Coursera: Data Science Final Capstone, Draft Submission

Introduction and Business Problem

With an ever-increasing wave of technologies aimed towards enhancing vehicle safety, the most notable of which being: self-driving autonomous cars, increased vehicle crash-test regulations, and vehicle collision detection systems; the automotive industry has progressively worked towards increasing vehicle occupant safety and decreasing the likelihood and severity of traffic accidents.

However, society is still far from achieving universal road safety. Through the analysis of the extensive collision records of the Seattle Police Department (SPD), over the period of 2004-2020, the present data-science investigation seeks to analyse the variables which contribute to traffic accidents, alongside the changing incident trends over the recent years, identifying potential areas for improvement and proposing solutions to enhance road-safety in the streets of Seattle.

Through data analytics, it is evident that although vehicle collisions have followed a decreasing trend, from their all-time high of 2015, significant improvements are still required, as annually, a mean of 17,638 motorists have been involved in collisions. Such rates of collision have drastic impacts on health and vehicle insurance providers, consequences to pedestrians and local residents, and ends lives prematurely.

Thus, identifying the researches stakeholders as the SPD, the Seattle Transportation Office, vehicle insurance providers, politicians, and active community residents, the present study seeks to propose solutions which can be implemented and acted upon to ensure a reduced impact to the local area.

Data Overview

The data consists of an extensive account with 180,965 inputs over the course of a 16-year period. Crashes are reported in detail, with variables such as the time of collision, location, invidivuduals, vehicle types, lighting and road conditions being reported, amongst others.

The data possesses numerous variables which allow for interesting analysis. Thus, the analysis considered correlations between the type of collisions and junction types, alongside the prevalence of collisions across junction types. In addition, analysis of the influence of lighting, time of day, and road conditions were considered when analysing accident prevalence.

Lastly, machine learning models and conditional probability were used for modelling and analysing the dataset and prevalence of collision types.

Data Cleaning and Analysis

The data possessed a particular high-frequency of non-numerical inputs, as with regards to vehicle collision type and junction type are present in the dataset. Due to their non-numerical nature, the frequency of such occurrences can be modelled and correlated using co-occurrence matrices, this made it particularly challenging when analysing the data and applying machine learning methodologies.

In addition, particular challenges were faced when cleaning the data and time of the incident, as not all incident date and times were recorded following the same methodology, this led to significant complications when cleaning the data and preparing for analysis. Challenges were also found when plotting geospatial representations, requiring the use of a sample of the data, due to the sheer amounts of data.

Overall, numerous plotting tools were used, amongst which were box plots, bar charts, co-occurrence matrices, geospatial representations, and scatter plots. Below are preliminary data-visualisation examples.

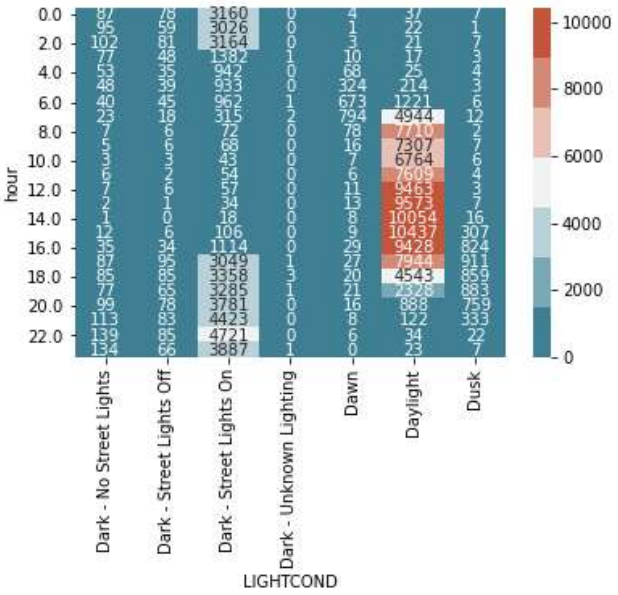

Figure 1 - Co-occurrence matrix correlating time of day and light condition at the time of vehicle collision. The data for the matrix addresses the entire data-set from 2004-2020

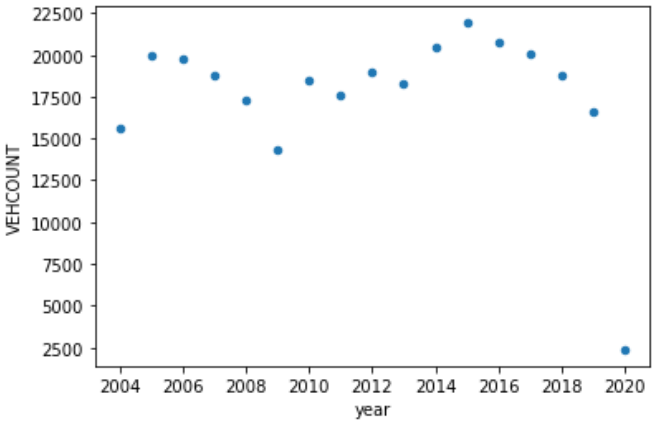

Figure 2 - Scatter plot indicating the total vehicle collisions per year and the trend from 2004 - 2020

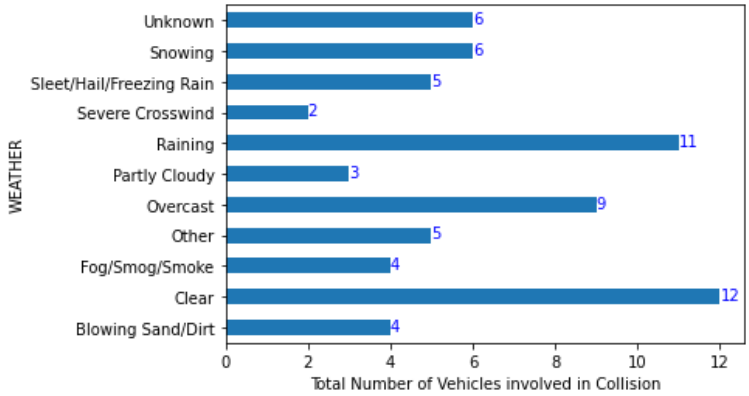

Figure 3 - Bar plot correlating the weather condition and maximum number of vehicles involved in the collision over the 2004 - 2020 period.