Final Report

Natasha Arokium, Jiffar Abakoyas, Thierno Diallo, Bethold Owusu

Title

Level-of-crime classification based on NYPD arrest data in New York City in 2019

Abstract

In this project we tested different classification models to predict the level-of-crime for arrests made in NYC by the NYPD. We first examine, understand and analyze the arrest data in New York City for the year 2019, looking at where the highest level of crimes committed occurred as well as the numbers of felonies or misdemeanors. We identified factors that would more likely contribute to an arrest leading to a felony charge, such as location, age and gender of the arrested. The classifiers we implemented included logistic regression, decision-trees(DT), random forest(RF), k-nearest neighbours (KNN). Our performance was identical for all models, only showing a slight increase from a baseline accuracy. Afterwards, we focused on the LR model further and applied PCA on the features before retraining it. Although PCA reduced the features by 60%, accuracy remained the same. We conducted gridsearchCV to search for the best parameters. We also conducted Randomforest feature importance to find features of importance and used those for classification, without significant improvement in accuracy. Our results conclude that the level-of-crime is not predictable from the information about the perpetrator recorded at the time of arrest.

Introduction

New York City is ranked 25th for most violent crimes in the United States of America, it is imperative in the name of safety to use our knowledge as aspiring Data Scientists to delve into the data to see what can be unearthed in trying to understand the who's, how and why.

With a population of approximately 6,068,009 persons over 18 years old in New York, there were 213,089 arrests made in 2019 (representing 3.5 % of the population). This information is useful for understanding the bigger picture of how to measure the crime situation on a scale from 1 to 10.

Background

After undertaking the task of cleaning the data as a part of the exploratory data analysis process, varying visualization tools were created to highlight and present important findings and summaries for all features and data. These visualizations will range from showing several bar graphs and charts laced with very important and interesting findings, that will be used to answer fundamentals of how, why, when and who questions. In addition to examining and trying to make sense of the data, we will also test several machine learning algorithms to ascertain the one with the best performance for making our prediction on which level of offence (felony, misdemeanor) is more likely to occur given characteristics such as age, race, gender, borough, and precinct of the arrestee.

The Data

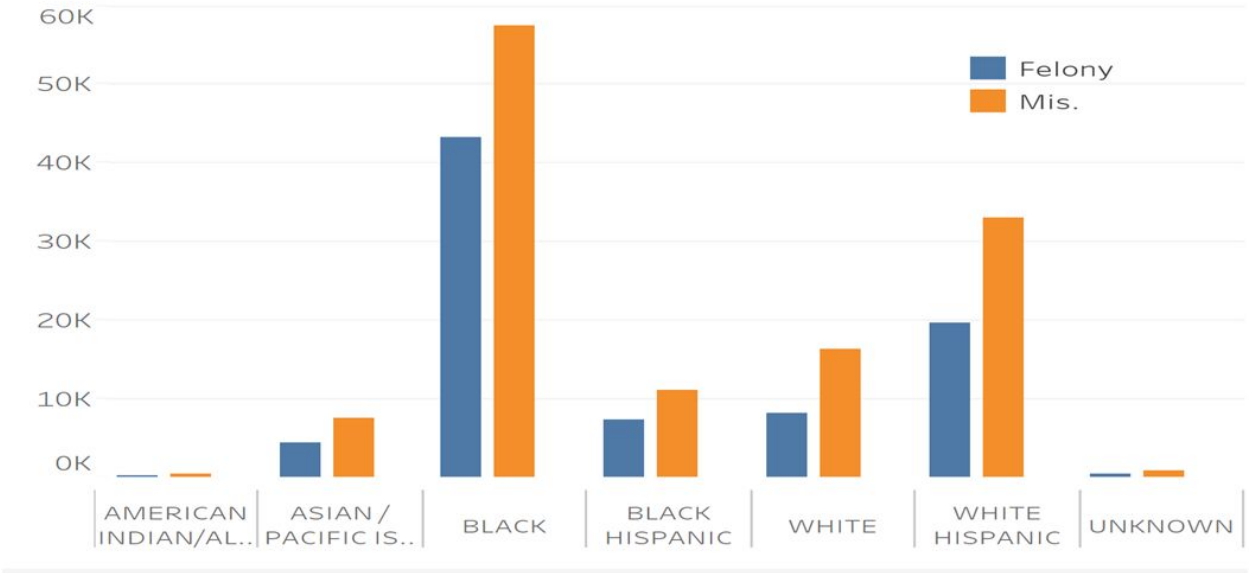
Each record represents an arrest affected in NYC by the NYPD and includes information about the type of crime, the location, date of the enforcement and the suspects demographics. There are 17 feature columns and the one target column[1]. The arrest data is collected from the five boroughs in New York, namely Brooklyn, the Bronx, Manhattan, Queens and Staten Island.

[1] https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc

Principal features for analysis

- Level of Offense (target feature)
- Borough of arrest
- Precinct where the arrest occurred
- Perpetrator's sex description
- Perpetrators' race description
- Date of the arrest
- Most frequently occurring arrests/locations on the map/month

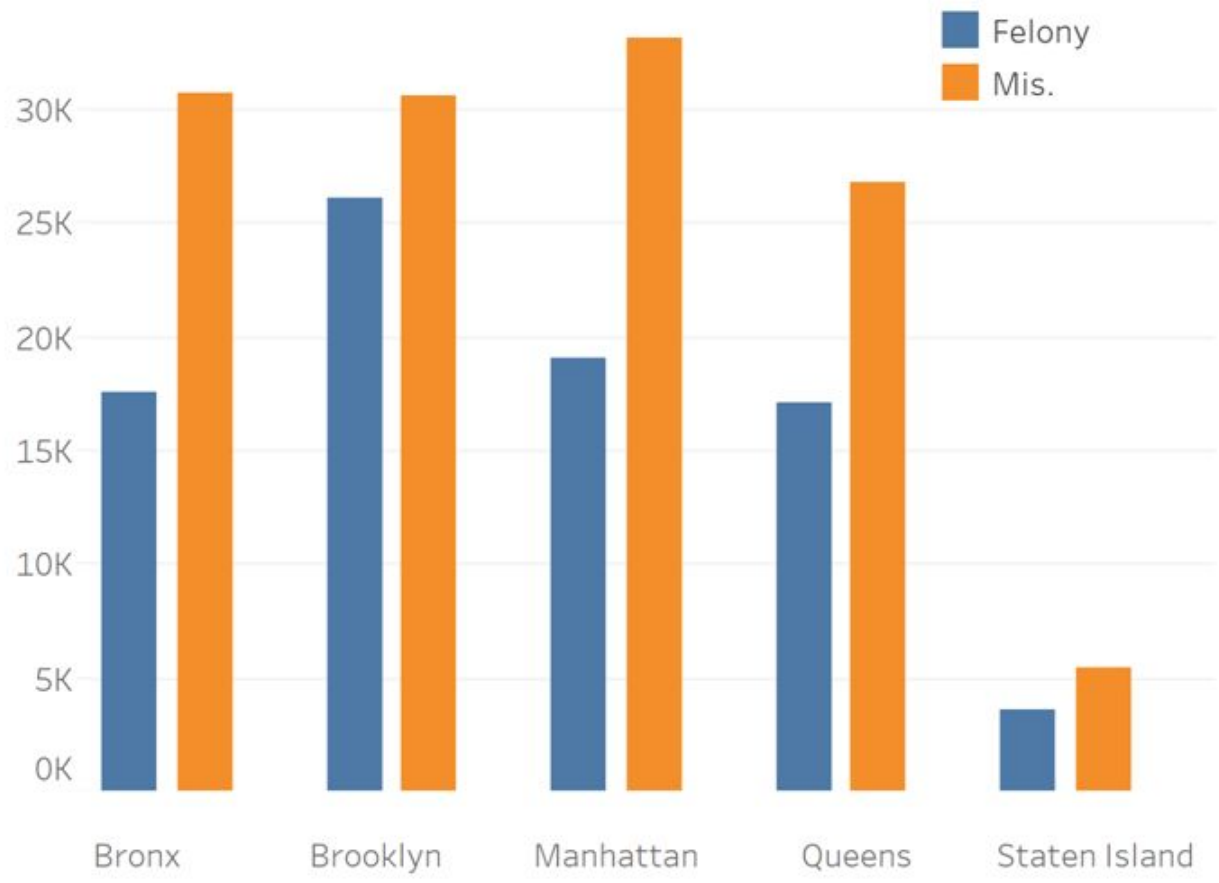
Exploratory Data Analysis



Above graph depicts black race with the highest Felony and Misdemeanor.

After normalizing the data to ensure that there were no biases or disparity between the proportions of race in conjunction with the Boroughs, it was found that 47.86 % of

blacks were arrested, followed by 25% white Hispanics during the year 2019. The borough of Brooklyn saw 27.25 % arrests, the highest numbers, with Manhattan following closely with 25 %, and the Bronx respectively with 23%. From the analysis Staten Island had the least amount of arrests, a mere 5%, which suggests that it was the safest borough during the year 2019.

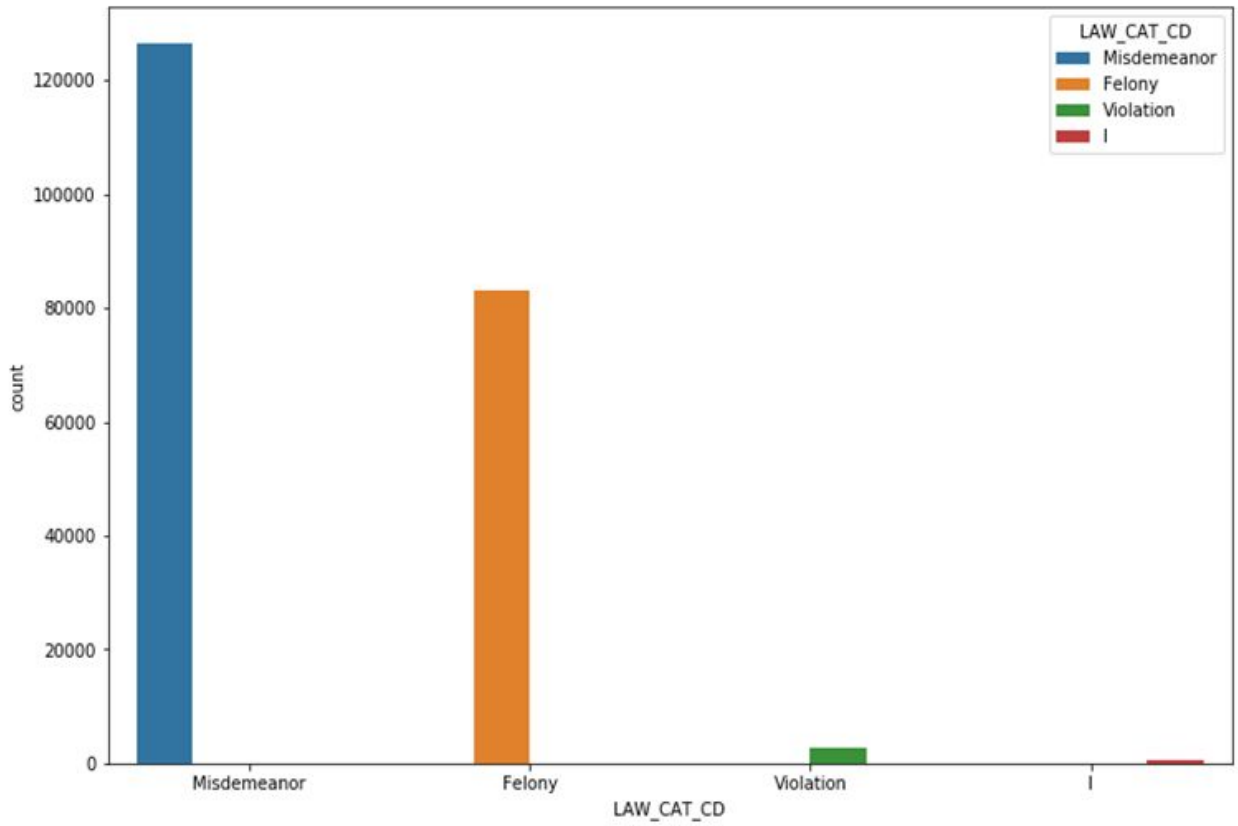


Above graph depicts the distribution of crimes across the boroughs.

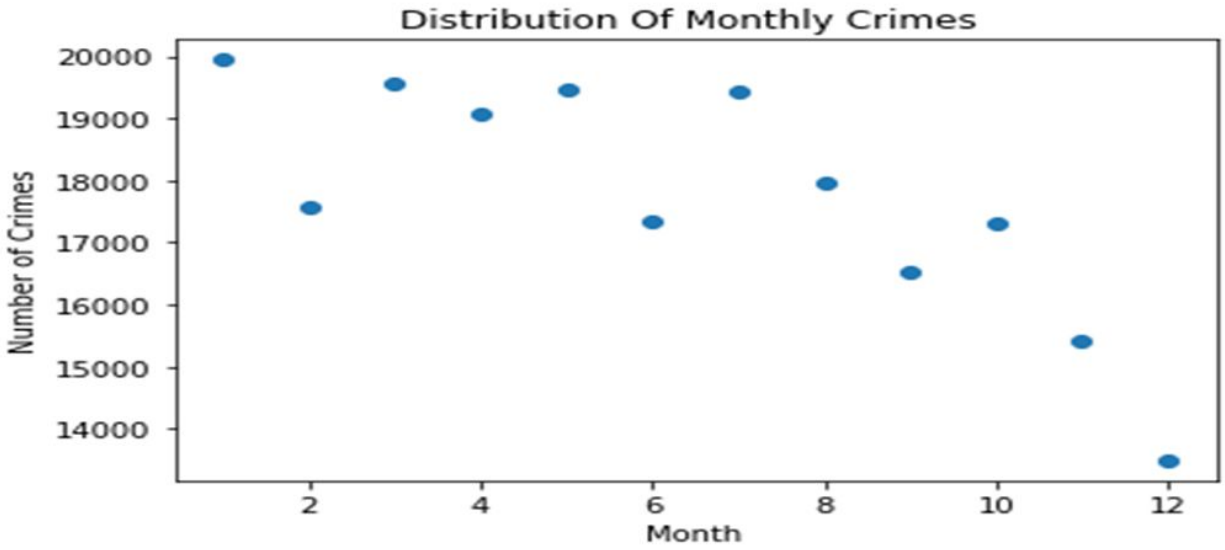
Preliminary analysis determined that Staten island had the least amount of crimes, and as such was deemed the safest borough in 2019. On the other hand, Brooklyn saw the highest number of arrests, with Misdemeanor having the highest ranking with reference to the level of offense.

Assaults and related offenses were the most frequently occurring level of offense, with Manhattan having the highest number of arrests in that category, followed by Brooklyn and Bronx. Overall, there were by far more misdemeanors (minor wrong doings), 59.41%, followed by felonies 39.07% (more violent crimes), then violations with 1.32%. To get a better understanding of why these disproportionate numbers came to be, a

wider range of data would have had to be collected, including education level, economic status, opportunity reach etc. in the relevant neighborhoods.



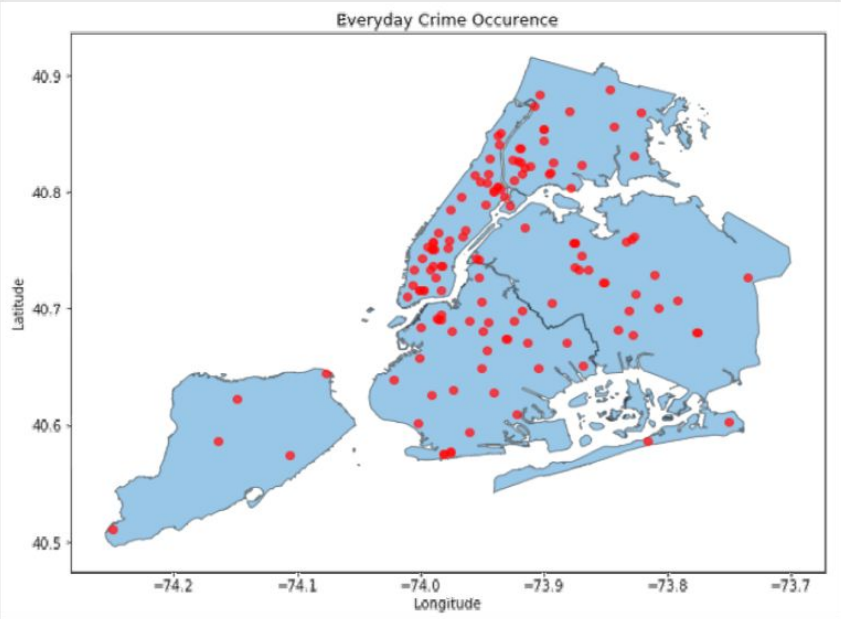
Above graph depicts distribution of crimes in our data. With Misdemeanor the highest.



Above graph depicts distribution of crimes across the year. January has the highest number of crimes.

Although the analysis shows there were more arrests in January, the worst single day in terms of number of arrests occurred on May 2nd, 2019, with 877 arrests. This finding was unexpected since one would expect more arrests to be made during the warm, sunny, and more comfortable months (June, July, August and early September), when more individuals are outdoors. Looking at weather.com data on that day, the day with the highest number of arrests, trying to find answers as to why, it showed that it was 77 degrees Fahrenheit and cloudy. It was however a more comfortable day overall compared to the previous days with 60, 59, and 55 degrees Fahrenheit respectively. In such a case, it is easy to assume that because of the reasonably lovely day, with all the excitement, more people came out, and more arrests were made.

Insightful Places for More Public Research



Algorithm to see where crimes occur daily

- ALBEE SQUARE (BK)
- QUEENS BOULEVARD EXPRESSWAY
- ROCKEFELLER CENTER
- FULTON MALL (BK)
- GATEWAY DR
- NOSTRAND SUBWAY STATION
- JFK AIRPORT
- TRANS MANHATTAN EXP
- BARTOW AVE (COOP CITY MALL)

The goal here was to find hotspots in our data where there were daily crimes. We implemented an algorithm to generate locations where crime occurred daily. We found there were 133 daily crime hotspots out of 120,000 unique locations in our 2019 crime data. The picture above shows the 133 locations where there are daily crimes. On the right depicts some popular locations.

This is very important because such locations can be brought to public and lawmakers' attention. This will call for researchers to further investigate why there are so many

crimes occurring daily in these regions. Thus, this will deepen our understanding in structuring our cities and ensuring its safety. Some research questions that could be further explored in the hotspot regions with daily crimes include education background, level of spirituality, aides for the individuals in these regions, policies governing these areas, history of these regions.

Methods

All features were categorical in nature[2] (although there were columns with number values, they were used as labels rather than a measurement), thus every column was one-hot-encoded before training with the classifier models.

We removed a number of columns for different reasons: the location columns were removed because they had significant variance (to a point of being almost unique) and thus would not have helped generalize. The crime descriptions were obviously removed because it would not be a prediction task to know the level of crime from its description. Other columns were removed for similarly obvious reasons that would give away the prediction task. We conducted feature extraction by taking the date column and extracting months out of it and one-hot-encoding it to twelve columns.

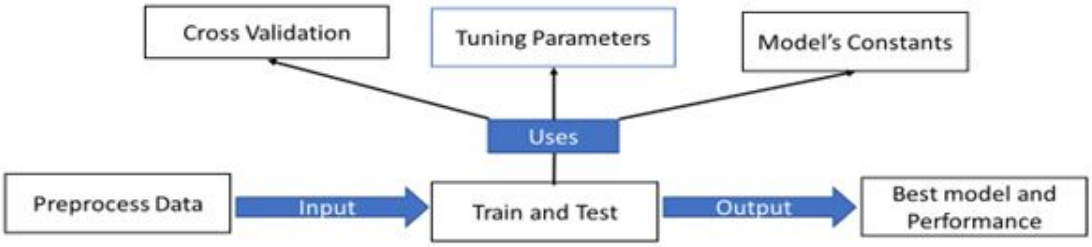
After preprocessing and one-hot-encoding, we had 126 columns.

Machine Learning Method

After each of us tried different ML classification methods, we built a classification workflow containing everything into a pipeline. See graph chart below.
This helps to just call one function to train and test and get the best model using cross validation. See the ML folder in the repository.

[2] https://data.cityofnewyork.us/api/views/uip8-fykc/files/62a746df-66ca-4603-aae4-46c02bac2972?download=true&filename=NYPD_Arrest_Incident_Level_Data_Footnotes.pdf

Our ML Classifier Workflow



1. User will preprocess the data to have a features data set and target column
2. User will call our train_test function by selecting model type
3. The function will call cross validation on the chosen model tuning parameters
4. User get back the best model trained for the optimal parameter and the performance on test set.

Results

After preprocessing the data by dropping rows with missing values and dropping columns such as code numbers and those with no correlation to our target values, we proceeded in implementing a couple of machine learning models to classify whether our target is Felony or Misdemeanor. We initially did one-hot encoding to the features (age, gender, borough, and race) and PCA, which reduced the one-hot encoding features from 126 columns to 45 columns. The machine learning models we implemented include Logistic Regression, Support Vector Machine, KNN and Random Forest Classifiers.

The Logistic Regression algorithm with 61% precision, 62% recall, and 57% F1 score was the best performing algorithm that predicted whether the level of crime was that of a misdemeanor or that of a felony, trained and tested with the features, age group, race, sex, borough of the arrested and precinct where the arrest occurred.

Dummy classifier Validation accuracy: 0.5235896824924048

	precision	recall	f1-score	support
Felony	0.40	0.40	0.40	13447
Misdemeanor	0.60	0.61	0.60	20127
avg / total	0.52	0.52	0.52	33574

Logistic Validation accuracy: 0.6207481980103652

	precision	recall	f1-score	support
Felony	0.57	0.21	0.31	13447
Misdemeanor	0.63	0.89	0.74	20127
avg / total	0.61	0.62	0.57	33574

Logistic Validation accuracy: 0.6174718532197534

	precision	recall	f1-score	support
Felony	0.57	0.18	0.27	13447
Misdemeanor	0.62	0.91	0.74	20127
avg / total	0.60	0.62	0.55	33574

Project Organization (see repository)

We organized our project as follows:

1. A EDA and ML Analysis Notebooks folder that contains the analysis done by each members (EDA and ML)
2. A ML folder that contains libraries and a pipeline to train and test a classifier (this is a combination of the models that each member tried in his/her ML analysis)
3. A Presentation and Final Report folder that contains the presentation and final report.
4. A Proposal folder that contains the project's proposal

Conclusion

All the classifiers we implemented could not perform well in predicting whether a target is Felony or Misdemeanor for both the training and validation sets. This is actually a good response because it is telling us that arrest is not based on one's gender, race, age group and borough. Specifically, if anyone trespasses the law to the point where he needs to be arrested, he will be arrested regardless of his race, gender, age group and the borough he lives in. This indicates that the NYPD is striving for a non-racist approach in the arrest procedures they conduct daily.

Team Contributions

Natasha Arokium (15 - 20 hours of work)

1. Created github account for the project
2. Wrote the 1st draft for the Proposal, title, abstract and Introduction and data for the final report.
3. Performed complete EDA on the data, including cleaning, and showing visualizations with approximately 21 different analyses.
4. Used K-means clustering to group data points together that minimizes differences between the data points in the same groups (Level of offence)
5. Prepared final updated Presentation slides in Powerpoint

Jiffar Abakoyas (35-hours: 15-hours coding, 5-hours organizing, 5-hours on presenting, 5-hours writing final report, 5-hours in group meetings)

1.Researched potential machine learning models and available features.
2. Wrote on methodology of proposal for different models.
3.Trained and tested different ML models (Logistic, KNN, SVC, Decision Tree) with and without PCA.
4. Documented results in notebooks for level of crime.
5. Contributed to the final report structure, methods.

Thierno Diallo(10-15 hours**)**

1. I created a notebook called thierno_analysis to analyse the data.
2. I used tableau to do visualization that we used in our presentation.
3. I attended three working sessions in zoom where we discussed how to tackle the project.
4. Then I created a ML pipeline to train and test models. The ML folder contains four essential .py files.
    a. Model_constants.py contains list of models and their tuning parameters
    b. Training_and_testing.py contains functions to use cross validation to train and pick the best model.
    c. Preprocess.py contains functions to preprocess data and return features data and a target column.
    d. Main.py to run the pipeline.

5. I used the ML pipeline to train a Random Forest model, a Logistic Regression, and a Support Vector Machine model.
6. Help to write the final report.

The ML can be used as a library to quickly find the best parameters for a chosen model.

Bethold Owusu (20-25 hours)

1. In parallel, I also did EDA in my repository, cleaning up data, dealing with missing values, grouping counts of crimes per borough,race and age
2. I used various Machine models such as Logistic Regression, KNN, SVC and Random Forest to predict race and also to predict Crime level after label encoding them
3. I wrote couple of scripts in my repository to help me perform quick EDA
4. I further investigated into the data by using Geopandas to map crimes on NY boroughs
5. I participated in all meetings as a team
6. I wrote an algorithm to detect locations where crimes occurs daily
7. I contributed to the final report