Car Accident Severity Prediction

Alex Lee

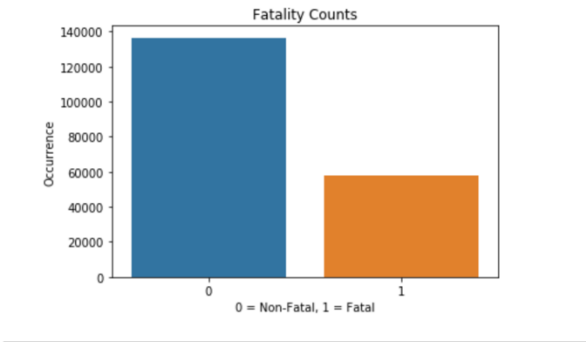
6 Oct 2020

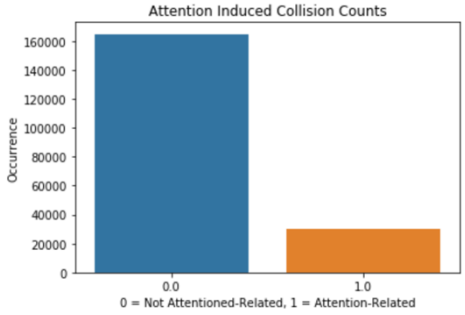
DATA UNDERSTANDING AND PREPARATION

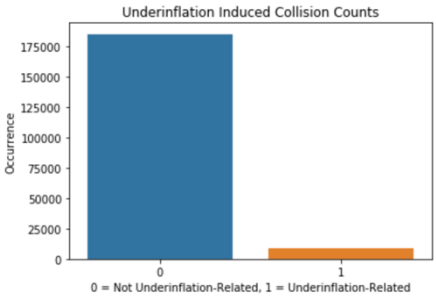
7. Data Source. The dataset used in this analysis is provided for, and is available as an open source download from the Seattle Police Department. The collisions data spans from 2004 to present.

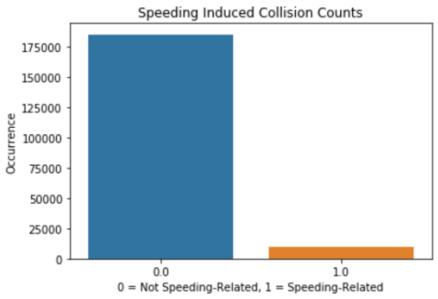
8. General Overview. The dataset consists 38 columns, and 194,674 rows of data. This translates to an average of 10,800 accidents per year. In the interest of this project, not all features are useful in the prediction of accident severity, and will be managed accordingly during the data preparation section.

9. Data Wrangling. Based on the 38 columns of data, a significant amount of it are present as strings, and will require conversion to numerical formats for correlation and model fitting. Data will significant amounts of missing fields will also be dropped as they will skew the outcome of the results.









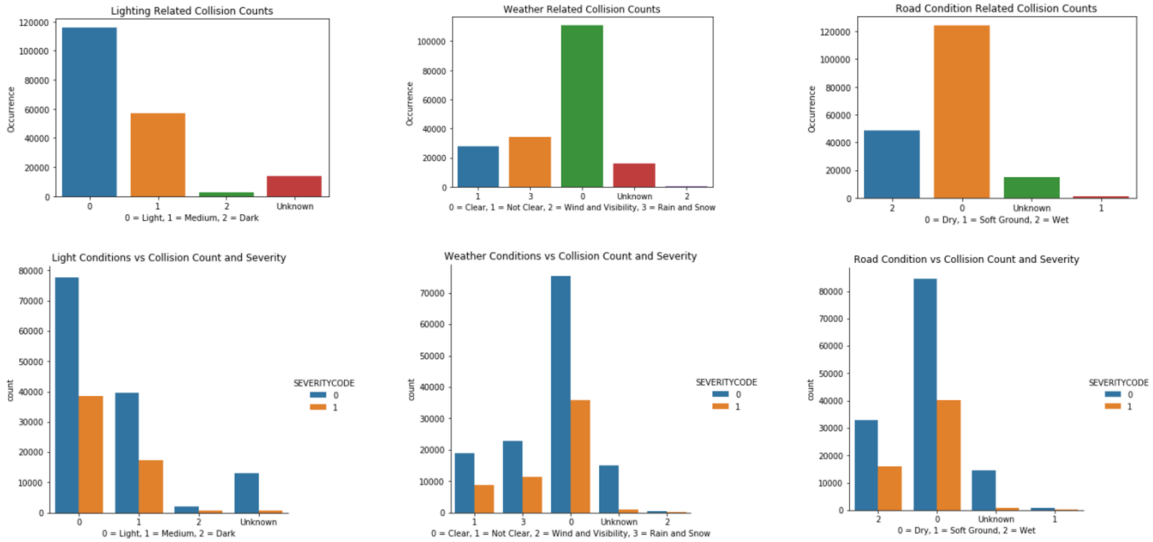
10. Fatality Counts against Contributing Factors. According to the four histograms plotted above, fatality count is significantly higher from attention-induced collisions than speeding and underinflation factors.

11. Collision Occurrences against Lighting, Weather and Road Conditions. The following observations could be made regarding the various contributing conditions.

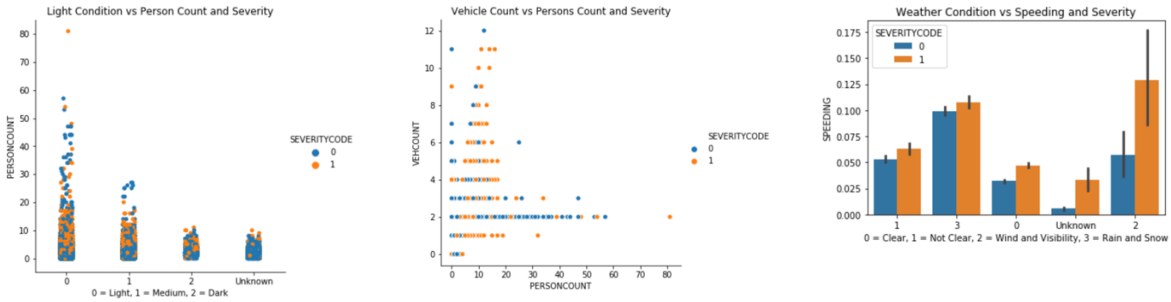
    a. Lighting Condition. Interestingly, most collisions happen during daylight hours, when lighting should be most optimum. Correspondingly, fatality rate was also higher during the day, contributing almost 50% of fatality for all daylight collisions. Collisions at night was significantly lower.

    b. Weather Condition. Most collisions happen when weather was clear, contrasting greatly with the number of collisions during foul weather.

    c. Road Condition. When road conditions were less than ideal, accident rates were much lower compared to when the road was dry.

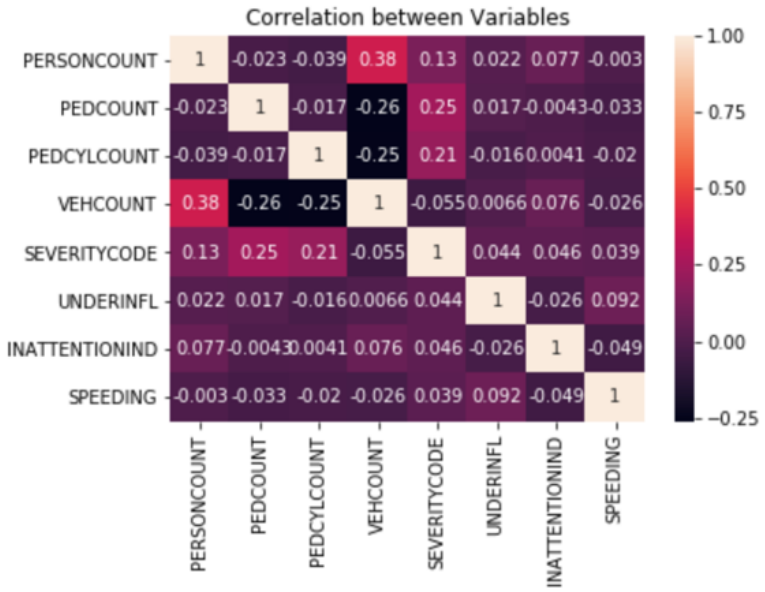


12. Severity of Collisions versus Multiple Variables. The accidents with the highest head count involved happened mostly in daylight hours, with vehicle counts hovering between 1 to 6. Speeding is a more significant contributor when road conditions are wet and soft, with a higher count of fatality.

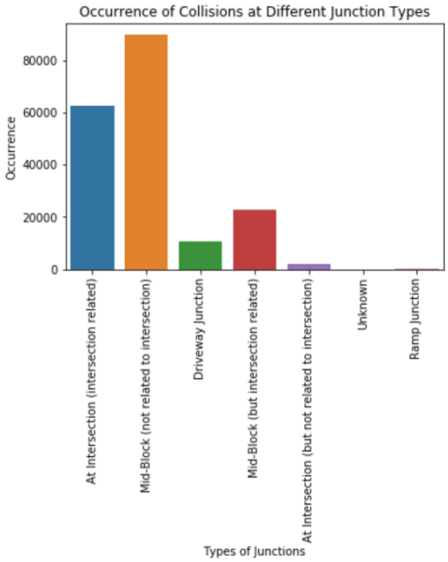


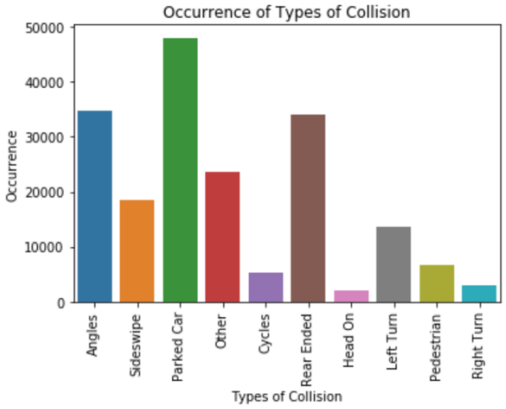
13. Correlation between Variables. Despite what seems to be a logical deduction, the data does not indicate strong correlation between the severity of accidents to the presence of inattention, underinflation, nor speeding. The strongest correlation between the variables

shown exist only with vehicle counts and person counts at 0.38, which is sensible, since more vehicles means more people involved.

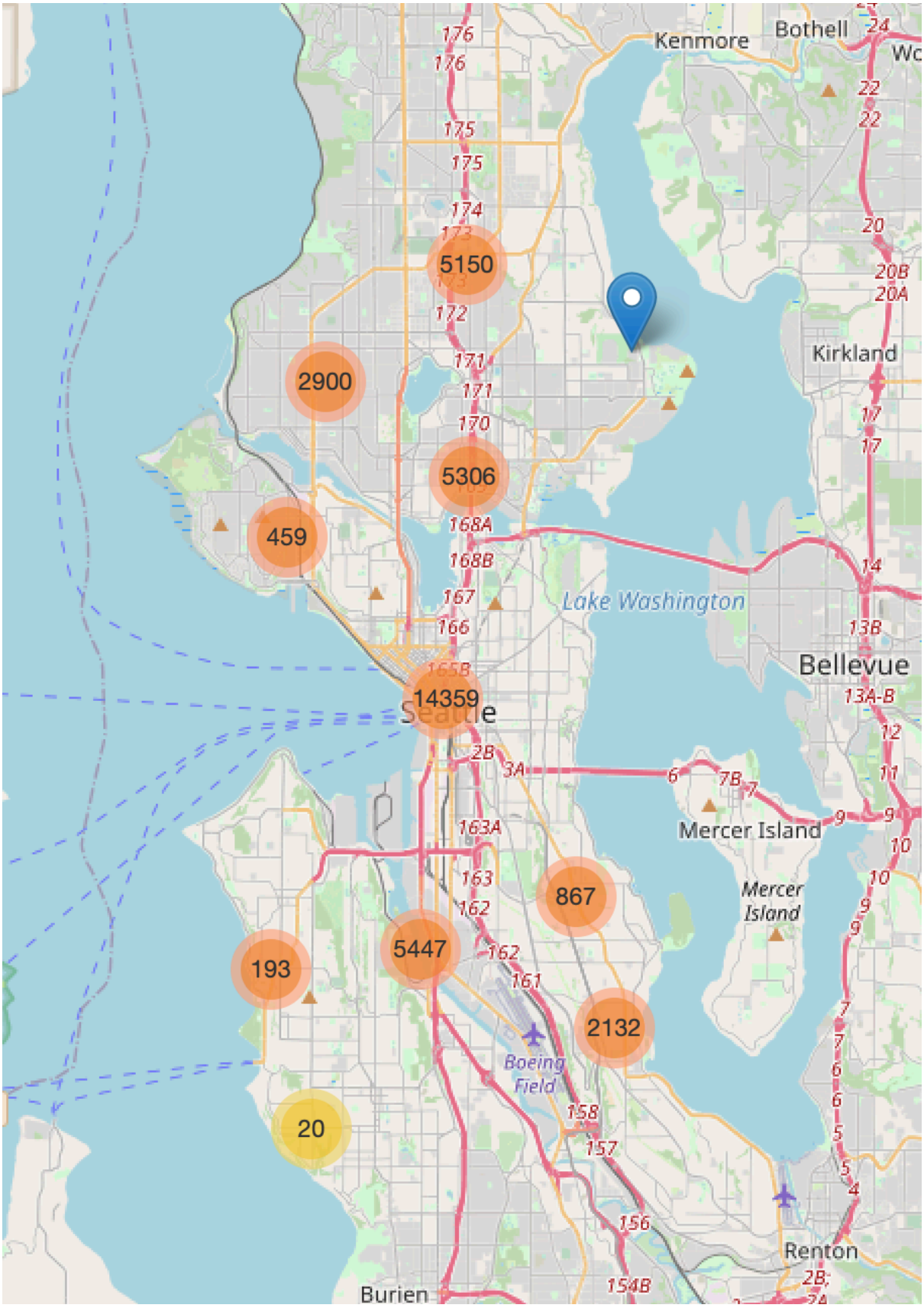


14. Categorical Visualisation of Data. From the three bar plots, most collisions happen at mid-block sections, with the highest occurrences involving parked cars.





15. Folium Mark Cluster Visualisation. Lastly, folium was used to visualise the occurrence of all 190,000 datasets on the map of Seattle to help build an appreciation of the locations of the collisions from 2004 till present.



16. It can be seen that a majority of the collisions occur along the freeway, which may be a consequence of the expected higher traffic volumes around those roads.