NYC Crime Analysis

Greg Kimatov, Mehedi Shohag, Zeal Patel

Introduction

The New York City Police Department (NYPD) is the largest and one of the oldest police departments in the United States. Like many major cities in the US, the NYPD provides access to non-personally identifiable data that helps researchers learn more about trends and patterns in criminal activity. In this blog post, we will discuss our analysis of the crime rate in New York City. The city is made up of five boroughs: Manhattan, Brooklyn, Bronx, Queens, and Staten Island. We will explore a dataset that includes all arrests made in NYC between 2006 and 2019. This data was collected through the NYC Open Data website and can be found on Kaggle.

The motivation behind exploring this dataset is to see if it has become more or less dangerous to live in NYC. Our goal is to predict the NYC crime rate in the near future by analyzing the crime rate trend over a recent 14-year time period.

About the datasets

The primary dataset we are going to use is called “NYC Crime Stats”. We took the data from Kaggle, and they acknowledge the data was collected from NYC Open Data. This dataset contains data collected through thousands of calls to the NYC Open Data platform between 2006 - 2019. We chose this dataset because as NYC residents, we want to see how safe our city is for tourists and residents alike.

The dataset has a total of 19 columns. However, we dropped some auto-computed columns. Please refer to the data dictionary below to get a better idea of what each column in the dataset means.

Column	Description
arrest_key	Randomly generated id for each arrest
arrest_darte	Exact date of arrest of reported event
pd_cd (specific classification)	Three digit internal classification code
pd_desc	Description of internal classification code
ky_cd (general classification)	Three digit internal classification code
ofns_desc	Description of internal classification
law_code	Law code charges corresponding to NYS penal law
law_cat_code	Level of offense: felony, misdemeanor, violation
arrest_boro	Borough of arrest
age_group	Perpetrator’s age within a category
perp_sex	Perpetrator’s sex description

perp_race	Perpetrator’s race
latitude	Latitude coordinate
longitude	Longitude coordinate

Supplementary datasets

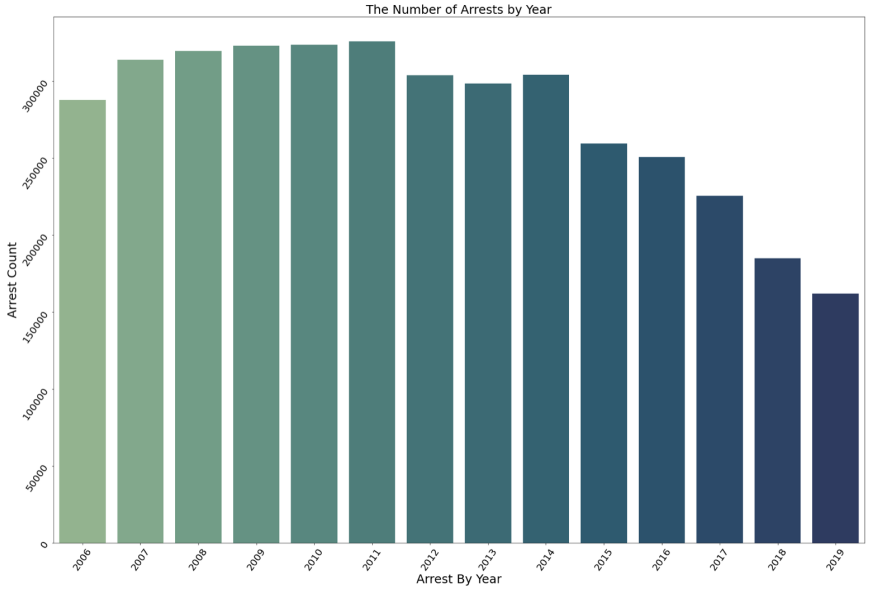
In addition to the crime dataset, we used two additional datasets on NYC Unemployment and annual percentage change in NYC GDP. The unemployment dataset contains data from the New York State Labor Department on the monthly unemployment rates across NYC. This dataset has unemployment rates from 1976 to 2021. The GDP dataset shows as a percentage by how much the GDP changed from the previous year within a certain county as well as across all 5 boroughs in the city. These datasets allowed us to use monthly unemployment rates and annual percentage change in NYC GDP as features in two linear regression models.

Visualizing the data

The cleaned dataset has approximately 3.89 million entries, which accounts for arrests made between 2006 and 2019. To make visualizations of the dataset, we used the seaborn library.

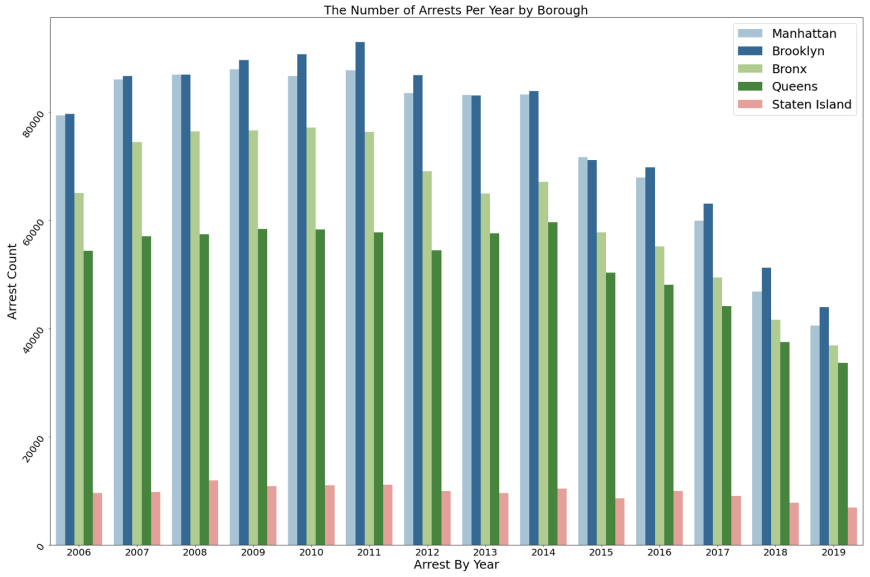
Number of arrests per year in the city

The graph below shows the total number of arrests across all five boroughs of NYC per year.



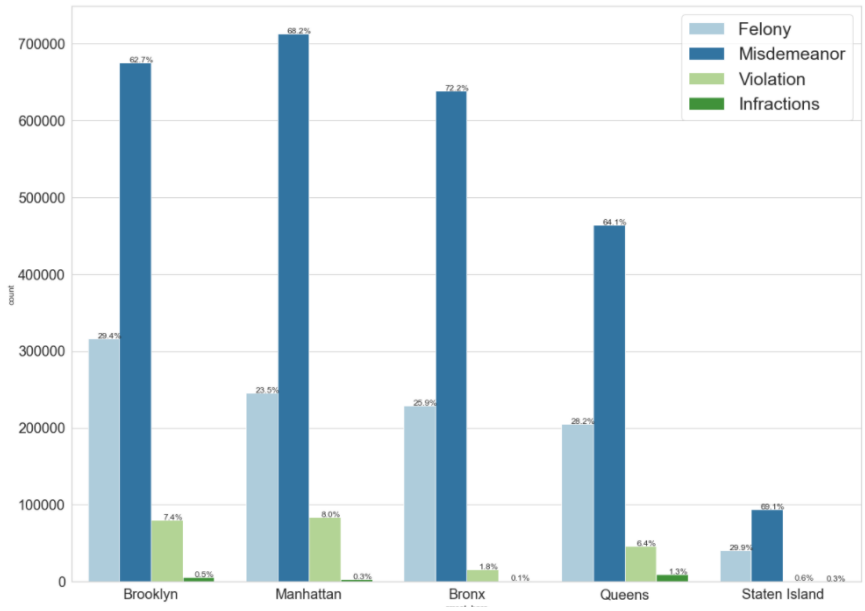
Number of arrests per year by borough

The graph below shows the number of arrests in each borough between 2006 and 2019. As we can see, in every year from 2006 to 2019, the arrest count is consistently highest in Brooklyn, followed by Manhattan, Bronx, Queens, and Staten Island.



Number of arrests by borough and offense level

Offense levels are classified into four different classes: misdemeanor, felony, violation, and infraction. As you can see in the graph below, the most common offense level is misdemeanor, followed by felony, violation, and infraction.



Number of arrests by offense type in each borough

There are 98 different types of offenses that were categorized out of 3.8 million arrests. Some of them are “Dangerous drug violations, Petit larceny, Felony assault, and Dangerous weapons.” The following offense types occurred most frequently across all 5 boroughs:

Offense Type	Arrest Count
DANGEROUS DRUGS	835598
ASSAULT 3 & RELATED OFFENSES	380510
OTHER OFFENSES RELATED TO THEFT	235898
OTHER STATE LAWS	187193
PETIT LARCENY	162330
VEHICLE AND TRAFFIC LAWS	154971
CRIMINAL TRESPASS	154807
FELONY ASSAULT	152031
DANGEROUS WEAPONS	148065
MISCELLANEOUS PENAL LAW	136695

These are the crimes that occurred most frequently within each of the 5 boroughs:

Manhattan:

Offense Type	Arrest Count
DANGEROUS DRUGS	195813
OTHER OFFENSES RELATED TO THEFT	89836
ASSAULT 3 & RELATED OFFENSES	71066
OTHER STATE LAWS	69289
PETIT LARCENY	59496

Brooklyn:

Offense Type	Arrest Count
DANGEROUS DRUGS	253454
ASSAULT 3 & RELATED OFFENSES	112091
OTHER STATE LAWS	65784
OTHER OFFENSES RELATED TO THEFT	60755
MISCELLANEOUS PENAL LAW	49260

Queens:

Offense Type	Arrest Count
DANGEROUS DRUGS	112127
ASSAULT 3 & RELATED OFFENSES	81506
OTHER TRAFFIC INFRACTION	38577
VEHICLE AND TRAFFIC LAWS	36655
OTHER STATE LAWS	35516

Bronx:

Offense Type	Arrest Count
DANGEROUS DRUGS	241916
ASSAULT 3 & RELATED OFFENSES	100271
OTHER OFFENSES RELATED TO THEFT	59422
CRIMINAL TRESPASS	55664
DANGEROUS WEAPONS	41153

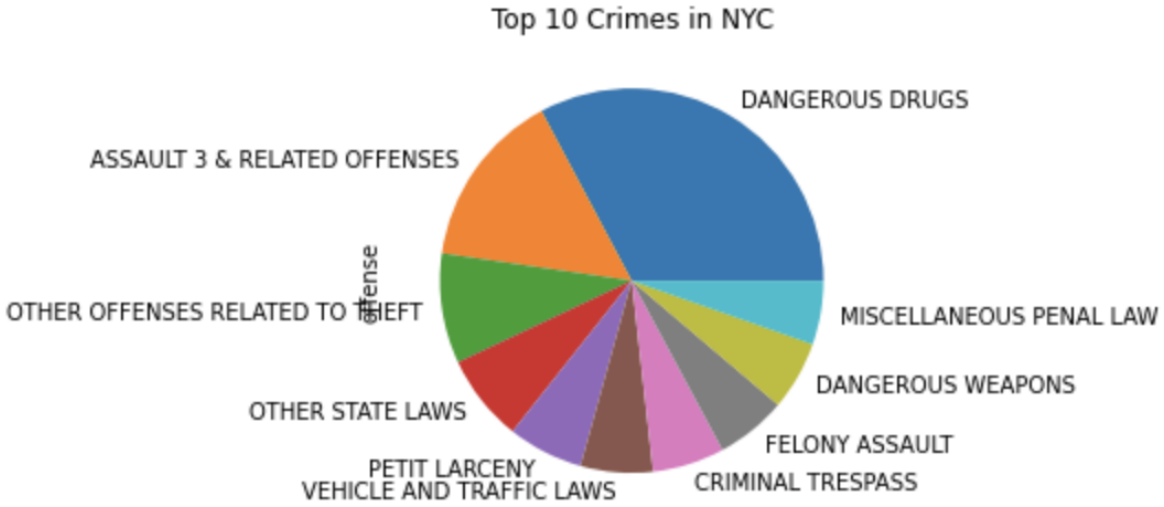
Staten Island:

Offense Type	Arrest Count
DANGEROUS DRUGS	32288
ASSAULT 3 & RELATED OFFENSES	15576
PETIT LARCENY	7300
VEHICLE AND TRAFFIC LAWS	7262
CRIMINAL MISCHIEF & RELATED OFFENSES	6333

In every borough, if we are to assume that arrest count is an indicator of criminal activity, drug-related crimes are most common. Crimes related to theft and assault 3 are the second most common offense types in Manhattan, Brooklyn, and the Bronx. Following drug-related crimes, traffic and driving violations are the most common crimes in Queens and Staten Island, which are not serious offenses. In the Bronx, the most significant crimes that occurred involved drugs, dangerous weapons, criminal trespass, and assault. Our findings lead us to conclude that the safest boroughs are Queens and Staten Island, and the most dangerous borough is the Bronx.

Top 10 types of offenses in NYC

The pie chart visualization shows the top 10 crimes committed in all of NYC. As you can see, “dangerous drugs”, “assault 3 & related offenses”, and “other offenses related to theft” were the top 3 types of offenses for which NYC residents were arrested.



Predicting crime rate in NYC

In this section, we will be looking at a couple of different models that attempt to predict what the crime rate (or arrest count) will be in the future by using the past 14 years' data.

Predicting the future yearly crime rate.

We choose to use the Linear Regression model to predict the future yearly crime rate. First, we calculated the yearly crime rate by accumulating the total number of arrests made in a year, then divided it by the population of NYC according to 2010 census data. Finally, we multiplied it by 100,000 to get the crime rate per 100,000 people. We did this for each year between 2006 and 2019 which produced the target. The corresponding year will be the feature.

Once we had the data, we split it into training and test data where 25% of the data is used to evaluate the model. After training the model, we get an accuracy of about 83%.

An accuracy of 83% looks relatively good so when we tried to predict the crime rate in 2030, we got approximately 1072 crimes per 100,000 people in New York. However, this is not a very good model as it does not take any other variables into account, and eventually, the model will hit 0. In fact, according to the model, by 2038 the predicted crime rate will be 31. This is highly unlikely, therefore in the next two models we looked at the unemployment rate & annual GDP changes to increase model accuracy

Predicting the future crime rate based on unemployment rate

We choose to use the Linear Regression model to predict the future crime rate. And for that, we are using the arrest date by month and the unemployment rate of that individual month for the features, and the target value is the number of arrests by month which implies the number of crime occurrences in each month. To get the unemployment rate data for NYC we needed an

additional dataset which we got from the New York state labor department then we did some cleaning to extract the features and the target values.

Then after we did the cleaning to get our expected features and target value columns we created the Linear Regression model. We have split our data for training purposes 80% of the data will use to train the model and 20% of the data to evaluate. And we've achieved an accuracy of 63%.

After we made our model we tried to predict some future predictions such as if the unemployment rate went up to 88% what will be the crime rate in 2045 and the result came 154920 crime in the whole year and if the unemployment rate is 1% then there will be 55668 crime in the whole year.

And by the result, we can conclude that the unemployment rate has a significant effect on the crime rate. If the unemployment rises the crime rate increases and if the unemployment decreases the crime rate also decreases.

Predicting the future crime rate based on annual GDP percent change in NYC

In a similar vein to the chosen features of the last linear regression model, we set the number of arrests as the target value, and set year and annual percent change in the GDP as our model's features. This allowed us to see how crime is influenced by the local GDP of a borough or city.

After doing some cleaning on the dataset to get the expected features and target value columns, we created a linear regression model. We allocated 80% of our data for training purposes and 20% of the data for testing. The result is an accuracy of 67%.

Future work

In the future, we can collect additional data to keep our datasets up to date since we were only able to collect crime data until 2019. We might also add additional data from before 2006 to have more data points for analysis purposes. Besides our primary data sets, the supplementary datasets that we used didn't have much data like "NYC Crime Stats". So, we will try to collect more data about unemployment and GDP. In addition, we will try to create different types of models rather

than just linear regression. That's how we will be able to see which type of model is working best for specific features.

Final Conclusions

Our exploratory analysis and observation of model results led us to the conclusion that crime has been consistently decreasing since 2011. Additionally, with a reasonable amount of accuracy, we can say that crime in NYC increases with higher unemployment rates and decreases with higher GDP.

One of us lives in the Bronx, and before moving there, I used to frequently hear that the Bronx is a dangerous borough, but I always wondered why people would say that. After analyzing the data, we were able to discover the answer. The high crime rate and the proportion of crimes that are violent and/or dangerous in the Bronx scare people away from the borough. This, in turn, likely causes fewer tourists to visit the borough, contributing less to the local economies of these communities, and ultimately causing their local GDP to decline over time. The result is often a never-ending cycle of crime in an area.