Collin Burns, Joey Chiarella, John LoBello
Prof. Kargin
MATH 457
December 8, 2023

Analyzing the Geography of NYC Crime

Section I: Introduction & Overview

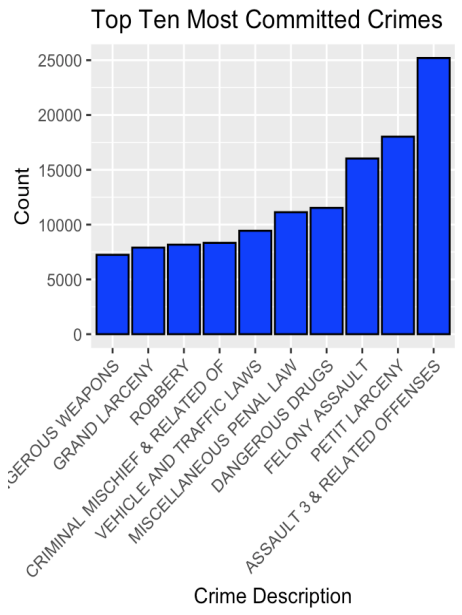
With NYC being one of the biggest cities in the United States, crime is an inevitable reality that residents have to deal with on an ongoing basis. Compared to other communities in New York State, NYC is in the 95th percentile in terms of crime occurrence. Additionally, the chance of becoming a victim of a violent or property crime in NYC is 1 in 40. Essentially, one can be a crime victim on virtually any block. Despite these generalizations, the occurrence and severity of different types of crimes differ based on location (borough, X & Y coordinates, lat-long, and neighborhood). In this project, we aim to predict the crime according to its severity and location and provide a variety of easy-to-understand maps to interpret crime in NYC. We would then train the data to predict possible crime in specified cluster points in the city, as well as to visualize it.

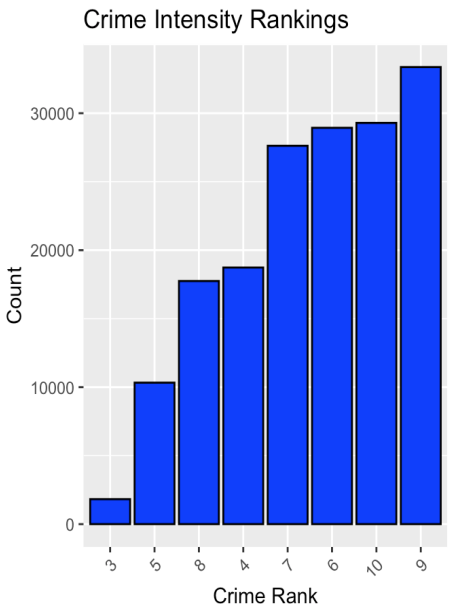
Section II: Data Overview & Preliminary Studies

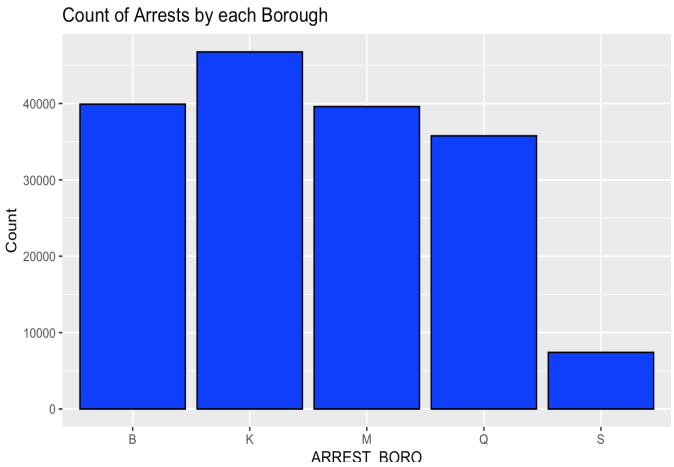
Our data was sourced from the NYC Open Data website and contains records from the NYPD concerning arrests in the city. The dataset has a total of 169,430 observations. Each observation represents a NYC arrest from 1/1/2023 to 9/23/2023. The 19 variables in the data are: arrest_key, arrest_date, pd_cd (granular classification code), pd_desc (description of crime), ky_cd (general classification code), ofns_desc (offense description), law_code, law_cat_cd (level of offense), arrest_boro (arrest borough), arrest_precinct, jurisdiction_code, age_group, perp_sex (sex of perpetrator), perp_race (race of perpetrator), x_coord_cd (x-coordinate), y_coord_cd (y-coordinate), latitude, longitude and new georeferenced column. A lot of the variables were irrelevant to our analysis, so a reduced dataset was constructed containing the variables: arrest_date, ofns_desc, arrest_boro, age_group, perp_sex, latitude and longitude.

In addition, we created a new column called crime_intensity. This new variable ranks how severe the crime is based on a short description of the crime. The offense description consists of 61 unique values. We assigned each crime description a score from 1-10. The rankings go from 1 (a non-threatening/minor infraction), to a 10 (heinous crime that's threatening to many). Crimes classified as a 10 include murder, arson, rape, while crimes classified as a 3 are gambling and traffic infractions. The lowest these rankings go is 3 as perpetrators are still getting arrested for these actions. Crimes classified as 1 or 2 were too small where people weren't being caught and prosecuted for them. This variable is essential to our KNN and K-means clustering data methods. Exploring this dataset after reducing the number of columns and creating the crime_intensity column, we wanted to see the frequency of crimes for each intensity rating, along with the ten most common crimes. From the bar graph, we can see that most crimes

committed were classified as severe, being a 9,10 or a 6. The most common crimes occurred were assault, petit larceny, then felony assault and dangerous drugs. We then looked at all boroughs and observed how many crimes occurred in each one, in preparation for our KNN and K-means clustering data. Brooklyn (K) had the most amount of crime, as shown.

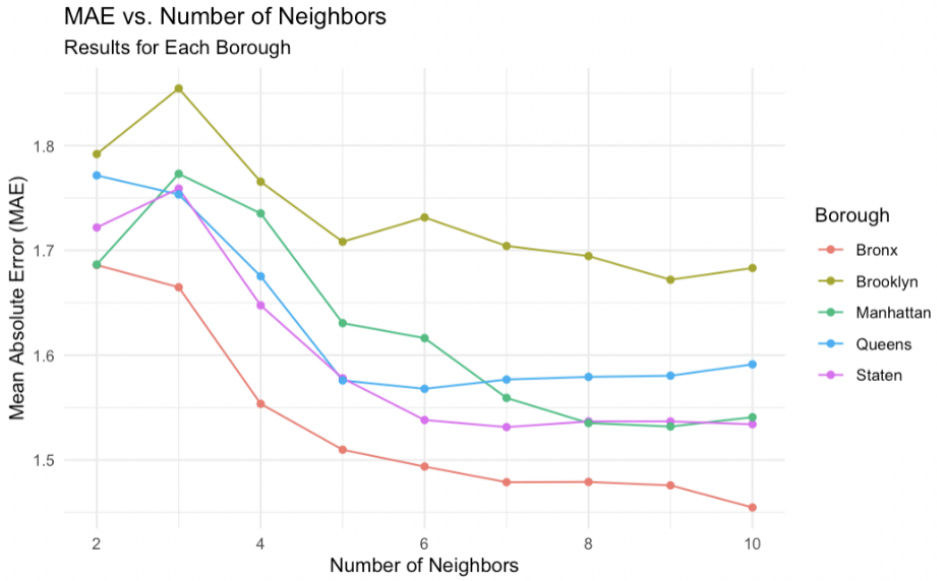






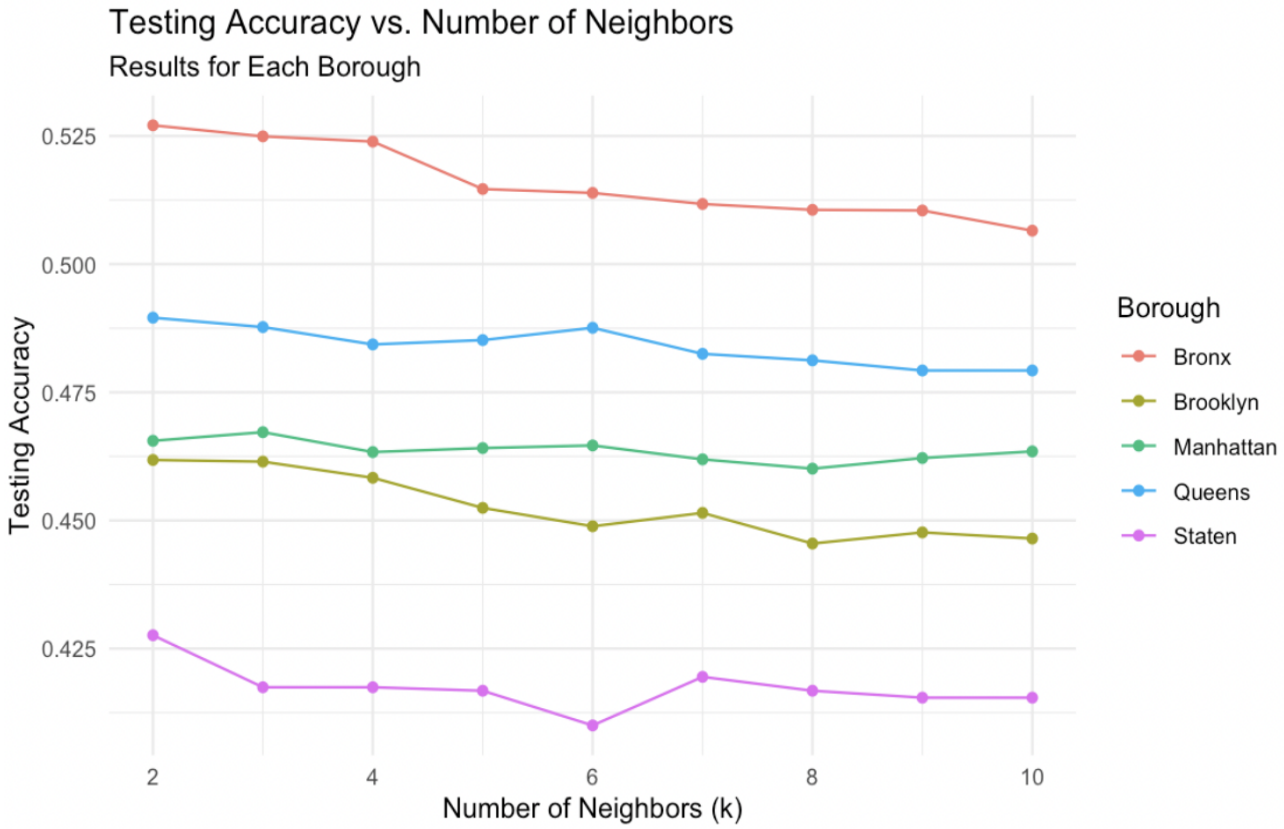
Section III: Methods and Results

The first method used was KNN. Our goal is to classify the crime in spatial clusters, therefore we felt that KNN was appropriate as it analyzed severity (class types) around a given point. For KNN, we partitioned an 80/20 split of the data with "latitude" and "longitude" as our selected variables, and trained the data with $2 \leq K \leq 10$ nearest neighbors. The trained KNN was then used to predict crime intensity for the remaining 20% of our data. We then calculated mean absolute error as *MAE = Actual Crime Intensity – Predicted KNN Values.* The results show that for all boroughs, MAE is greatest at K = 2, except for Staten Island and Queens. Our testing accuracy ranges from 0.412 to 0.525. This isn't ideal as it carries about the same accuracy as a coin flip.



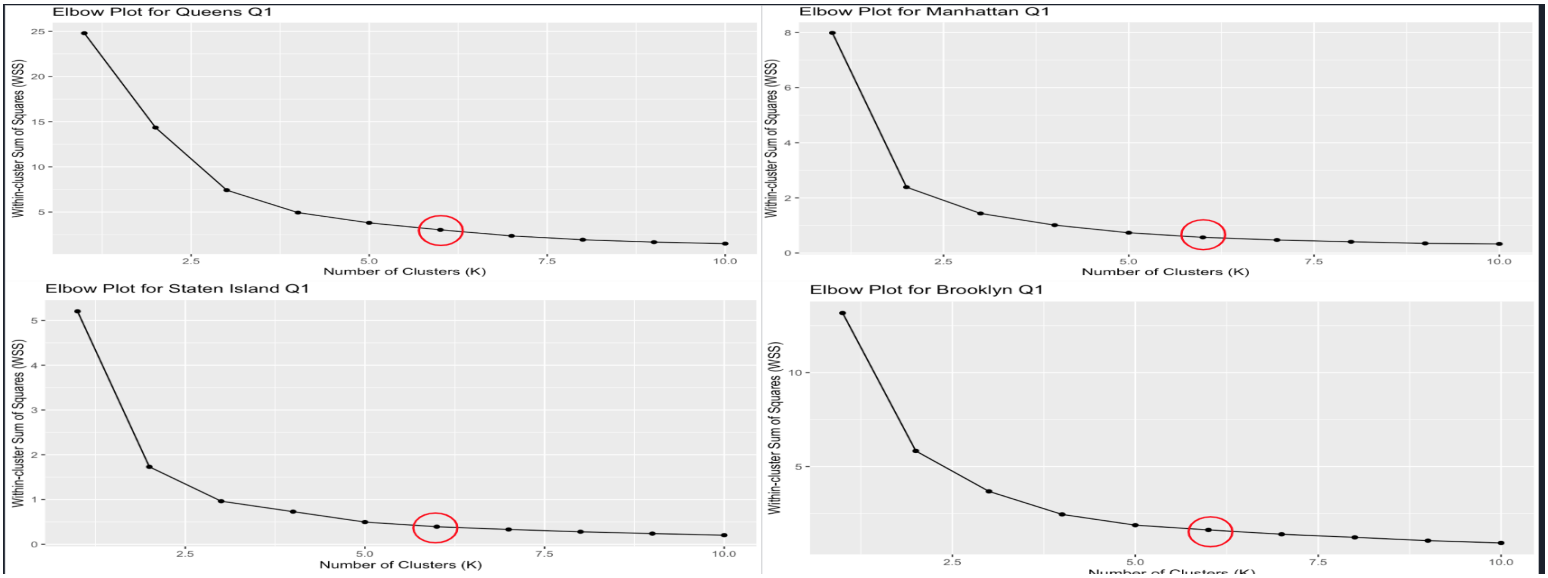
Example of Results:

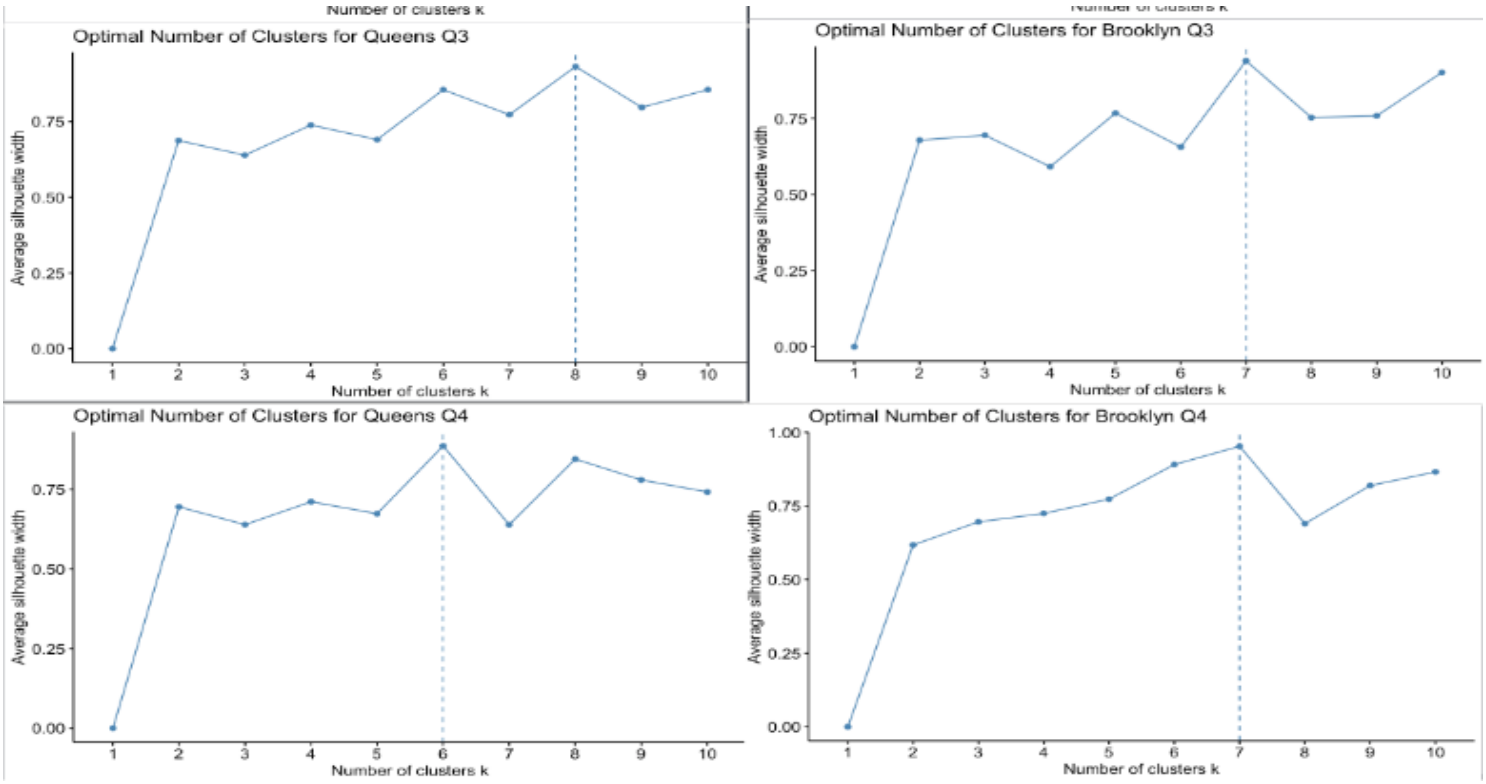
Borough	**MAE (Using K = 9)**
Manhattan	1.55
Bronx	1.45
Staten Island	1.55
Queens	1.58
Brooklyn	1.66
	AVERAGE = 1.558



Using KNN may be inefficient in this context for many reasons. First, we approach KNN as a classification method, yet that doesn't align with our objective. We aim to predict crime on a spatial map, which is a regression question. In addition, outcomes might be biased in assigning crime ratings solely based on a single crime occurrence in a specific location, even when that area may be particularly safe. The overall safety of a neighborhood isn't accounted for. The lack of ample differentiation is also evident, as the minimum score observed is 3, a frequency approximately 20 times lower than the more prevalent score of 9. As mentioned before, data with lower severity scores tend to go unnoticed and therefore unrecorded.

The second method used to classify crime intensity scores is K-means clustering. Our goal using K-means is to visualize and model spatial data to use for data visualization. To predict the number of clusters, we used elbow plots for every quarter of each borough. This yielded similar results, where the optimal number of clusters was 6. We wanted to dive further into finding the optimal number of clusters as well as seeing how well-defined the clusters were. We did this by exploring silhouette scores for each quarter and analyzing them based on a -1 to 1 scale. Our results from silhouette scores ranged from 0.6 to 0.9 and showed the optimal number of clusters was around 6-8 clusters per quarter. Silhouette scores of 0.6 to 0.9 indicate that crimes in each cluster were well-assigned to their cluster and not well-assigned to other clusters. We chose the optimal number of clusters to be 6 due to the results from silhouette scores and elbow plots.





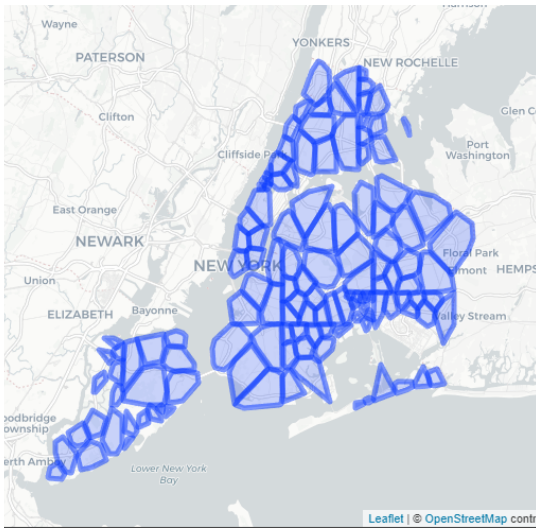
We have defined well-separated clusters for each quarter. For each borough, we will now analyze how well this clustering model performed based on the accuracy and mean absolute error (MAE). The mean absolute error for each quarter on average approximately 1, meaning that on average our prediction model for each cluster differed by a score of 1 based on our crime intensity score scale. In addition to the MAE, we use accuracy scores to evaluate our model and compare it to KNN. We set a threshold of 1.25 to assess the

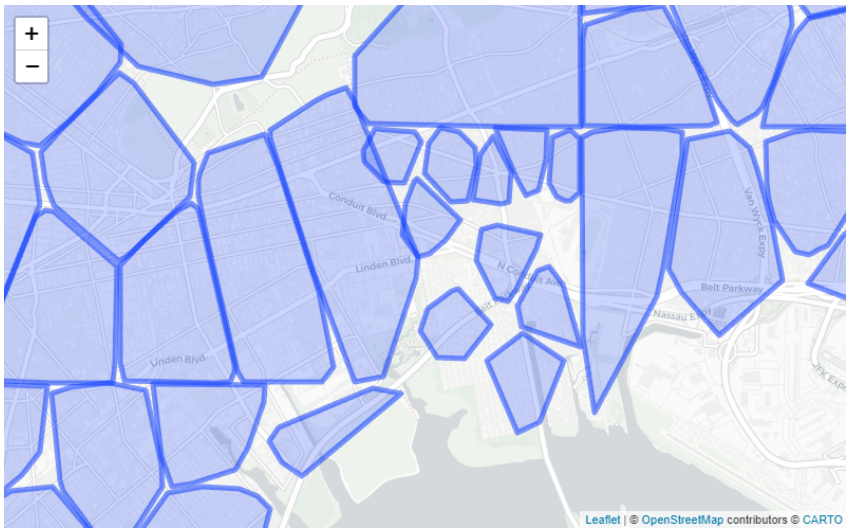
Borough_Quarter	Accuracy
Bronx_Q1	0.6586054
Bronx_Q2	0.8631153
Bronx_Q3	0.6630195
Bronx_Q4	0.5656494
Staten_Q1	0.8134685
Staten_Q2	0.8197227
Staten_Q3	0.8782609
Staten_Q4	0.8444444
Queens_Q1	0.8108301
Queens_Q2	0.7337893
Queens_Q3	0.7620397
Queens_Q4	0.7635811
Manhattan_Q1	0.6586054
Manhattan_Q2	0.8631153
Manhattan_Q3	0.6630195
Manhattan_Q4	0.5656494
Brooklyn_Q1	0.8975894
Brooklyn_Q2	0.8118660
Brooklyn_Q3	0.7523040
Brooklyn_Q4	0.6448256

accuracy of our clusters to see how many crime intensity predictions are within this threshold by quarter. 18 out of 20 (90%) clusters had accuracies between 65% to 90% accuracy with 2 quarters out of 20 (10%) quarters having at accuracy 56%. This model performed significantly better than KNN and we will use these results to visualize this data in the following section.

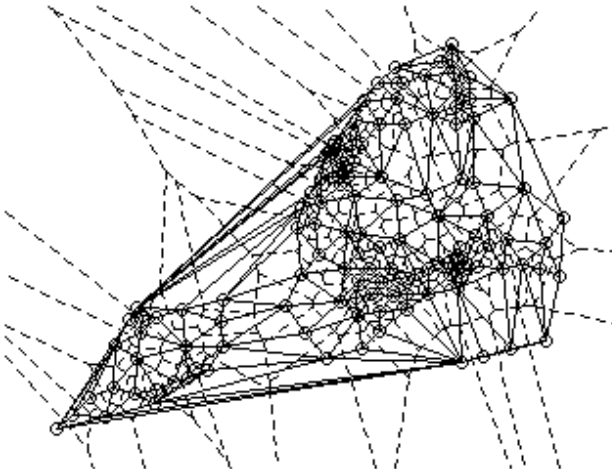
Section IV: Visualization

One of the major objectives of our project is to create an interpretable map for a normal individual to understand when looking at New York City. To do this, we originally simply compiled the KMeans Clusters into polygons by the points within them. The turnout of this was not exactly what we were hoping for:

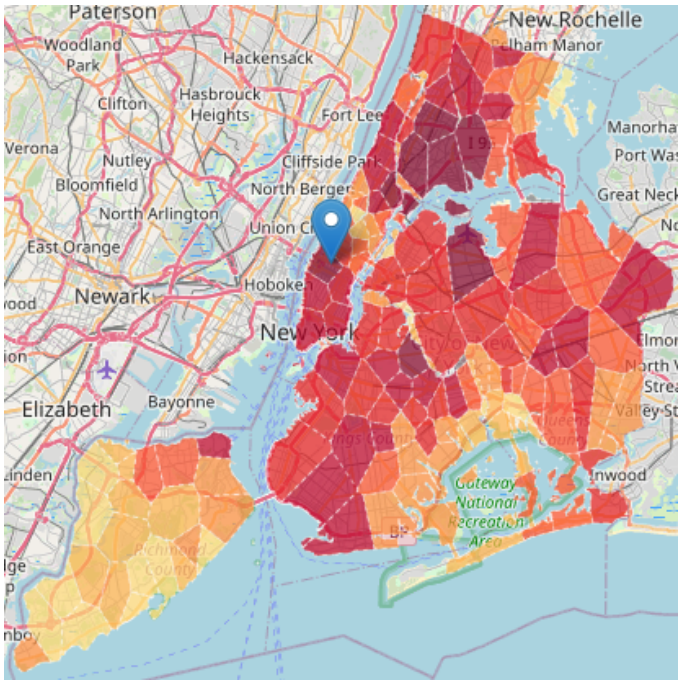




The polygons did not fill all of the area of NYC and had quite erratic shapes. This would kill the interpretability of the map because not all of the areas someone might want to look at would be visualized within a polygon. This led to the utilization of Voronoi tessellation. Tessellation helped better fit the polygons to the available space in NYC to be summarized. We used the K-mean cluster centroids as points for the Voronoi tessellation, which ended up accomplishing exactly what we wanted:



Then using an SF map of the NYC boroughs, we constrained the tessellation to only include the area bounded by NYC. The output was a perfect overlay of our clusters created through tessellation on the map of NYC. However, the aforementioned issue with the lack of variability in our data caused an issue with the interpretation of the map. Using our standard crime scores, which we assigned at the beginning, for the visualization was not enough to make an easily understandable map of the city. We then decided to scale the crime intensity by the log of the count of arrests within the polygon. Then, we divided by the max value of that calculation to create our scaled score. The scaled score created a much more interpretable map to the naked eye, which also allowed for point prediction at a specific lat, long value:



Section V: Discussion/Conclusion

The objective of this project was to train NYPD arrest data in order to predict the severity of crime in different sections of NYC. To accomplish this, we used a plethora of different spatial diagrams, some of which include polygon maps and its updated version improved by Voronoi tessellation. Our methods included KNN, which predicted the severity of surrounding crime relative to a point, and K-means clustering, to model the spatial data for data visualization purposes. Although we specifically focused on crime classification, the creation of other GIS models such as COVID-19 occurrence maps, weather maps and urban planning all use the same processes described in this project. From everyday city NYC commuters, to foreigners who are only visiting the city temporarily, these models can be beneficial to their safety and help them stay clear of high crime areas going forward.

Works Cited

Aynat, Ascen. "The Ultimate Guide to the 5 Boroughs of New York City + MAP." *Capture the Atlas*, 27 Aug. 2023, capturetheatlas.com/5-boroughs-of-new-york-city.

New York City, New York Population 2023. worldpopulationreview.com/us-cities/new-york-city-ny-population.

"New York Struggles With Rise in Violent Crime Amid COVID-19." *PBS NewsHour*, 20 May 2022,www.pbs.org/newshour/show/new-york-struggles-with-a-sharp-rise-in-violent-crime-amid-covid-19.

NYC Crime | NYC Open Data. 8 Nov. 2023, data.cityofnewyork.us/Public-Safety/NYC-crime/qb7u-rbmr.

NYPD Arrest Data (Year to Date) | NYC Open Data. 8 Nov. 2023, data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc.

Sachinsoni. "K Nearest Neighbours — Introduction to Machine Learning Algorithms." *Medium*, 11 June 2023, medium.com/@sachinsoni600517/k-nearest-neighbours-introduction-to-machine-learning-algorithms-9dbc9d9fb3b2.

S7. chiller, Andrew. "New York, NY Crime Rates." *NeighborhoodScout*, 25 Sept. 2023, www.neighborhoodscout.com/ny/new-york/crime#description.

Sidewalk | NYC Open Data. data.cityofnewyork.us/widgets/vfx9-tbb6?mobile_redirect=true.

What Is GIS? | Geographic Information System Mapping Technology. www.esri.com/en-us/what-is-gis/overview.

What Is the K-nearest Neighbors Algorithm? | IBM. www.ibm.com/topics/knn.