CIS 5930-001: Projects for Data Science

A2: Data Analysis

Fatema Tabassum Liza, Veera Ganesh Ponugoti & Sajib Biswas

For the first part of this assignment, we have selected a dataset from the NYC Open data repository and propose some analysis using a dataset on “Motor Vehicle Collisions & Crashes”. The dataset can be accessed using the following URL:

https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

Previously, we looked at some of the interesting properties of the crash data such as Zip Code, contributing factors to the crash, time-of-day, time-of-the-year, vehicle type etc. The data we have has a shape of (1000, 29) which means 1000 rows and 29 attributes. Some of the attributes have null values for all of the instances. These attributes are not very useful and they provide very less information. So we removed these attributes from the data in the feature selection step and focused on the attributes that provide significant information.

We want to find out what contributes to the most number of accidents, so we count the occurrences according to the contributing factors. We found out that **“Driver Inattention/Distraction”** is the most frequent cause for vehicle crashes in the city. Another interesting information can be the type of vehicles that are most frequently present in crashes. We can show that **“Sedan”**s are the most likely to be present in a road accident from this data.

For today's class, we have looked at more graphic visualization to get some insights. We have shown the Number of crashes for each vehicle type with the help of a simple line graph. We have also plotted the line graphs for the number of crashes for each month of the year and each hour of the day. From these graphs we can see during which month or time of the day it is more likely that a vehicle crash might occur. We see that the month of April contributes the highest number of crashes. Additionally we have also generated the graph for visualizing which time of the day contributes to the highest number of crashes. We can see that as evening approaches the number of crashes also increases. At 6PM the highest number of crashes occurs.

We have showed correlations between different columns of the dataset. We have generated visual representations that have given us more insight by combining information of two or more columns. From the graph we can see that the attributes “number_of_motorist_injured” and “number_of_persons_injured” are correlated most. So we have dropped one of the attributes since only one of the attributes is enough to provide the information. This way we can select important features for applying Machine Learning Methods further.