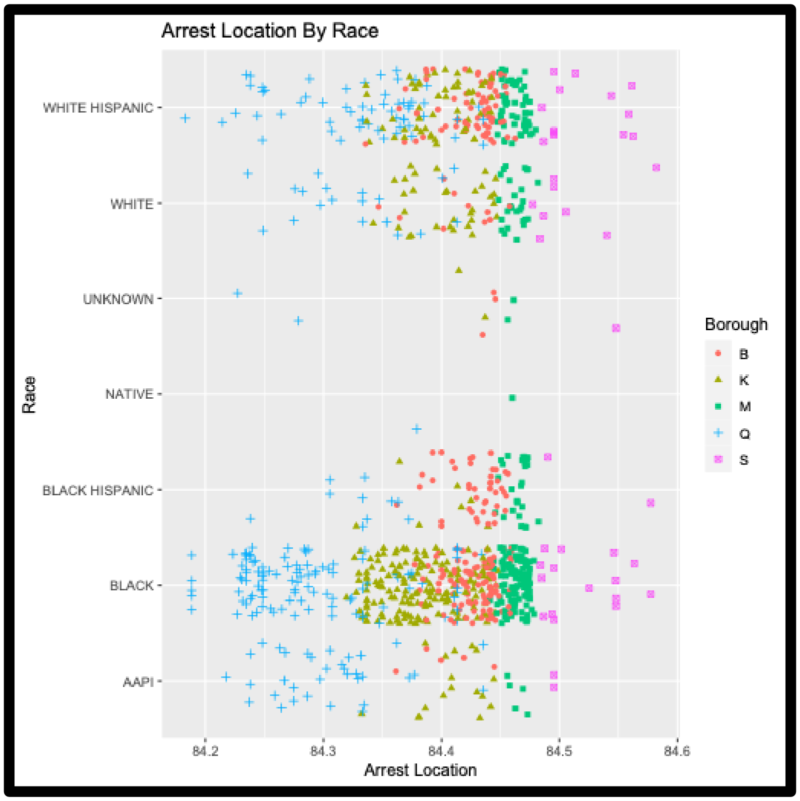
Analysis of Arrest Data for the New York City Police Department

by Sarah Weinflash



Introduction

I wanted to learn more about the issues around policing and racial segregation, and therefore selected for analysis the most up-to-date NYPD Arrest Dataset[1], which reports data on individuals and their arrests in New York City between January 1 and March 31 2023. I supplemented this with population data pulled from the US Census.[2] Through this analysis, I hope to uncover the impact of demographics and location on arrest numbers, as well as the differences between arrest statistics and overall population statistics. The dataset had 54,577 entries and 19 variables. To make the data more manageable, I removed four columns and randomly selected 1000 rows to analyze. Descriptions of the remaining variables are below.

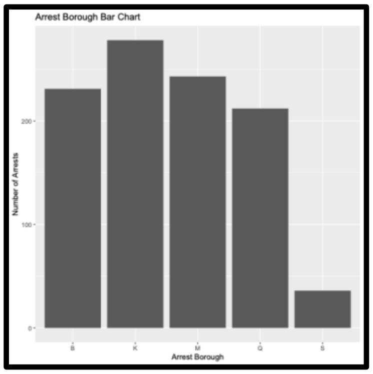
Variable Name	**Description**	
ARREST_KEY	Index	
ARREST_DATE	Arrest Date	January 1 thru March 31, 2023
PD_CD	Police Code	more granular than Key Code
PD_DESC	Description of Offense	
KY_CD	Key Code	more general than Police Code
OFNS_DESC	Description of Offense	
LAW_CAT_CD	Law Category Code	M (Misdimeanor); F (Felony); V (Violation); I (Unclassified)
ARREST_BORO	Arrest Borough	B (Bronx); M (Manhattan); K (Brooklyn); S (Staten Island); Q (Queens)
ARREST_PRECINCT	Arrest Precinct	between 1 and 123
JURISDICTION_CODE	Jurisdiction Code	0 (Patrol); 1 (Transit); 2 (Housing); 3 (Non-NYPD)
AGE_GROUP	Age Group	<18; 18-24; 25-44; 45-64; 65+
PERP_SEX	Perpetrator Sex	M (Male); F (Female); U (Unknown)
PERP_RACE	Perpetrator Race	AAPI; BLACK; BLACK HISPANIC; NATIVE; UNKNOWN; WHITE; WHITE HISPANIC
LATITUDE	Latitude	40.51 to 40.89
LONGITUDE	Longitude	-74.25 to -73.73

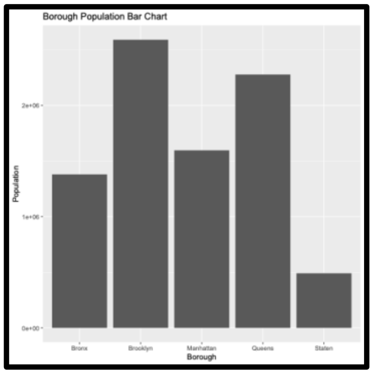
[1] New York Police Department. (2023, April 27). *NYPD arrest data (year to date)*. NYC Open Data. https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc

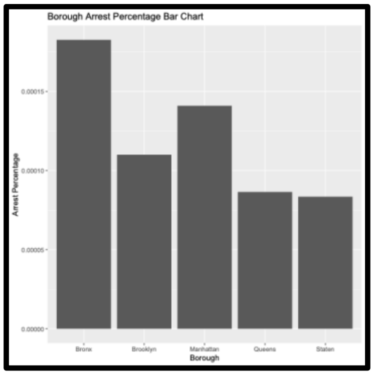
[2] *U.S. Census Bureau Quickfacts*. United States Census Bureau. (2022, July 1). https://www.census.gov/quickfacts/fact/table/richmondcountynewyork,newyorkcountynewyork,queenscountynewyork,kingscountynewyork,bronxcountynewyork/PST045222

Overview

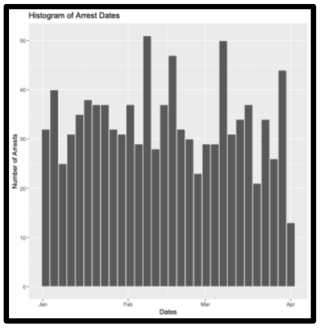
I began with an examination of Arrests by Borough, which are visualized on the bar charts below. The first shows the number of arrests by borough. As you can see, there are a similar number of arrests in each borough, with the exception of Staten Island which shows relatively few arrests. Next, I found the population of each borough, which I plotted on the second chart. On the third chart, I found the number of arrests as a percentage of the borough's population; the Bronx has a relatively high number of arrests given its population. Manhattan also has an unusually high number of arrests per resident. Brooklyn, Queens, and Staten Island have relatively similar ratio of arrests to population.

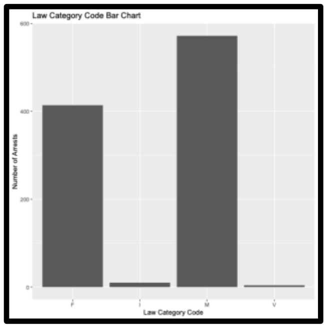

Arrest Borough Bar Chart

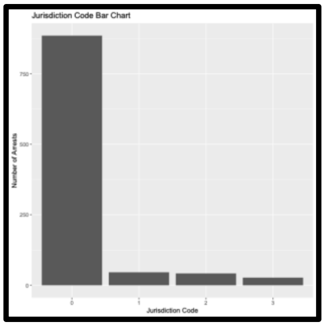

Borough Population Bar Chart

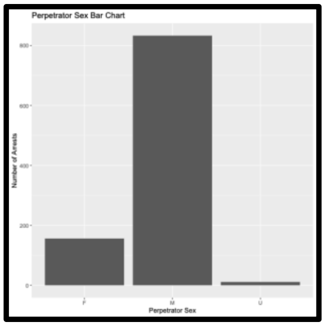

Borough Arrest Percentage Bar Chart

Next, I created a histogram showing the number of arrests by date; there seem to be more arrests in warmer months, but the pattern is not clear, and the data is limited to the first three months of the year. The bar chart for Law Category Code indicates the majority of violations were Misdimeanors (M) and Felonies (F). The bar chart of Jurisdiction Code shows that most of the arrests were made by patrol officers. I also found the sex of individuals arrested, which indicates that over four times the number of males have been arrested this year than females. Very few individuals whose sex is unknown or does not fit into the male/female binary were arrested in this time period.

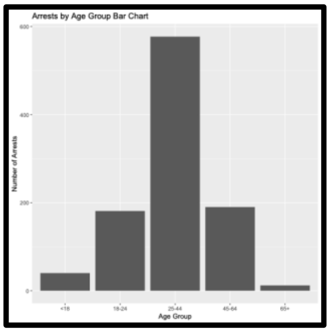

Histogram of Arrest Dates

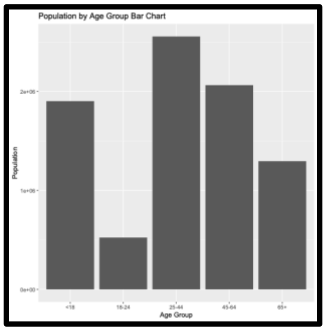

Law Category Code Bar Chart

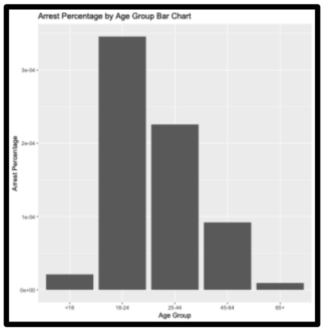




I then looked at arrests by age group, which are depicted in the bar charts below. The first shows the number of arrests by age group, which indicates a very high number of arrests of individuals aged 25-44. However, the age group 18-24 only captured six years, whereas the others each encompassed around 20 years. On the second chart, which shows the population of each age group, I expected 18-24 year olds to be about a third smaller than the other categories, but it seems like they represent even less. On the third chart, I found the number of arrests as a percentage of the age group's population. It is clear that, although 18-24 year olds represent a relatively small portion of the population, they are arrested at much higher rates.

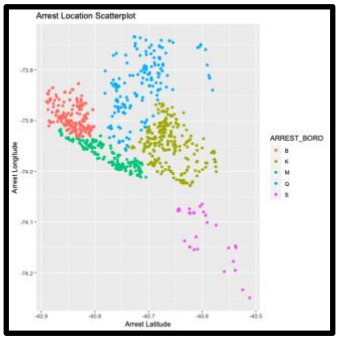


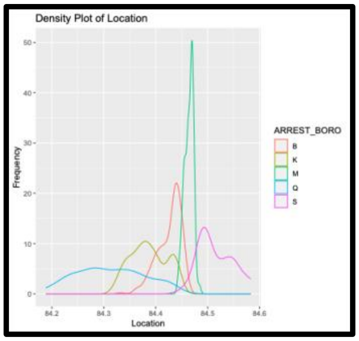




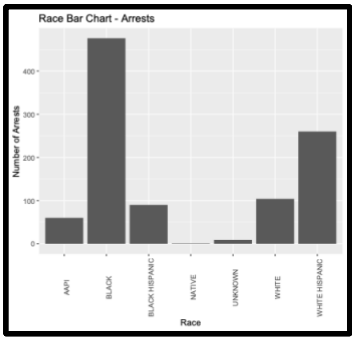
I also created a scatterplot of the Latitude and Longitude of arrests. Plotting maps is outside of my skillset, but this was an easy way for me to visualize the location of arrests. The Latitude had to be reoriented as a negative value, because Latitude moves from East to West, making the map (which plots from left to right, or West to East) appear in reverse. I also colored the points by borough, to further orient myself.

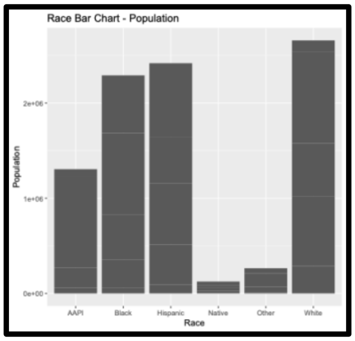
I realized that, in order to perform regression analyses surrounding location, I would need to create a single variable which captured both the Latitude and Longitude. I used Pythagoras's Theorem to create a new variable entitled "Location". I plotted the densities of each variable in the second plot below. Manhattan is a small but populous island, so arrests necessarily take place near each other - you can see this in the green spike around 84.48. Queens is larger, so its arrest locations are more spread out, which is reflected by a softer curve.





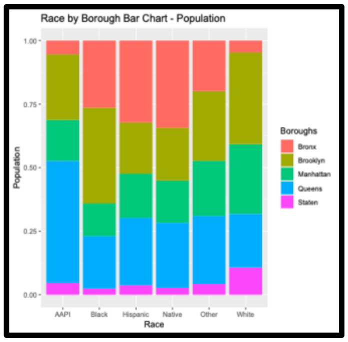
I wanted to know the races of individuals arrested, which is the first plot below. I compared that with the races of the population of New York City, visualized in the second plot. As you can see, while White New Yorkers represent the majority of the population, they represent a very small percentage of individuals arrested. On the other hand, Black New Yorkers are arrested at much higher rates than their population levels would predict.

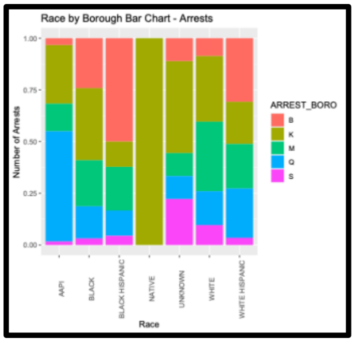


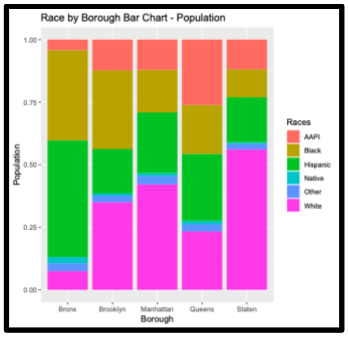


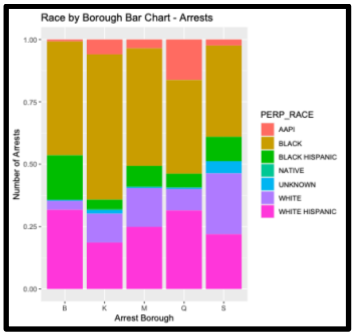
The four charts below show similar ideas, but in different ways. The first chart shows where people of each race live. For example, approximately half of all AAPI New Yorkers (the first bar) live in Queens (the blue segment). The next chart shows the same thing, but instead of population, it shows where individuals of that race are arrested. The most obvious difference here is that, although New Yorkers of Native American heritage (the third bar) are fairly evenly distributed throughout the city (as seen in the first chart), all of their arrests occurred in Manhattan (the olive green segment, as seen in the second chart).

The third chart shows the racial breakdown of each borough. The last chart shows the racial breakdown of arrests in each borough. Although Staten Island (the last bar) is shown in the third chart to have about 50% White residents (the magental segment), they only account for about 20% of individuals arrested (as seen in the last chart).









I am very curious about this dataset's amenability to regression analysis. I am excited to explore the findings, but I worry, since the variables are largely categorical, that this will limit my ability to produce statistically significant findings.

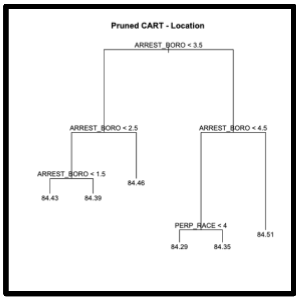
Classification and Regression Tree (CART)

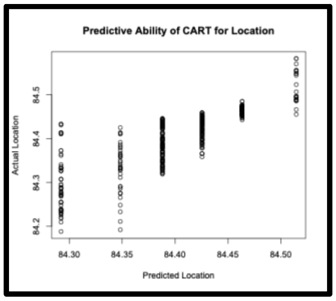
To determine which to utilize as the response variable, I created CARTs for each variable in the dataset; most of these produced single-node CARTs, indicating a single prediction regardless of other variables. For example, the predicted Race of the person arrested was always Black, regardless of Location, Borough, Age, etc. The only variables that produced multi-node CARTs were Location (predicted by Borough and Race) and Borough (predicted by Location). I decided to move forward with predicting Location using Borough and Race, as multi-variable predictors with a continuous outcome seemed more interesting. Because the outputs of the CART were difficult to read utilizing named categorical variables, I reconfigured the variable categories as numbers; they are listed below.

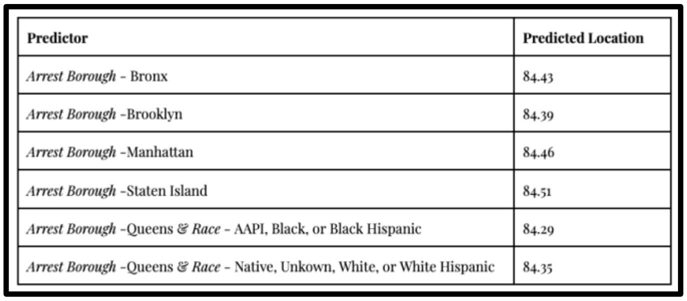
Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Arrest Borough	Bronx	Brooklyn	Manhattan	Queens	Staten Island	-	-
Perpetrator Race	AAPI	Black	Black Hispanic	Native	Unknown	White	White Hispanic

I created a chart describing the output of the CART with a little more clarity, which is below, along with the CART output. As you can see, the locations were predicted primarily by Borough, with the exception of Queens, which is quite large. In that case, the CART formula segmented Queens into two parts by race: AAPI, Black, and Black Hispanic in one area, and Native, Unknown, White, and White Hispanic in another area.

I then performed a prediction test using the CART formula, and plotted the predictions against the actual values. Because the location values were continuous, I found it difficult to really understand whether the technique was successful. I therefore calculated the Mean Square Error, which was 0.0013. This is quite small, but in the context of the extremely limited range of this data (the Bronx and Brooklyn are predicted to be only 0.04 Location points away from each other), it didn't give me as much information as I would have liked on the success of the prediction. I therefore transformed the outcomes into discrete variables by rounding to one decimal point. Using this test, the formula was found to be 79.6% successful in making predictions about location.







Predictor	**Predicted Location**
Arrest Borough – Bronx	84.43
Arrest Borough –Brooklyn	84.39
Arrest Borough –Manhattan	84.46
Arrest Borough –Staten Island	84.51
Arrest Borough –Queens *& Race* – AAPI, Black, or Black Hispanic	84.29
Arrest Borough –Queens *& Race* – Native, Unkown, White, or White Hispanic	84.35

Regression Techniques

While the 79% success rate from the CART prediction is strong, I wanted to explore what other models might predict. I created model matrices to stand in for the x variables utilizing the training and testing data created earlier. Least Squares Regression utilizes a Linear Model; the process to create this model is described on pages 11-12 of this report. The model generated by this process is below.

Least Squares Regression: $Location \ = \ 84.136 + (1.522 \times 10^{-5} \times \ Date)$
$+(3.426 \times 10^{-4} \times Law\ Category) + (-7.205 \times 10^{-3} \times Borough) + (-5.849 \times 10^{-3} \times Jurisdiction)$
$+(3.827 \times 10^{-3} \times Age\ Group) + (-1.003 \times 10^{-2} \times Sex) + (2.593 \times 10^{-3} \times Race)$

Ridge Regression minimizes the sum of squared error just like the Linear Model, but it also minimizes ß (the slope). While Least Squares Regression creates an unbiased estimator of ß, Ridge Regression creates a biased estimator in an attempt to reduce the variance. It requires the sum of the squared ß values to be less than a pre-determined constant, often limiting the size of the ß values.[3] Utilizing a cross validation function on the Generalized Linear Model, I found the ideal constant (0.0625). The model generated by the Ridge Regression function is below. As you can see, the values of ß are smaller than those in the Least Squares Regression equation.

Ridge Regression: $Location \ = \ 84.233 + (1.118 \times 10^{-5} \times \ Date)$
$+(2.717 \times 10^{-4} \times Law\ Category) + (-4.444 \times 10^{-3} \times Borough) + (-3.195 \times 10^{-3} \times Jurisdiction)$
$+(2.746 \times 10^{-3} \times Age\ Group) + (-6.180 \times 10^{-3} \times Sex) + (1.622 \times 10^{-3} \times Race)$

Lasso Regression (an acronym for Least Absolute Shrinkage and Selection Operator) is similar to Ridge Regression, but limits the sum of the absolute values of ß (rather than their squared values).[4] This allows ß to shrink down to zero, thus creating a model with fewer predictors when those predictors are inefficient, but without limiting the ß values as severely as the Ridge Regression might.[5] Utilizing a cross validation function on the Generalized Linear Model, I found the ideal constant (0.000928) to restrain the ß values. The model generated by the Ridge Regression function is below. The values of ß are smaller than those in the Least Squares Regression equation, but larger than the equation for Ridge Regression.

Lasso Regression: $Location \ = \ 84.415 + (-0.004 \times Borough) + (0.001 \times Race)$

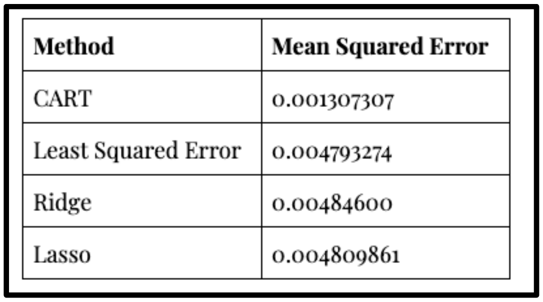
[3] Li, J. (n.d.). *Ridge Regression*. Applied Data Mining and Statistical Learning. https://online.stat.psu.edu/stat857/node/155/

[4] Li, J. (n.d.). *The Lasso*. Applied Data Mining and Statistical Learning. https://online.stat.psu.edu/stat857/node/158/

[5] Le Menestrel, Thomas. (28 April 2022). *Lasso and Ridge Regression: An Intuitive Comparison*. Towards Data Science. https://towardsdatascience.com/lasso-and-ridge-regression-an-intuitive-comparison-3ee415487d18

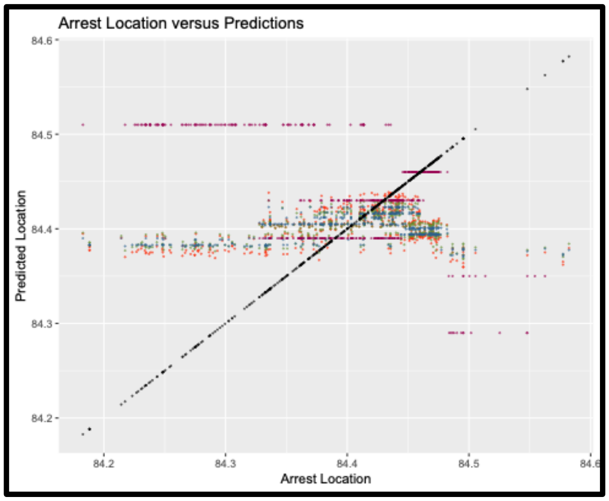
Mean Squared Error

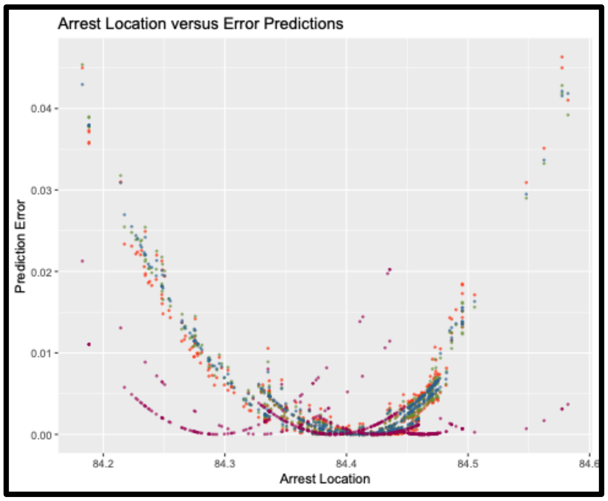
The outcomes of the Mean Squared Error values are below. These are used to determine the average amount of error (that is, the difference between the predicted and actual Location values). As you can see, CART had the lowest Mean Squared Error by quite a lot. The Least Squared Error, Ridge, and Lasso Regression functions all showed approximately the same amount of error, with the Least Squared Error model showing the least amount of squared error, followed by the Lasso Regression and then Ridge Regression.



Method	**Mean Squared Error**
CART	0.001307307
Least Squared Error	0.004793274
Ridge	0.00484600
Lasso	0.004809861

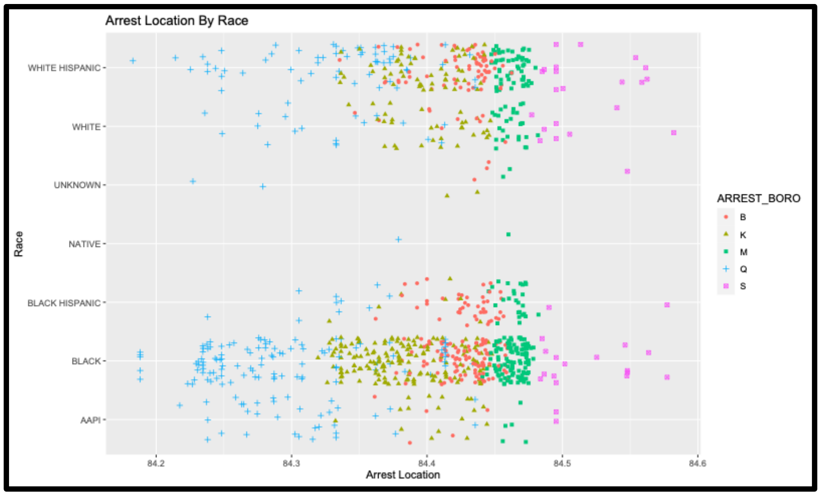
In the charts below, the purple dots represent the CART predictions, the orange dots represent Least Squared Errors predictions, the green dots represent the Ridge Regression predictions, the blue dots represent the Lasso Regression predictions, and the black dots represent the actual Locations. What I notice first is that most of the predictions except the CART tend to hover between 84.35 and 84.45. This is where the majority of the Locations are, but not *all* the Locations - and this is key. In the next chart, I plotted the prediction error against the actual Location. You can see that for the Arrest Locations between 84.35 and 84.45, the Lasso, Ridge, and Least Squared Error Regression predictions are the most accurate. However, in all other Locations, the CART predictions are more accurate. This really helped me elucidate exactly why the CART predictions had significantly less error than the other models' predictions.





Conclusion

I'd like to conclude with a final graph which helped me visualize some of the outcomes I explored, specifically those which found some interaction between the Race of those arrested and the Borough and Location in which they were arrested. This is a jittered scatterplot, with Location on the x axis, Race on the y axis, and Borough indicated by the color and shape of the points on the graph. The jittering spreads the points out, and helps to conceptualize the density of the data.



Black and White Hispanic individuals are arrested at much higher rates than any other Race. This is true across most Boroughs and Locations, though where their arrests occur is interesting: specifically, Black individuals are arrested in a different part of Queens than White Hispanic individuals are. Using the Location differences outlined in the CART model, I found the corresponding Latitudes and Longitudes leading to that Location in Queens. I discovered that the area of Queens in which higher rates of White Hispanic individuals were arrested is 49.5% Hispanic.[6] This explains why they may be arrested at higher rates, though my earlier finding showed that population demographics are not necessarily an indicator of the demographics of individuals who are arrested.

While I was not able to uncover any new truths about systemic racism, the prison industrial complex, or even how policing works in New York City, the patterns I discovered in the dataset were very interesting to me. Specifically, the ones that didn't follow population patterns, such as age, race, and borough. Unfortunately, I'm not sure that the Location variable I created was the best one to analyze this data. I hope to return to this material in a few weeks and find a new way to transform Latitude and Longitude; I believe that would enable me to make more accurate predictions about arrest location to be performed.

[6] New York University. (n.d.). *Elmhurst/Corona Neighborhood Profile*. Furman Center for Real Estate and Urban Policy. https://furmancenter.org/neighborhoods/view/elmhurst-corona