Human life is finite, precious and needs protection, unfortunately an average of 16 Americans for every 100,000 inhabitants lose theirs to road fatalities every year. This reality calls for concern and constitutes the propelling force and direct the course of this project. In an attempt to palliate this problem, New York City is used as a case study with the hope that the successes recorded will viewed as reference tools of all other States in United States.

New York City's demographics show that it is an ethnically diverse metropolis and most populous city of the United States with an estimated 2018 population of 8,398,748 and a density of 26,403 people per square mile. The high population density combined with the high number of commuting vehicles makes the New York population more vulnerable to road accidents. This study investigates the destructive effect of car crashes or collisions on daily activities of the population on one hand and economic losses on the other. This data was pulled from the New York City (NYC) Open Data website.

The Motor Vehicle Collisions crash data is made up of details on crash occurrences. Each row in the table represents a crash event. The Motor Vehicle Collisions data tables contain information from all police-reported motor vehicle collisions or crashes in NYC where police report MV104-AN is required for disclosure of whether someone is injured or killed, or at least $1000 worth of damage suffered.

This project attempts to harness the full potential of the The Motor Vehicle Collisions crash data sized at 337 MB, by investigating the level of vulnerability of the population of New York City in the United States of America and providing a trained model with a tuned hyper parameters using k-fold cross validation in order to improve performance and accurately predict the likelihood of a motor crash or collision occurrence for local, state or federal authorities to take proactive measures and prevent such occurrences.

Python 3 technologies, was used to acquire, explore and clean and analyse these data. While going through these data science life cycle steps, it was discovered that a total of 1859 persons were killed in crashes from January 01/01/2013 to December 12/31/2018 with the observation that passenger vehicles were the most involved in those incidents. The highest number of deaths per crash incident amounted to 8 (6 pedestrians and 2 cyclists) and was recorded on 10/31/2017 in the borough of MANHATTAN with the Zipcode of 10014 followed by QUEENS with the deaths sequence of 5, 4, 4 respectively recorded in the communities with the zipcodes of 11418, 11105 and 11354 respectively. The number of repeated events show the highest crash frequency in QUEENS borough and its communities, and make their population more exposed to crash-related vulnerabilities.

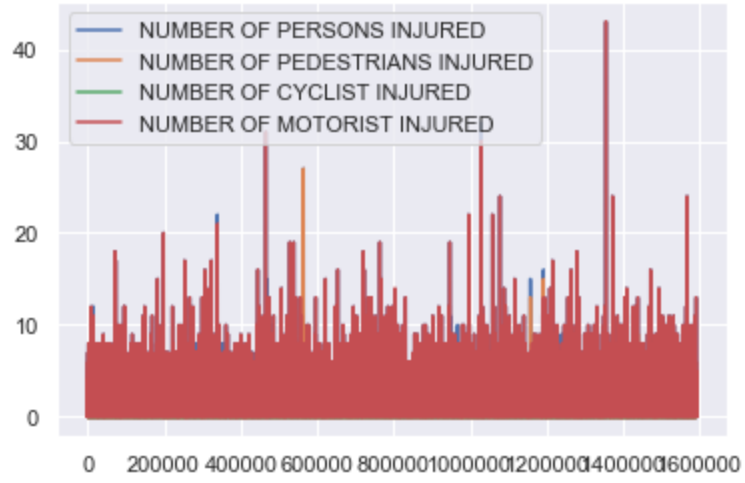
The highest number of pedestrians injured amounted to 27 and was reported on 05/18/2017 at 11:54 EST in MANHATTAN's community with the zipcode of 10036 followed by another record of 13 pedestrians and 2 motorists in the same community on 08/05/2014 at 15:26 EST, and 15 pedestrians with 1 motorist in QUEENS's community with the zipcode of 11103 on 06/28/2014 at 6:30 AM EST. It was also reported that 4 cyclists were injured on 07/19/2014 at 17:2 EST in BROOKLYN, zip code 11230 while 10 additional zip codes recorded 3 cyclists' injuries each in motor crashes.

From 01/01/2013 to 12/31/2018 a total of 417416 persons were injured while 1859 were killed in New York City by motor collisions of crash events, this averages the cumulative motor crash-related registered injured and dead victims statistics at 69569 and 310 individuals per annum respectively. The death statistics average of a six-year data acquisition period of 310 is greater than the NYC average of the same parameter of 222 and 200 in 2017 and 2018 respectively. While the downward trend of these cumulative statistics is noticeable, there is still a need for greater effort deployment to drive accident-related deaths to an abysmal level and we have a strong conviction that the proposed model will effectively address this need.

Plots

1-

Plotting of “NUMBER OF PERSONS INJURED”,”NUMBER OF PEDESTRIANS INJURED”, “NUMBER OF CYCLISTS INJURED”, “NUMBER OF MOTORISTS INJURED” in function of “COLLISION_ID”



2-

Plotting of “NUMBER OF PERSONS KILLED”, “NUMBER OF PEDESTRIANS KILLED”, “NUMBER OF CYCLIST KILLED” and “NUMBER OF MOTORIST KILLED” in function of “COLLISION_ID”.