OYEKANMI OLUWATOBA

FINAL REPORT

ANALYSIS OF MOTOR COLLISIONS & CRASHES IN NEW YORK

Problem Statement

Motor Vehicle traffic crashes are the leading cause of injury related death for New York residents. This project aims to analyse and predict vehicle collisions in New York CIty.

Background

There are about 6 million vehicle crashes in the United States. Many of those happen right here in New York city and it’s evident in the amount paid in vehicle insurance yearly. Every month, the NYPD publishes car crash data for all five boroughs. Stationing NYPD's, EMTs and FDNY strategically before accidents occur can help prevent accidents before they occur or reduce fatalities. This project aims to take a data oriented approach to establishing facts on vehicle collisions and predict accidents counts by borough.

Datasets

The City of New York Datasets website contains data of NYC vehicle collisions which is updated monthly by the FDNY. The data can be accessed either by downloading in multiple formats or extracted using APIs. The dataset was extracted using APIs and it contained 29 features and 10,000 rows but was wrangled to 18 features. Some of the features are:

- Crash dates and time
- Location: borough, street name, latitude, longitude, zip code
- Vehicle type
- Contributing factor
- Number of casualties and fatalities

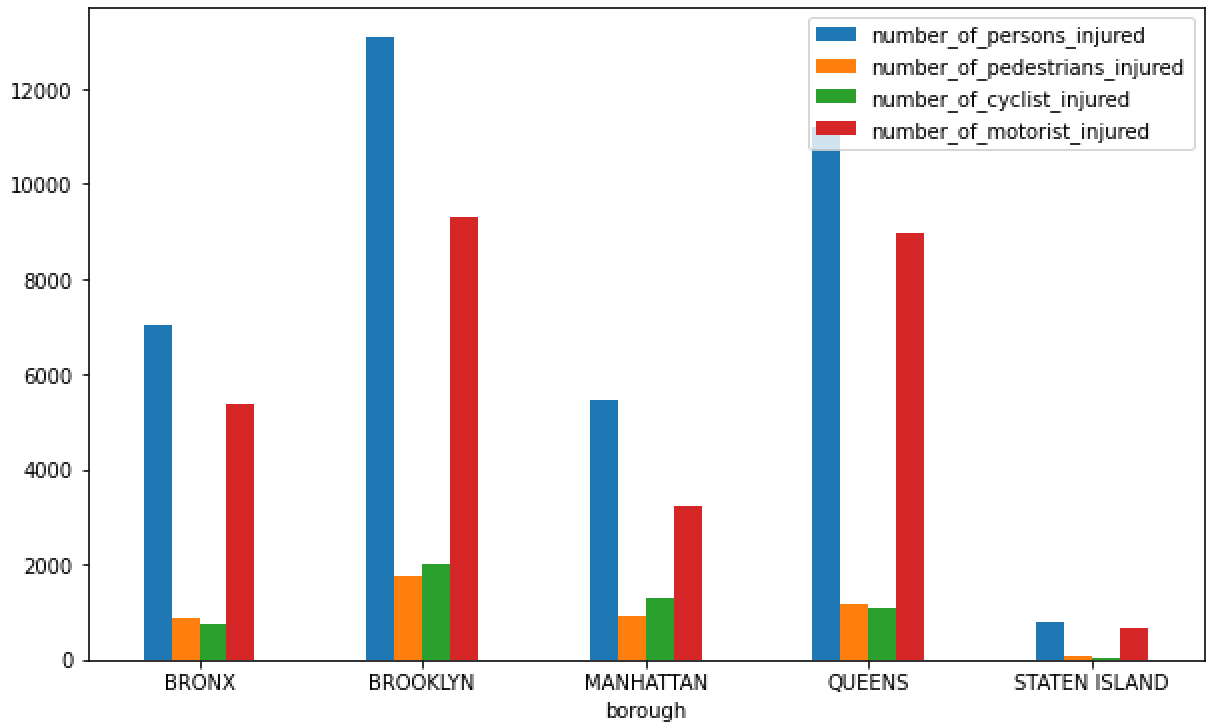
Data Wrangling

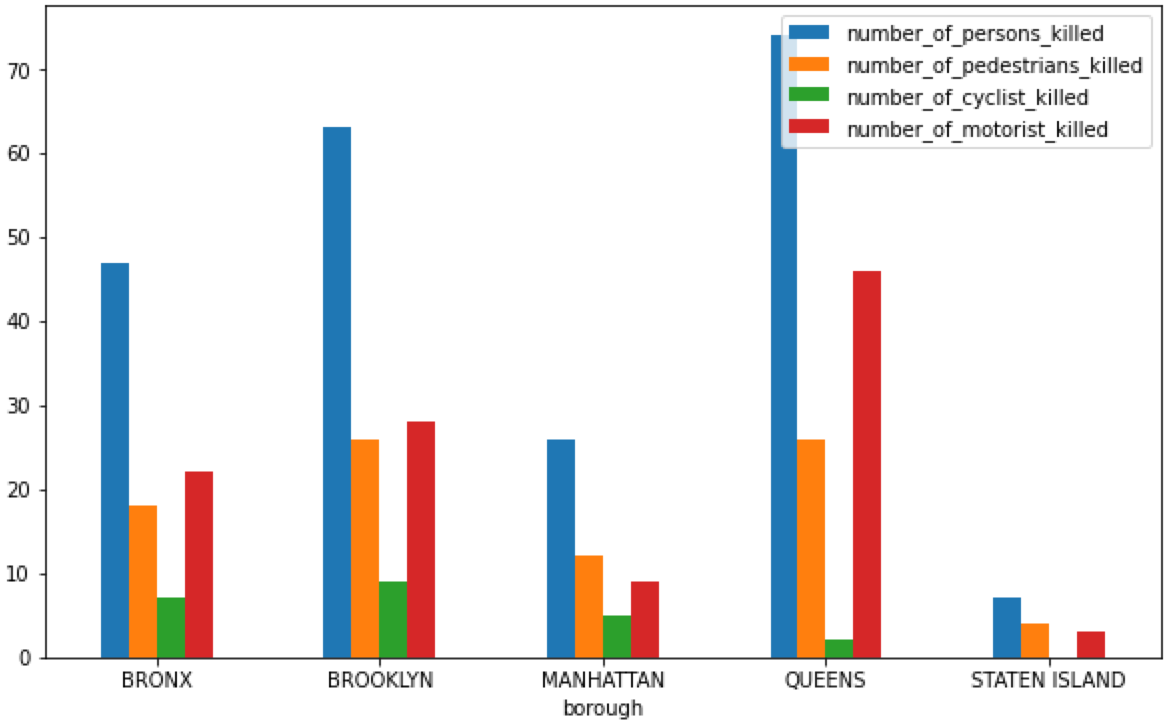
The raw extracted dataset from the NYC database contained 10,000 rows with 29 columns, but was wrangled to 8925 rows and 18 features. The columns dropped were either irrelevant to our goal or had too many null values. We basically dropped all columns with over 50% null values, most of which were irrelevant to our problem statement. The missing values in important columns such as zip code and borough columns were obtained by looking up it’s longitude and latitude features using uszipcode library. The data types were further put in check to ensure data serves its purpose. The dates and time were converted to datetime types as well it’s floats and categorical features. Further and final wrangling resulted in a data shape of 8860 rows and 18 columns.

Exploratory Data Analysis

At this stage, using univariate and bivariate plots, we analysed and explored multiple facets of our data to gain a better understanding of our data.

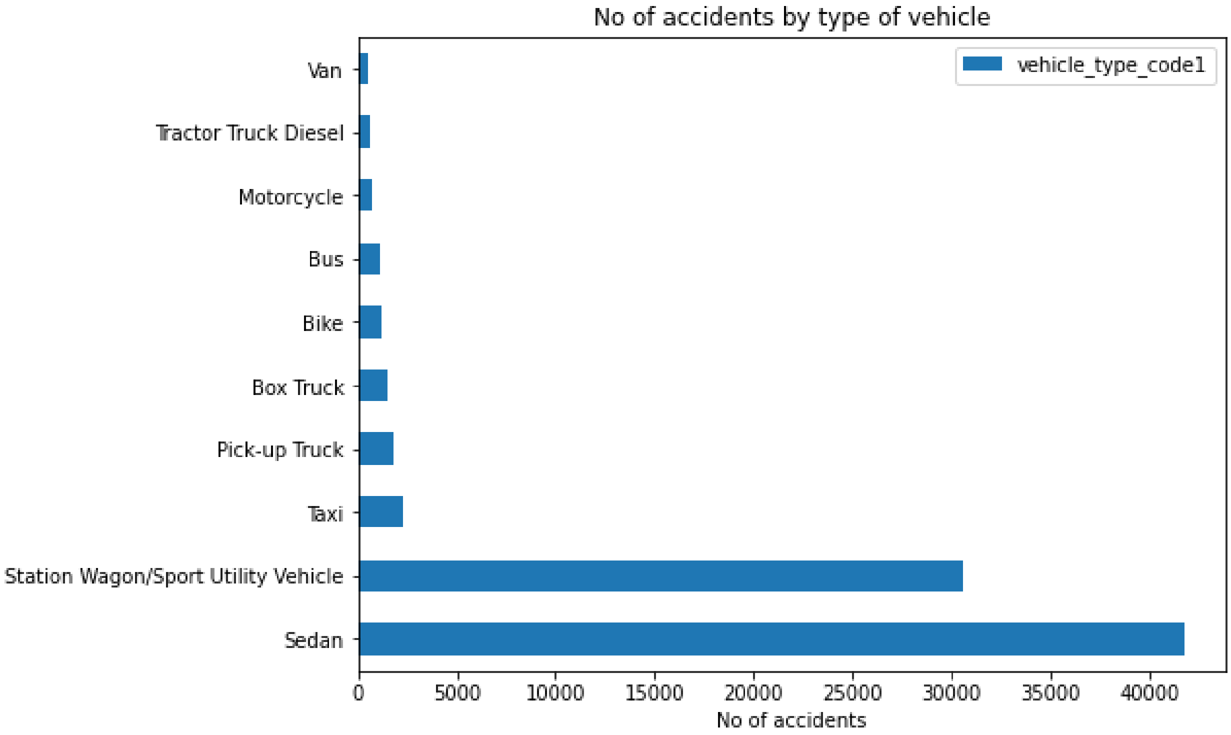
Severity and fatalities of accidents were analysed using hued bar plots by location and we saw Broklyn has the highest number of non-fatal accidents 13089 but Queens had the most fatal accidents 74. Staten Island had the lowest number of accidents across all kinds of accidents.



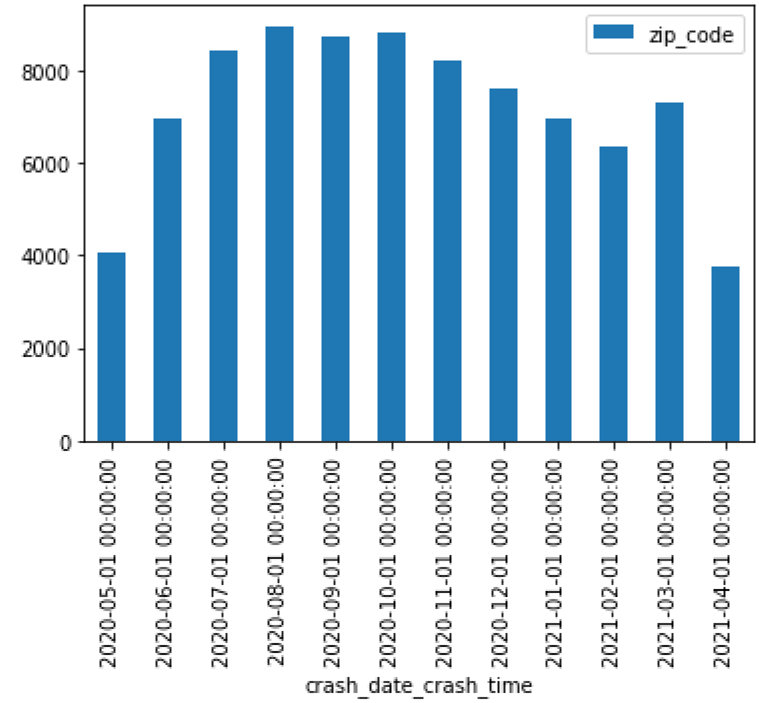


We further analysed and ranked the causation of accidents. As we might expect, the majority of accidents are caused by Driver Inattention/ Distraction with a whopping 21,510 accidents. The runner-up Failure to yield right-of-way had about 5260 accidents. Surprisingly, Alcohol involvement ranked low with just 1346 accidents.

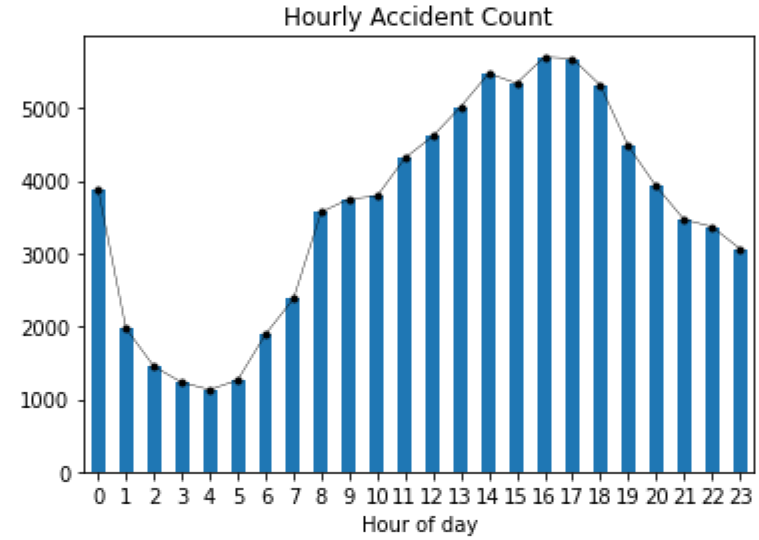
It came as no surprise that the majority of vehicles involved in accidents were Sedans. It makes sense since inexperienced and newbie drivers are more likely to drive sedans thus, increasing the likelihood of crashes in Sedans. The runner up was Sport Utility Vehicles.



Number of accidents were further explored by months and we deduced that accident counts spiked between May and November. These months are summer periods and people are more likely to drive carelessly in this season.



Also, we observed that crashes trended upward from 8am and downwards after 8pm.



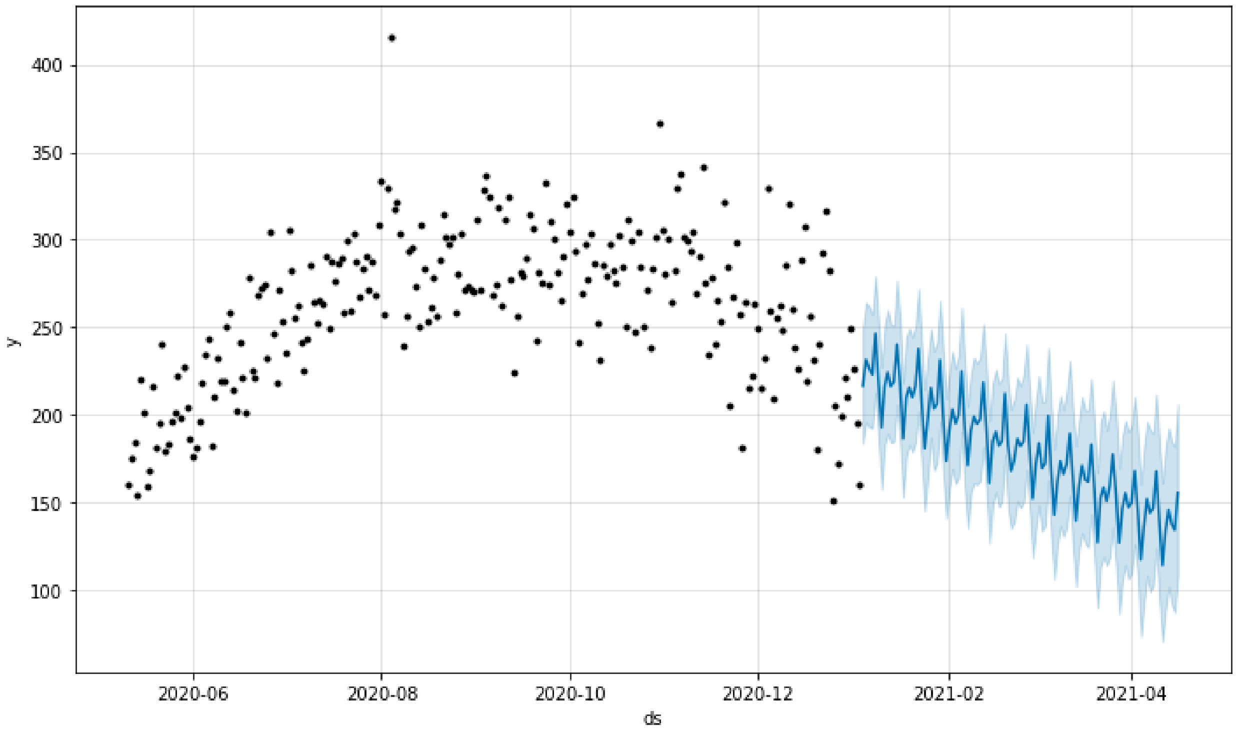
At this stage, we have a strong grasp of the relationships between different features of our data. We prepare our data for modelling using feature engineering by further cleaning and creating dummy variables for our categorical variables. We chose a target variable in counts of accidents and it was dependent on all other features.

Modelling

After the preprocessing stage, our data was thoroughly clean and model ready. Our attempt was to predict the total number of accidents per borough in a day. Our data spanned across 2 years. We tested multiple models to identify the best option and the metrics for ranking and selecting the models are R^2, MAE and RMSE.

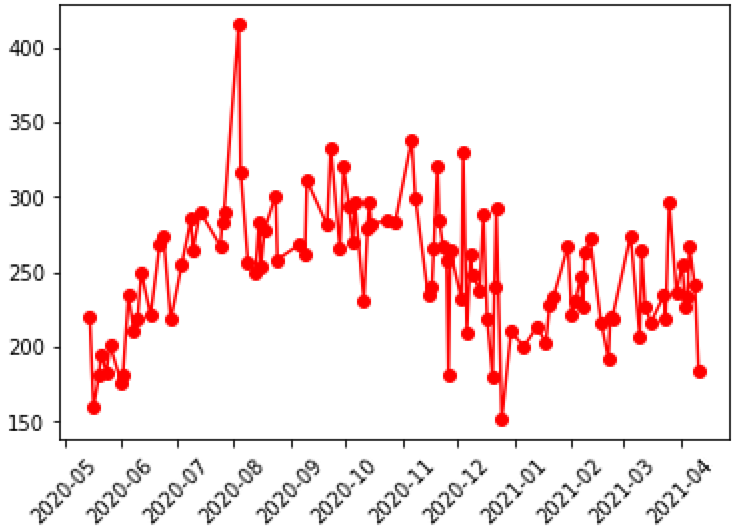
We basically used out-of-the-box estimators and selected the best two models with the best metrics. The best two selected models were further optimised using gridsearchcv hyperparametization techniques.

We had high hopes for FbProphet for good reasons being that it identified multiple phases of trends but the it's rmse score was way too high to be taken seriously. It failed to identify the upward inflexion in the trend.



After attempting multiple models; Linear Regression, RidgeCV, ElasticNet, DecisionTree, Lasso, KNN among several other models consistently performed very poorly. Decision Tree with a training score of 1.0 and a test score of 0.2 obviously shows overfitting.

The models with the best metric scores are RandomForestRegressor and Gradient Boosting Regressor. We further did some hyperparameter tuning for the Random Forest and GradientBoosting model using GridSearchCrossValidation and the RandomForest outperformed with a slight rmse difference. Below is a scatterplot of predictions made using the randomForest model.



Key Takeaways

Still on the sole purpose of predicting the number of accidents in NYC per day by borough. We learnt out-of-box Random forest performed best among all models attempted with a RMSE value of 28.73 and also supported by R^2 metric score of 0.56.

Margin of Error: The average number of accidents in NYC is 151. Thus our model's margin of error is 11.379%.

Future Works

Though this project is done but the NYC accidents problem is far from complete as it can be further extended on many sides:

- Analysis and development of Hospitals and Treatments Centres around accident hotspots with high severity in NYC. This work could help reduce drastically the time for accident victims to get help.
- Analysis of traffic tickets and road safety in NYC. We would all agree that the state extorts money from the people using lame traffic tickets as excuses which have no correlation with road safety. A data-driven approach would be a fairer way to traffic tickets.
- Analysis and predicting of factors affecting accidents severity in NYC.
- Exploratory analysis of weather effects on accidents in NYC