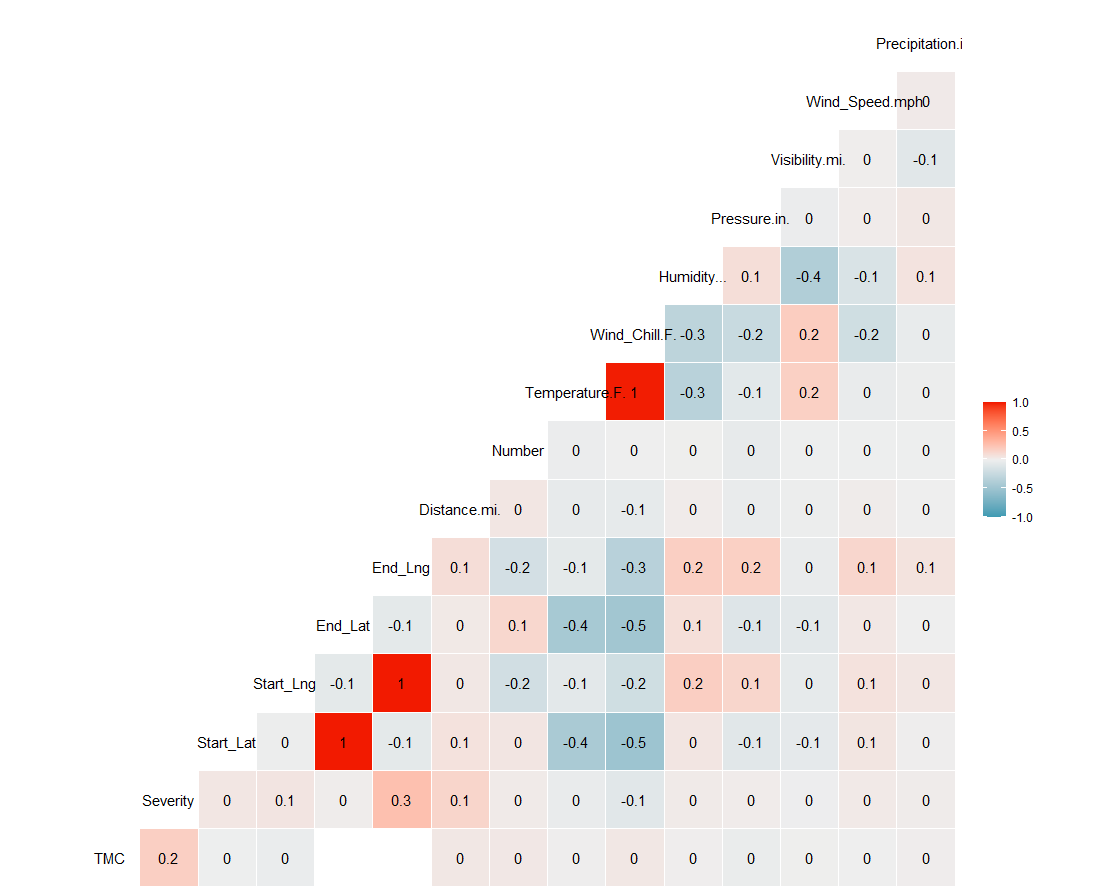
Numerical attributes like Number, End\_Lat, End\_Lng , TMC have been replaced by computing the median of the data.

Categorical attributes like Weather\_Condition, Weather\_Timestamp , City have been replaced by mode of the data.

**3.1.2 Data Formatting**

Converting the information to satisfy standard specifications is called data formatting. In the dataset only few records are in the format of zipcode-areacode. All such data is converted into the 5 number zipcode to make the data into a uniform format.

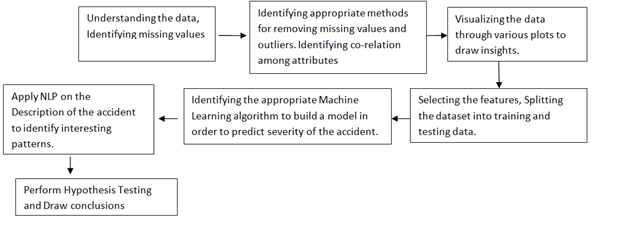
**3.1.3 Correlation Analysis**



**Fig 3.** Represents the Correlation Matrix

We observe that no attribute is highly correlated with the prediction variable, i.e, severity. Only location attributes and {Wind\_Chill.F. and Temperature.F.1} are positive high correlated attributes.

**3.1.4 System Architecture**



**Fig 4.** System Architecture

**3.1.5 Software and Hardware Platform**

**Software Requirements**

Operating System : Windows 7 or Higher

Software used :Jupyter Notebook, Tableau, R, PostgreSQL

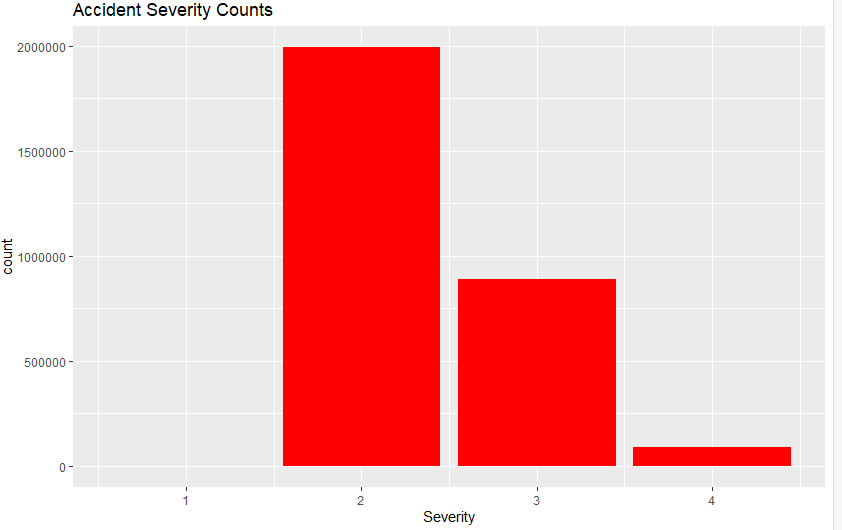
Programming Language : Python, R , SQL

Database : US Accidents

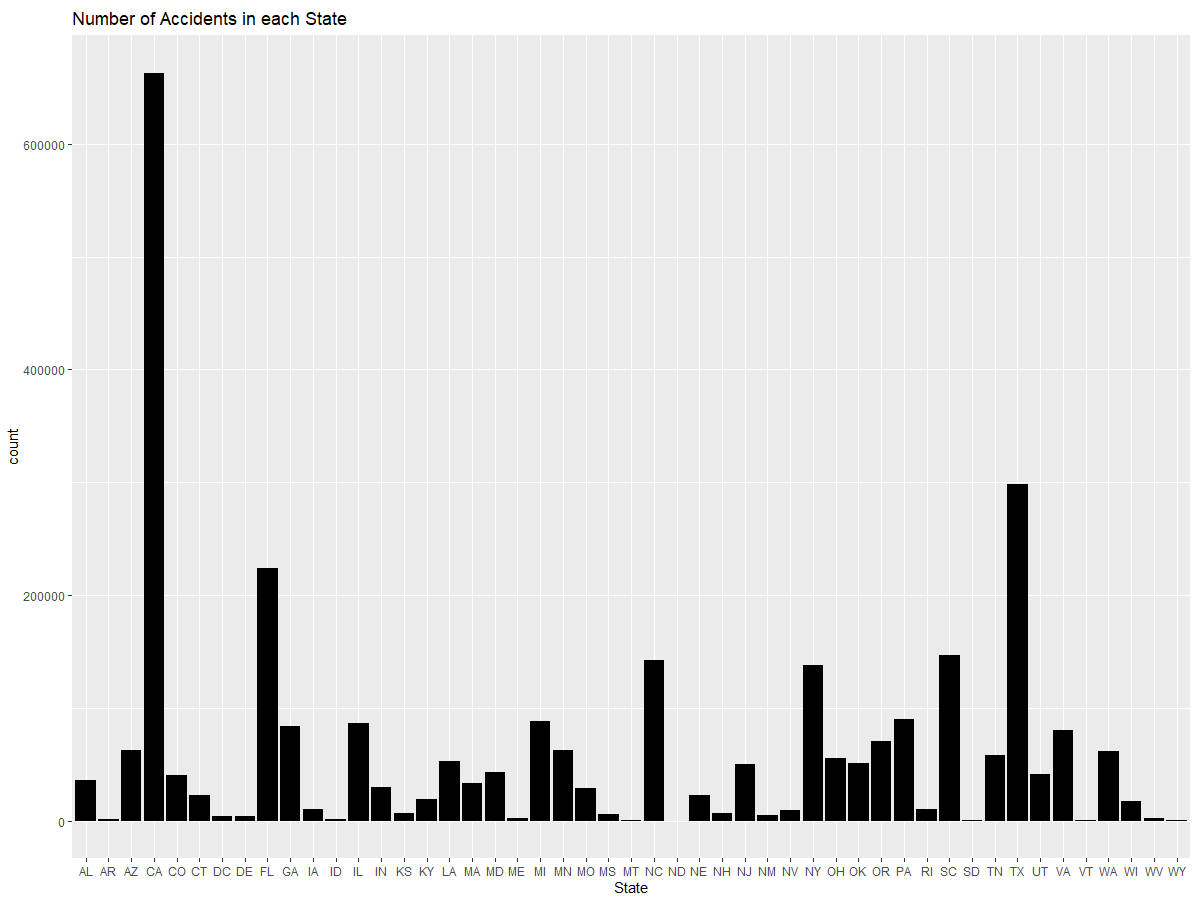
**Hardware Requirements**

System : Intel core I7

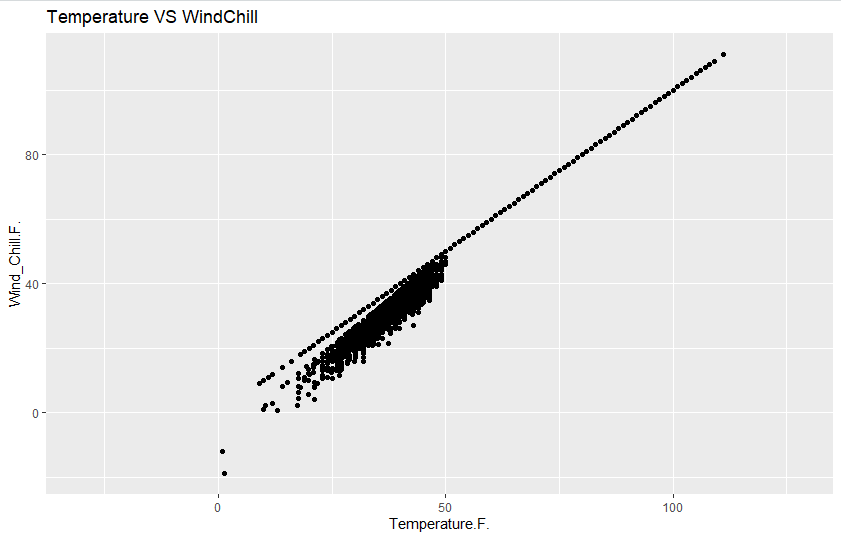
RAM :8GB

1. **RESULTS**
   1. **Visualizations using R**

**Fig 5.** Accident Severity Counts

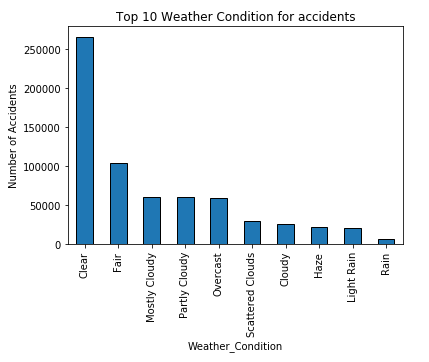
The figure 5, shows that most of the accidents that occur in USA are of severity 2.

**Fig 6.** Number of accidents in each state

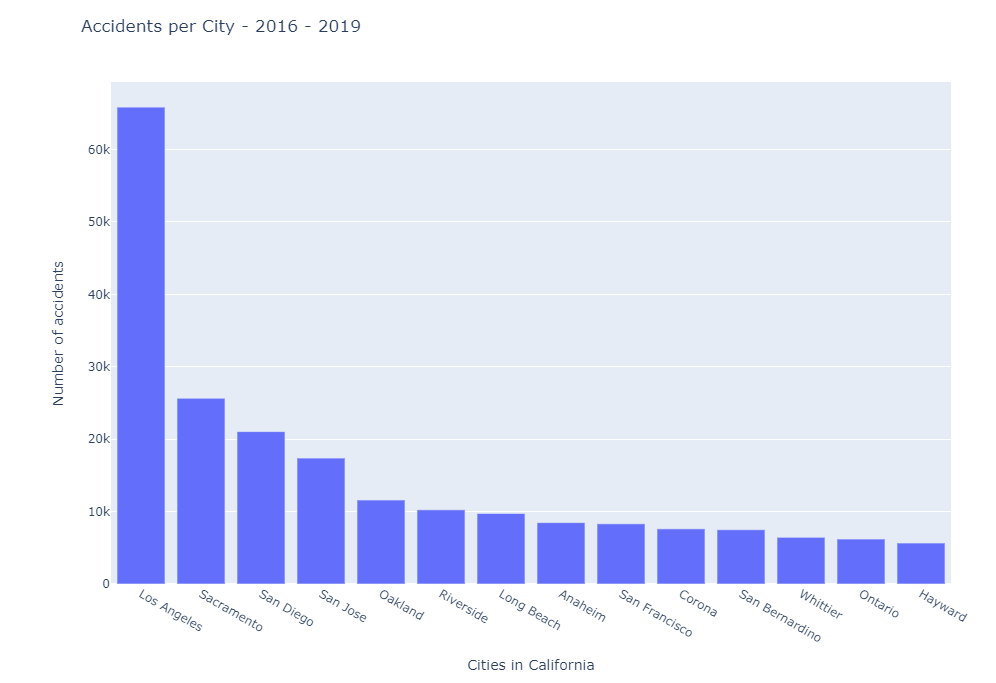
In fig 6, we observe that more than 6,00,00 accidents happened in California. California has the highest number of accidents followed by Texas. Performed a detailed analysis on California, to find out the reason behind its highest number of accidents in US.

**Fig 7.** Scatter plot between Temperature and Wind\_Chill.F

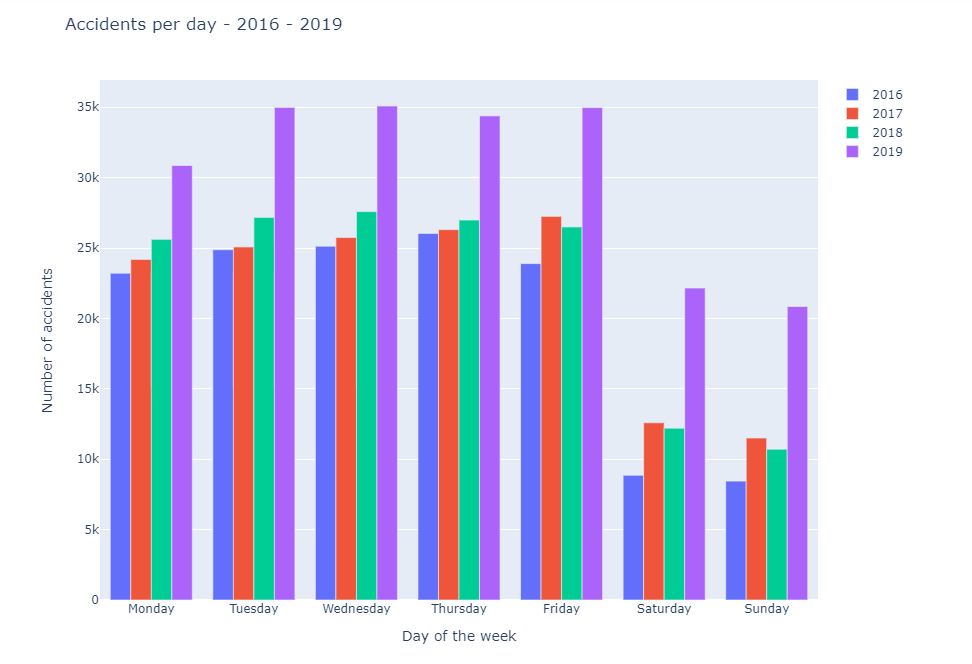
From fig 7, we observe that Temperature and Wind\_Chill.F are positive and correlated highly.

**4.2 Visualizations using Python**

**Fig 8.** Top 10 Weather condition for accidents in California

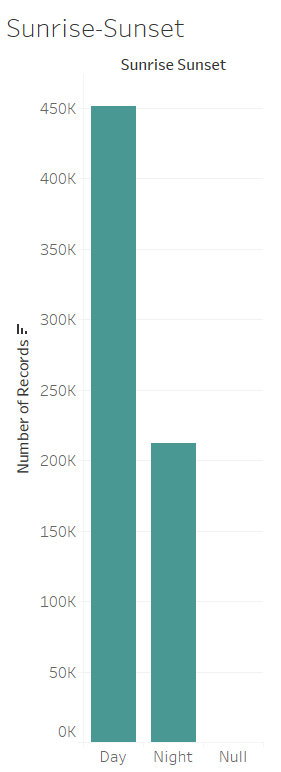
From Fig 7, we observe that most of the accidents in California happen during a ‘Clear’ weather. Second most common weather condition in California for accidents is ‘Fair’ weather.

**Fig 9.**Accidents in Cities of California from 2016 – 2019

From fig 8, we observe that in California maximum number of accidents have occurred in Los Angeles, followed by Sacramento and San Diego.

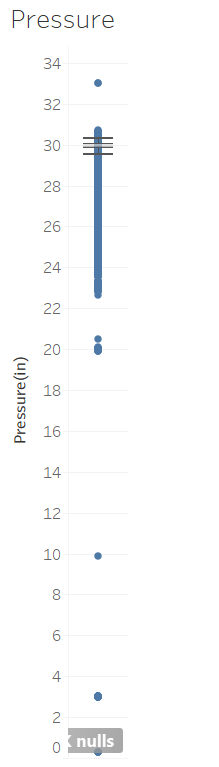
**Fig 10.**Accidents per day in California from 2016 - 2019

From fig 9, we observe that the accidents are increasing monotonically every year in the period of 2016 – 2019. It is observed that most of the accidents happen during the working days of the week, the number of accidents during the weekend is very less.

* 1. **Visualizations using Tableau**

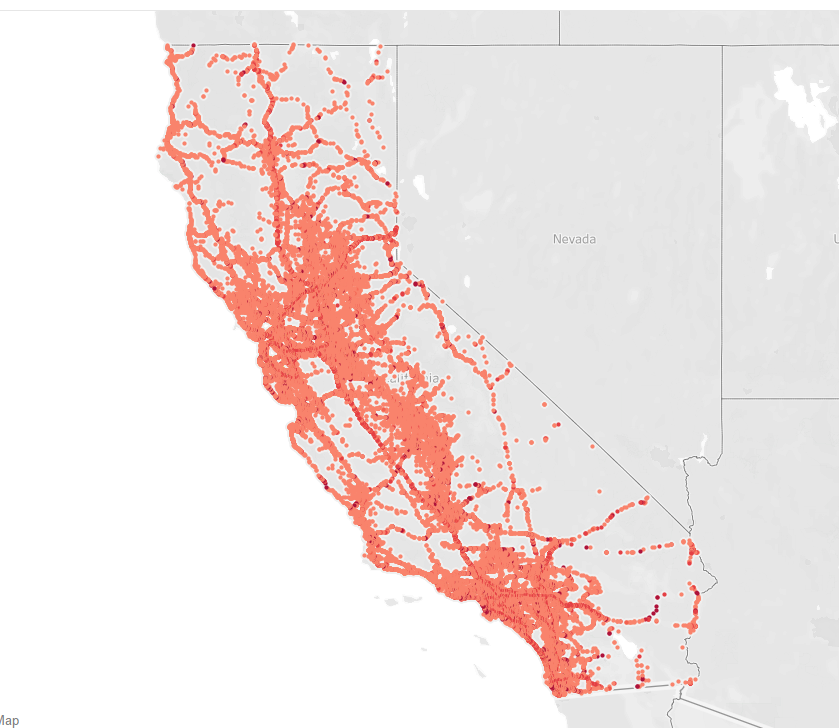
**Fig 11.** Count of Accident records during Day and Night

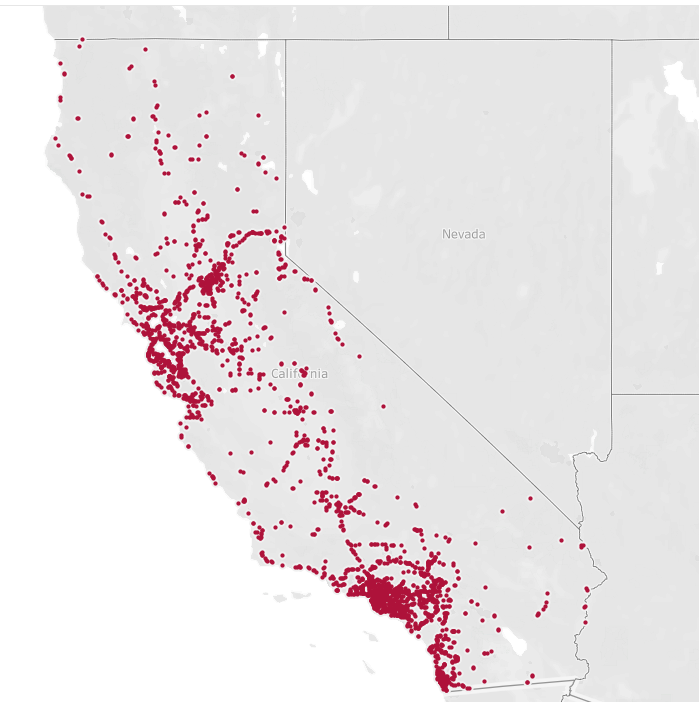
From fig10, we observe that most of the accidents in California have been recorded during the day. The number of records during night is half the number of records during day.



**Fig 12.**Boxplot of Pressure(in)

From fig 11, we can observe that the pressure values mostly lie in the range of 29.84 and 30.3. The median is observed to be 29.94(in).

**Fig 13.** Accident Hotspot Locations in California

From fig 13, we observe that most of the parts of California are accidental prone areas.

**Fig 14.** Accident Hotspots in California with Severity 4

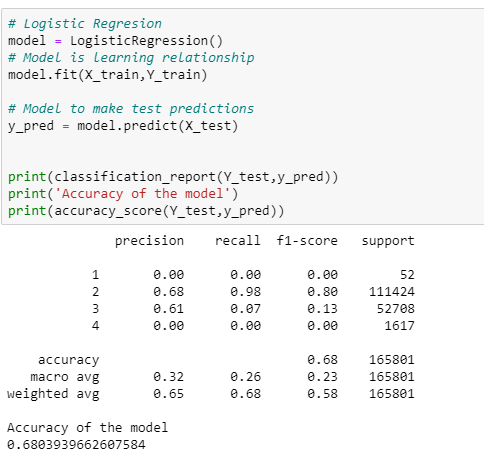
From fig 14, we observe that most of the places in California are with severity 4 of accidents.

**4.4CLASSIFICATION**

Classification also known as supervised learning, is training the data with class attribute. The test set is used to calculate the accuracy of the model. Classifier is the algorithm that performs Classification. In python we need to import sklearn package. Pre-processed data is fed to the classifier in order to train the model.

**4.4.1**Logistic Regression

Logistic Regression can be performed when the class variable is categorical attribute. The logistic regression output is computed based on the values of all other attributes. This regression can also be used for feature extraction. One of the major drawbacks of logistic regression is it cannot handle too many categorical variables.

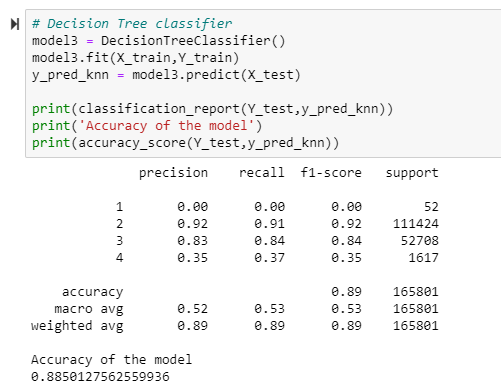


**Fig 15.** Output of Logistic Regression

The Logistic Regression model showed an accuracy of 68%.

**4.4.2** Decision Tree Classifier

The Decision Tree is a supervised learning algorithm. The decision tree can be used for classification and regression, so it works well with categorical and numerical attributes. Algorithms like Gini Index and Information Gain are used for splitting of nodes. The leaf nodes in the decision tree, show the class label to which the given node is classified.[3]

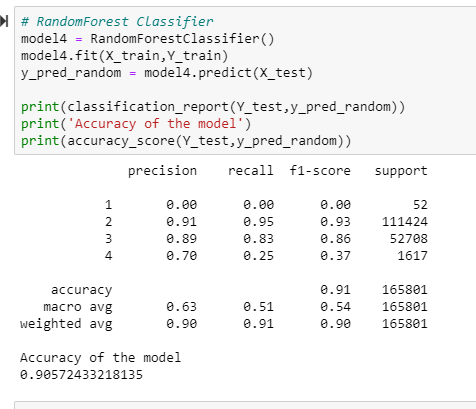


**Fig 16.** Output of Decision tree

The Decision Tree algorithm showed an accuracy of 88%.

**4.4.3** Random Forest Classifier

Random Forest Classifier is an ensemble learning algorithm. This algorithm creates many decision trees randomly and aggregates all of the decision trees to provide the output of the algorithm.[5]



**Fig 17.** Output of Random Forest Classifier

The Random Forest model showed the highest accuracy,i.e, 90.54. It may be because it combines the output from many trees.

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **Logistic Regression** | **68.03** |
| **Decision Tree Classifier** | **88.50** |
| **Random Forest Classifier** | **90.54** |

**Table 1:** Model name and the accuracy

**4.5** Natural Language Processing

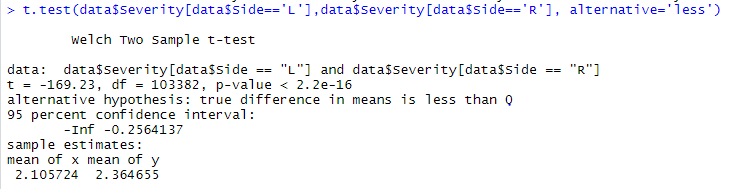
Natural Language Processing is the subset of Computational Intelligence that converts the human language into an understandable form so it can be processed by the computers. In the NLP part, converted the description of accidents into a list, then made them into tokens, removed punctuation and stop words from the text and then found the most frequently used words.[2]

**Fig 18.** Top 200 most frequently used words in Accident Description

From the above figure we can say that

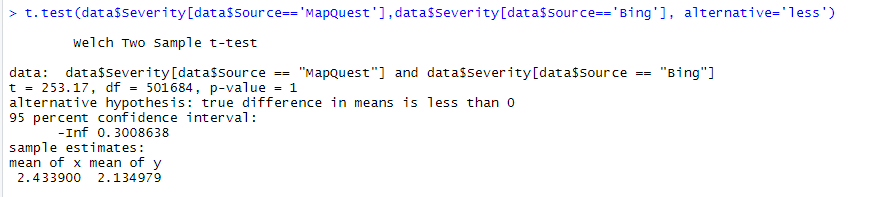
* Most the accidents occurred when the lane was blocked.
* Many accidents occurred at valley and highways.

**4.6** Hypothesis Testing

In Hypothesis Testing, the main assumption is the population parameter, which may be true or false. There are two types of hypothesis: Null Hypothesis denoted as H0 and Alternate Hypothesis denoted as Ha.

**Fig 19.** Result of Hypothesis Test 1

When the p-value is less than 0.05, we can reject null hypothesis and accept alternate hypothesis. The accidents which took place at the right side of the road have more severity than the accidents which took place at left side of the road.

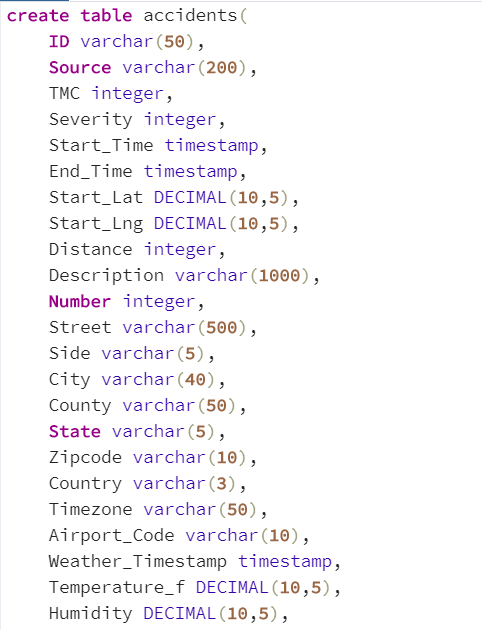


**Fig 20.** Result of Hypothesis Test 2

When the p-value is greater than 0.05, we accept null hypothesis and reject alternate hypothesis. The accidents which had the source MapQuest had more severity of accidents compared to the ones with source Bing.

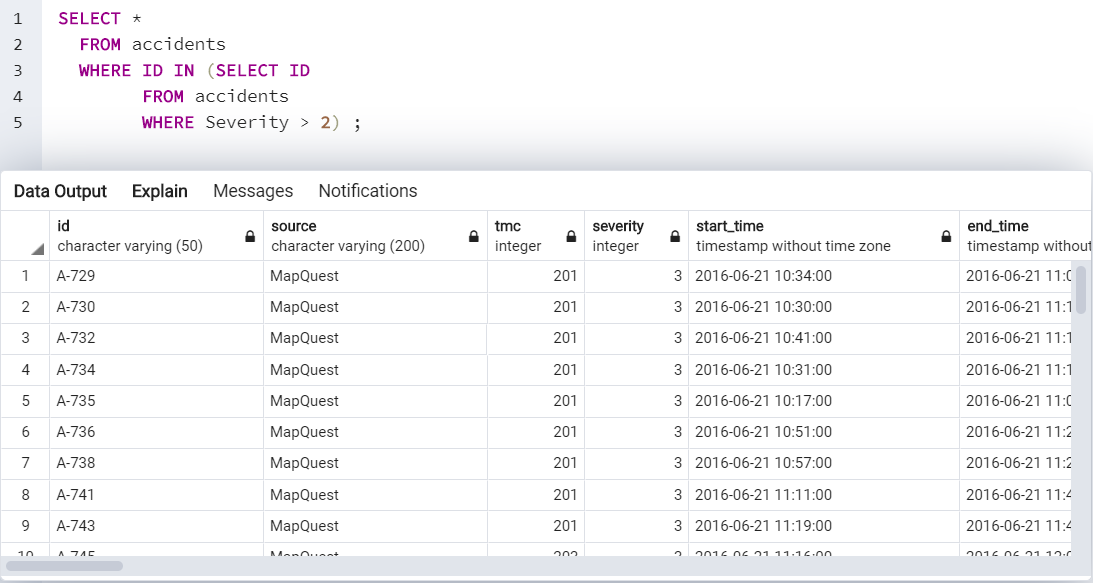
**4.7** SQL

SQL is Domain specific language which is mainly used for data query and data manipulation in database systems.[4]

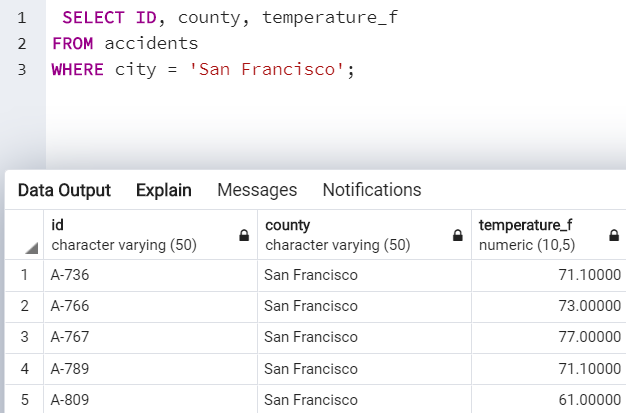


**Fig 21.** Create table accidents command

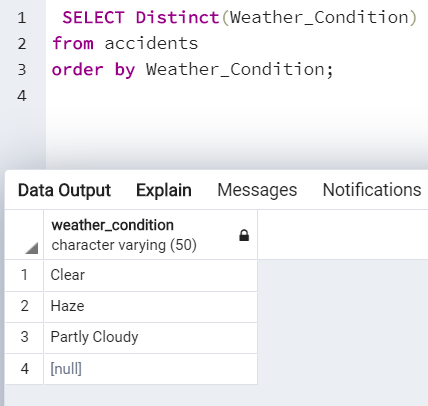
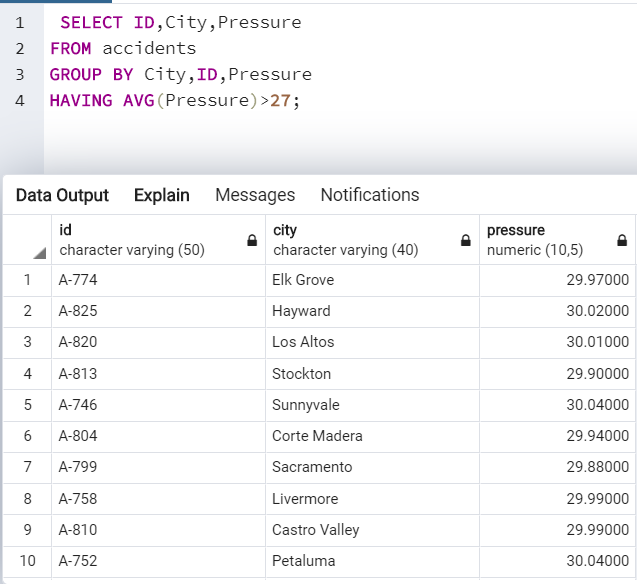
Created a table named accidents and imported the .csv file to this table.



**Fig 22**. A Nested query to find the records with severity greater than 2



**Fig 23.** A query to find the accident records of the city ‘San Francisco’

**Fig 24.** A query to find records with average pressure greater than 27 in California

**Fig 25.**A query to find distinct weather conditions in California

1. **LIMITATIONS& FUTURE RESEARCH**

**5.1 LIMITATIONS**

The US Accidents is a very large dataset with 3 million records. The results could have been more efficient if there was no missing data in this dataset. It is really difficult to predict the latitude and longitude position for the missing records.

**5.2 FUTURE RESEARCH**

This data can be made real time, i.e, updated daily in order to draw real insights and analysis on accident prone areas. Based on the Textual Analysis done using NLP and the Accident hotspot locations can help self- driving cars, such that these cars can be cautious in those areas.

1. **CONCLUSIONS**

* California is the most accident-prone state in US.
* Los Angeles, Sacramento and San Diego are the most accident cities in California. Majority of the accidents take place in Los Angeles.
* Most of the areas in California are having accident severity 4.
* Most of the accidents happen in California during a Clear and Fair weather condition.
* The number of accidents in California has been increasing monotonically from the year 2016 to 2019.
* Most the accidents happen during weekdays, especially Tuesday, Wednesday, Thursday and Friday, accidents during the weekend are comparatively less. So, most accidents happen during working days because people are rushing for work.
* In Classification Models, Random Forest Classifier gave the highest accuracy, i.e, 90%.
* From NLP we could tell that
* Most the accidents occurred when the lane was blocked.
* Many accidents occurred at valley and highways.
* From Hypothesis Test we could tell that
* The accidents which took place at the right side of the road have more severity than the accidents which took place at left side of the road.
* The accidents which had the source MapQuest had more severity of accidents compared to the ones with source Bing.

1. **REFERENCES**

[1] Moosavi, Sobhan . "US Accidents (2.25 million Records) - A Country Wide Traffic Accident Dataset (2016-2019) ", <https://www.kaggle.com/sobhanmoosavi/us-accidents>, Accessed on 29th March, 2020.

[2] Geitgey, Adam. "Natural Language Processing is Fun", Medium, <https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e>, Accessed on 20th April, 2020.

[3] Sehra, Chirag. "Decision Trees Explained Easily", Medium, <https://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248>, Accessed on 25th April, 2020.

[4] “8.5. Date/Time Types.” *PostgreSQL*, www.postgresql.org/docs/9.5/datatype-datetime.html, Accessed on May 9th, 2020.

[5] Patel, Savan. "Chapter 5. Random Forest Classifier", Medium, <https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1>, Accessed on May 5th, 2020.

[6] “Working with Missing Data¶.” *Working with Missing Data - Pandas 1.0.3 Documentation*, pandas.pydata.org/pandas-docs/stable/user\_guide/missing\_data.html, Accessed on April 15th,2020.