PREDICTING ACCIDENT SEVERITY

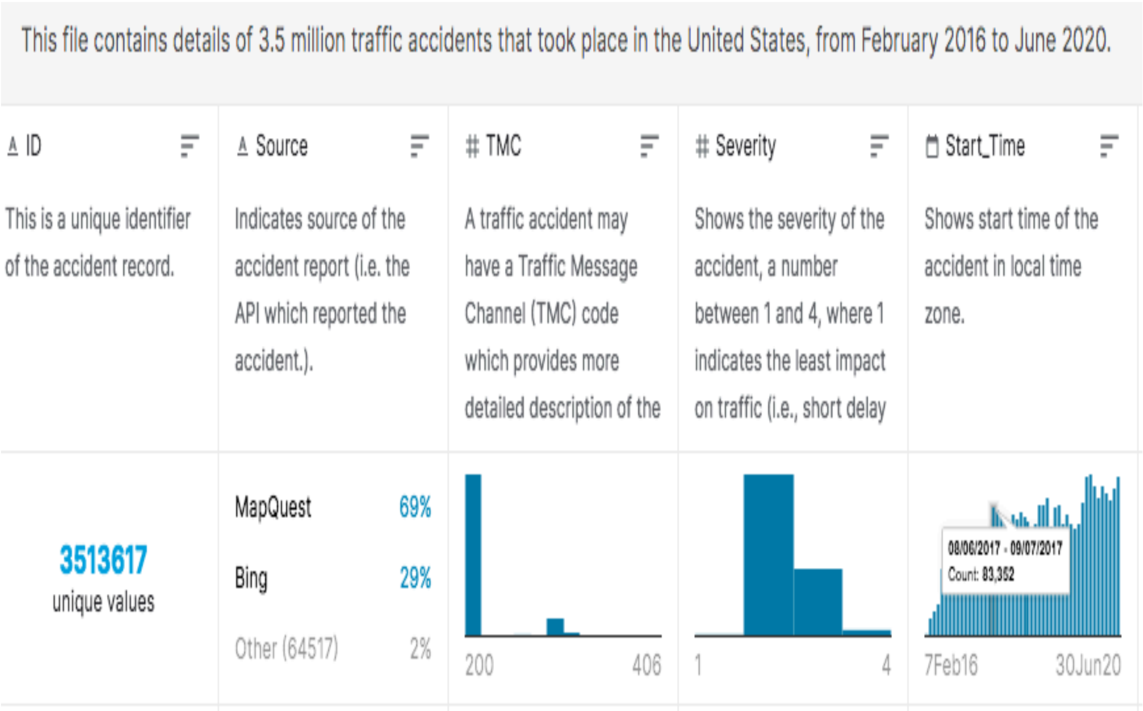
INTRODUCTION

- Road accidents are a severe concern for most nations worldwide because accidents can cause severe injuries and fatalities and substantial economic losses.

- In this project, CRISP-DM (Cross Industry Process for Data Mining) methodology will be used to predict any road's accident severity by training an efficient machine learning model with the help of existing accidents.

DATA USED

- I'll be using US accidents dataset acquired from Kaggle.com. the dataset covers 49 states of the USA. The accident data are collected from February 2016 to June 2020.



This file contains details of 3.5 million traffic accidents that took place in the United States, from February 2016 to June 2020.

ID	Source	# TMC	# Severity	Start_Time
This is a unique identifier of the accident record.	Indicates source of the accident report (i.e. the API which reported the accident.).	A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay	Shows start time of the accident in local time zone.
3513617 unique values	MapQuest 69% Bing 29% Other (64517) 2%	200 – 406	1 – 4	08/06/2017 - 09/07/2017 Count: 83,352 7Feb16 – 30Jun20

DATA WRANGLING

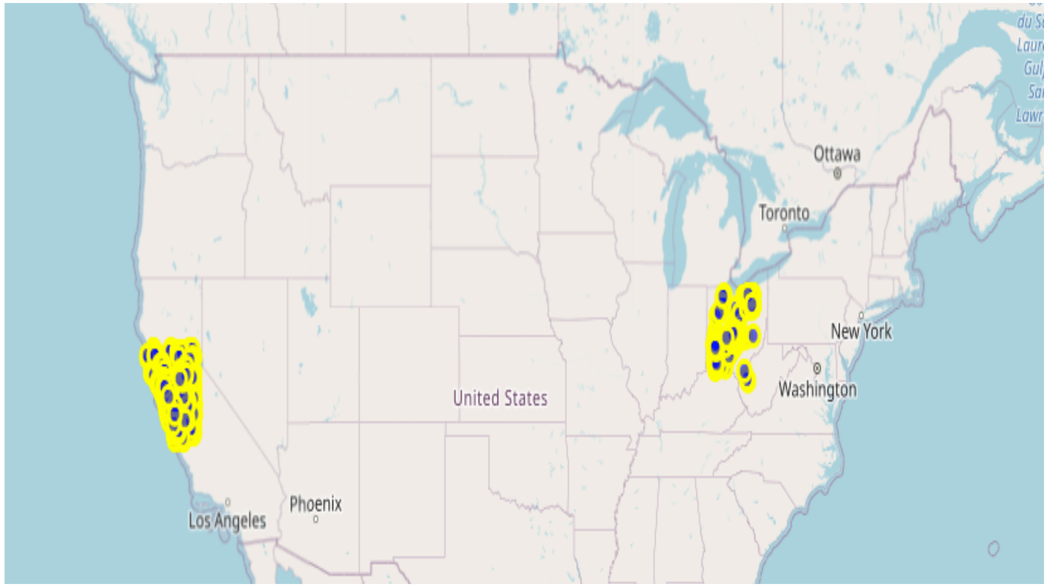
- Had to limit the data to only 10000 rows.

- I dropped the unwanted columns. Then, I replaced the NaN values with the mean for some columns and deleted the rows that contain the NaN values

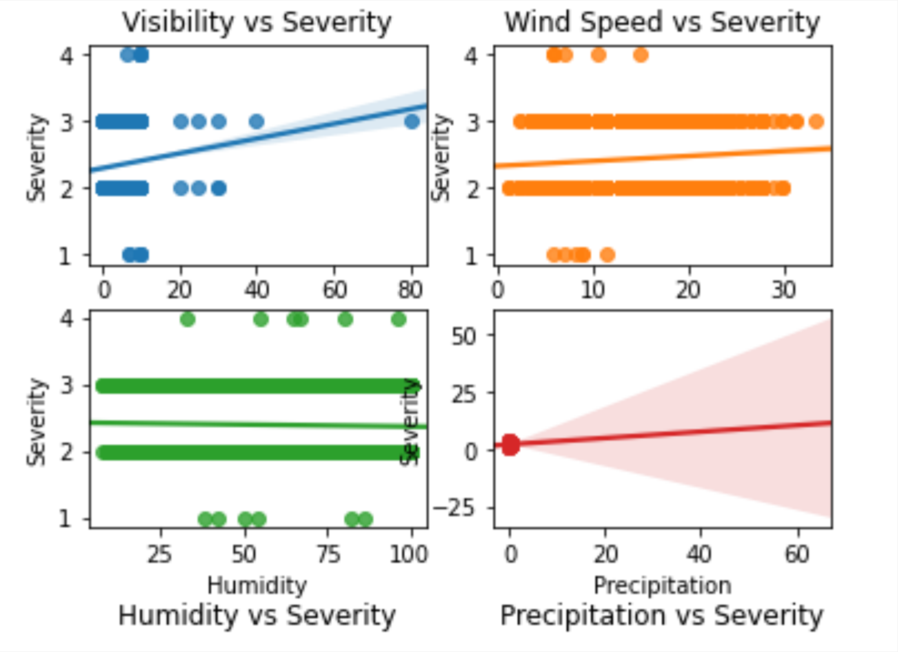
EXPLORATORY DATA ANALYSIS

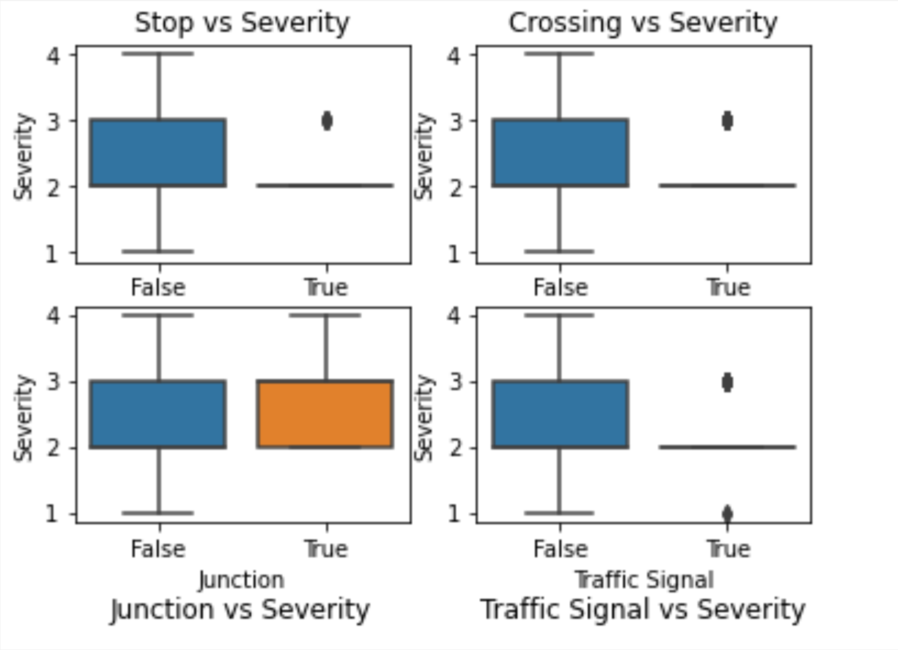
- Used folium to create a leaflet map.

- Used the correlation function to get the correlation between the attributes.



EXPLORATORY DATA ANALYSIS CONT.





MODEL DEVELOPMENT AND EVALUATION

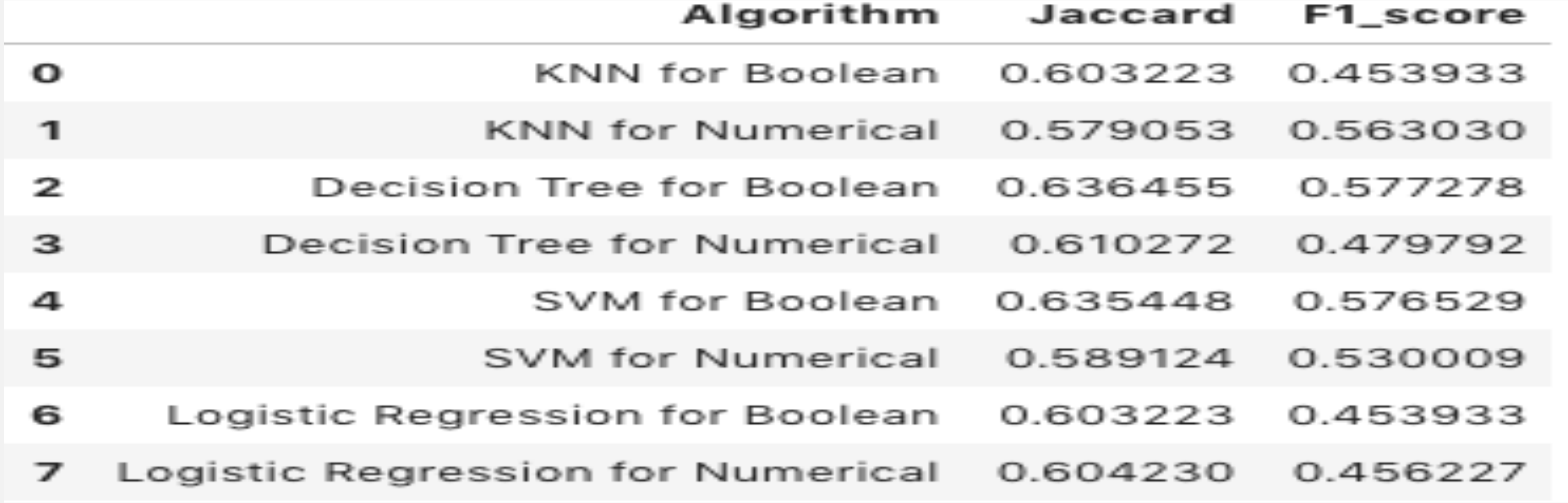
- Used Supervised Learning to develop my model, as I am using fully labeled data

- Used the train test split and used 10% of the data for testing

- Used different classification techniques (K Nearest Neighbor, Decision Tree, Logistic Regression, Support Vector Machine)

- Used different evaluation techniques (jaccard similarity score, f1 score).

RESULTS



	Algorithm	Jaccard	F1_score
0	KNN for Boolean	0.603223	0.453933
1	KNN for Numerical	0.579053	0.563030
2	Decision Tree for Boolean	0.636455	0.577278
3	Decision Tree for Numerical	0.610272	0.479792
4	SVM for Boolean	0.635448	0.576529
5	SVM for Numerical	0.589124	0.530009
6	Logistic Regression for Boolean	0.603223	0.453933
7	Logistic Regression for Numerical	0.604230	0.456227

DISCUSSION

- I think we need more data attributes with higher correlation coefficient to be able to get more accurate model