Coursera Capstone Project: Seattle Car Accident analyses 2004-2020

The City of Seattle through its Department of Transportation provides data free of charge to the public for all accidents and collisions since 2004. The purpose of this open data program has been to increase the safety of traffic, reduce accidents and increase the quality of living of its inhabitants.

The aim of the analyses is to first understand accident hotspots and their circumstances, then predict the severity of accidents in the future so that the City of Seattle can act upon it to decrease accidents and injuries. We will focus on 2004 to 2020 data.

The data set is rich in details with own geocoding, street names, severity degrees, the weather, light and road condition when the accidents happened. As it covers nearly 200'000 accidents, focus needs to be laid in the analyses as the City of Seattle and its inhabitants benefit more from insights into hotspots and key reasons to change behaviour or infrastructure on Seattle streets.

Downloading and Loading the Data: The table counts 194'673 accidents of different severity between 2003 and 2020, detailed in 37 + 1 index columns
Internal Police codes are meaningless for this analysis and will be dropped. "SEVERITYDESC" is a verbal repetition of SEVERITYCODE 1 or 2. After deep thought, we decided to drop PEDCYLCOUNTAll to focus on original accidents and the available SEVERITYCODE 1 and 2 descriptions for every accidents.

Incidents were inattention column was filled out occured 29'805 times incl NaN. Speeding was documented 9'333 times incl NaN.

1=injury; 2= property damage. Nearly twice as many injuries as property damages occured. For AI / ML purposes, the data set containing number of 1 and 2 severity codes will be balanced to a total of 58'188 each in the next step.

The dataset has now 116'376 rows equal to 2 times 58'188 values for severity code 1 and 2 incidents.

In most reported accidents, 2 vehicles were involved

In 109'836 reported accidents, no pedestrian was hurt.

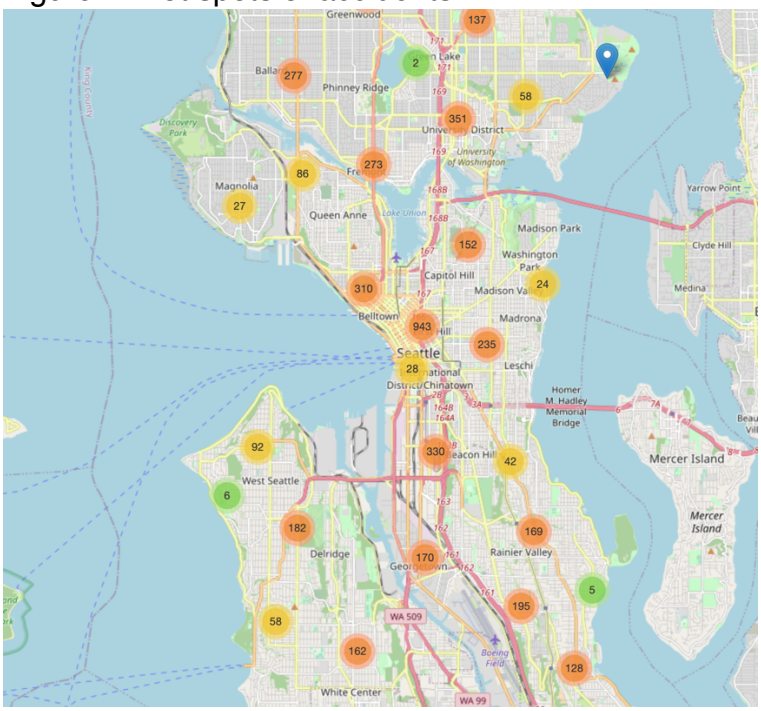
Surprisingly, most accidents happened on dry roads. Unfortunatly, in 6'931 cases no road condition was recorded.

67'856 accidents happened in clear weather conditions. Unfortunately, in 6'965 cases no weather data was recorded.

71'530 incidents happened during daylight.

There are clear "hot spots" where accidents happened, ie Battery St Tunnel, N Northgate Way, Aurora Ave, etc. This is demonstrated with the following graph overlaying the accident locations with the Seattle map.

Figure 1: Hot spots of accidents



The total of 116'376 incidents where spread across 19'634 spots in Seattle taking latitude data. There were 35 key spots with >= 75 incidents per degree of latitude across Seattle in the analysed timeframe.

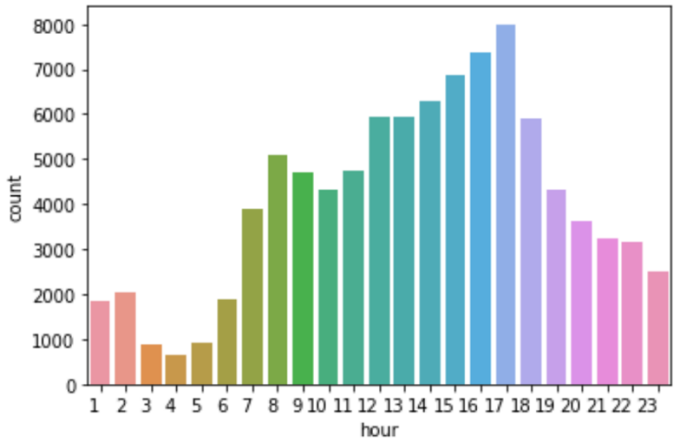
5'946 times alcohol or drugs were recorded in the balanced dataset.

Most accidents occured during day light. Approximately 1/3 in low light conditions

Speeding was reported in 6'048 incidents while the large majority were involved in accidents without prior speeding.

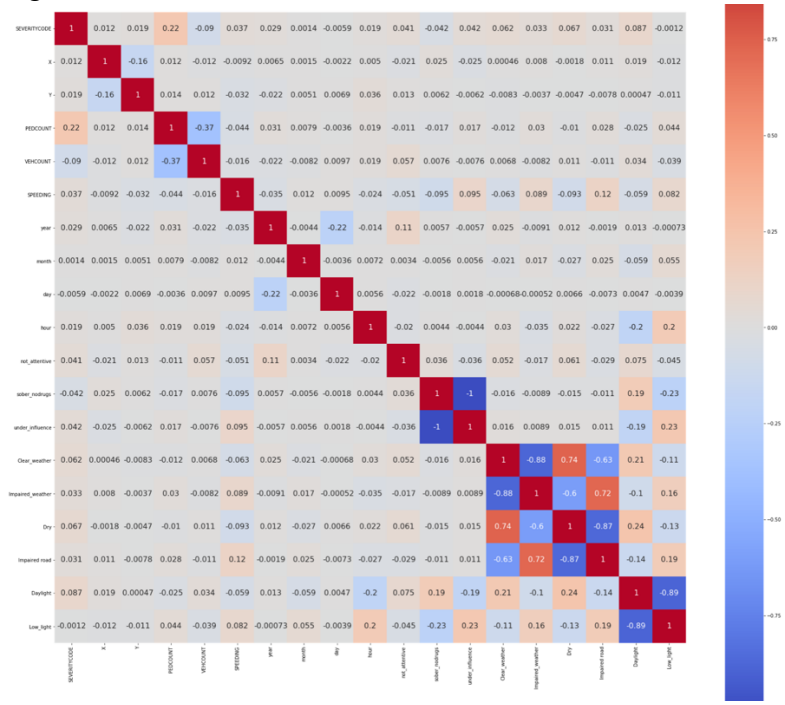
Most accidents occurred during rush hour between 4pm and 6pm

Figure 2: accidents by hour of day



Most attributes are solely weakly correlated with one another as the following graph shows.

Figure 3: correlation matrix



Due to sensitivity of correlation and Machine Learning algorithms used later on, all data get scaled with StandardScaler.

4 supervised learning methods were used to train and test the underlying datasets:
Logistic Regression
KNN
Decision tree classifier
Support vector model

In Conclusion:

The accuracy rates of the tested methods differed by a small range. Logistic regression achieved an accuracy of 0.578. KNN achieved 0.603 with n=7, the Decision Tree Classifier 0.604 and SVM achieved 0.579 after a long calculation time. In terms of precision to categorize injuries the SVM model was best with 0.61. The analysis showed that many factors contributed to the 116'376 incidents reported over 2004 to 2020. The model is ready to be applied to special focus areas such as locations with high number of incidents as we have seen earlier, detailed analysis by hour as most accidents happened during "rush hour" between 16:00h and 17:00h or in special weather conditions. Fortunately, accidents due to drivers tested for being under influence was small.