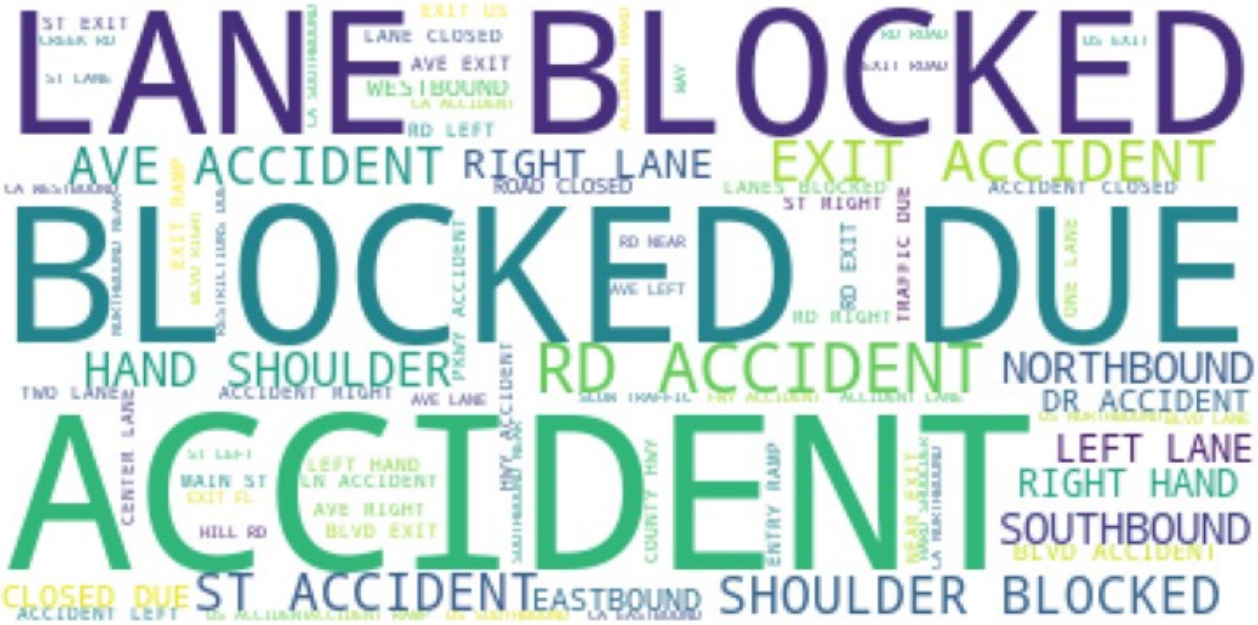


Midterm Project Update

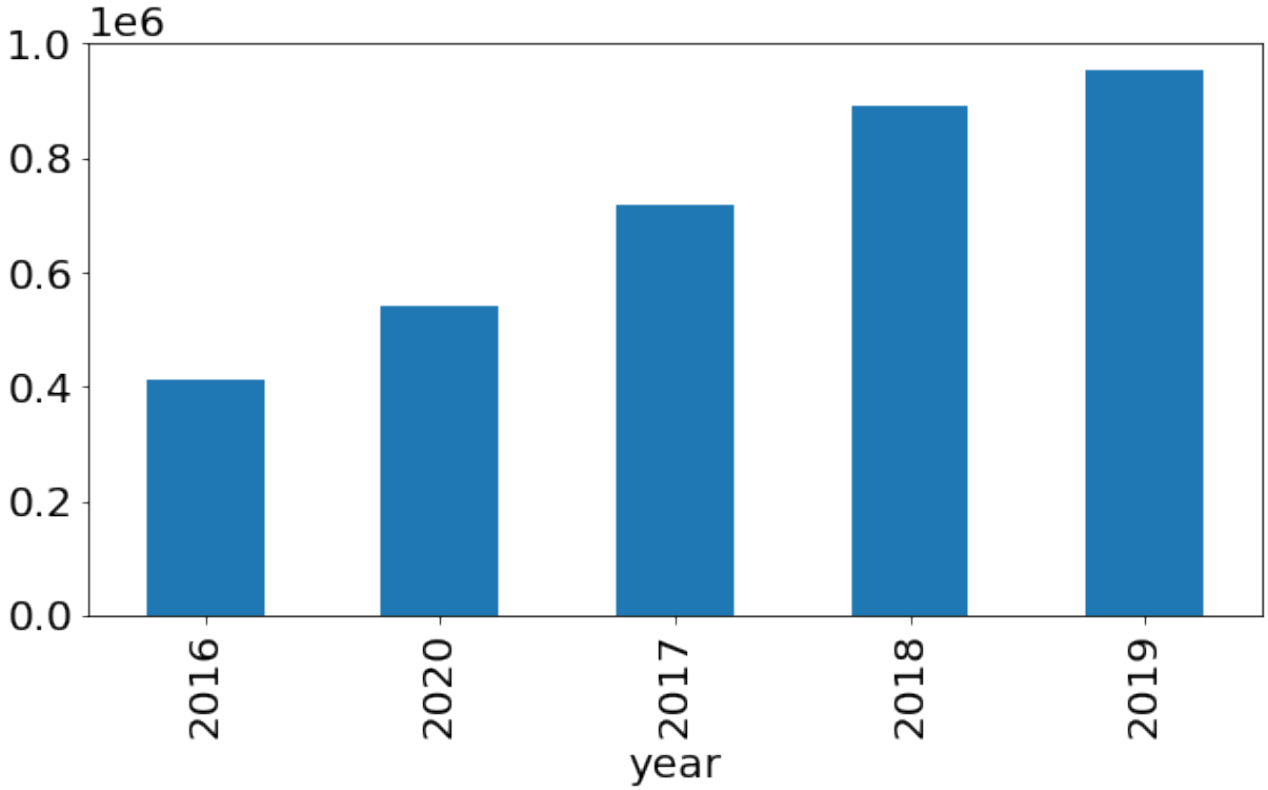
Accident Severity Predictor (ASP)

by
Ishaan Shrivastava
Rajat Bansal
Amber Singh
Rohit Raj
Sajal Saini

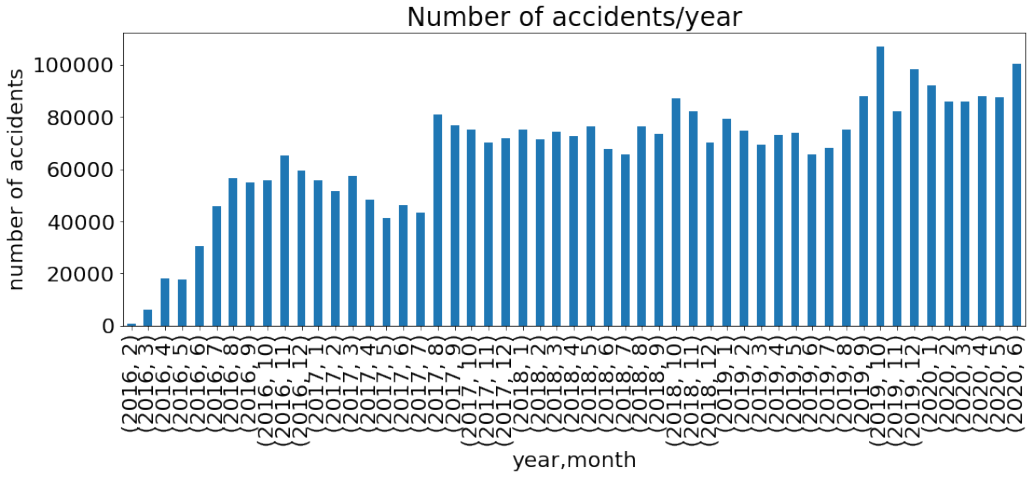
Progress on Data Collection, Cleaning and Visualization

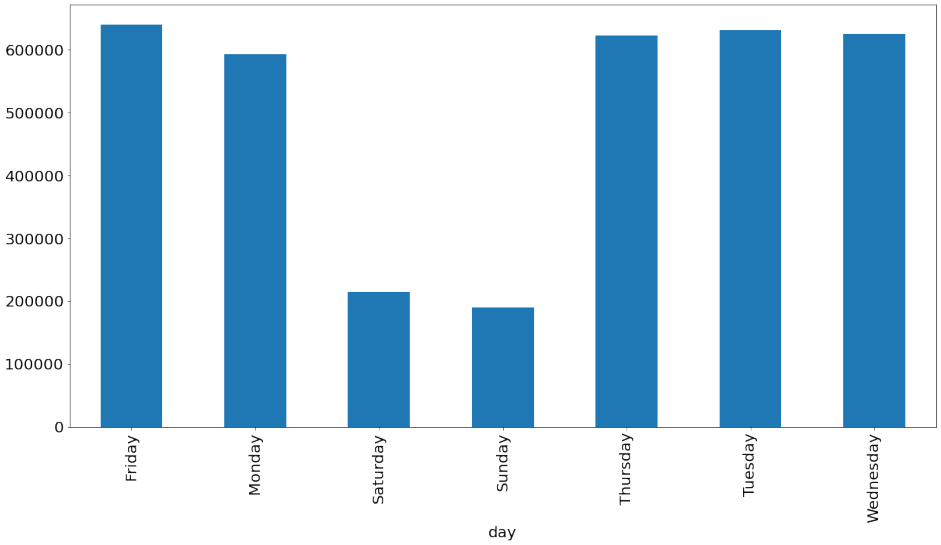
Since we are using a very diverse dataset from Kaggle on US Accidents, identification of attributes to be used in modelling data is of crucial importance. Our priority is to identify what not to take while modelling. Also, one the advantages of having such diverse data is that the data visualization techniques can create vivid stories centred on ‘Accidents’ while adeptly illustrating various dependencies which help in eliminating redudant attributes.

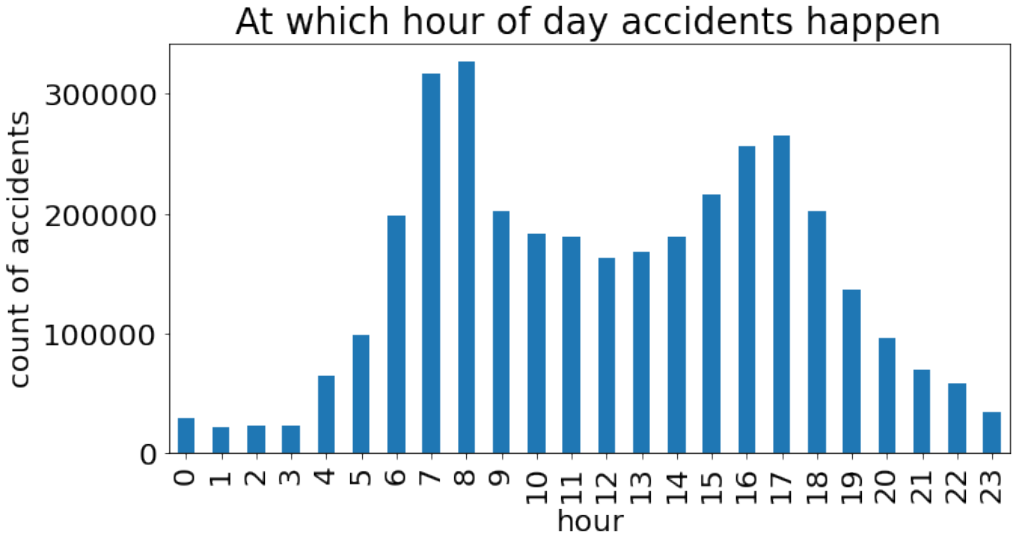
Some of the observations and visualizations:



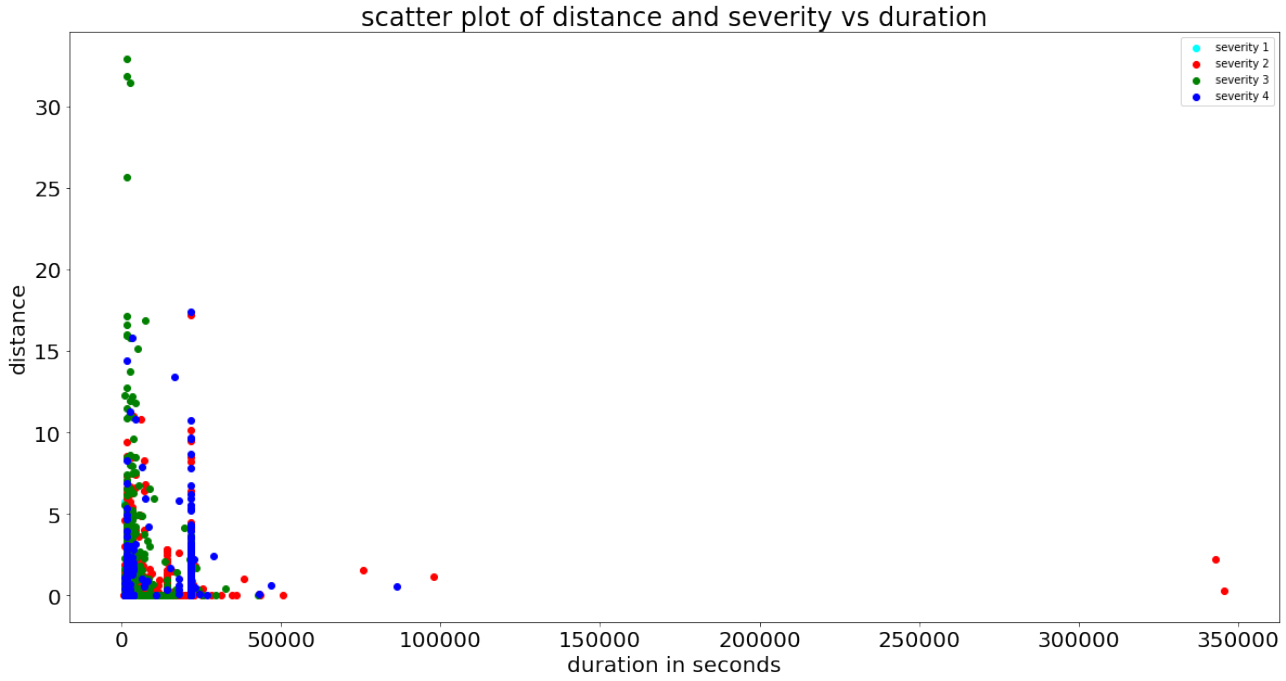
Here we can see that although 2020 has pandemic situation and the data has records till May/June 2020, the number of accidents has already surpassed that of 2016. Although this graph doesn’t help us in modelling, it gives us the hang of how our data is.



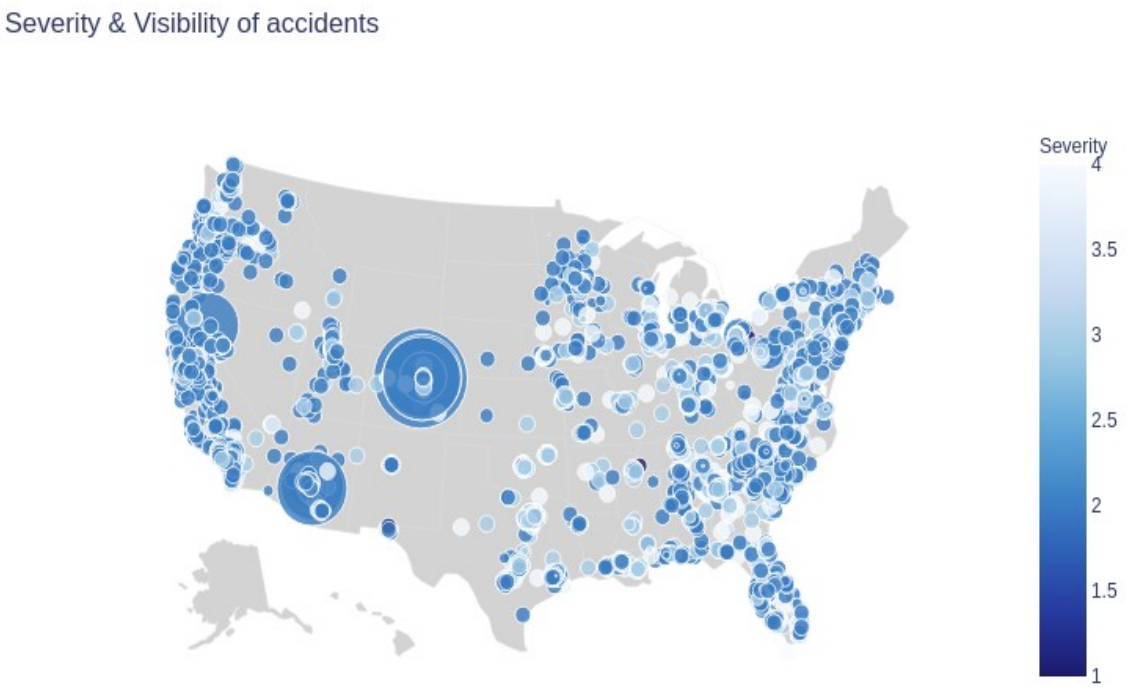




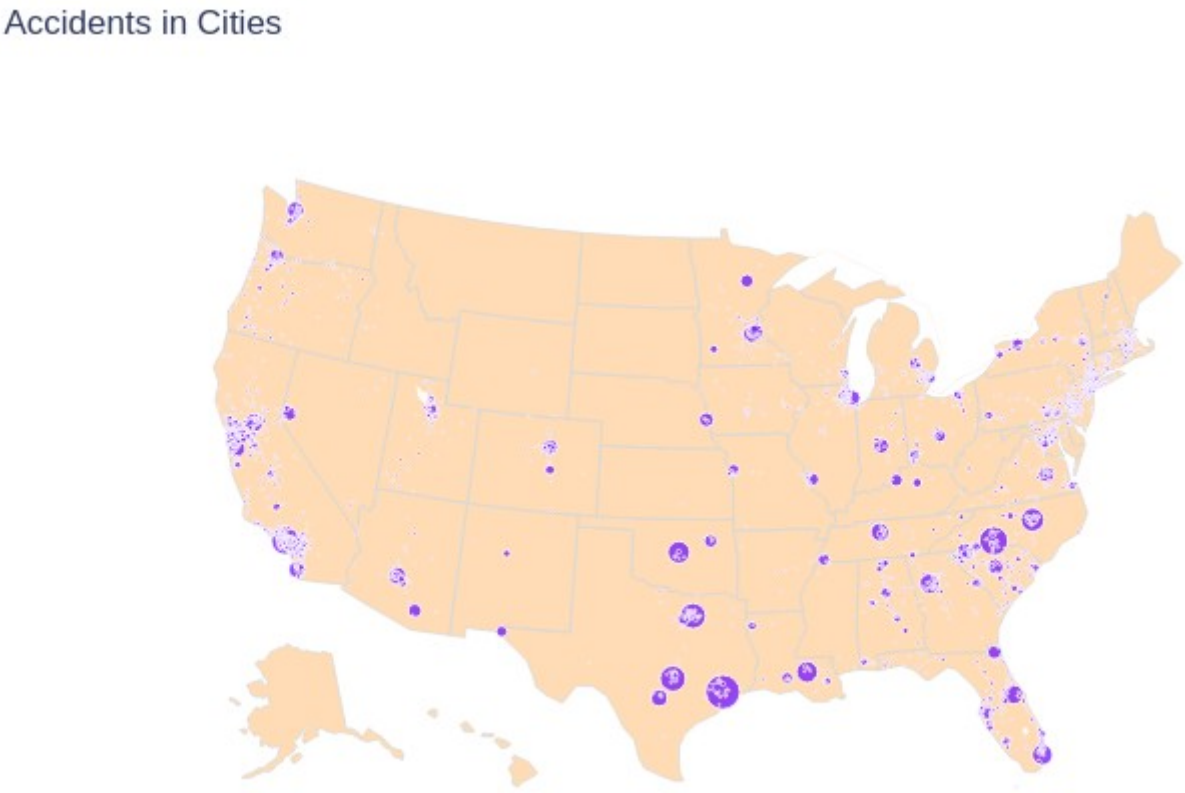
The above given time series bring out various interesting trends on accidents. Also, while generating these, the time attribute is cleaned in the usable format of hours, years, days, etc hence can be used in the model as attributes.



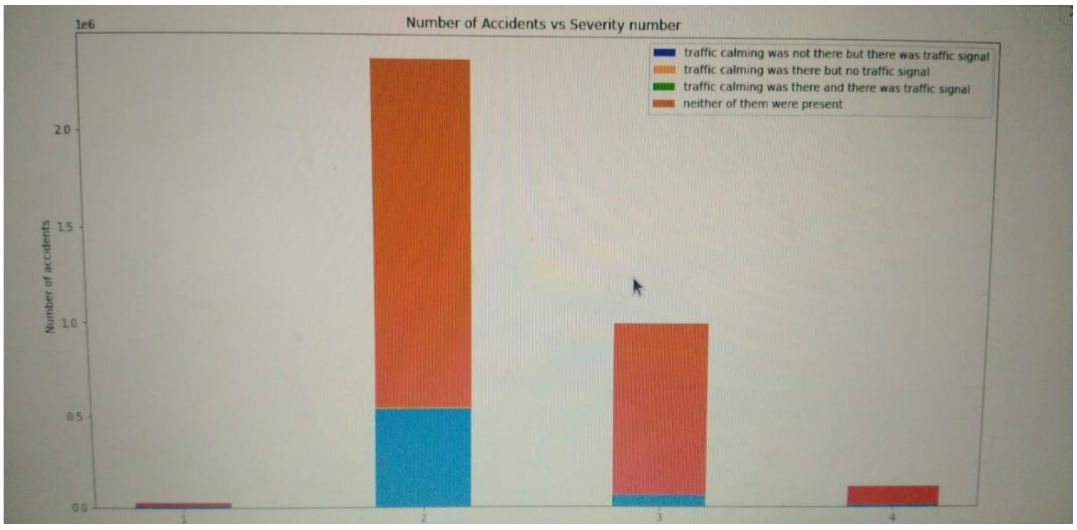
This graph was made to address whether Severity of an accident can be predicted with given delay it has caused in the traffic and the distance of traffic affected by the accident. As can be seen clearly, there are no apparent clusters and hence the dependency is very low. Hence the corresponding columns are dropped (data cleaning) and any possibility of predicting distance and duration based on severity is also dismissed.



This visual shows location, visibility(radius of circles) and severity. We can observe that accidents of severity>=3 occur mainly in areas of low visibility. Also, a clustering of accidents can be seen on the left and right side of the map suggesting how important location is for frequency and severity.



Cluster of cities can be seen as hotspots of accidents. Larger the city larger number of accidents and possibility of high severity if we compare with earlier graph.



This graph shows the importance of traffic calming measures and presence of traffic light to reduce the severity of accident. Also, the graph brings out skew nature of data as severity 4 and severity 1 accidents occur rarely. This emphasizes the model needs to take skewness in account in order to generate usable results.

Progress Summary and current Challenges

Data cleaning is complete for modelling and redundant columns are removed. Preliminary data visualization is also complete.

Although we are in search of a correlation metric that works on boolean data as majority of the attributes are boolean in nature.
Most of the python implementations of neural networks are not multi-classification. Which leaves on a crossroad, either to break the problem in two classification problem i.e. {1,2} {3,4} and then {1} {2} {3} {4} or implement the algorithm ourselves. The feasibilities of both are not known to us. If both won't be possible for us, we might have to move to simpler algorithms like kNN or multi-class SVM.
Another problem with us is that of to account for skewness in the data (which is evident in the last figure in the report.)

Link for our Notebook: https://www.kaggle.com/ishaanshrivastava/ds250-project
(Note: This might not contain our latest revision, as we are not using Kaggle for version control. Kaggle notebook is solely used for generating results as using dataset from the site itself is very handy)