Frederick Tantowi

Yang Yang

Jianru Shen

Jenny Zhang

**Housing and Crime: Group Research Project**

***Research Question***

The findings from the literature review showed a significant relationship between crime rates and housing. Specifically studies showed a negative relationship between crime and housing demand, as measured by rent prices. We wanted to expand upon the existing research by conducting our own analysis on the relationship between crime and housing.

Our main research question was, is there a consistent relationship between crime and housing projects in New York, across the five boroughs? If there was, we aimed to dive deeper and assess if the relationship was stronger for certain construction types, such as new construction or preservation. Our hypothesis was there would be fewer new construction in boroughs with higher numbers of crime.

***Data Preparation***

We collected New York crime data from the New York open data resource. We wanted to gather crime data over several years. However, after exploring several options we decided it was best to merge a historical dataset with a Year-To-Date arrest dataset.

We wanted to create a new dataset with longer periods that match the years in the housing project dataset. We dropped the rows containing variables that we deem not required for our study and turned categorical variables into dummy variables, such as a borough, construction type, etc. We also removed some irrelevant crime types in the offense description column in the “Arrest Dataset”, such as “abortion”. Additionally, we found that most missing values are in the categorical variable columns, such as the “offense description” in the arrest dataset and the “project starting date” in the housing dataset. So to avoid the inaccuracy that might be raised by imputing incorrect information into those missing cells, we decided to remove the rows containing missing values.

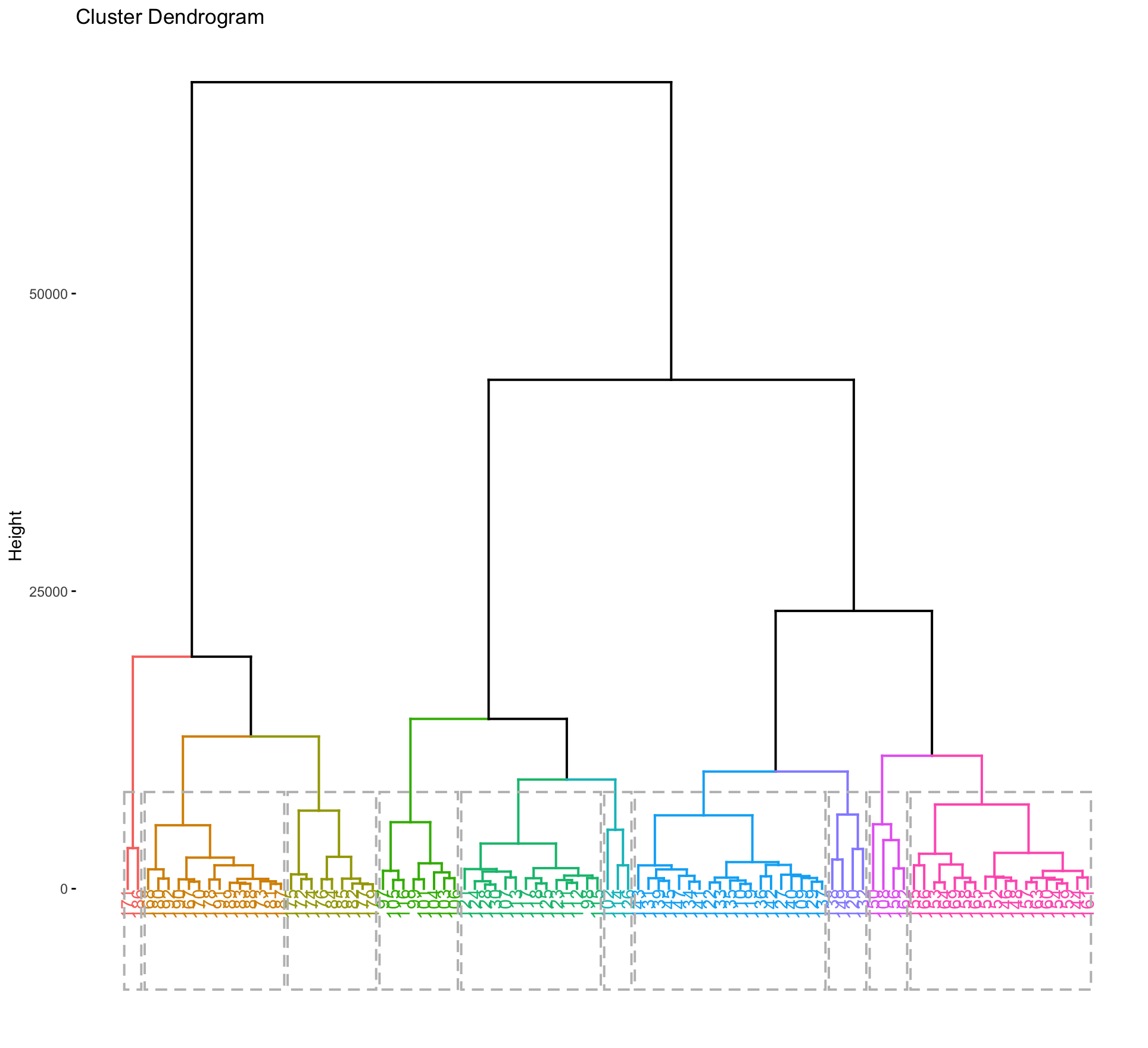
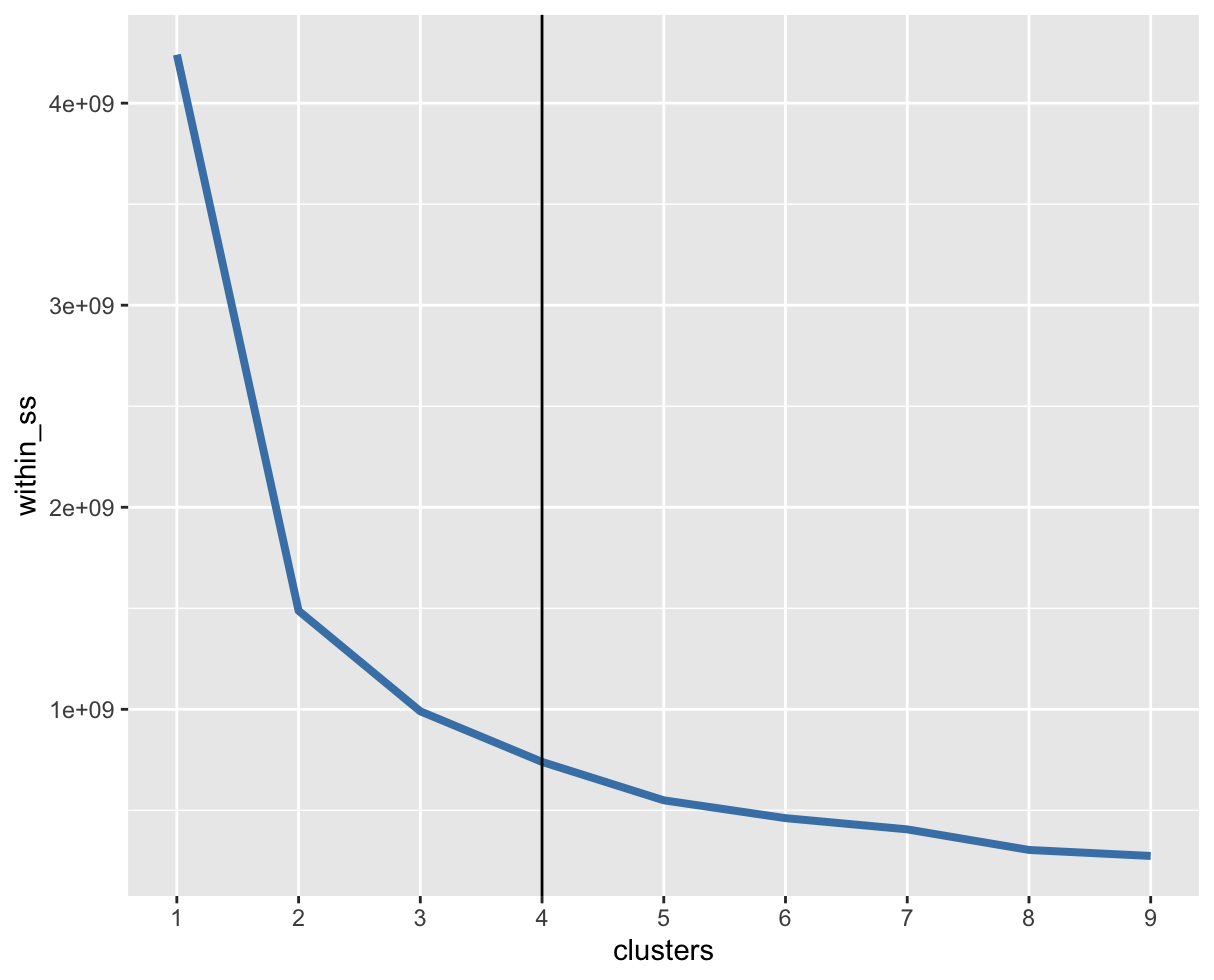
***Data Visualization***

After cleaning the crime dataset, we use R to visualize the proportion of each type of crime in the cleaned dataset. From the chart below, the top 20 crimes are displayed in descending order. It is clear that the top 3 crime types are dangerous drugs, assault and related offenses and other offenses related to theft. The cleaned dataset simply gives us an intuition of the ranking of each type of crime happening across the borough. In order to further study the relationship between the reported crime and housing project, we would like to use clustering and spatial analysis by using the combined datasets.

***Clustering Analysis***

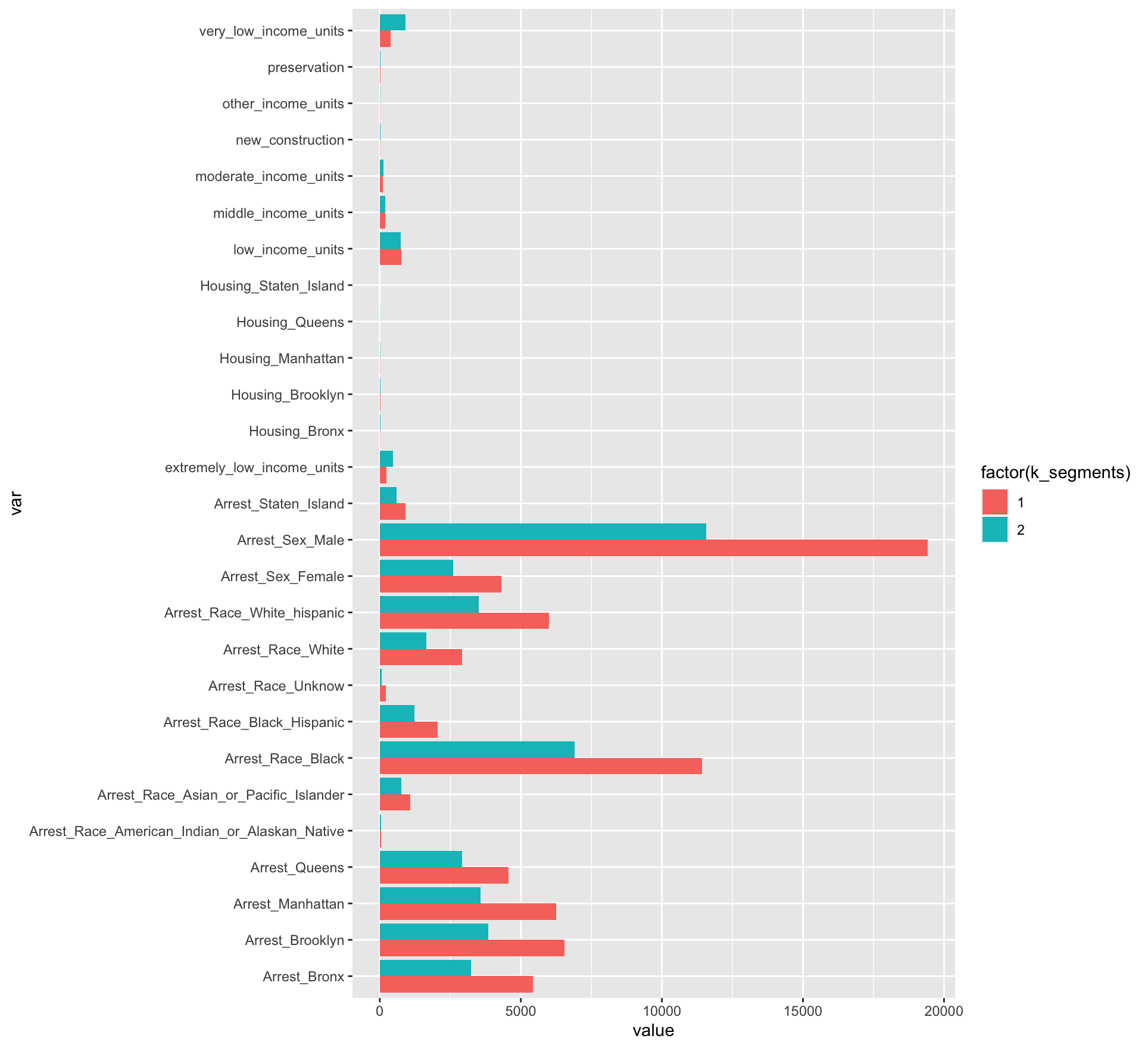
We used cluster analysis, an unsupervised learning technique, to visualise the data and assess any similarity groupings in the crime and housing data. First, we created a dendrogram to assess how many clusters to group the data by. We also used the “elbow” method via the within sum of squares plot to assess the optimal number of clusters. Both showed a two cluster solution was ideal.

Figure 1: Dendogram on Combined Dataset Figure 2: Total Sum of Squares Plot (“Elbow” Plot)

Next we used the K-Means clustering technique to group and visualize the data. After running the 2-cluster solution we attempted to extract unique identifiers for the two segments by plotting the output (figure 3).

Figure 3: Attributes of the Segments

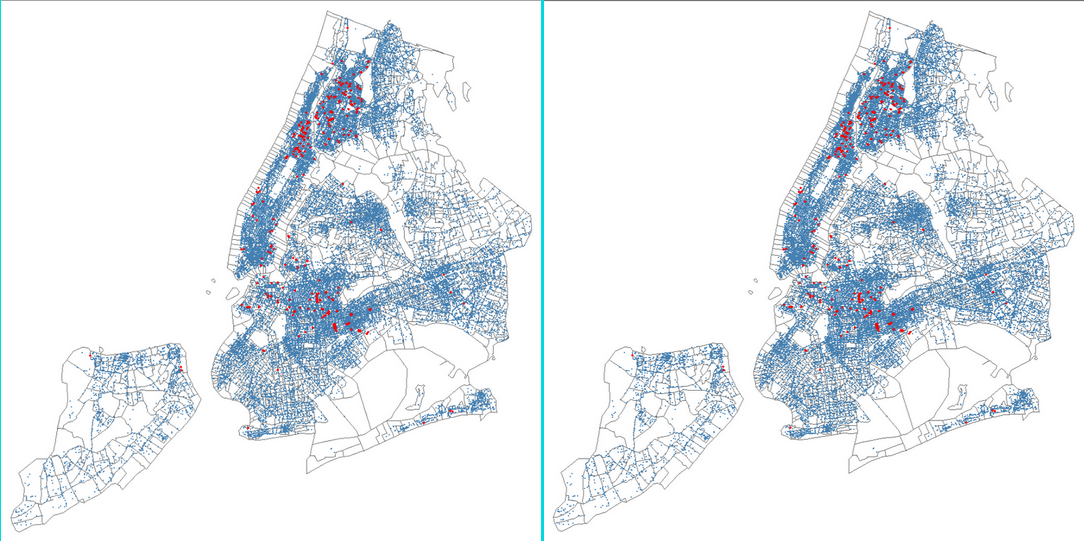


We discovered that while clustering can provide a useful visualization to assess any groupings by similarities within the data, it did not show any key differences in this case and did not help address any of our questions. Hence we used other analysis techniques to assess the relationship between crime and housing.

***Spatial analysis***

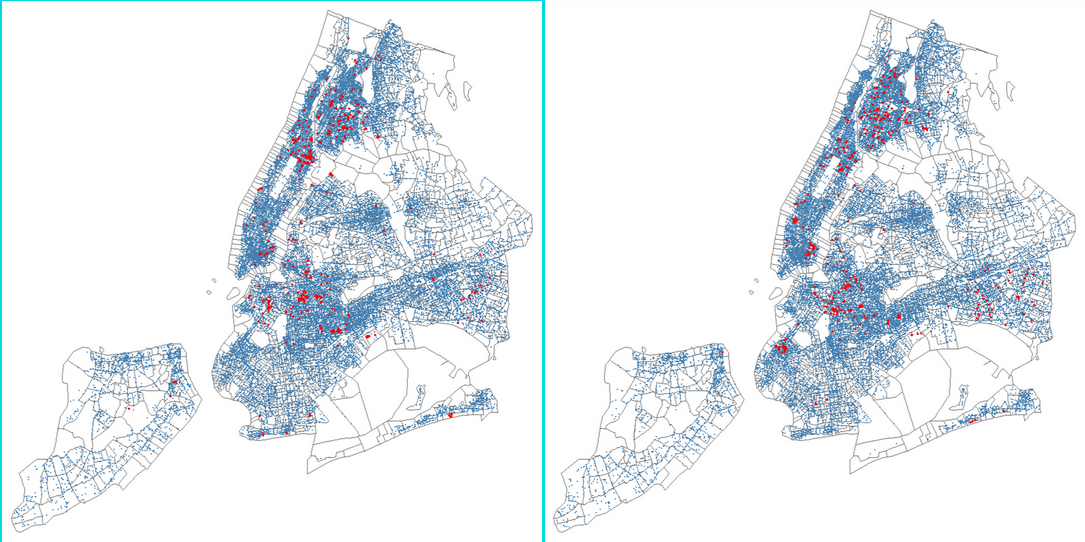
Our arrest and housing datasets came with coordinate values in the form of longitude and latitude. The coordinate values are mapped out and filtered by year to visually represent the locations of arrests and the location of housing projects that happened during the period of time. We used the US Census Shapefiles from the TIGER/Line geodatabase along with the Tigris R library to map out the city of New York and used tracts() to add to all values of New York City’s counties to make a map template. The arrests are colour coded as steel blue and the housing projects are colour coded as red.

2014 and 2015



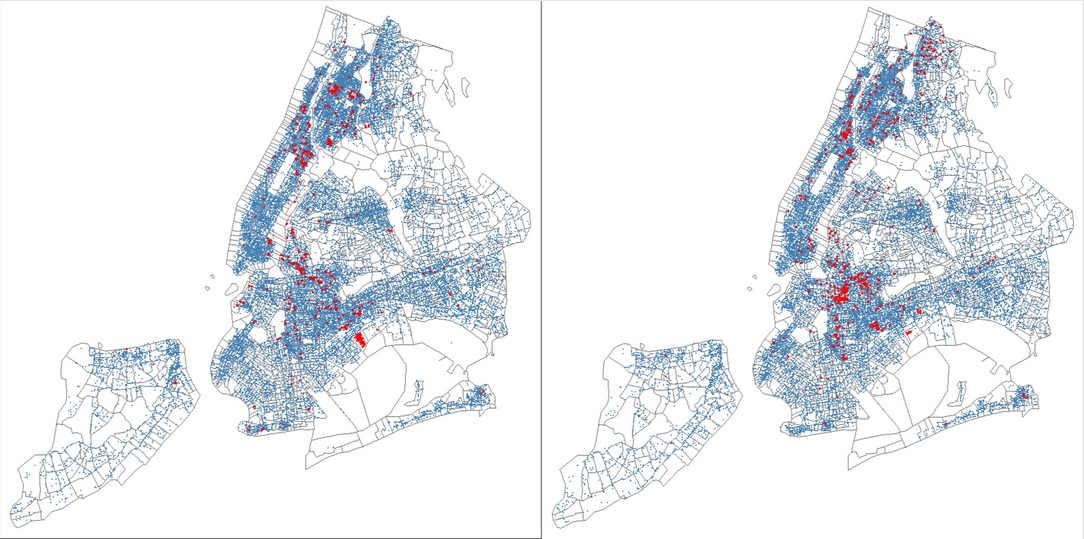
In 2014 and 2015, arrests are dense in Brooklyn, Manhattan and the Bronx. Housing projects are also dense in those areas.

2016 and 2017



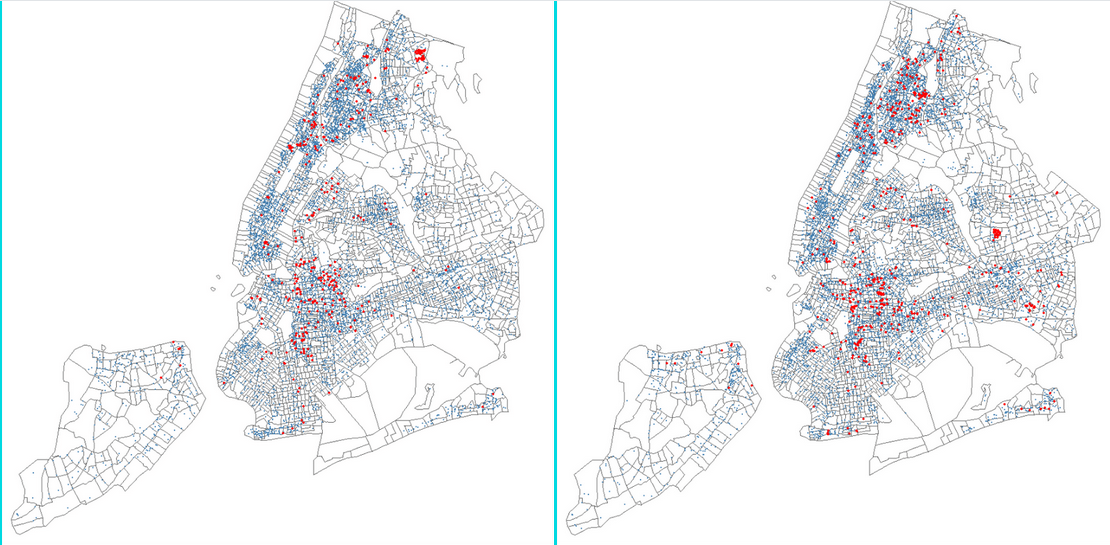
In 2016 and 2017, there is a visible downward trend in arrest numbers and an upward spike in housing areas in Brooklyn, Bronx and Manhattan. Queens do not have as many housing projects as the other boroughs but there is a noticeable increase in 2017. More housing projects are created and spread out towards less crime-populated areas.

2018 and 2019



There is a substantial decrease in arrests in the entirety of New York City since 2016. The housing projects count has skyrocketed in Brooklyn, Manhattan and The Bronx. There is a visible increase in housing projects, especially in the regions that are not as densely populated with arrests (the white areas of the map).

2020 and 2021



2020 is an anomaly due to the global pandemic that affected New York severely, therefore we can see a large decrease in the number of arrests that happen in the city. Housing projects also took a toll, but there is still a large number of housing projects created, and we can also observe the red dots in lesser steel-blue areas of the map. In 2021, the arrest numbers barely increased, but the housing project number increased to numbers parallel to those of 2018 and 2019.

So there is a visible relationship between housing projects and the number of arrests that happen in an NYC borough. Therefore, we decided to explore our hypothesis further.

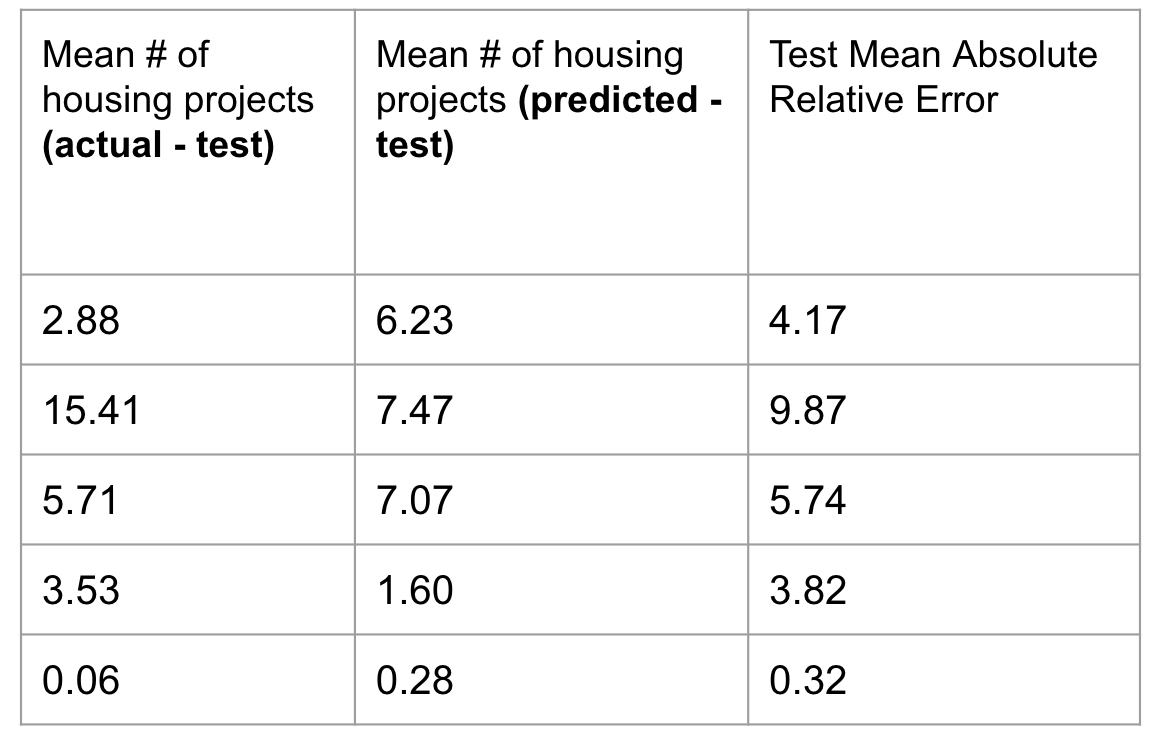
Our initial plan for spatial analysis was to create heat maps based on the number of crimes and the number of housing projects started aggregated in a zip code and within a period of time. The first step was to reverse geocode both datasets and create a new column called zip code. This is where we hit a wall. Our geocoding algorithm requires the use of the Google Maps API. We were able to acquire the key for it and receive $300 credit worth. However, due to the nature of our large dataset (5 million rows of data in the arrest dataset alone), we are able to successfully reverse geocode less than half of our dataset before the credit ran out. We eventually found another method to reverse geocode using the revgeo() library in R with the Photon method. However, there was an error with the Photon method in the package. It was an unsuccessful attempt as the zip codes came up as NULL. After more research, there was another attempt at reverse geocoding with a package from GitHub under the same name, revgeo(), which fixes the domain error that parses the zip code data into our dataset. This is where another problem became apparent. It took approximately 8 minutes to reverse geocode 1000 lines of data. From that calculation, we are able to estimate that it will take over 18 hours to do 155,000 lines of data (the year 2021 of the arrest dataset alone) and about 28.5 days to reverse geocode the entire 5 million rows of the dataset. This is where we decided to use spatial visualization instead.

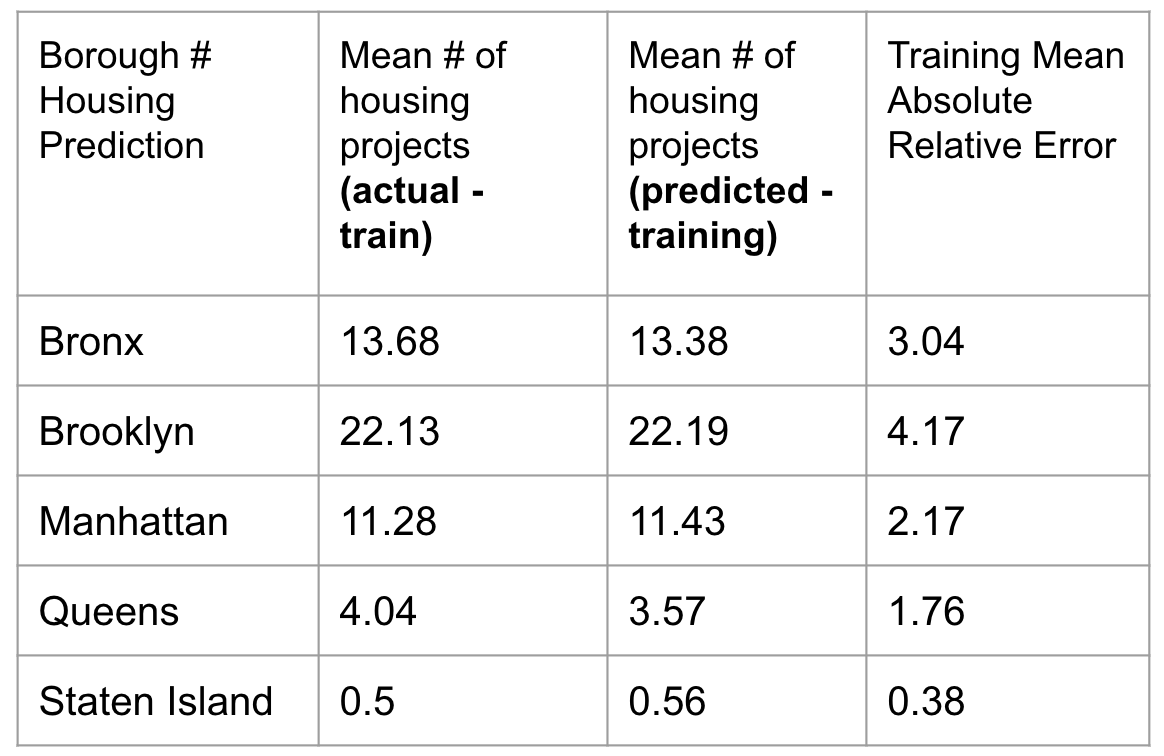
Our original plan included aggregating the count of different crimes and demographic variables in the arrest dataset and aggregating the count of different housing project types which will be joined by zip code and month and year. But we withdrew this idea as the time factor and computing power were insufficient between our members.

***Random Forest***

Next we looked into modelling techniques that may be useful in predicting housing developments. We first examined random forest models as the cuts on the tree could be indicative of the importance of crime variables in predicting housing projects. We ran 5 random forest models, one for each borough, to predict the number of housing projects in a given month.

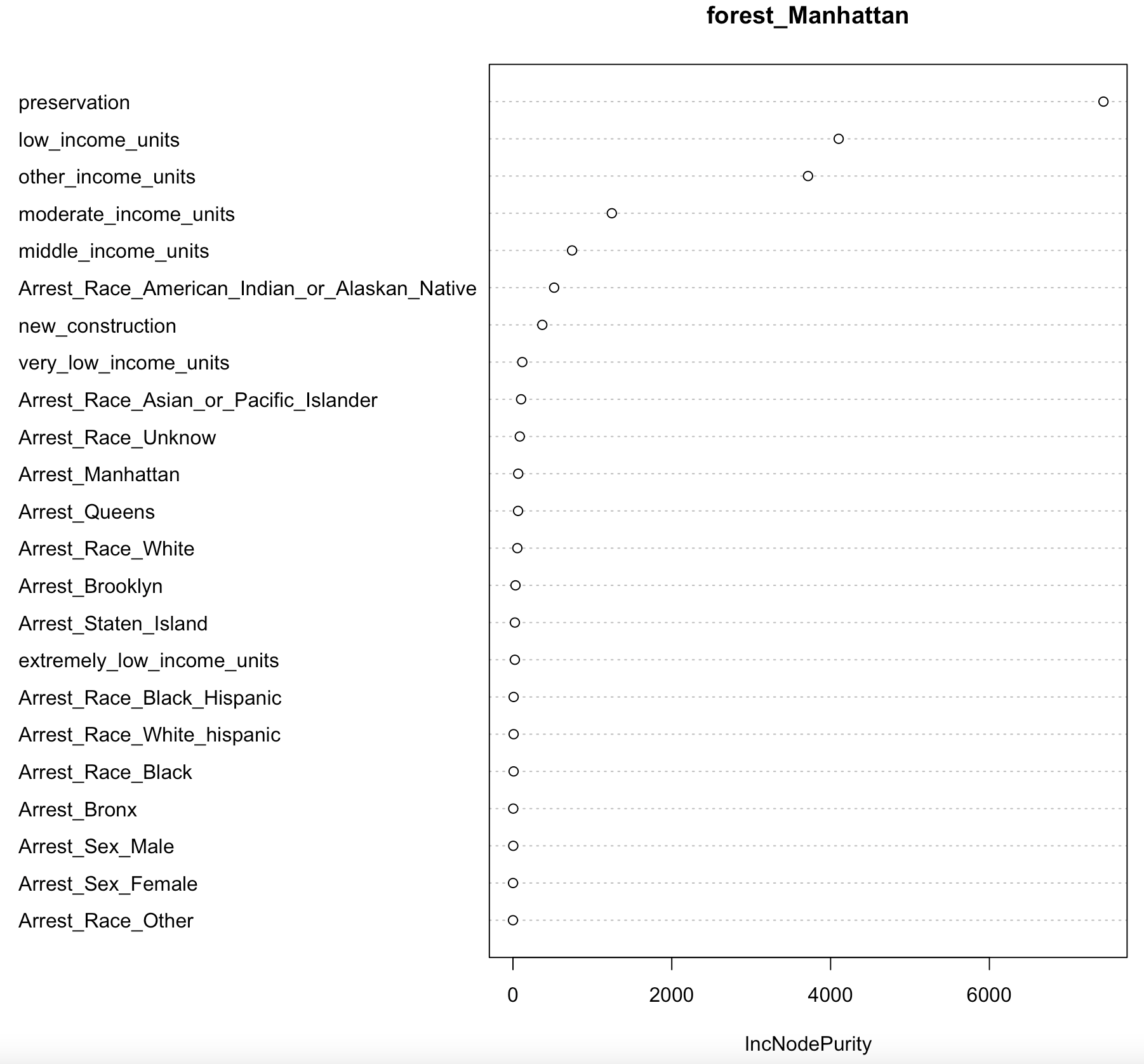
We could not use traditional accuracy measures as there were months with no housing projects. As a substitute for accuracy, we looked at a combination of the mean number of housing projects and the relative prediction error for each borough (Figure 4).

Figure 4: Random Forest Model Results



Results show that while our random forest models were able to generally predict the housing for the training data, we were unable to generalize the random forest models to new data. This shows the random forest models can not accurately predict housing projects with the crime variables. Additionally, we assessed the importance of the crime data on the number of housing projects via the variable importance plots.

Figure 5: Variable Importance Plot for Manhattan Random Forest Model



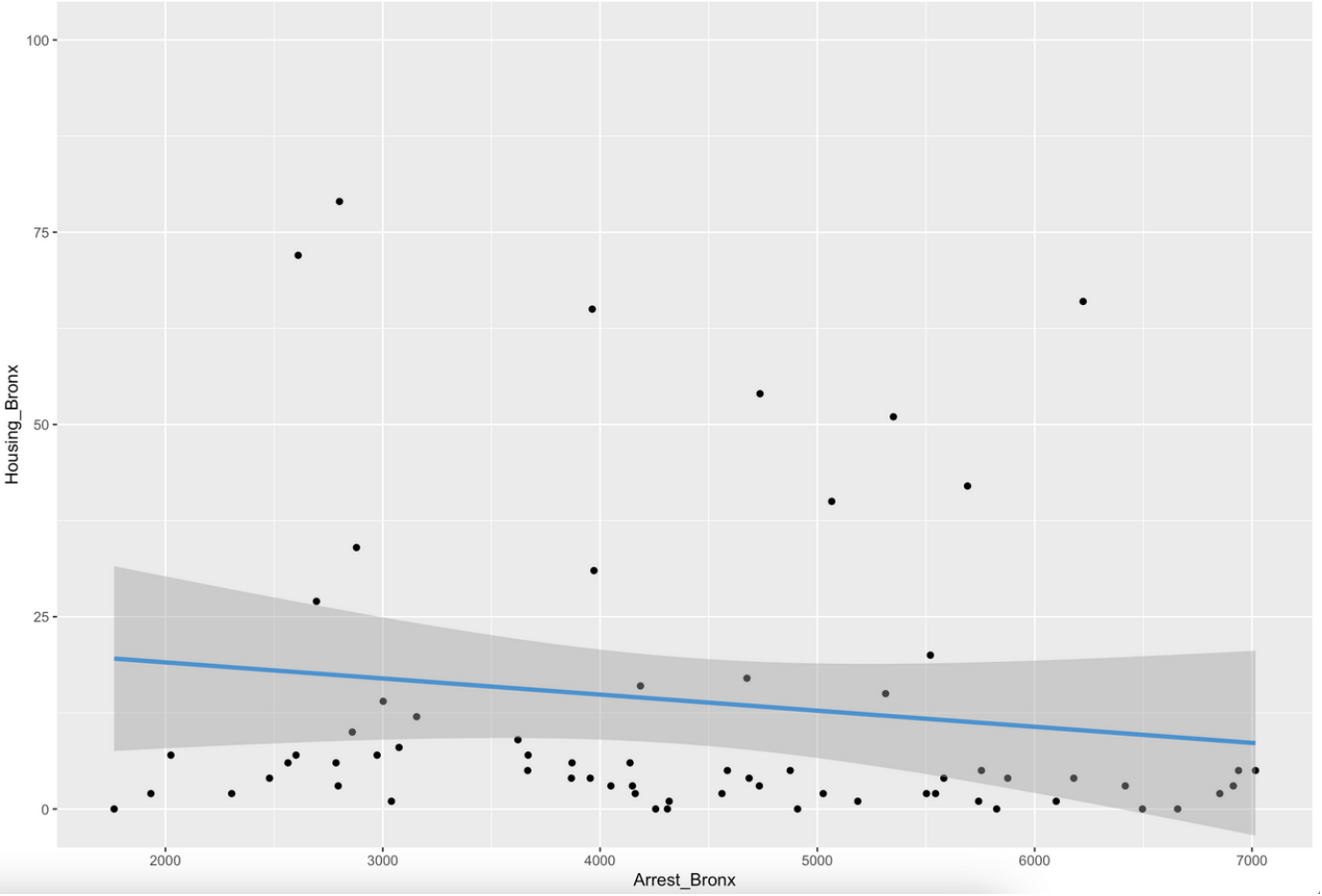
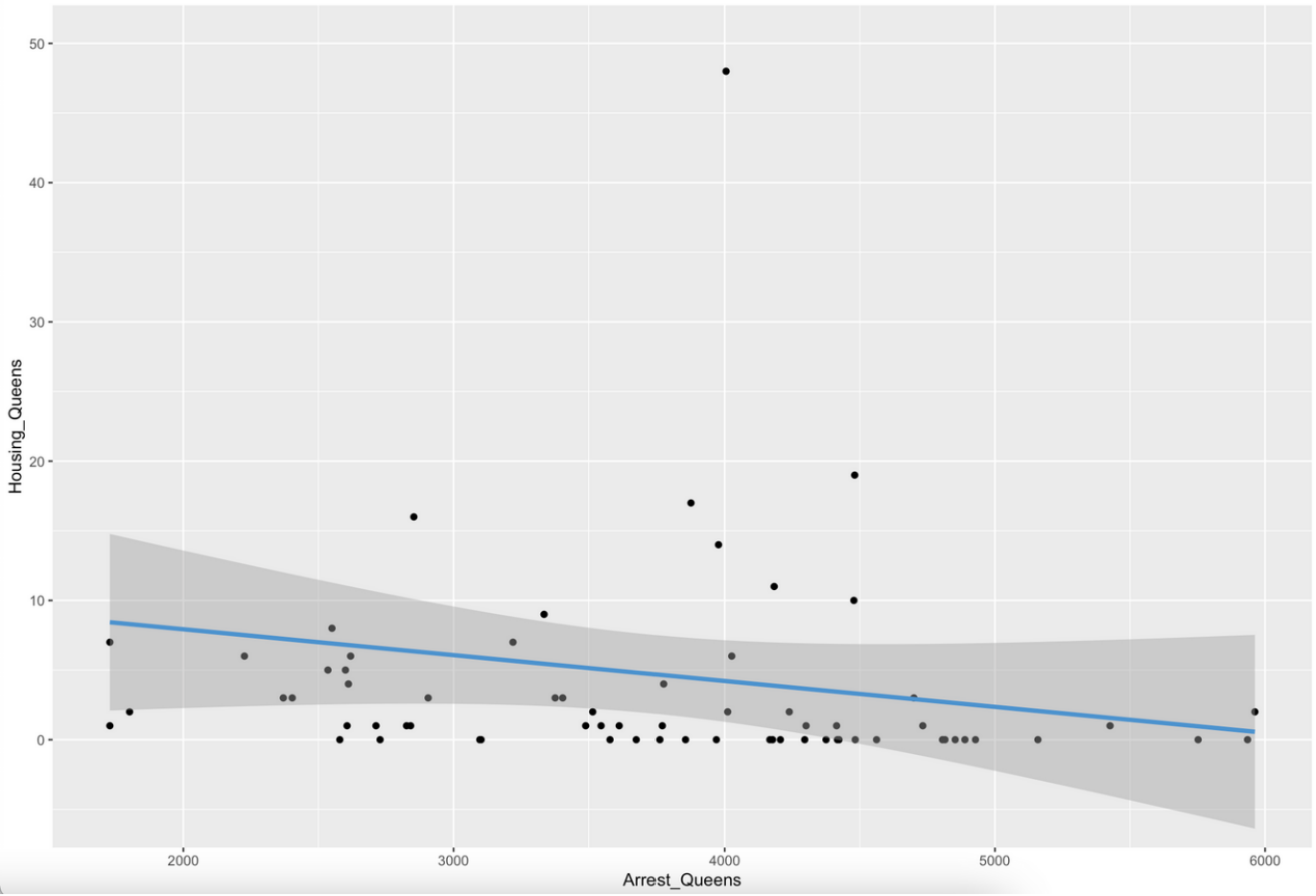
The variable importance plots showed the variables that had the greatest impact on the accuracy of the models were the housing variables, not the crime variables (figure 5). This was consistent across the random forest models for each borough. Thus, we were unable to conclude a relationship between housing projects and the number of crimes.

***Logistic regression analysis***

Logistic Regression Analysis: Housing = Coefficient \* Arrest + Intercept

|  |  |  |  |
| --- | --- | --- | --- |
| Borough/Statistics | Coefficient | Intercept | P-Value |
| Bronx | -0.00008789 | 2.879 | 0.000523\* |
| Brooklyn | -0.00001324 | 3.248 | 0.373 |
| Queens | -0.0004029 | 2.974 | 0.000001\* |
| Manhattan | -0.00004181 | 2.593 | 0.0624 |
| Staten Island | -0.001976 | 0.590636 | 0.0597 |

We also performed a logistic regression analysis to further investigate the relationship between crime rate and the creation of housing projects. We combined the statistics results into the above table. The p-values are only statistically significant for the coefficients in two boroughs, which are Bronx and Queens. The coefficients are both negative and nearly 0. That means, for the housing projects in the Bronx and Queens, the crime rate has a negative impact on housing creation, however, the impacts are very small with the crime rate alone as the independent variable. For the other 3 boroughs, their P-values are > 0.05 threshold, which means our sample did not provide sufficient evidence to conclude that the effect exists.

** **

We also conducted scatter plots to visualize the correlation between crime rates and the creation of housing projects in those two boroughs(See graphs above). Both regression lines slope slightly down to the lower right and their correlation coefficients respectively are r = -0.1263449 for the Bronx and r = -0.157342 for Queens, which are both negative and are also considered small on the scope of 0~1.

***Results & Conclusion:***

Our results show that the arrest data we procured had a relatively low impact on predicting the presence of housing projects in New York City. Our regression analysis model showed a weak linear relationship between the number of crimes and the number of housing projects in the Bronx and Queens while showing no correlation in Brooklyn, Manhattan and Staten Island. Our team concluded that the linear correlation is inconsistent across the boroughs, therefore needing further investigations to find more answers. While there is no clear picture that indicates a strong relationship between the crime rate and housing projects, the evidence we gathered is still not strong enough to reject the assumption that there is no relationship between the crime rate and housing projects.

***Recommendations & Improvements***

One area of improvement is to break out the number of crimes by type and conduct our analysis with different crime types versus an aggregate view of all crimes. The count of total crime in each borough resulted in an insignificant relationship, however, it might be the case that more violent crimes will have a stronger relationship to housing projects. In the iteration of this project, our group will look at crime and housing by the different types of crime. Another area of improvement is to incorporate other macro variables into consideration. For example, we could incorporate COVID infection rates as well as changes in median household income.

Lastly, we ran into issues with the large dataset while conducting spatial analysis. The code kept crashing thus we had to sample a small portion of our data to run our analysis. Thus for future spatial analysis, we hope to use a machine with more computing power to fully reverse geocode and carry out our initial plan. Additionally, we may also find alternative ways to aggregate the number of crimes that happen in the proximity (e.g 0.5 miles) of the housing projects and map out the differences for the next iteration of our spatial analysis.

**Bibliography**

**Literature Review References**

Ge, Jinging. “Crime In New York City: linking to income and public housing.” Medium, 23 Nov. 2020,https://jingjing-ge.medium.com/crime-in-new-york-city-linking-to-income-and-public-housing-c81cfcc29a4b. Accessed 28 Feb, 2022.

Kelly, Jack. “New Yorkers are Leaving in Droves: Here’s Why They’re Moving And Where They’re Going.” Forbes, 5 Sept. 2019, <https://www.forbes.com/sites/jackkelly/2019/09/05/new-yorkers-are-leaving-the-city-in-droves-heres-why-theyre-moving-and-where-theyre-going/?sh=22729db541ac>. Accessed 28 Feb, 2022.

Renthop. “How Much Does Crime Affect Rent Prices.” Renthop, 22 June. 2017, <https://www.renthop.com/studies/national/how-much-does-crime-affect-rent-prices>. Accessed 28 Feb, 2022.

Schwartz, Amy. Susin, Scott. Voicu, Ioan. “Has Falling Crime Driven New York City’s Real Estate Boom?” Furnman Center, 2013. <https://furmancenter.org/files/publications/Has_Fallen_Crime.pdf>, Accessed 2 March, 2022.

**Dataset references**

NYPD Arrests Data (Historic) <https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u>

NYPD Arrest Data (Year to Date)

<https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc>

Housing New York Units by Building

<https://data.cityofnewyork.us/Housing-Development/Housing-New-York-Units-by-Building/hg8x-zxpr/data>