Project 5

LUIS SALINAS,
JOHN LAWLESS,
JOE ROSENBLUM,
TARA CELESTA,
LINH UNDERWOOD

Problem Statement

The costs and effects of crime touches everyone to some degree. With the increased wave of crime seen within the city of NYC, the city is trying to reduce these costs and effects for its population. They hypothesize that different community programs might help, however they do not know which programs will create the most positive impact.

We have been hired to analyze NYC population demographics and crime data to provide helpful information for the city so they can better examine the social impact of crime in their city. Knowing this social impact, will help them better determine how to allocate their resources amongst these community programs and create a healthier NYC.

Data Acquisition

NYC Crime 2018-2019

- Source: NYC OpenData

ZIP Codes

- Source: NYC OpenData SHAPE FILE

Neighborhood

- Source: NYC Dept of City Planning SHAPE FILE

Precinct Shapefile

Population/ Income:

- Model: Data World(2013)
- ArcGIS: Enrich (2020)

Data Frame Size: (916841, 30)

Methodology

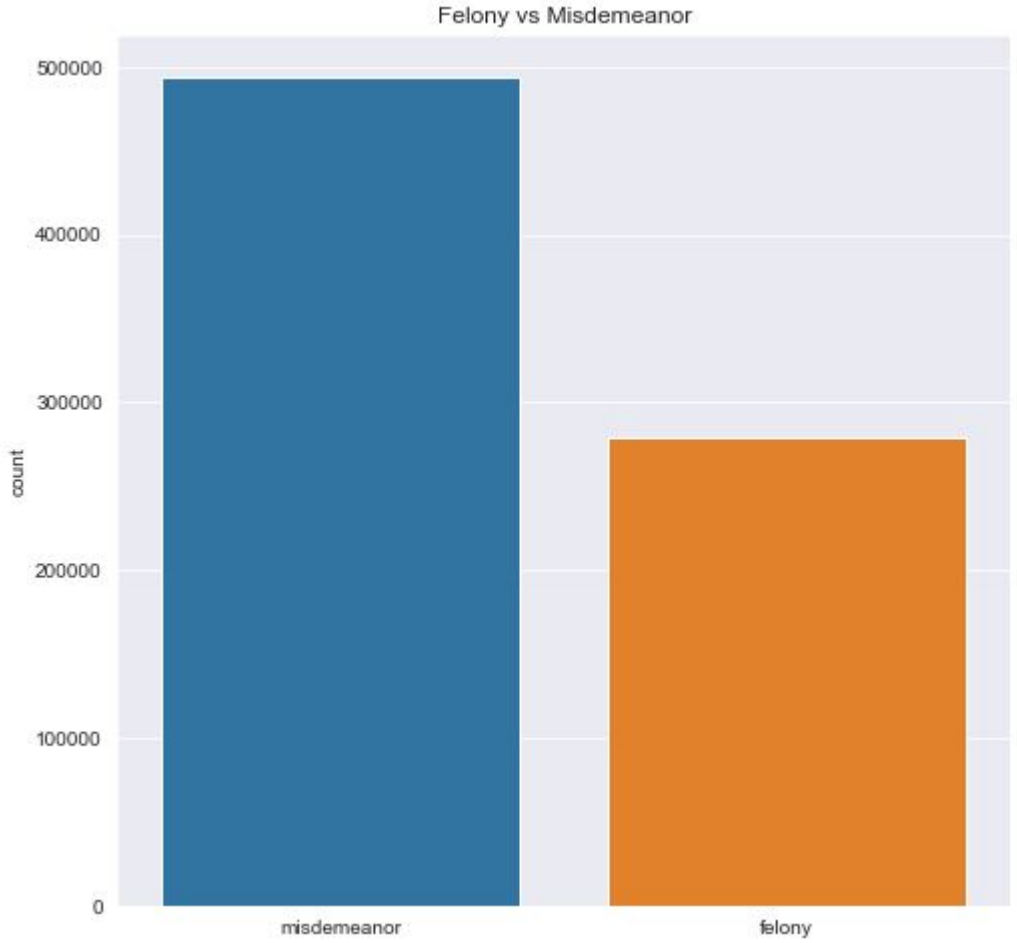
Data Preprocessing:

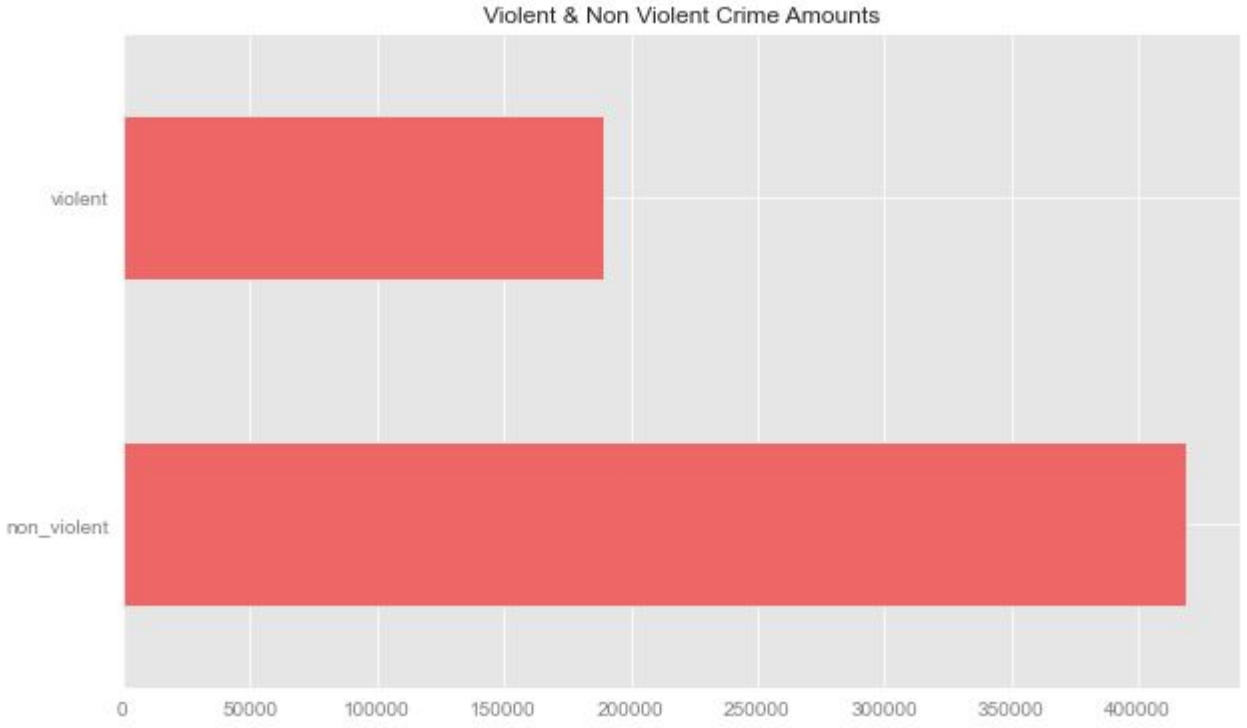
- Combined zip codes, boros, arrest, and income with NYC Complaints
- Eliminated unneeded columns and nulls

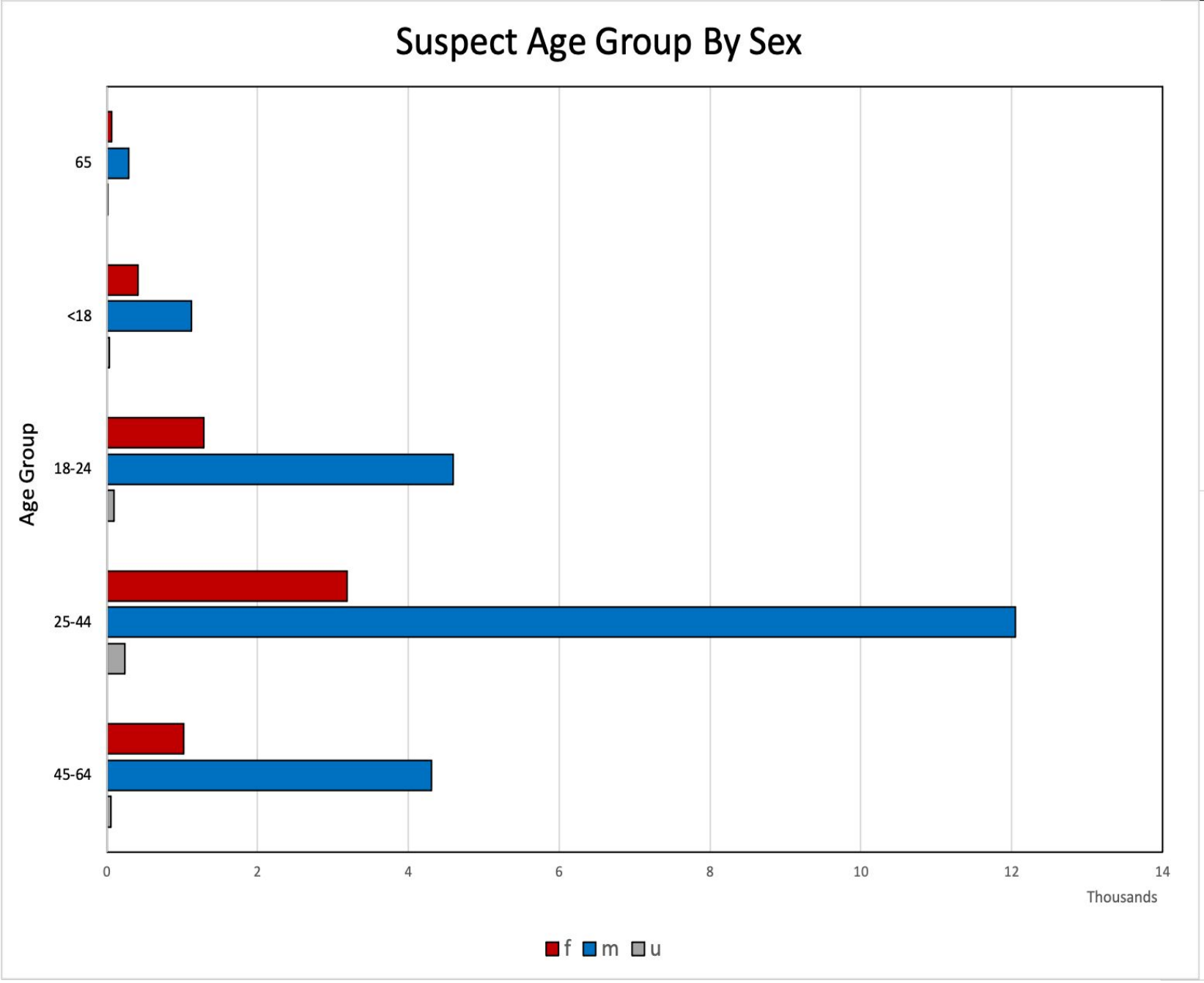
EDA

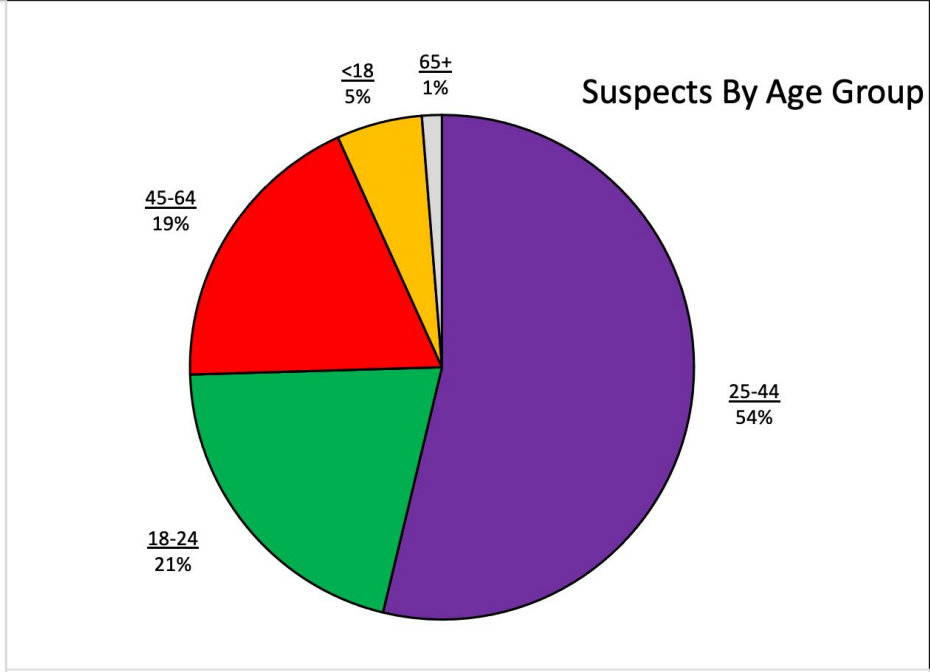
- Demographics: Population, Income
- Time, and Crime Frequency
- Crime by Precinct
- Offense Type

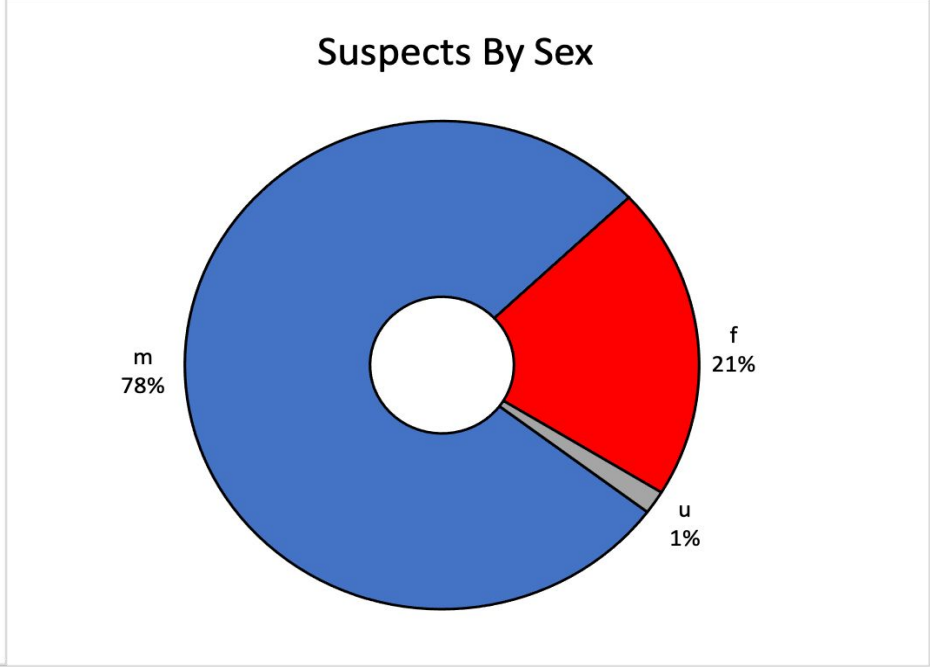
NYC Crime Overview





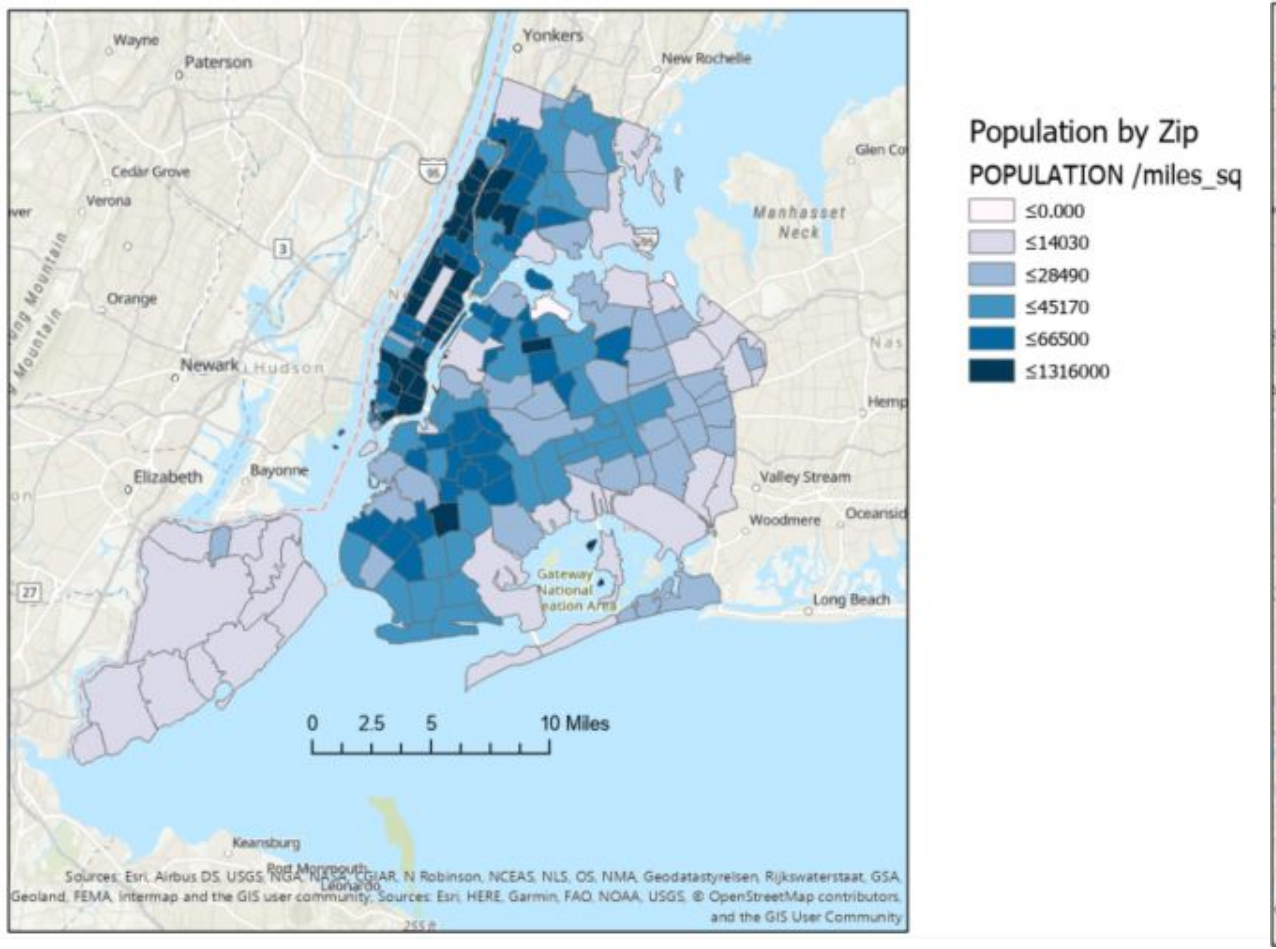

Suspect Age Group By Sex
Age Group
65
<18
18-24
25-44
45-64
0
2
4
6
8
10
12
14
Thousands
f
m
u


Suspects By Age Group
<18
5%
65+
1%
45-64
19%
25-44
54%
18-24
21%


Suspects By Sex
m
78%
f
21%
u
1%

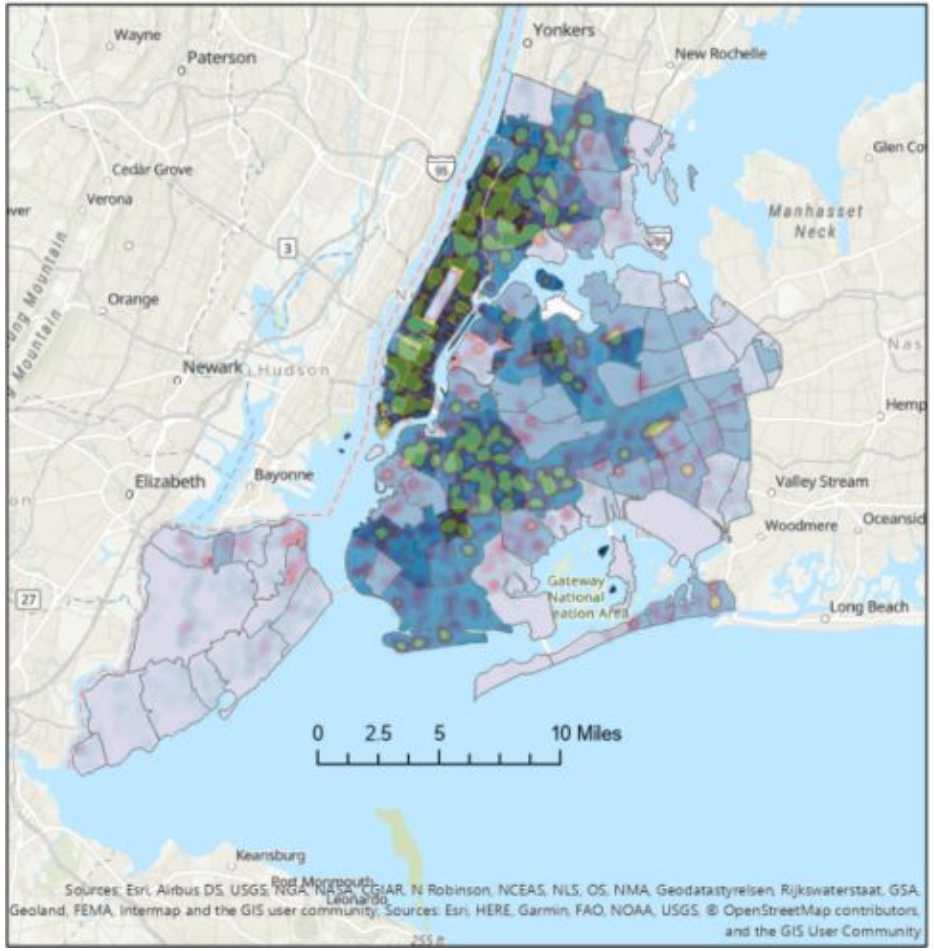
Demographic: Population Density

Population by Zip Codes



Population by Zip
POPULATION /miles_sq
- ≤0.000
- ≤14030
- ≤28490
- ≤45170
- ≤66500
- ≤1316000

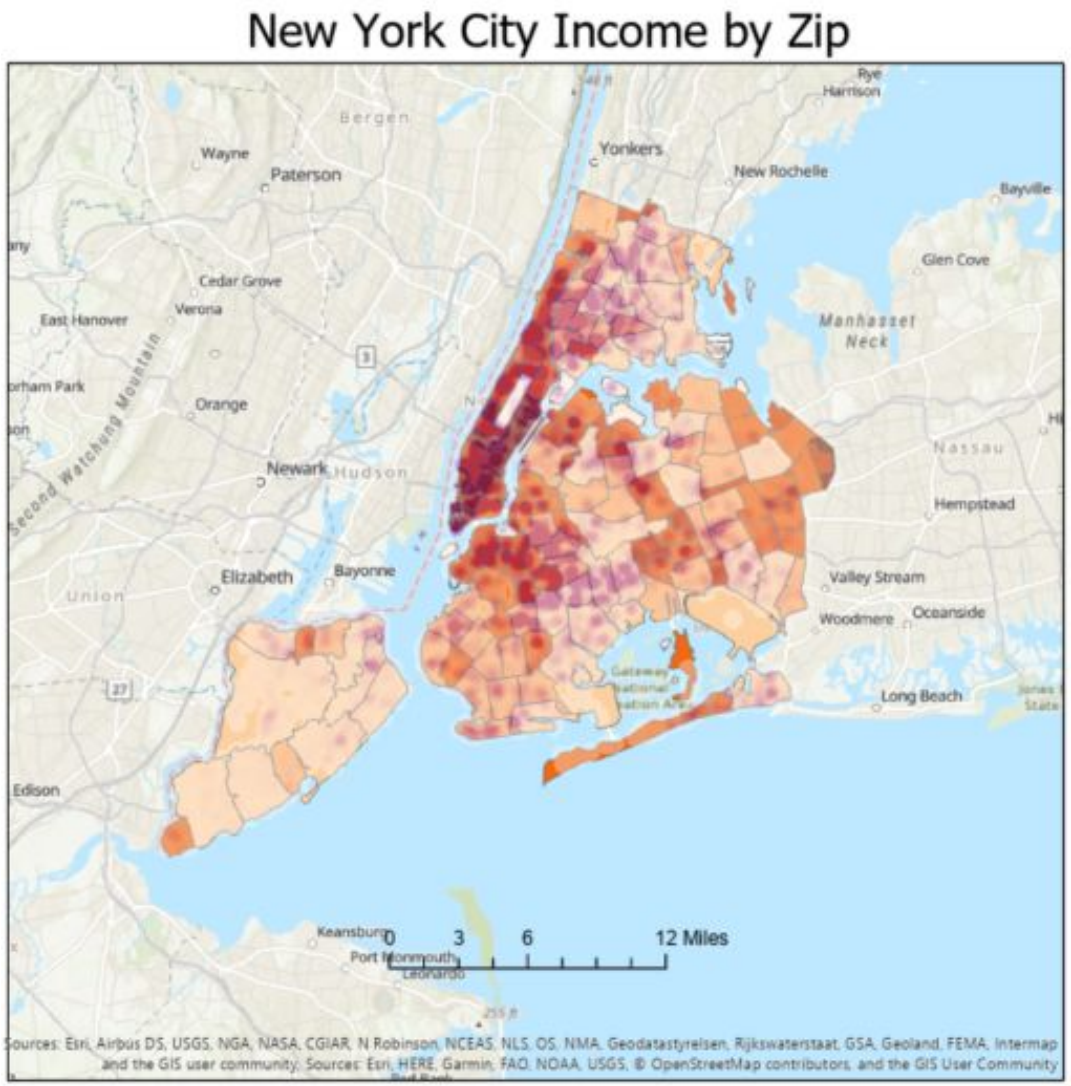
Population and Correlation to Crime

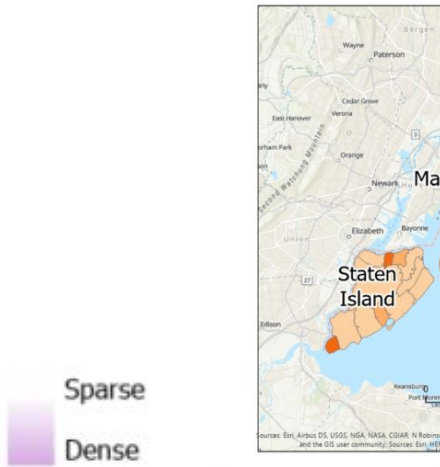


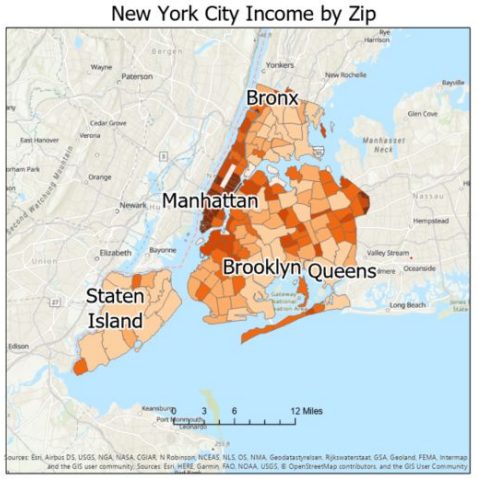
Sparse

Dense

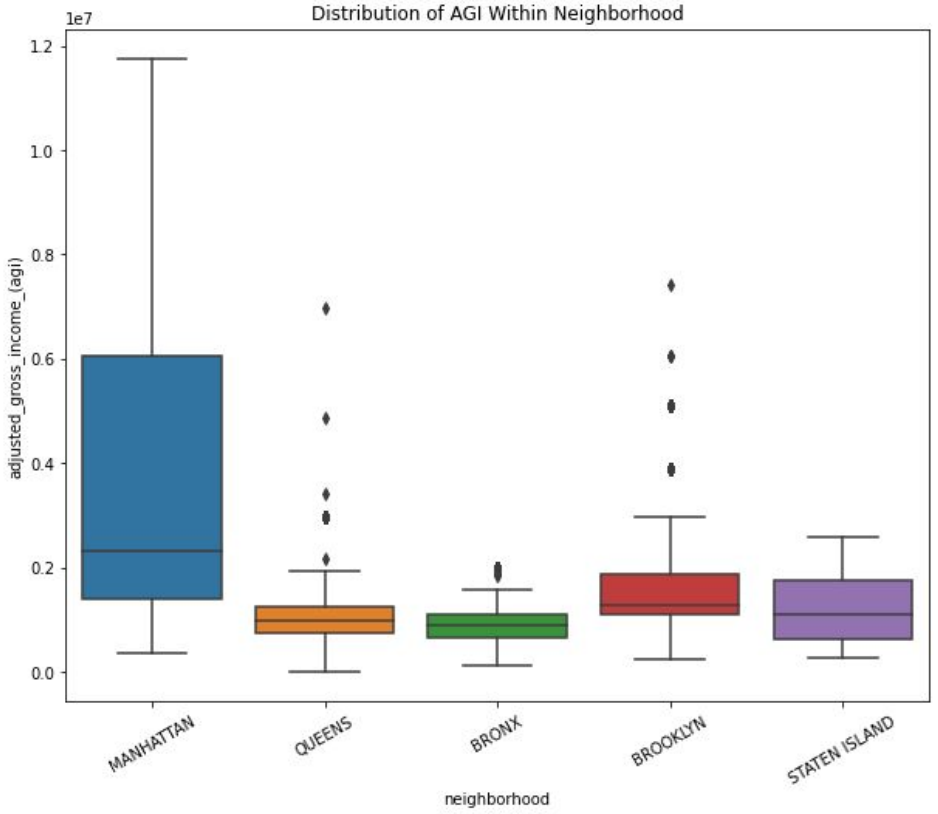
Population by Zip
POPULATION /miles_sq
- ≤0.000
- ≤14030
- ≤28490
- ≤45170
- ≤66500
- ≤1316000







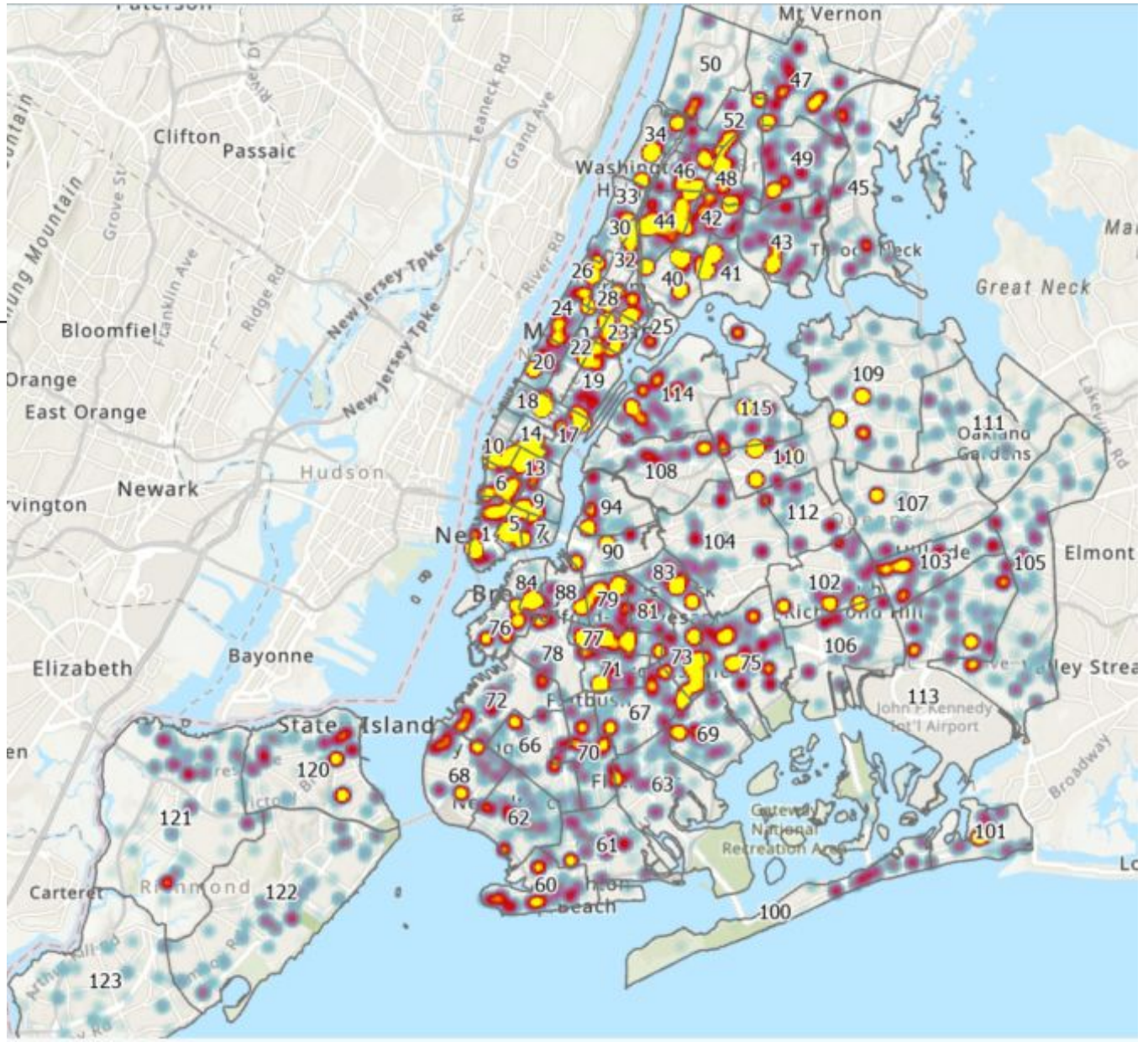
- Huge difference in income from Manhattan area compared to Queens, Bronx, Brooklyn, and Staten Island.
- Crime is in high income and low income areas.
- Important to understand that crime does not always happen in the same area a person is from. Johnny could live in Staten Island but commit a crime in Manhattan.

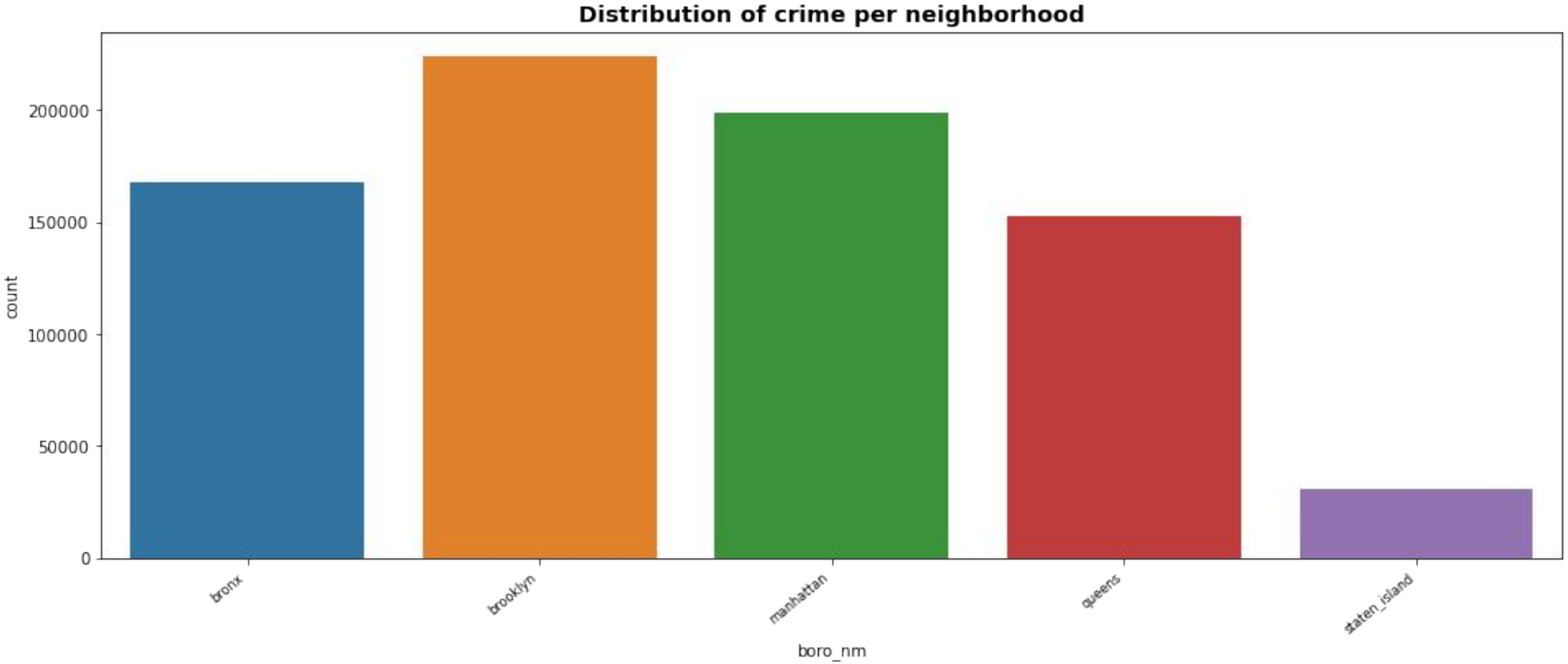


Crime by Precinct

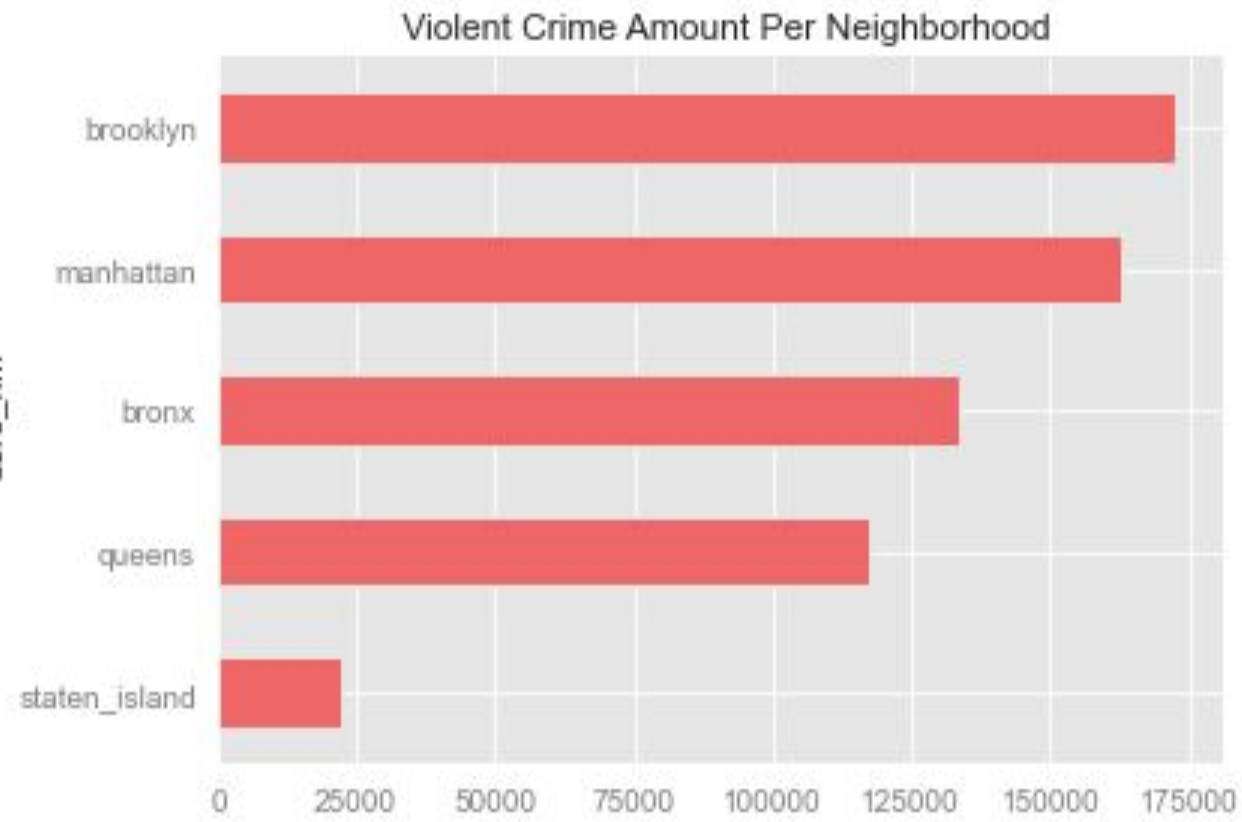
There is an uneven distribution of crime occurring throughout the city.

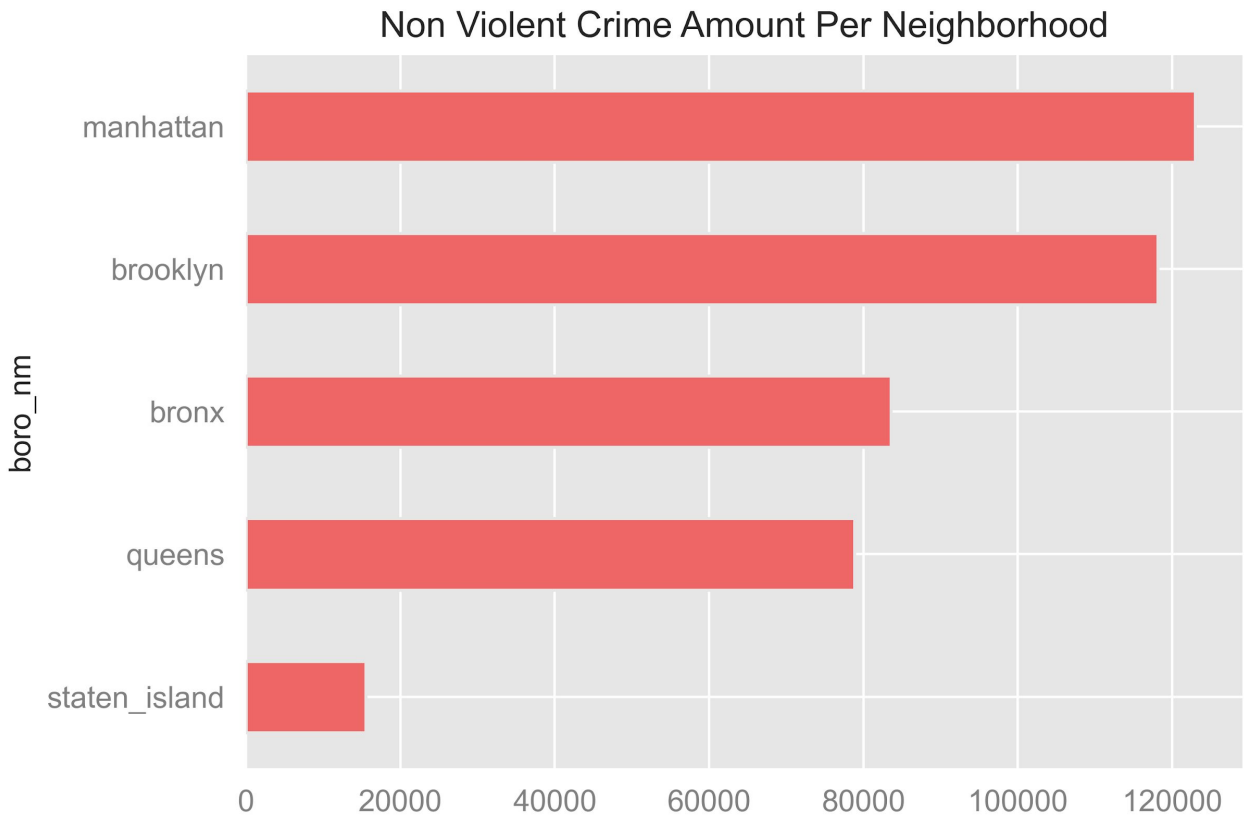
By looking at the map, we understand which precinct we need to allocate more patrol and funding. This narrowed down the area of focus more compared to zipcodes and boros.



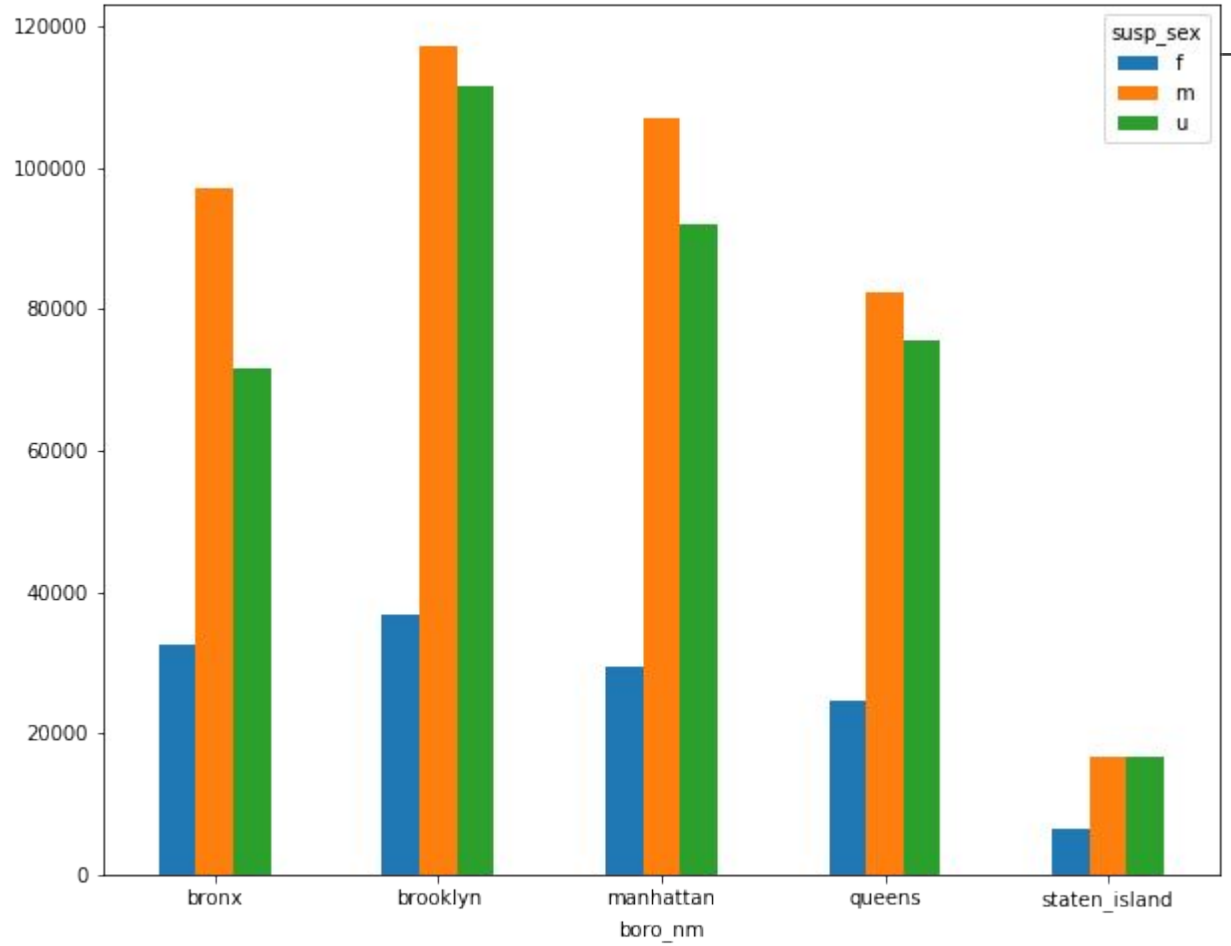

Distribution of crime per neighborhood
200000
150000
100000
50000
0
count
bronx
brooklyn
manhattan
queens
staten_island
boro_nm

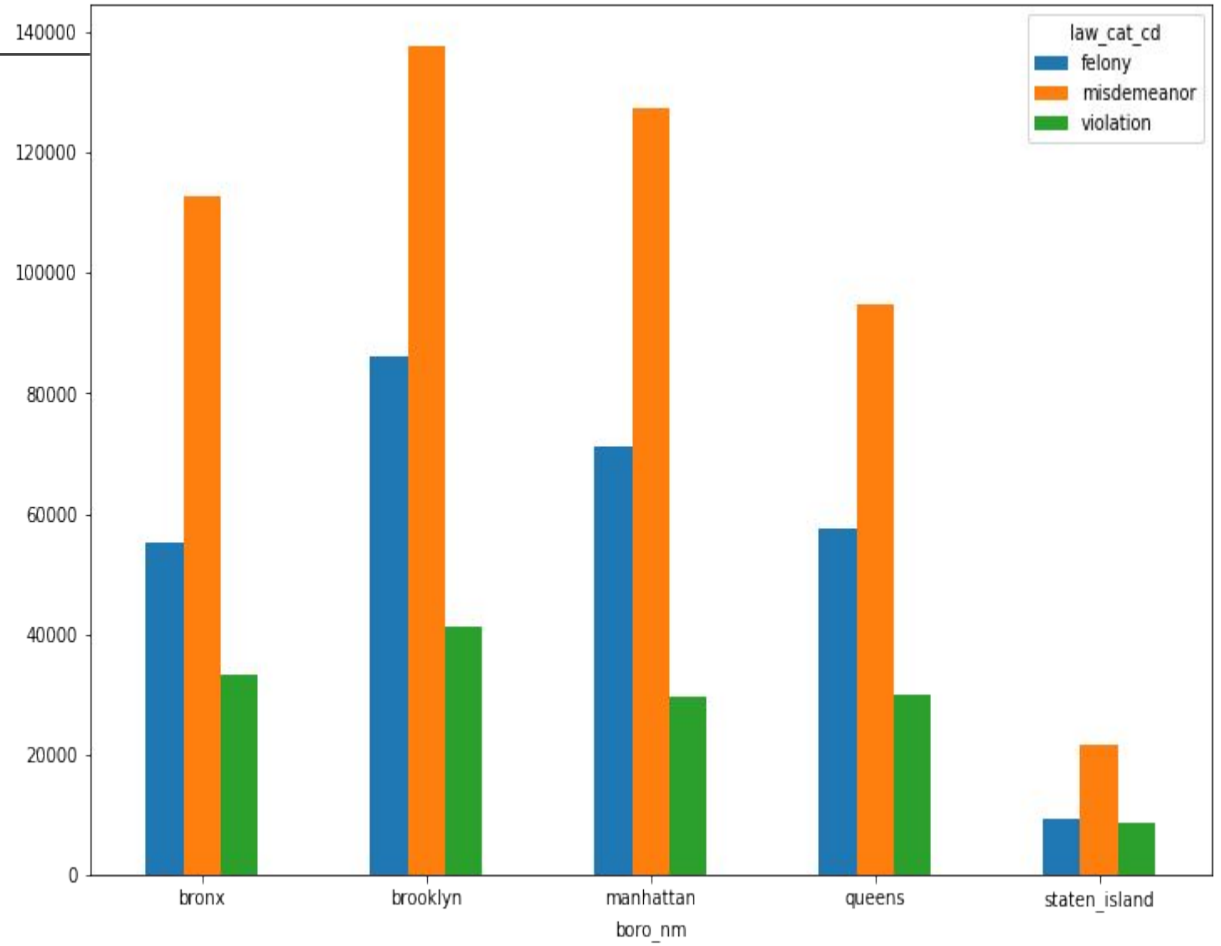
Violent vs Non Violent Crime
by neighborhood



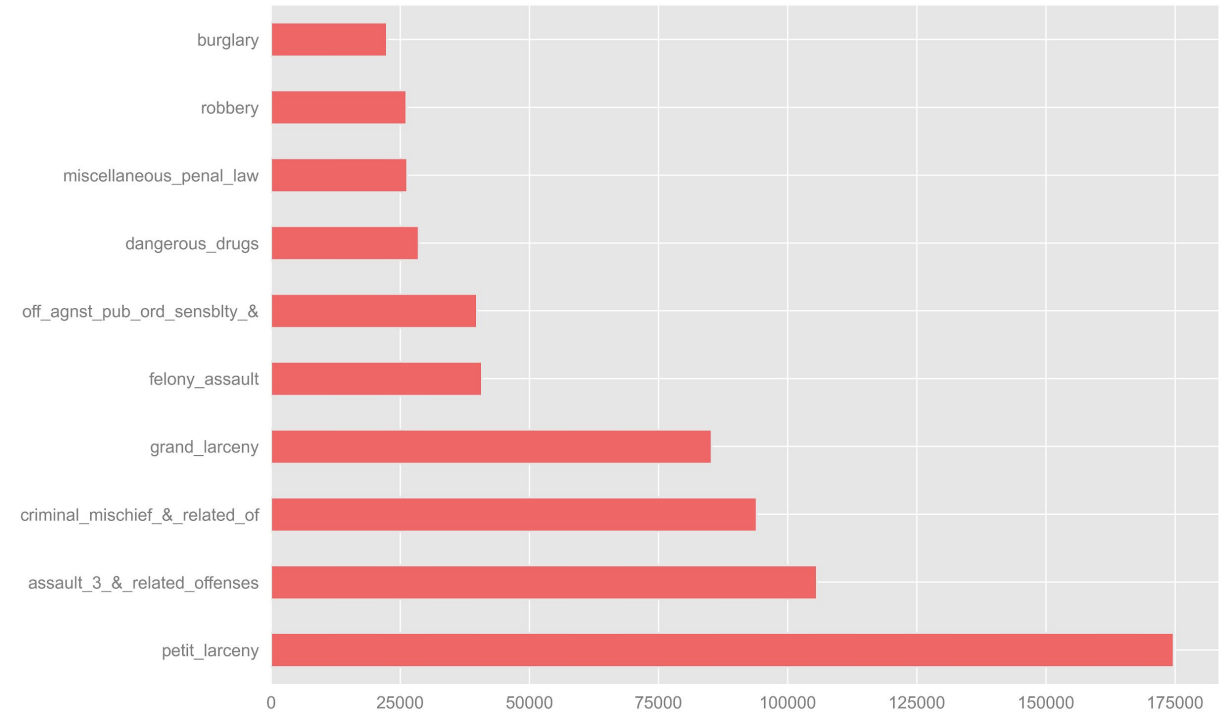


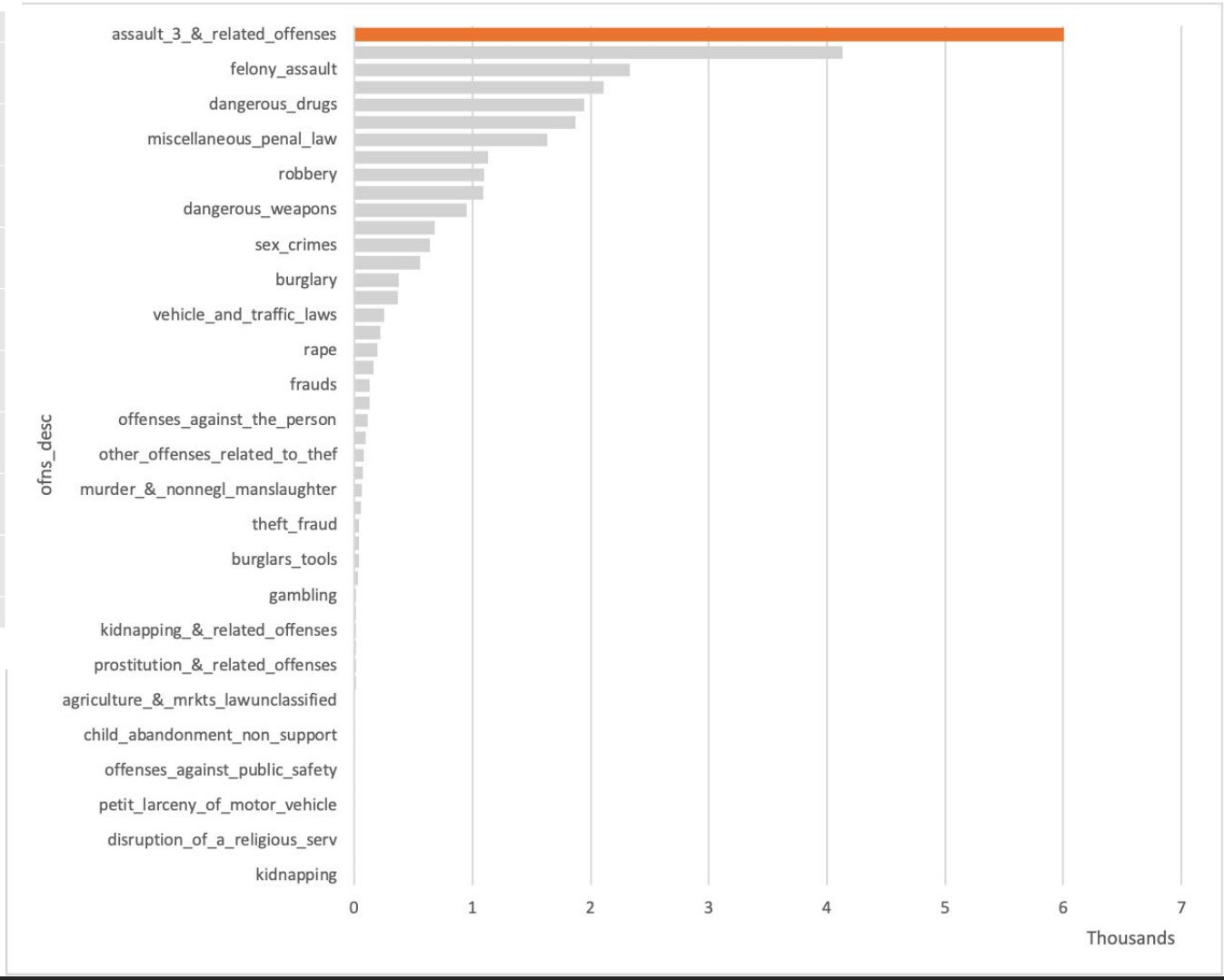
Suspect Demographics By Neighborhood

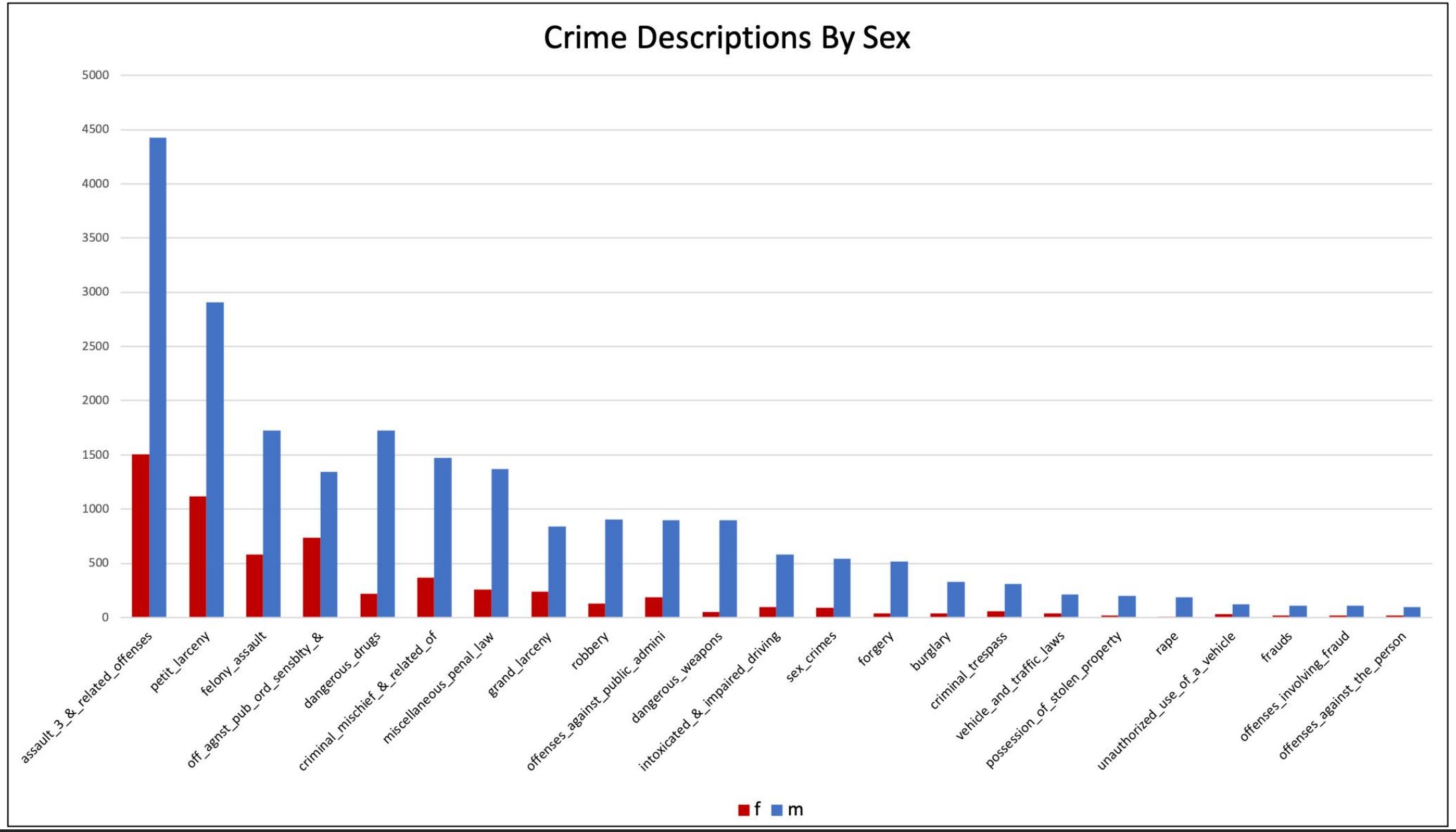




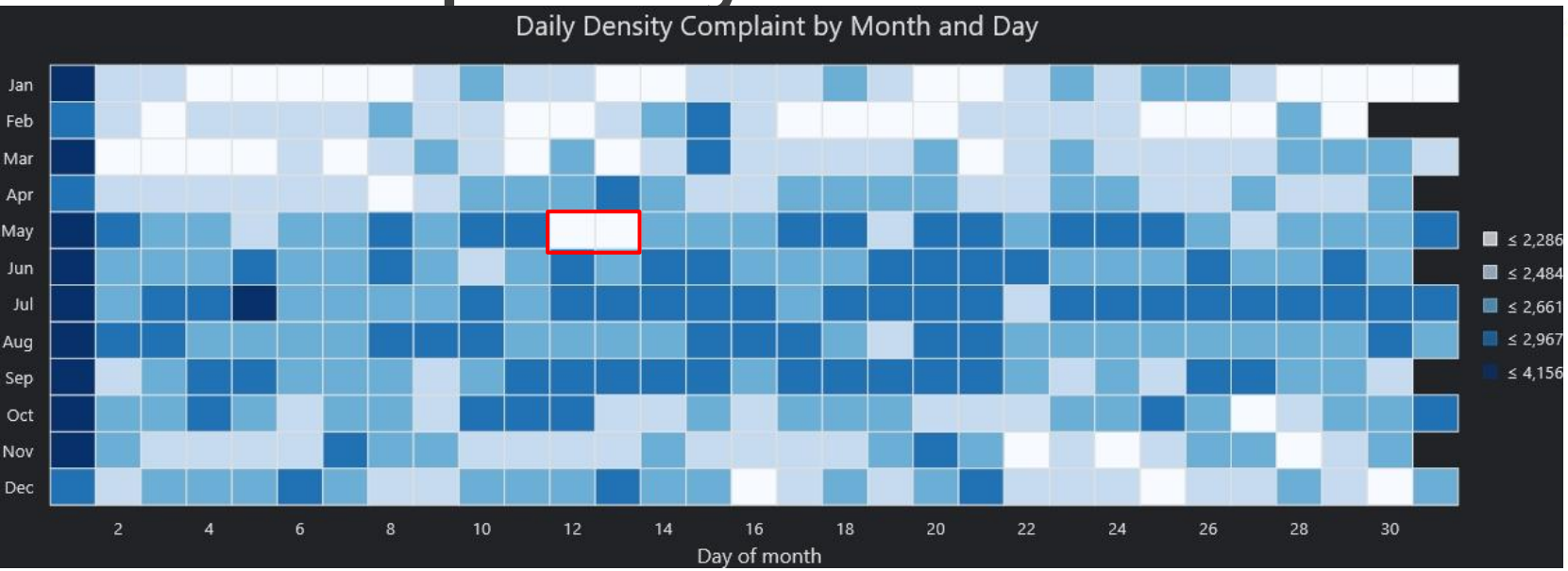
Crime Amounts By Type

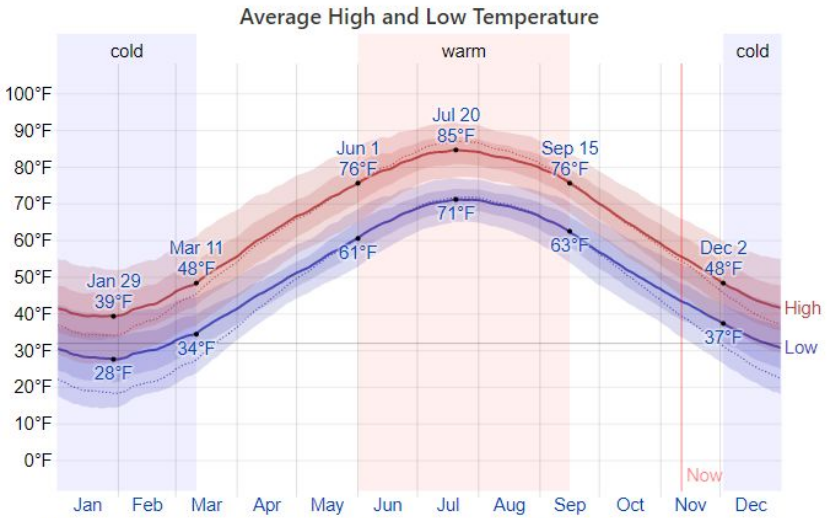




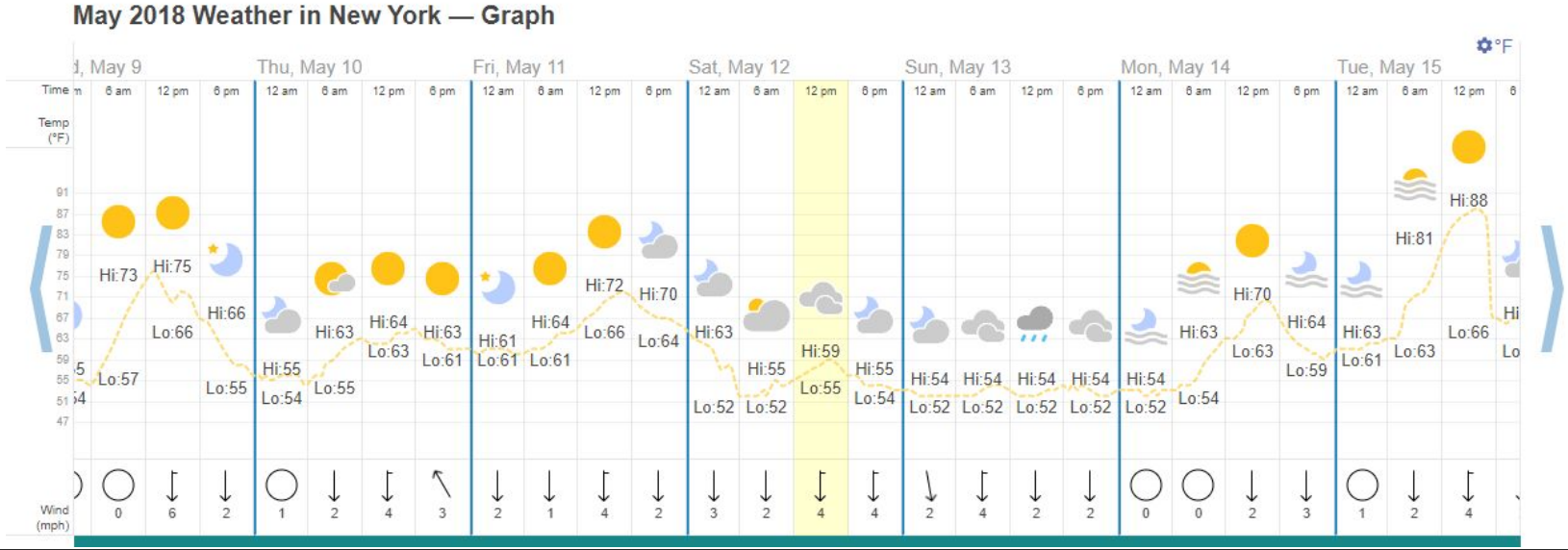

Crime Descriptions By Sex
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
assault_3_&_related_offenses
petit_larceny
felony_assault
off_agnst_pub_ord_sensblty_&
dangerous_drugs
criminal_mischief_&_related_of
miscellaneous_penal_law
grand_larceny
robbery
offenses_against_public_admini
dangerous_weapons
intoxicated_&_impaired_driving
sex_crimes
forgery
burglary
criminal_trespass
vehicle_and_traffic_laws
possession_of_stolen_property
rape
unauthorized_use_of_a_vehicle
frauds
offenses_involving_fraud
offenses_against_the_person
f
m

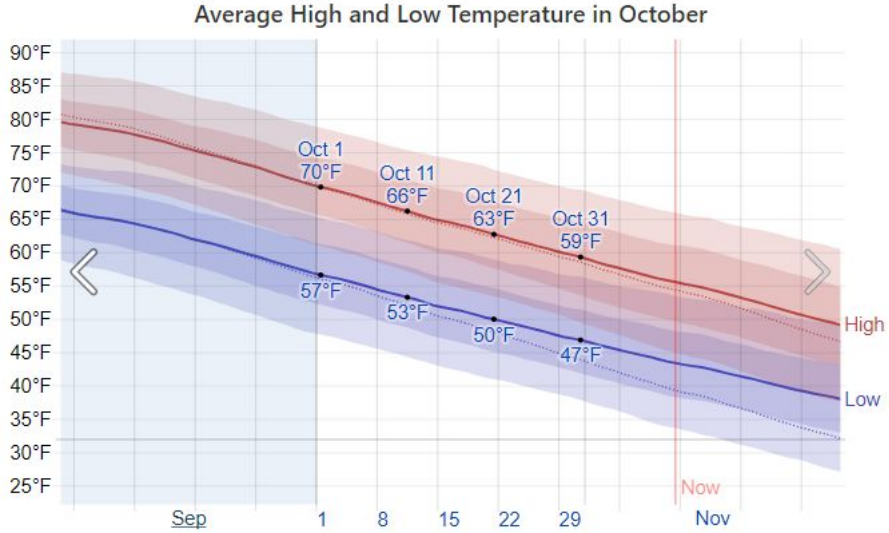
Frequency of Crime





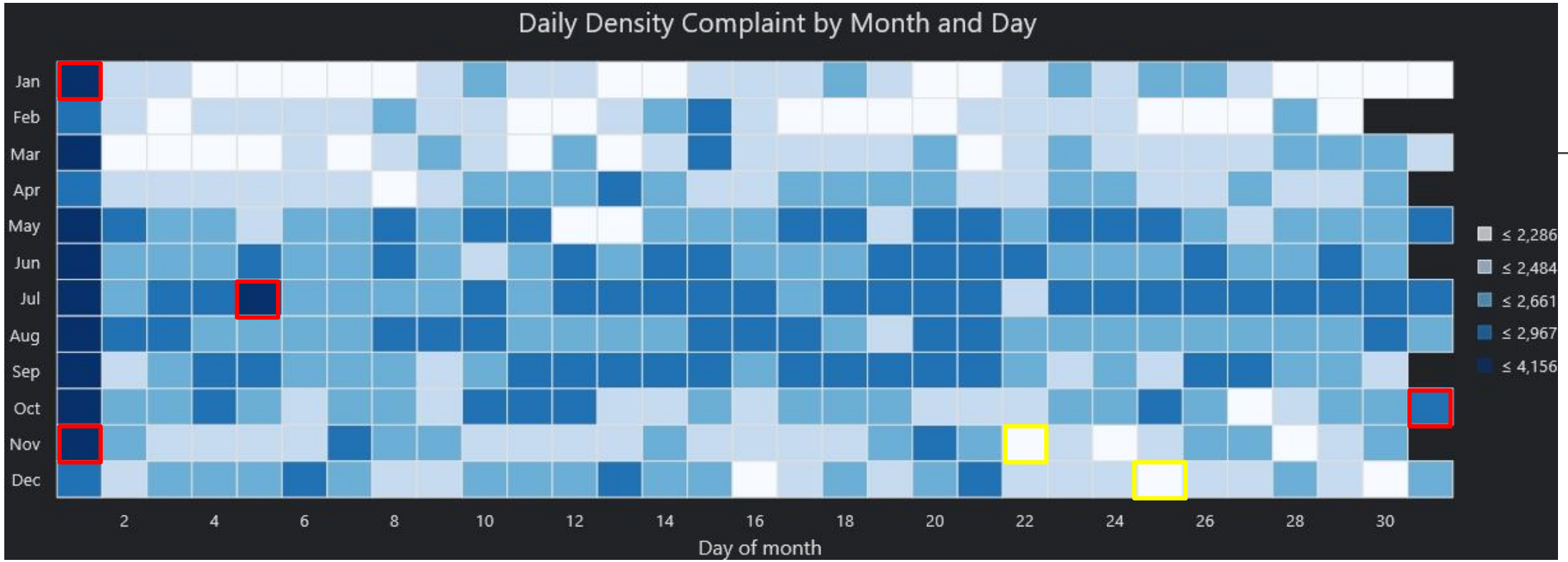
The daily average high (red line) and low (blue line) temperature, with 25th to 75th and 10th to 90th percentile bands. The thin dotted lines are the corresponding average perceived temperatures.





The daily average high (red line) and low (blue line) temperature, with 25th to 75th and 10th to 90th percentile bands. The thin dotted lines are the corresponding average perceived temperatures.

Frequency of Crime



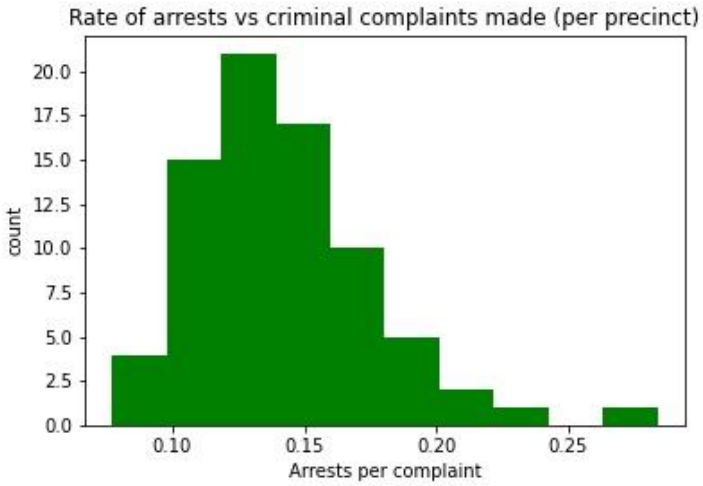
<u>Increased Crime Holidays</u>
January 1 - New Years,
July 4/5 - Independence Day
October 31/November - Halloween

<u>Decreased Crime Holidays:</u>
November 22 - Thanksgiving
December 25 - Christmas

Good Indication of when you want more patrol, vacation time for officers, and when to allocate more funds

Arrest Modeling: Methods

- The intent of these models is the prediction of arrest count by using complaint data broken down by crime type, full census population demographics, and average income by precinct.
- A linear regression was chosen for the highest interpretability and gridsearching through the parameters of elastic net regularization was used to decrease bias.
  - Alpha: 4.53
  - L1 Ratio: 0.9
- Arrests / Complaints were created to be used as a target metric, to measure any noticeable difference of crime occurrence vs law enforcement
  - Arrests / Complaint ranges between 0.077 to 0.2839, so complaints alone do do not exclusively account for arrest numbers.

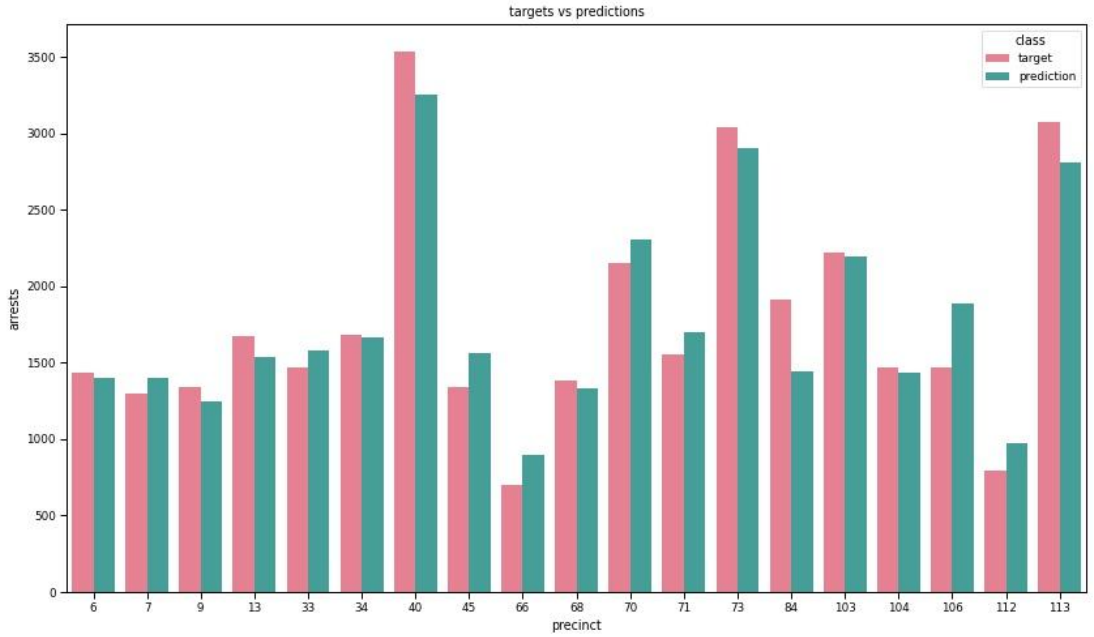


Arrest Modeling: Results

Using both a train/test split and cross-validated scores:
Training set = 515 columns, 57 rows (19 rows used to test)

- Train R2 score: .9999 Test R2 score: .7206
- Cross-validated without train/test split score: .9999

This model would benefit from a data set that spans multiple years to increase the number of samples and includes further details within each record.



Neighborhood Modeling: Methods

- Several models were utilized in an attempt to classify these criminal complaints by the borough they originated in
- Models used: Logistic Regression, Random Forest, a Gradient Boosted Tree Model, and a Neural Network
- Baseline accuracy - 28%
- The sheer size of the data made train times very long, so grid searching was not a viable option.
- The data contained a large number of weak learners, so the Gradient Boost and Neural Network were the priority
- To pick up on signal, models prioritized complexity
- For example, neural network had 7 layers, with an input of 8000 neurons

Neighborhood Modeling: Results

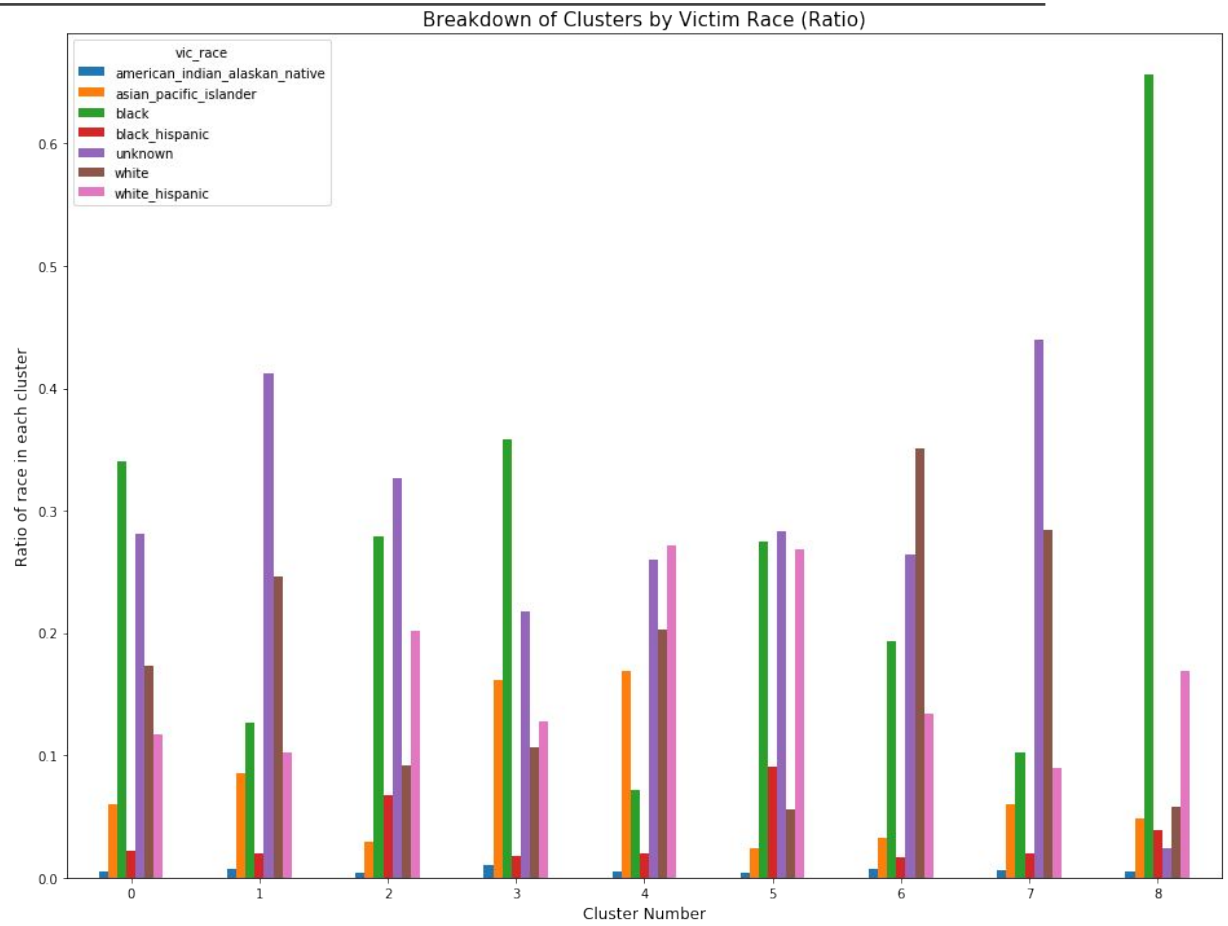
- All 4 models had very similar performance in terms of accuracy
- There was very low variance in all models
- Performance of the models was between 42% and 43% on both training and testing data
- This score reflects a score of about 15% higher than the baseline accuracy

Unsupervised modeling methods

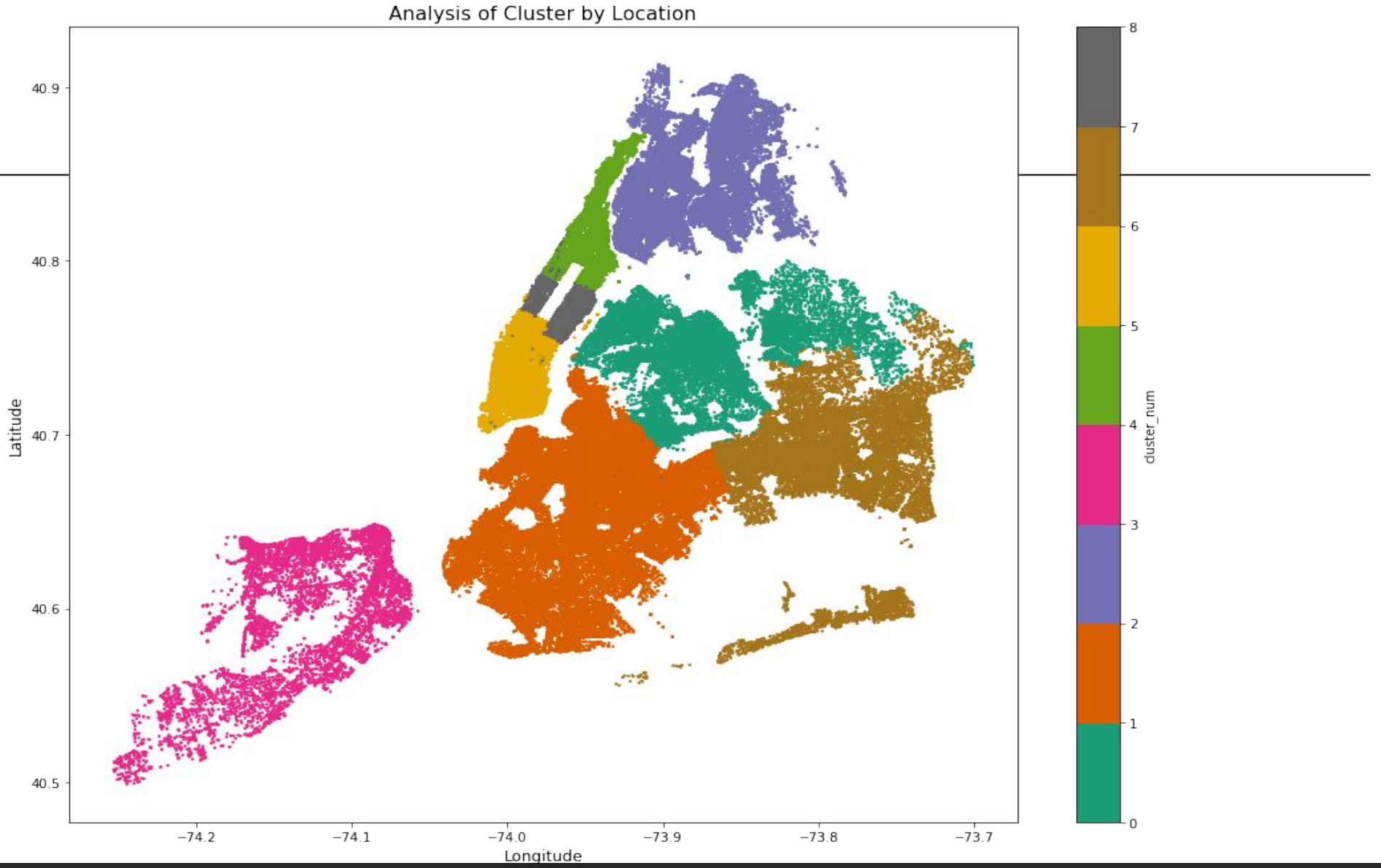
- Given poor accuracy performance in models, this model was set to identify trends for further study
- By finding clusters, we can further analyze what can be learned by crime data in the city of New York
- Again given the challenge of immense data and time constraints, this model was fit to 35% of the entire dataset
- Optimal method - a principal component analysis, reducing the data to 10 features, which was then passed into a K Means cluster
- NOTE: This is purely an exploratory study meant to guide the direction of future data collection, and not meant to be interpreted on its own

Unsupervised modeling results

- The optimal PCA to KMeans model was set to 9 clusters, decided by applying the elbow method
- This model had a silhouette score of .436, which was a significant improvement compared to a KMeans or DBScan alone
- EDA conducted on the clusters created showed differences in suspect and victim demographics, as shown here
- Combining these features led to an interesting localizing of the data clusters, as shown in the next slide



Visualization of Clusters



Conclusion - Supervised Models

We can use the coefficients created from our linear regression models to make some tentative observations regarding current law enforcement practices. For example, as this data was taken from 2018, it's no surprise that marijuana possession complaints were used by the model to assess how many arrests were made. The publicly available arrest data can confirm that this crime constituted a high number of arrests. In particular, it may be said that these types of complaints may be made less frequently without already being attached to an arrest.

We also have some data that seems to have an inverse effect on the expected numbers of arrests, such as complaints made for motorcycle theft. This is also expected, as there are frequently reports made for certain crimes without arrests to go with them.

Population demographic data is also included within, which may show some indication of how the demographics of the people that live in an area can affect the rates of law enforcement, independent of the frequency of the occurrence of crime.

Feature	Coefficient
complaints	75.812924
forgery_etc_misd_complaints	40.021874
ny_state_laws_unclassified_fel_complaints	35.648988
P0020034	33.9882
P0010069	33.887338

Conclusion - Unsupervised Models

- Looking at the similarity of boroughs with the total count of complaints in New York, there appears to be unique features that separate the clusters that our model created.
- Stronger than this is the Latitude and Longitude coordinates, which clearly show the separation of each cluster. Given that many of these clusters are touching, I do think it likely that some other features separate these clusters clearly along location lines.
- Despite the relatively poor performance of the classification model in the Neighborhood Analysis Notebook, this result is a clear indication that location is an important feature to identify differences in the impact of crime in New York - specifically, according to the map created above.

Important Caveat to these findings - due to processing and time limitations, this study had to be done on a small sampling of the entire data. This analysis would ideally be done on the entirety of the dataset, to hopefully find even stronger correlations.

Next Steps

- Sadly, the income information that was drawn for this project was merely an estimation that was too general to yield helpful results. However, it is still highly suspect to be a strong predictor of the number of and type of crimes committed. Perhaps following more recent Census and IRS data to the specific regions given above, we can analyze these differences to evaluate their effects on crime.
- This data is clearly showing differences based on these geographic locations. I would recommend a thorough survey of each of these 8 regions displayed (as cluster 8, the 9th cluster, is practically not visible on the plot). An analysis of their specific traits for comparison would be helpful here. In addition, a study of each region's specific history, local laws/leaders, current events, and even a consultation with individuals knowledgeable about the areas of each cluster would lead to more advanced insight, and should likely be the primary direction of a follow-up study.
- A more complete weather dataset would be included as a factor in additional analysis used to predict occurrence of crime.