Final Project 2

Sophie Poole

2024-11-27

Abstract

For my project, I created a linear model for predicting the total number of hate crimes in the New York City boroughs. I combined data that included the total number of hate crimes, the number of offenders, and number of victims given the borough and year (from 2010 to 2022) with election data (percentage of people who voted and percentage of people who voted Democrat in that year's most recent primary election), racial demographics (estimated count from an annual American Community Survey), and median and mean income (also from the American Community Survey) for each borough. I used backwards elimination and stepwise regression to determine which predictor variables gave the best fit for this model. I used boxcox to transform the data, then used the ncvTest and Shapiro-wilks test to check the diagnostics and compare both models. The model obtained by stepwise regression had a larger R-squared value, and both models showed constant variance and were not normally distributed. I checked the fit of the model by analysing the Mean Squared Error (MSE) which closest to 0 for the smallest boroughs. I compared this to a model re-fit to data excluding leverage points, outliers, and influential points; this model had a smaller MSE for three of the five boroughs (which suggests a better fit). Lastly, I used pairwise comparison to compare the boroughs if they were the only predictor for the number of incidents; it showed that there was no significant difference in total incidents between the Bronx and Staten Island and between Manhattan and Brooklyn. From the model summary, the county the hate crime is committed in has the largest affect on predicting the total number of hate crimes.

Introduction

New York City, the largest city in the United States, is made up on 5 boroughs: the Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Each borough has their own cultures and personalities. Manhattan is one of the world's most famous business and financial hubs. Brooklyn and Queens have the largest populations of the boroughs and are also the most diverse racially. Lastly, the Bronx and Staten Island are the smallest and are home to many low income and working class families.

While these cultures can form strong communities, they can also form biases and stereotypes. A crime motivated by bias against "race, color, religion, national origin, sexual orientation, gender, gender identity, or disability" is considered a hate crime. Hate crimes could be committed against people or property, it could have multiple people committing it, or have multiple people victimized by it. In such a large city, such as New York, it's important to understand what factors contribute to the number of hate crimes.

These factors could be the borough's population, its political standing, and even election years (since election years tend to be very stressful). From this project, I hope to understand which factors are most significant and how it affects the number of predicted hate crimes in a certain borough.

Data set

The main data set used was taken from data.ny.gov. This data set is for all the counties in the state of New York from 2010 to 2022. However, I will just be looking at New York City's 5 boroughs: Bronx, Brooklyn, Manhattan, Staten Island, and Queens. For each borough and year (which is represented as the rows), it includes the Crime.Type (Crimes Against Persons or Property Crimes), the bias for the hate crime (ex. based on gender, age, race, religion, ethnicity, sexuality, disability), the total number of incidents, total number of offenders, and total number victims. I have called this data set nycCrime.

I have combined that data set with the "Annual Population Esimates for New York State and Counties" (also downloaded from data.ny.gov). This data set covers all the counties in New York State, so I've filtered it to only include New York City's boroughs. The distribution of the population amongst the 5 boroughs throughout the years I'm analyzing is shown in Figure 1.

I found the election percentages and number of votes for the 2008, 2012, 2016, and 2020 election from the New York Times and Politico, shown in Figure 2. The election data from 2008 has been added to years 2010 to 2011 in the nycCrime, election data from 2012 has been added to 2012 to 2015, and the pattern follows for election data for 2016 and 2020.

Lastly the race demographics for each borough, shown in Figure 3, and the income median and mean,shown in Figure 4, were taken from the United States Census Bureau. The race demographics are from the American Community Survey (ACS) Demographic and Housing Estimates (DP05). The income data is from Income in the Past 12 Months (S1901), also from ACS.

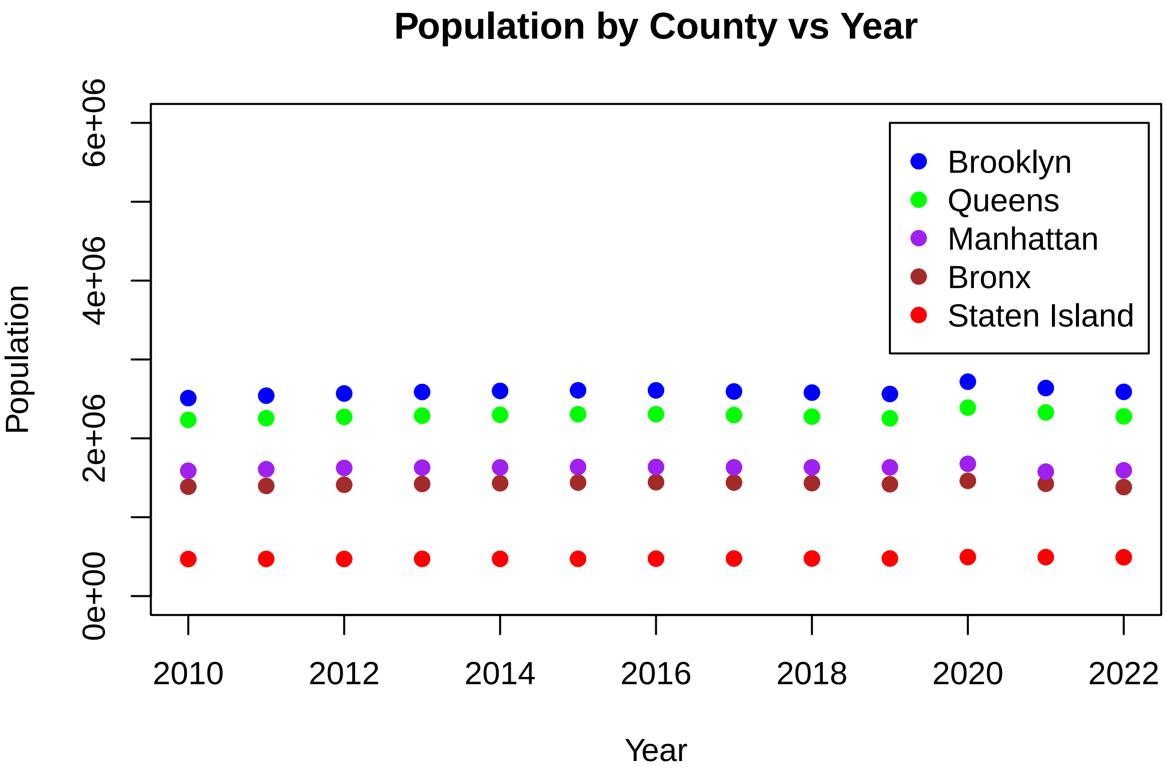

Figure 1: Population vs Year color cordinated by borough

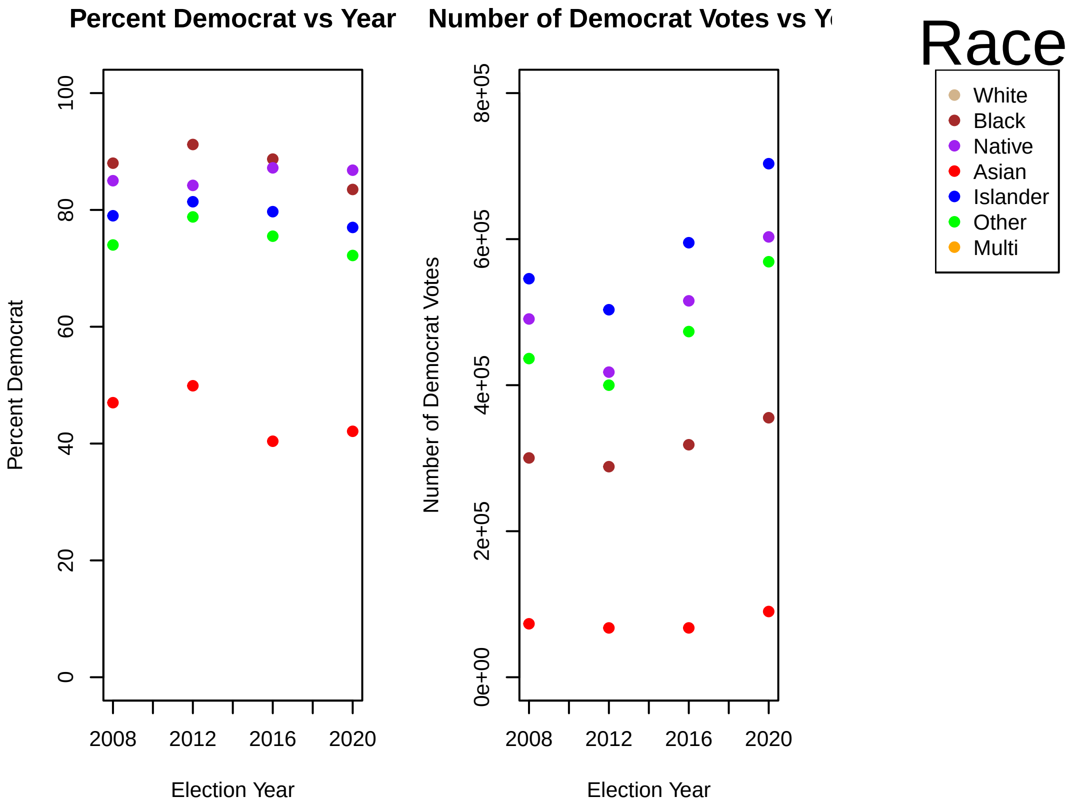

Figure 2: Left: Percent Democrat by county vs Election Year, Right: Number of Democrat Votes by county vs Election Year

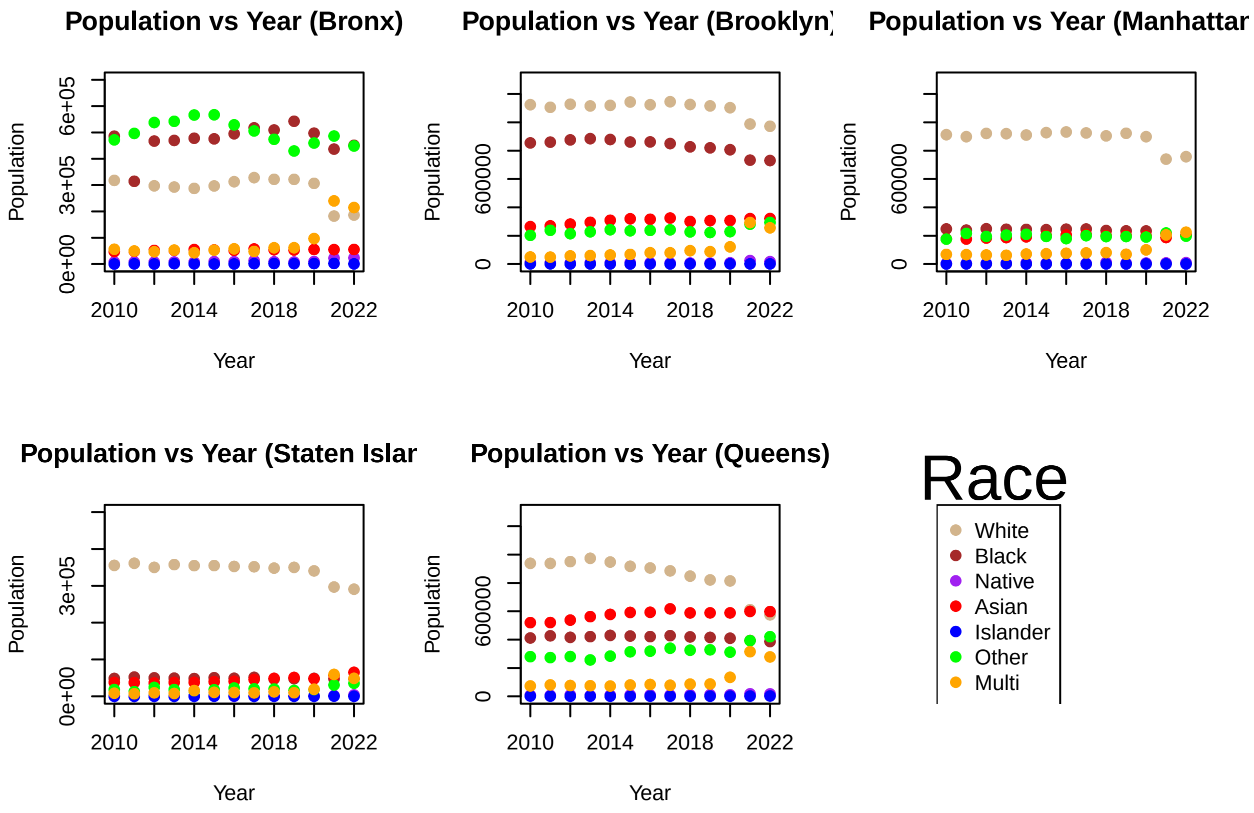

Figure 3: Population vs Year for each borough color cordinated by race

Analysis

For each initial model, we will use the linear model function to predict our response variable with the following as the predictor variables:

- County (Categorical: Brooklyn, Queens, Manhattan, Bronx, or Staten Island)
- Ratio (Total.Offenders/Total.Victims)
- CountyPop (county's population)
- ElectionYear (boolean where TRUE if that year was an election year)
- Perc.Dem (percentage of voters that voted democrat in the most recent election for that year)
- PercVoted (percentage of the county's population that voted in the most recent election for that year)
- White.pop (White population)
- Black.pop (Black or African American population)
- Native.pop (American Indian or Alaskan Native population)
- Asian.pop (Asian population)
- Islander.pop (Native Hawaiian or Pacific Islander population)
- Other.pop (Other race population)
- Multi.pop (Two or more races population)
- Income.med (Median income)
- Income.mean (Mean income)
- Year

The model will be constructed from 95% of the original data, this will become the train set. The test set is from the other 5% of the data which we can use later on to determine how well our model did at predicting the response variable.

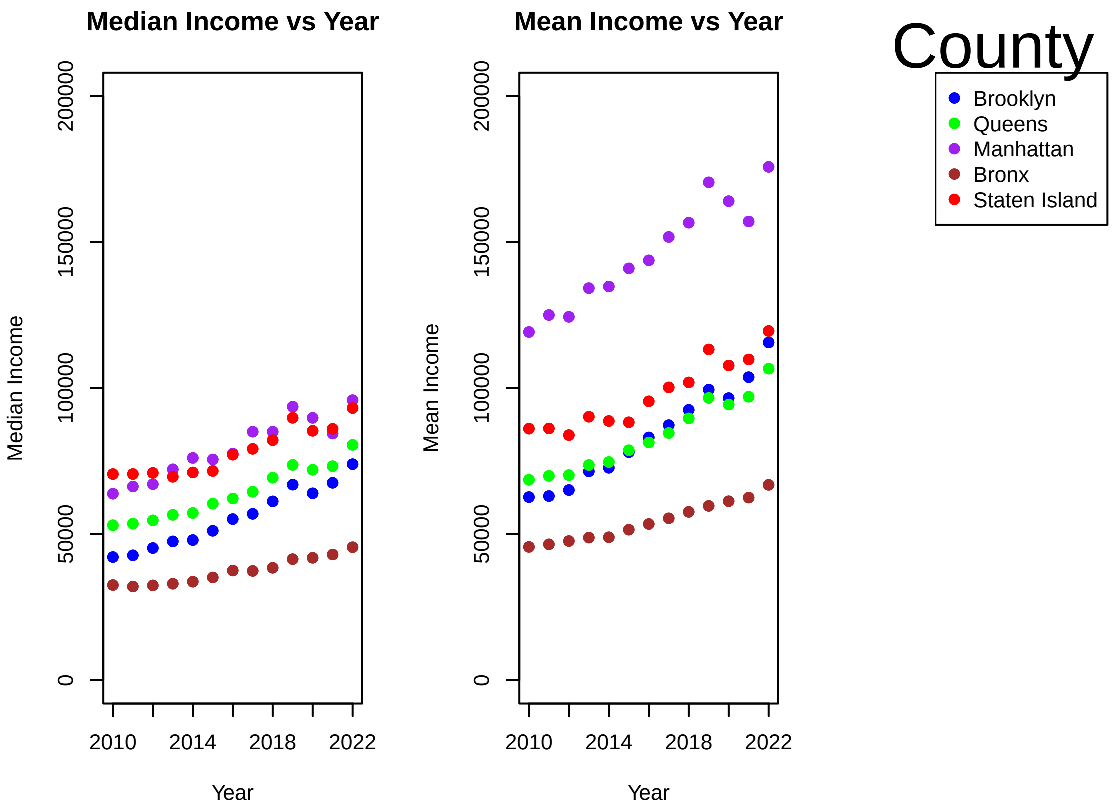

Figure 4: Left plot: Median Income by county vs Year, Right plot: Mean Income by county vs Year

We will use backwards elimination and stepwise regression (with AIC) to determine which of the predictor variables are best to use for the model. Then we use box-cox to determine if a transformation is appropriate for that model.

We will then perform regression diagnostics which include checking assumptions, finding leverage points, outliers, and influential points, and determining whether including these points makes the model fit better or worse. Lastly, we can check for collinearity of the predictor variables. This is important when analyzing the individual affect each predictor's estimated coefficient has on the predicted value.

Linear Model for Total.Incidents

Below is the summary of this model. It has a significant p-value and an Adjusted R-squared of 0.7455 The closer the R-squared value is to 1, the more the model represents a linear relationship. It also shows that many of the predictor variables are not significant. Next we can use backwards elimination, removing the predictor variable with the highest p-value above 0.05 until we're left with all significant predictor variables.

```
## [1] "Summary of the full model"

##
## Call:
## lm(formula = Total.Incidents ~ County + Crime.Type + Ratio +
##     CountyPop + ElectionYear + Perc.Dem + PercVoted + White.pop +
##     Black.pop + Native.pop + Asian.pop + Islander.pop + Other.pop +
##     Multi.pop + Income.med + Income.mean + Year, data = train.nycCrime)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.445  -8.595  -1.035   6.306  48.795
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.408e+03  2.945e+03   0.478 0.633557
## CountyBrooklyn             3.446e+02  9.066e+01   3.801 0.000245 ***
## CountyManhattan            2.164e+01  5.081e+01   0.426 0.671032
## CountyQueens               2.413e+02  8.815e+01   2.737 0.007307 **
## CountyStaten Island       -2.184e+02  8.107e+01  -2.693 0.008272 **
## Crime.TypeProperty Crimes -2.843e+00  3.350e+00  -0.849 0.398074
## Ratio                     -1.692e+00  8.164e+00  -0.207 0.836234
## CountyPop                 -2.804e-04  7.421e-05  -3.779 0.000265 ***
## ElectionYearTRUE           7.829e+00  4.065e+00   1.926 0.056860 .
## Perc.Dem                   6.511e+01  8.399e+01   0.775 0.440020
## PercVoted                 -1.588e+01  7.372e+01  -0.215 0.829887
## White.pop                 -4.326e-05  7.699e-05  -0.562 0.575399
## Black.pop                  6.240e-05  6.067e-05   1.028 0.306185
## Native.pop                -5.392e-04  7.626e-04  -0.707 0.481115
## Asian.pop                  8.812e-05  1.474e-04   0.598 0.551356
## Islander.pop              -2.623e-03  3.328e-03  -0.788 0.432435
## Other.pop                 -4.534e-05  8.709e-05  -0.521 0.603723
## Multi.pop                  1.387e-04  9.304e-05   1.490 0.139208
## Income.med                -2.756e-03  1.478e-03  -1.864 0.065191 .
## Income.mean                2.279e-03  9.321e-04   2.445 0.016192 *
## Year                      -5.299e-01  1.480e+00  -0.358 0.720996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.44 on 102 degrees of freedom
## Multiple R-squared:  0.7872, Adjusted R-squared:  0.7455
## F-statistic: 18.86 on 20 and 102 DF,  p-value: < 2.2e-16
```

The order in which the predictor variables were removed with backwards elimination are: Ratio, Perc.Voted, Year, Other.pop, White.pop, Income.med, Asian.pop, Native.pop, Islander.pop, Black.pop, Crime.Type, and ElectionYear. This leaves us with a model using predictor variables: County, CountyPop, Perc.Dem, Multi.pop, and Income.mean. Like the full model, it also has a significant p-value and a little larger Adjusted R-squared value.

```
## [1] "Summary of the model obtained by backwards elimination"
```

```
##
## Call:
## lm(formula = Total.Incidents ~ County + CountyPop + Perc.Dem +
##     Multi.pop + Income.mean, data = train.nycCrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.667  -8.185  -0.271   7.158  56.497
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.956e+02  8.294e+01   2.358 0.020072 *
## CountyBrooklyn       3.424e+02  5.817e+01   5.887 4.04e-08 ***
## CountyManhattan      5.391e+01  1.401e+01   3.847 0.000197 ***
## CountyQueens         2.327e+02  4.372e+01   5.323 5.18e-07 ***
## CountyStaten Island -1.868e+02  5.379e+01  -3.472 0.000731 ***
## CountyPop           -2.503e-04  4.996e-05  -5.009 2.01e-06 ***
## Perc.Dem             1.588e+02  6.159e+01   2.578 0.011198 *
## Multi.pop            1.176e-04  3.098e-05   3.796 0.000237 ***
## Income.mean          4.772e-04  1.410e-04   3.386 0.000975 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.32 on 114 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.7493
## F-statistic: 46.59 on 8 and 114 DF,  p-value: < 2.2e-16
```

Diagnostics

We can use box-cox to determine if a transformation is appropriate for the model. Figure 5 shows the λ interval. It seems like from the plots, we can use $\lambda = .3$ for our transformation.

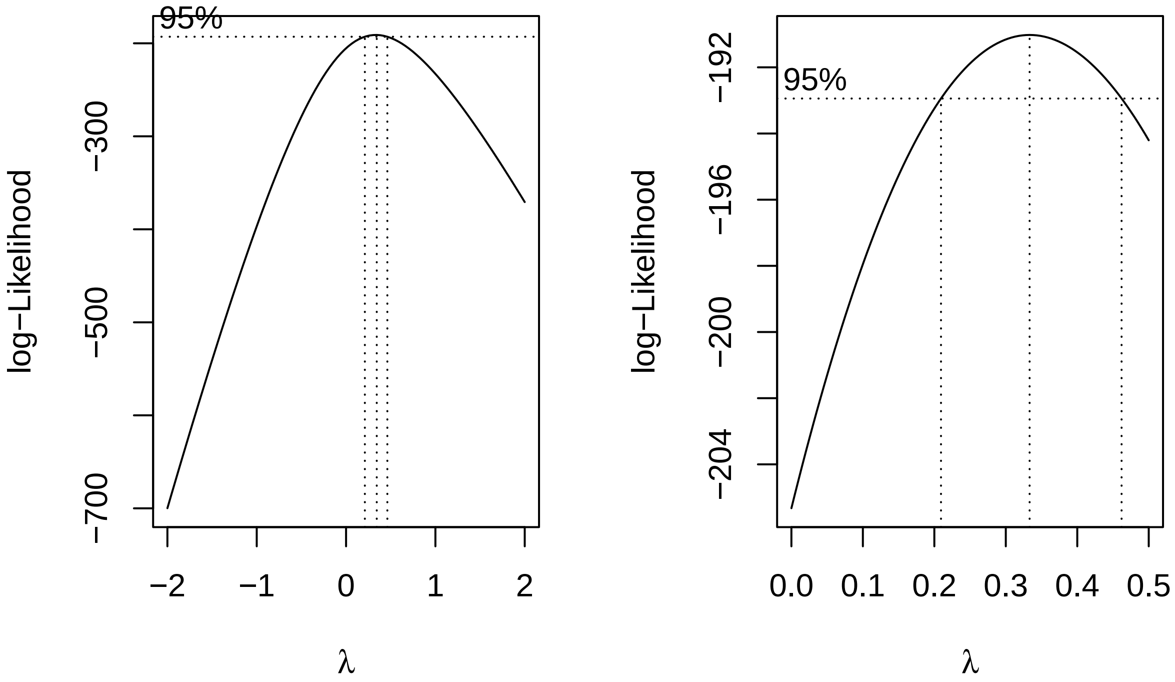

Figure 5: Boxcox Total.Incidents with predictor variables picked from backwards elimination. Left plot shows original boxcox and right plot shows a narrowed range.

Below is the summary using that transformation. We see that it has a better Adjusted R-squared value and Perc.Dem and Income.mean is no longer considered significant at the 0.05 level.

```
## [1] "Summary of the model obtained by backwards elimination with the transformation"

##
## Call:
## lm(formula = (Total.Incidents)^(0.3) ~ County + CountyPop + Perc.Dem +
##     Multi.pop + Income.mean, data = train.nycCrime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87098 -0.18049 -0.01943  0.22007  0.97672
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.694e+00  1.882e+00   2.495 0.014037 *
## CountyBrooklyn       5.849e+00  1.320e+00   4.432 2.16e-05 ***
## CountyManhattan      1.401e+00  3.179e-01   4.409 2.36e-05 ***
## CountyQueens         3.956e+00  9.917e-01   3.989 0.000117 ***
## CountyStaten Island -2.836e+00  1.220e+00  -2.324 0.021889 *
## CountyPop           -3.848e-06  1.133e-06  -3.395 0.000946 ***
## Perc.Dem             2.732e+00  1.397e+00   1.955 0.052980 .
## Multi.pop            2.546e-06  7.029e-07   3.622 0.000438 ***
## Income.mean          6.056e-06  3.197e-06   1.894 0.060768 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3476 on 114 degrees of freedom
## Multiple R-squared:  0.803,  Adjusted R-squared:  0.7892
## F-statistic:  58.1 on 8 and 114 DF,  p-value: < 2.2e-16
```

However, looking at just the model's p-value and R-squared value isn't enough to determine if this model is a good fit for estimating Total.Incidents. We can also check our our residuals, which we assumed to have constant variance. The plots to test these are shown in Figure 6.

We can also use the ncvTest; it has a p-value of 0.65508, and since this is larger than 0.05, it suggests that this model has constant variance. The Shapiro-Wilks test, used to test normality, has a p-value (0.5437) greater than 0.05. This means we fail to reject the null hypothesis and the residuals are not normally distributed.

```
## [1] "ncvTest: "
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.1995615, Df = 1, p = 0.65508
```

```
## [1] "Shapiro-Wilks test:"
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lmod$residuals
## W = 0.99031, p-value = 0.5437
```

AIC

We can compare these results to a model obtained by using AIC, stepwise regression.

It suggests that we use the predictor variables County, CountyPop, ElectionYear, Perc.Dem, Multi.pop, Income.med, and Income.mean. It has the same predictor variables as backwards elimination but with ElectionYear and Income.med included. The summary for this model is included below. The predictor variables hat are not considered significant (since they have a p-value less than 0.05) is CountyManhattan (which we cannot take out since it's a category of County), Perc.Dem, and Income.med. Despite that, it has

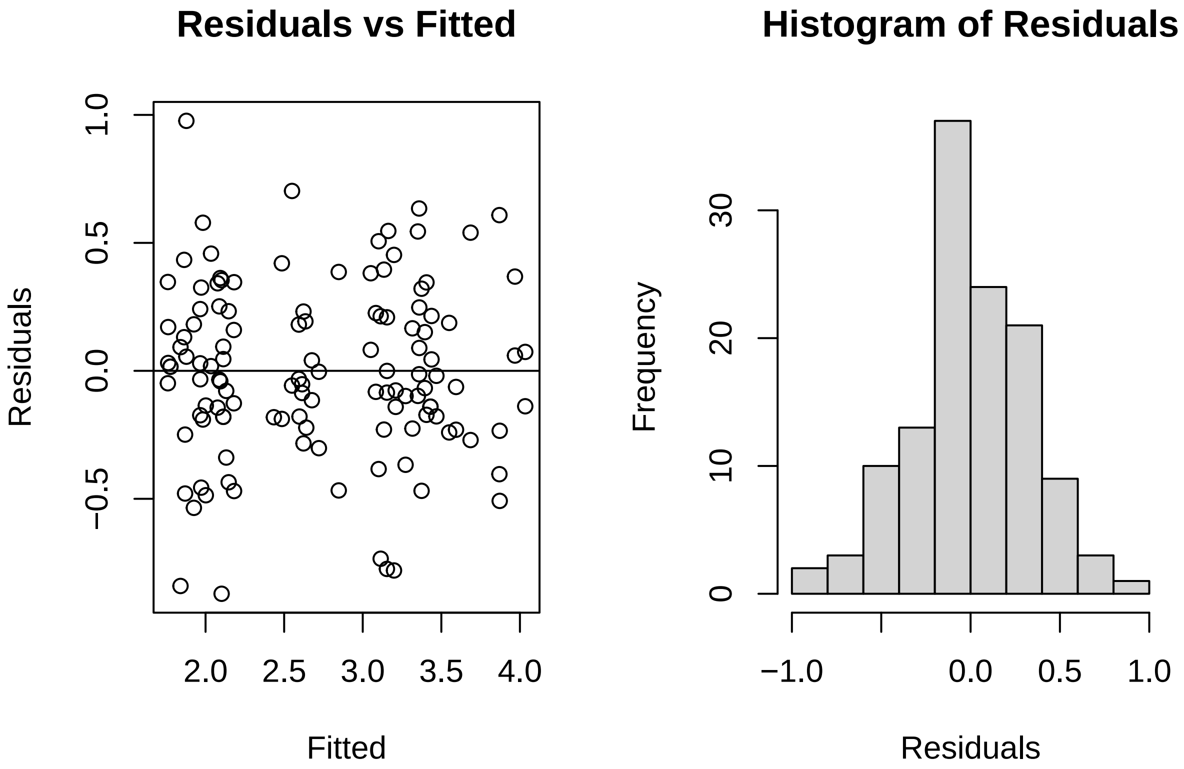

Figure 6: Plots show the Residual vs Fitted values (left) and Histogram of the residual values (right) for the Total.Incidents model obtained from backwards elimination and transformation suggested by boxcox

a better R-squared value than the previous model's. Lastly, the boxcox suggested the same transformation, with $\lambda = 0.3$ as shown in Figure 7.

For this reason, I'll be using the AIC model for predicting the Total.Incidents of the test cases.

```
## [1] "Summary of the model obtained by stepwise regression (AIC)"

##
## Call:
## lm(formula = Total.Incidents ~ County + CountyPop + ElectionYear +
##     Perc.Dem + Multi.pop + Income.med + Income.mean, data = train.nycCrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.004  -8.687  -0.579   6.075  52.996
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.864e+02  8.942e+01   3.203 0.001772 **
## CountyBrooklyn       3.761e+02  6.328e+01   5.943 3.22e-08 ***
## CountyManhattan      2.564e+01  2.487e+01   1.031 0.304839
## CountyQueens         2.718e+02  4.701e+01   5.782 6.77e-08 ***
## CountyStaten Island -2.123e+02  5.773e+01  -3.677 0.000364 ***
## CountyPop           -2.838e-04  5.423e-05  -5.234 7.84e-07 ***
## ElectionYearTRUE     7.573e+00  3.673e+00   2.062 0.041535 *
## Perc.Dem             1.136e+02  6.724e+01   1.690 0.093886 .
## Multi.pop            1.150e-04  3.091e-05   3.719 0.000314 ***
## Income.med          -2.007e-03  1.238e-03  -1.621 0.107912
## Income.mean          1.770e-03  7.870e-04   2.249 0.026440 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.05 on 112 degrees of freedom
## Multiple R-squared:  0.7779, Adjusted R-squared:  0.758
## F-statistic: 39.22 on 10 and 112 DF,  p-value: < 2.2e-16
```

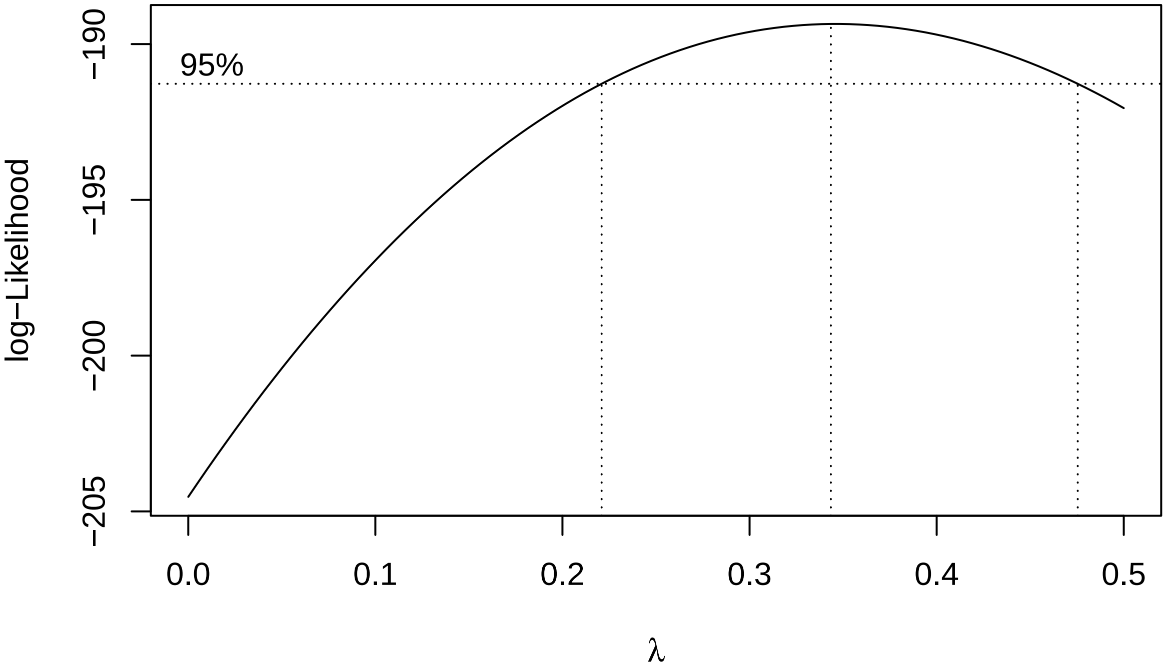

Figure 7: Boxcox Total.Incidents with predictor variables picked from AIC. The plot shows the narrowed range.

However, similarly to the model obtained by backwards elimination, the p-value (0.50755) is larger than 0.05, so it suggests constant variance; but the Shapiro-Wilks test has a p-value (0.6439) that suggests that the residuals are not normally distributed.

```
## [1] "ncvTest for model obtained by AIC: "

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.4391051, Df = 1, p = 0.50755

## [1] "Shapiro-Wilks test for model obtained by AIC: "

##
##  Shapiro-Wilk normality test
##
## data:  lmod$residuals
## W = 0.99137, p-value = 0.6439
```

Predictions and Leverage, Outlier, and Influential points

Figure 8 shows what the predicted Total.Incidents are for the test observations compared to the actual Total.Incidents. The model seems to have smaller differences between the actual and predicted values for smaller boroughs (like the Bronx and Staten Island) then the larger boroughs.

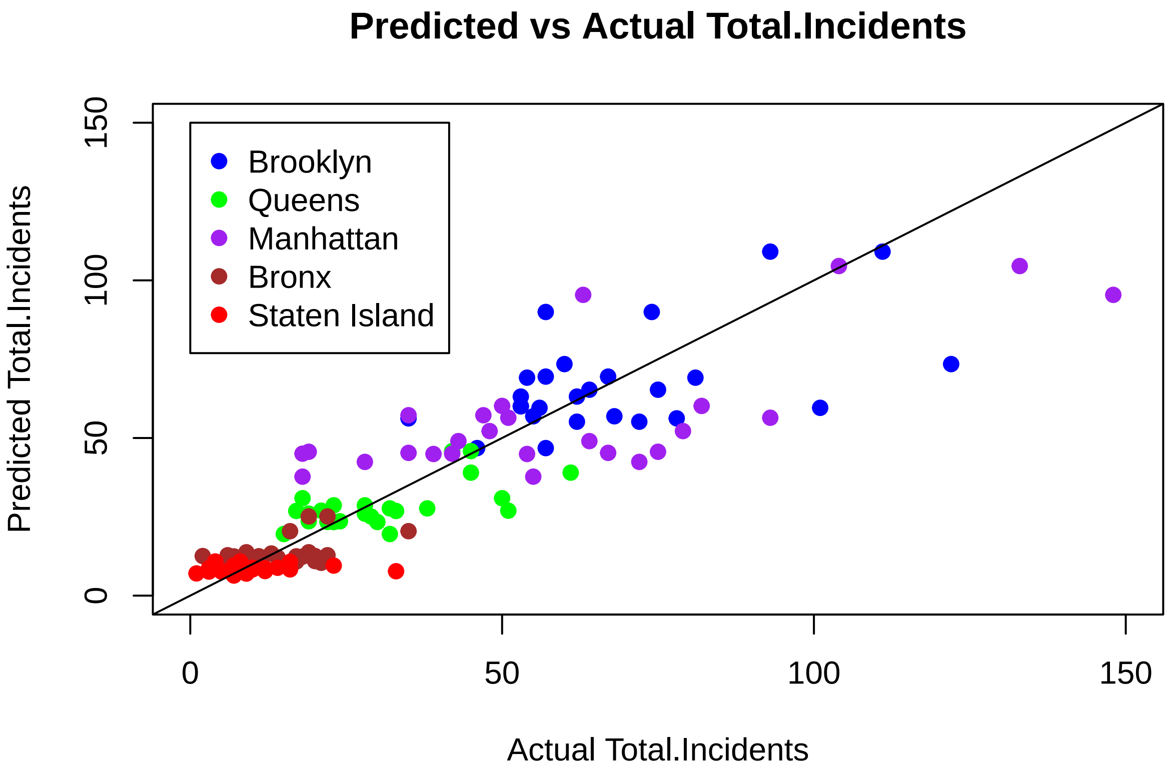

Figure 8: Predicted vs Actual Total.Incidents by county for the model obtained from backwards elimination and transformation suggested by boxcox

This is also shown in the MSE (Mean Squared Error). The MSE measures the average distance between the actual and predicted value. As stated before, the Bronx and Staten Island have the smallest difference between the actual and predicted values and they also have the smallest MSE. Brooklyn and Manhattan's predictions seem to be the farthest from the predicted values in the plot, and they also have the largest MSE's. This shows that the model might not be sufficient for predicted the Total.Incidents for all the boroughs, but it's best for the Bronx and Staten Island.

```
## [1] "Brooklyn MSE:"

## [1] 315.0249

## [1] "Queens MSE:"

## [1] 89.28301

## [1] "Manhattan MSE:"

## [1] 497.8696

## [1] "Bronx MSE:"

## [1] 39.14954

## [1] "Staten Island MSE:"
## [1] 45.59739
```

We can check for which observations might be influencing the fit of the model by finding leverage points, outliers, and influential points (using hat values, rstandard, and cooks distance).

Leverage Points: 37, 115
Outliers: 6, 48, 60, 62, 72
Influential points: 1, 6, 12, 19, 31, 45, 48, 60, 62, 63, 72, 100, 108

Below is the summary for when the outliers are taken out of the train data. The Adjusted R-squared is actually closer to 1 than the model we've been using.

```
## [1] "Summary of model without outliers: "

## 
## Call:
## lm(formula = (Total.Incidents)^(0.3) ~ County + CountyPop + ElectionYear + 
##     Perc.Dem + Multi.pop + Income.med + Income.mean, data = train.nycCrime_noOutliers)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.53666 -0.18486  0.00025  0.19069  0.62051 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          5.439e+00  1.894e+00   2.872 0.005012 ** 
## CountyBrooklyn       5.714e+00  1.438e+00   3.975 0.000136 ***
## CountyManhattan      1.634e+00  5.060e-01   3.228 0.001701 ** 
## CountyQueens         3.716e+00  1.079e+00   3.446 0.000843 ***
## CountyStaten Island -3.215e+00  1.258e+00  -2.556 0.012136 *  
## CountyPop           -3.807e-06  1.241e-06  -3.066 0.002807 ** 
## ElectionYearTRUE     9.261e-02  7.229e-02   1.281 0.203206    
## Perc.Dem             1.912e+00  1.305e+00   1.465 0.146034    
## Multi.pop            2.429e-06  6.386e-07   3.803 0.000250 ***
## Income.med           6.372e-06  2.453e-05   0.260 0.795623    
## Income.mean          7.576e-07  1.563e-05   0.048 0.961432    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2733 on 97 degrees of freedom
## Multiple R-squared:  0.8685, Adjusted R-squared:  0.855 
## F-statistic: 64.07 on 10 and 97 DF,  p-value: < 2.2e-16
```

However, when comparing the MSE with and without the outliers, the MSE decreases for Brooklyn (by 26.492), the Bronx (by 1.09882), and Staten Island (by 0.48946). However, the MSE increases for Queens (by 9.3847) and Manhattan (by 49.1686). This suggests that removing the outliers improves the fit of the model for the boroughs where MSE was decreased.

```
## [1] "Brooklyn MSE:"

## [1] 315.0249

## [1] "Brooklyn MSE (no outliers):"

## [1] 288.5329
## [1] "Queens MSE:"

## [1] 89.28301

## [1] "Queens MSE (no outliers):"

## [1] 98.66771

## [1] "Manhattan MSE:"

## [1] 497.8696

## [1] "Manhattan MSE (no outliers):"

## [1] 547.0382

## [1] "Bronx MSE:"

## [1] 39.14954

## [1] "Bronx MSE (no outliers):"

## [1] 38.05072

## [1] "Staten Island MSE:"

## [1] 45.59739

## [1] "Staten Island MSE (no outliers):"

## [1] 45.10793
```

Correlations

It is important to check correlations between the predictor variables because it's more difficult to determine a variable's individual contribution to the prediction when it has a high correlation with another variable. Below shows the pairs of numerical predictor variables that have an absolute correlation of 0.5 or greater.

The numerical predictor variables we used in our model is CountyPop, Perc.Dem, Multi.pop, Income.med, and Income.mean, The pairs of predictors that have a correlation coefficient greater than 0.5 is CountyPop and Perc.Dem (0.6270585) and Income.med and Income.mean (0.8762079).

```
## [1] "CountyPop and Perc.Dem"
## [1] 0.6270585
## [1] "CountyPop and White.pop"
## [1] 0.8107687
## [1] "CountyPop and Black.pop"
## [1] 0.8388867
## [1] "CountyPop and Native.pop"
## [1] 0.5801376
## [1] "CountyPop and Asian.pop"
## [1] 0.7567654
## [1] "CountyPop and Islander.pop"
## [1] 0.5236037
## [1] "Perc.Dem and Black.pop"
## [1] 0.5693543
## [1] "Perc.Dem and Native.pop"
## [1] 0.5291612
## [1] "Perc.Dem and Other.pop"
## [1] 0.7666706
## [1] "PercVoted and Black.pop"
## [1] -0.5126586
## [1] "PercVoted and Other.pop"
## [1] -0.5464466
## [1] "PercVoted and Income.med"
## [1] 0.7867493
## [1] "PercVoted and Income.mean"
## [1] 0.7712091
## [1] "White.pop and Black.pop"
## [1] 0.5667383
## [1] "White.pop and Asian.pop"
## [1] 0.6576144
## [1] "Black.pop and Native.pop"
## [1] 0.5189201
## [1] "Black.pop and Income.med"
## [1] -0.5793646
## [1] "Native.pop and Islander.pop"
## [1] 0.5037438
## [1] "Native.pop and Other.pop"
## [1] 0.6301609
## [1] "Native.pop and Multi.pop"
## [1] 0.8192702
## [1] "Other.pop and Income.med"
## [1] -0.722687
## [1] "Other.pop and Income.mean"
## [1] -0.5156514
## [1] "Income.med and Income.mean"
## [1] 0.8762079
```

Results

```
## [1] "Final model (model obtained by stepwise regression with transformation and no outliers when fit
```

```
##
## Call:
## lm(formula = (Total.Incidents)^(0.3) ~ County + CountyPop + ElectionYear +
##     Perc.Dem + Multi.pop + Income.med + Income.mean, data = train.nycCrime_noOutliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53666 -0.18486  0.00025  0.19069  0.62051
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.439e+00  1.894e+00   2.872 0.005012 **
## CountyBrooklyn       5.714e+00  1.438e+00   3.975 0.000136 ***
## CountyManhattan      1.634e+00  5.060e-01   3.228 0.001701 **
## CountyQueens         3.716e+00  1.079e+00   3.446 0.000843 ***
## CountyStaten Island -3.215e+00  1.258e+00  -2.556 0.012136 *
## CountyPop           -3.807e-06  1.241e-06  -3.066 0.002807 **
## ElectionYearTRUE     9.261e-02  7.229e-02   1.281 0.203206
## Perc.Dem             1.912e+00  1.305e+00   1.465 0.146034
## Multi.pop            2.429e-06  6.386e-07   3.803 0.000250 ***
## Income.med           6.372e-06  2.453e-05   0.260 0.795623
## Income.mean          7.576e-07  1.563e-05   0.048 0.961432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2733 on 97 degrees of freedom
## Multiple R-squared:  0.8685, Adjusted R-squared:  0.855
## F-statistic: 64.07 on 10 and 97 DF,  p-value: < 2.2e-16
```

In the end, we used County, county population, percentage who voted Democrat, population of people who are two races or more, and mean income as the predictor variables.

The intercept represents the transformed predicted number of Total.Incidents if it happened in the Bronx and all the other (numerical) predictors had a value of 0; so it predicts about 282.9634 hate crimes. The model predicts that incidents in Brooklyn, Queens, and Manhattan increases the number of incidents by 333.5294, 79.47899, and 5.138545 respectively. When the hate crime is in Staten Island, it is predicted that the number of hate crimes decreases by 49.04596. However, it's important to note that, using pairwise comparison, the pair Staten Island and the Bronx and the pair Manhattan and Brooklyn are not significantly different when only County is used to predict Total.Incidents (this is shown in Figure 9).

The model predicts that each unit increase in CountyPop (the county's population increases by 1) will decrease the Total.Incidents by $(-3.807e-06)^{1/0.3}$ (which is a very small amount). But just looking at the sign of the coefficient, it suggests that as the county's population increases, the predicted amount of total incidents decreases.

ElectionYear and Perc.Dem are not considered significant. But the positive estimated coefficients says that when it is an election year or the percent of people who vote democrat increases, there is an increase in the predicted number hate crimes (by 0.0003593571 and 8.675469 respectively). It is also difficult to analyze CountyPop and Perc.Dem inidivudally since they were shown to have a high correlation with each other.

The model predicts that each unit increase in Multi.pop (the population of people who are two or more races) will increase the Total.Incidents by $(2.429e-06)^{1/0.3}$ (which is a very small amount).

Income.med and Income.mean both aren't considered significant and they were considered highly positively correlated with each other, which make them difficult to analyze the estimated coefficients individually. They both have a positive estimated coefficient, which suggest that as the median or mean income increases, the predicted number of total incidents increase as well.

Overall, compared to the initial case of a hate crime occuring in the Bronx in a non-election year with the numerical variables equaling 0, the predicted number of hate crimes increases if in it occurs in Brooklyn, Manhattan, Queens, it's an election year, there's an increase in the percentage of people who voted Democrat in the most recent election, the percentage of people who are multi-racial increases, or if the mean or median income increases. And there is decreased number of predicted hate crimes if it occurred in Staten Island or the county's population increases.

Limitations and Conclusion

By using county, population, income, racial demographics, and election results for each borough, we've created a model to predict the total number of hate crimes in New York City. We found that the borough that the hate crime is committed in has the most influence on the predicted number of hate crimes.

I also tried to fit 5 models, one for each borough (it used the same predictor variables as the full model

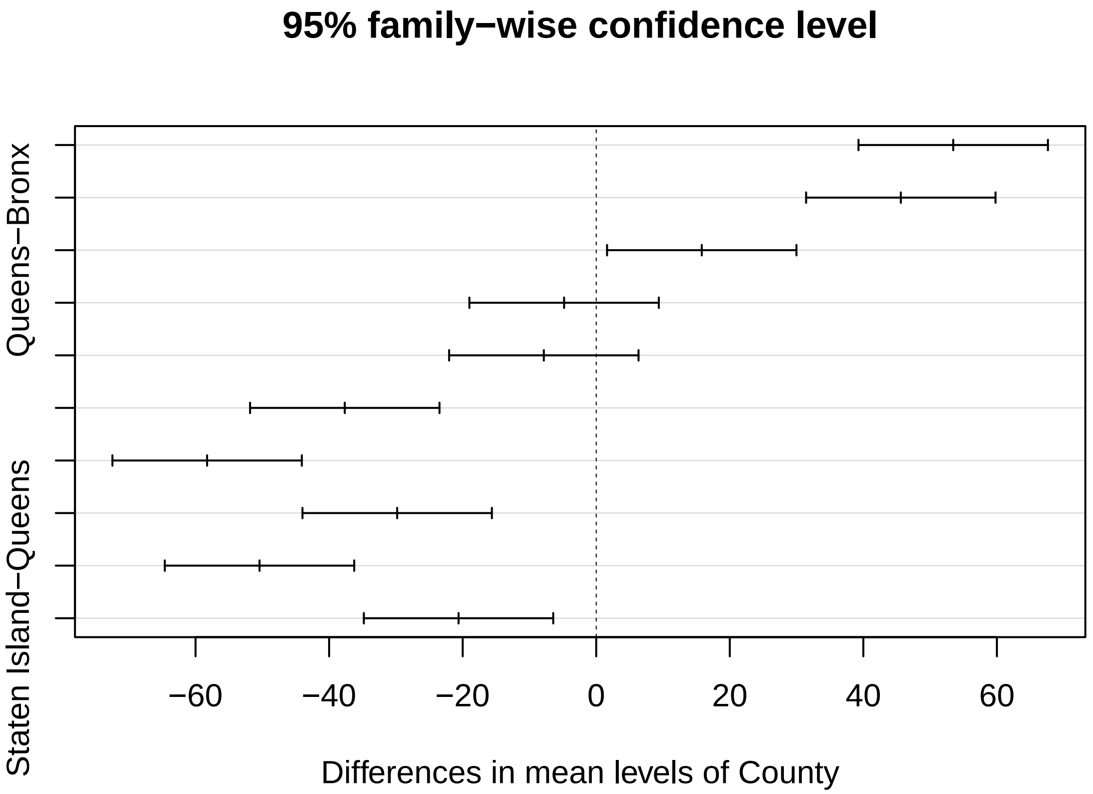

Figure 9: Pairwise comparison of County for predicting Total.Incidents. It shows that the pair Staten Island and the Bronx and the pair Manhattan and Brooklyn are not significantly different in terms of Total.Incidents.

except for County). Then I used AIC to choose which variables to use; the predictor variables for each borough's model is shown below:

- Bronx: Crime.Type, Ratio
- Brooklyn: CountyPop, Perc.Dem, PercVoted, White.pop, Native.pop, Asian.pop, Other.pop
- Manhattan: Ratio, CountyPop, ElectionYear, Black.pop, Asian.pop
- Staten Island: Crime.Type, CountyPop
- Queens: Crime.Type, Ratio

Crime.Type, Ratio, CountyPop were each included in 3/5 of the borough's models. This differs from the all-boroughs model (which I'm going to call the final model) which did not include Crime.Type or Ratio. The model for the Bronx, Staten Island, and Queens were non-significant. This was interesting because those 3 boroughs had the best (smallest) MSE in the final model. Brooklyn and Manhattan (which were shown to have no significant difference in Total.Incidents) had a significant p-value and an R-squared value of 0.3257 and 0.627 respectively. Both of these R-squared values are smaller than the final model's R-squared value, which suggests that they have a worse linear fit than the final model. Therefore, the final model is best when trying to predict Total.Incidents for all the boroughs.

Some limitations appear when trying to fit a model for hate crime a of specific bias. For example, Anti.Jewish, Anti.Gay.Male, and Anti.Black were the most prevalent types of hate crimes in this data set. While a model can still be constructed with the predictor variables I used in this project, it might also be useful (and make the model more accurate) to have sexual orientation and gender demographics for Anti.Gay.Male or religion and ethnicity demographics for Anti.Jewish. However, these limitations could be overcome with future research on these demographics.

Bibliography

City population data:
https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-cities-and-towns.html

Hate Crime definition:
https://www.justice.gov/hatecrimes/learn-about-hate-crimes

Main data set:
https://data.ny.gov/Public-Safety/Hate-Crimes-by-County-and-Bias-Type-Beginning-2010/6xda-q7ev/about_data

County to Borough names:
https://portal.311.nyc.gov/article/?kanumber=KA-02877

NY population data set: https://data.ny.gov/Government-Finance/Annual-Population-Estimates-for-New-York-State-and/krt9-ym2k/about_data

VIF additional information:
chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://tspace.library.utoronto.ca/bitstream/1807/106198/1/cjfr-2020-0257suppla.pdf

2008 election by NYC County:
https://archive.nytimes.com/www.nytimes.com/elections/2008/results/states/president/new-york.html?scp=62&sq=100&st=Search

2012 election by NYC County:
https://www.politico.com/2012-election/results/president/new-york/

2016 election by NYC County:
https://www.politico.com/2016-election/results/map/president/new-york/

2020 election by NYC County:
https://www.politico.com/2020-election/results/new-york/

ACS Demographic and Housing estimates (DP05):
https://data.census.gov/table?q=dp05&t=Populations%20and%20People&g=060XX00US3608160323

Income in the past 12 months (S1901): https://data.census.gov/table?q=s1901&g=060XX00US3608160323

Cities by Population: https://www.britannica.com/topic/Whats-the-largest-US-city-by-population