New York City Housing Index

K. Bhargava, T. Rebhan, L. Barto, R. Rosebrook, & G. Leibovich

Part I: Quality Index

Research Question:

1. Since the 1970s, housing quality has improved dramatically; however, some sectors of the housing stock continue to face poor conditions and some specific maintenance deficiencies continue to show higher prevalence. Create a housing quality index for the NYCHVS that enables a view of the housing conditions faced by residents. Contestants may consider the relative importance of different conditions now and/or how the prevalence of these issues has shifted over time.

Given the vast amount of data available in an increasingly globalized and developing world, the concern on the quality of housing for various groups of individuals in the most populous city in the most developed country in the world is a pressing and pertinent issue. Our group looked into the research topic focused on the examination of disproportionate quality-of-housing improvements over the years from 1991 to 2017 for various sectors of New York City.

In order to address this research problem, we took the ten datasets and concatenated them into a single data frame. We then combined similar variables to reduce redundancy and create a more manageable set of variables. Additionally we refactored several of the variables to make them consistent across all of the survey years. For example, we combined X_d1 and X_d2 since both referenced different issues with the walls that could be appropriately described with a single column. In fact, in the 2017 survey these questions were consolidated into a new variable X_d12 which combines the issues listed in X_d1 and X_d2. Another example of an improvement on the "quality of life" transformation of the data to streamline and make the data consistent would be refactoring variables like X_25c to have the same types of answers displayed in the 2017 data. We also renamed variables and changed the values from integers to strings so that it was easier to understand without consulting the codebook. Once the data frames were refactored and cleaned up we joined them together, removed the unused columns and exported the new data frame into a CSV file, "*housingAggregation.csv*". Having all the data in one file made it much easier to create summary statistics and graphs. Currently several of the columns, mainly those with continuous values such as the price of t he home and rent paid have not been refactored.

The next step was to create a statistical index which reduces these variables into a single measure of housing quality. As a launching point, we used the Poor Quality Index (PQI), which was created by HUD for analysing the American Housing Survey. This survey has many of the

same questions and observations as the NYC data we are analysing, so it could be adapted fairly easily. The index weights external structural issues higher than other issues such as presence of pests, because these likely have a more immediate effect on quality of life.

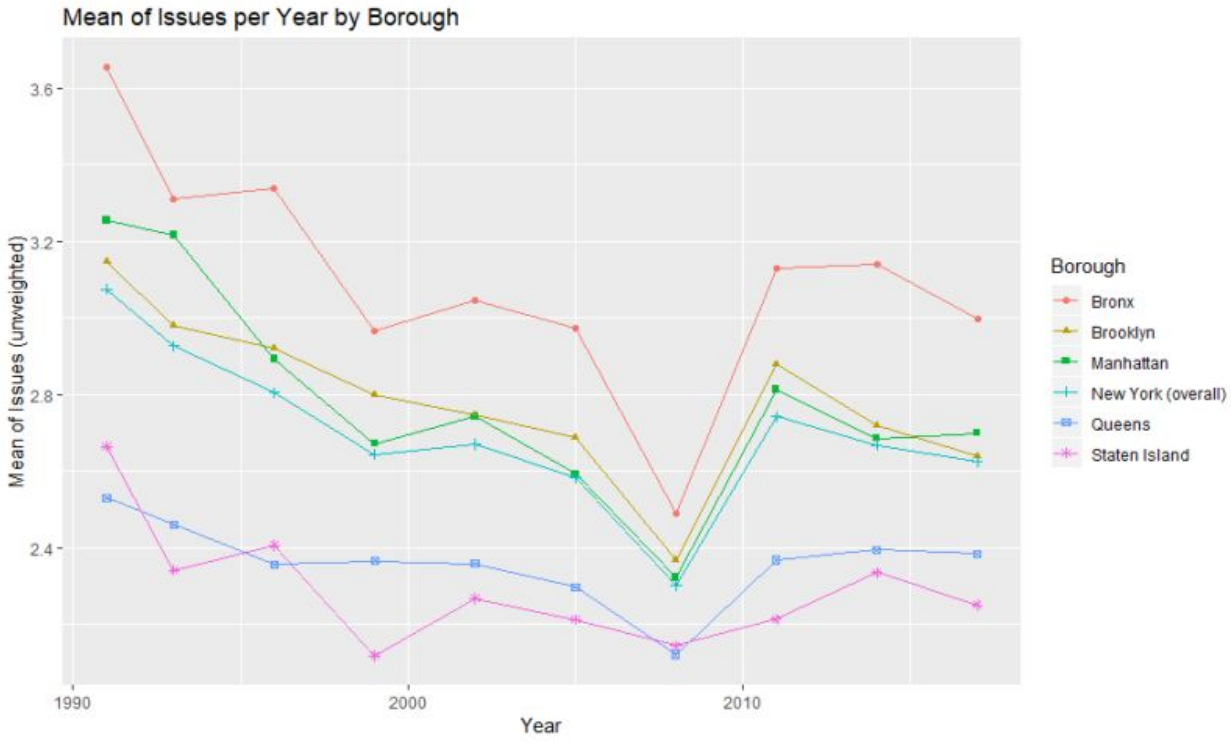
We created two different graphs from the data after wrangling it. The first graph plots the average reported problems about the building conditions over time (weighted). The graph displays the average number of issues reported broken down by borough and shows an overall descending number of problems. This seems to be consistent with the premise of the research question. Since the problems reported are decreasing, it is reasonable to believe that the cause of fewer problems being reported is due to building conditions improving. However, the average problems reported per building in the Bronx are still substantially higher than any of the other boroughs in NYC, and is even significantly higher than the average number of problems reported per building in the entire city of New York.

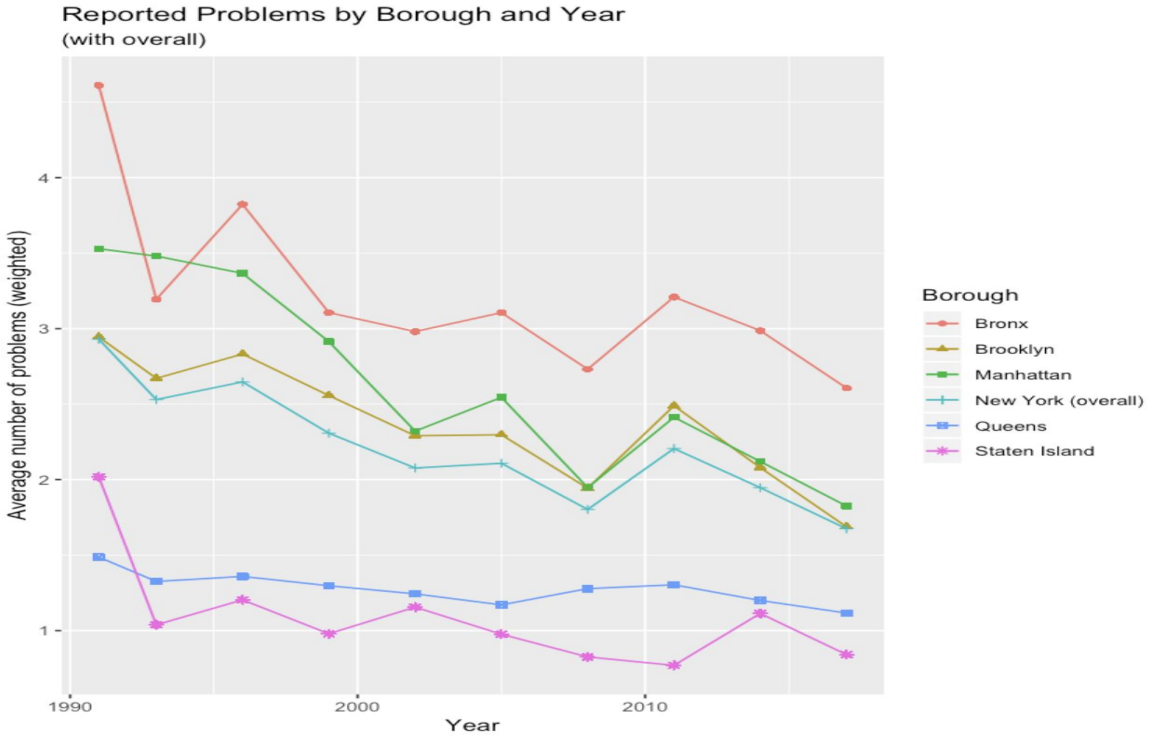
To compare the results of our weighted data, we contrasted this graph with reports of problems in buildings in each borough without weights to see how much of a significance the weights caused. The weighted graph shows a smoothing of the reported problems when compared with the unweighted graph. There is less volatility in the average problems reported on the weighted data. This implies that while there were fewer problems reported from 1995-2000 and 2008-2010, there were still more structural issues reported during those time periods.

Buildings in NYC based on the index of the avg # of problems reported (weighted)

Year	Mean	Median	Standard Deviation
1991	2.929875	2	4.733919
1993	2.530253	0	4.171418
1996	2.647202	0	4.500677
1999	2.307498	0	4.140150
2002	2.077146	0	3.460923
2005	2.108148	0	3.525571
2008	1.802488	0	3.154266
2011	2.205662	0	3.458278
2014	1.946688	0	3.241130
2017	1.674710	0	2.988205

The table above shows the mean, median, and standard deviation of the index we used to determine the average number of problems reported. Most surprising is that the median is 0 for most of the years. This implies that at least half the buildings reported no problems for most of the years the survey was conducted. This would then mean that the buildings that did report problems reported increasingly more severe problems and would also imply that the differences between “sound” and “deteriorating” buildings is significant and would lend to the argument of an increasingly sharp contrast in the quality of buildings.


Mean of Issues per Year by Borough
Mean of Issues (unweighted)
3.6
3.2
2.8
2.4
1990
2000
2010
Year
Borough
Bronx
Brooklyn
Manhattan
New York (overall)
Queens
Staten Island


Reported Problems by Borough and Year
(with overall)
Average number of problems (weighted)
4
3
2
1
1990
2000
2010
Year
Borough
Bronx
Brooklyn
Manhattan
New York (overall)
Queens
Staten Island

Part II: Predictive Modeling

In our initial approach to data wrangling we wrote a script that would read in the data dynamically, with the goal of creating our housing quality index. In preparation of building our models, however, we used a simpler process. First, we created two separate data frames for units that were rented and unites that were owned. Second, we used only the three most recent datasets: 2011, 2014, and 2017. The latter change was to mitigate the impact of inflation on the differences in home values and rent prices since time was not a considered in any of our models. In future models, we would like to consider adjusting the economic values to account for inflation so that we may use all datasets and include time as a factor. Finally, we focused primarily on columns including numeric data for all of our Gini indices, linear regression, logistic regression, and machine learning models. In addition to dropping non-numeric columns we also dropped rows consisting of incomplete cases.

The Gini Index shows the purity of a group and how separable the data is within a group. It is scored between 0 and 1. The closer the score is to 1, the more "impure" and difficult to separate the data within the group. The closer to 0 the score is, the more easily separable the data is and the more "pure" the data within the group is. When comparing owned homes to rented homes, it was easier to group the data based on number of rooms, number of tenants under 6, and number of tenants under 18.

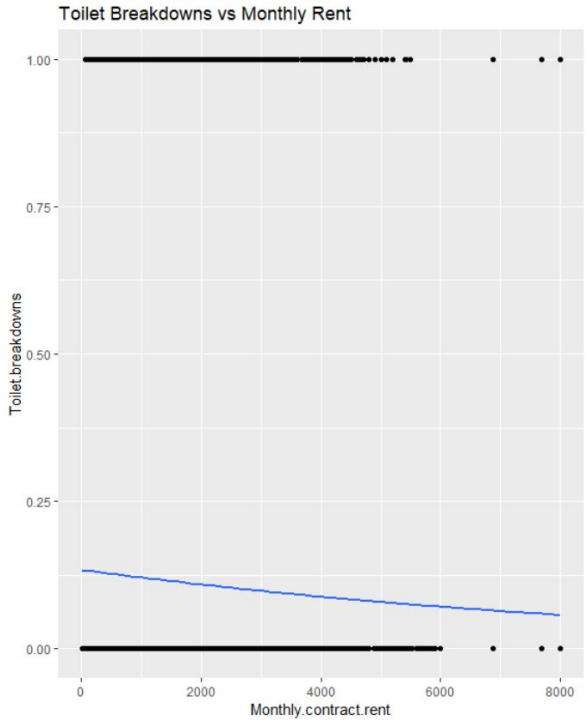
Interestingly enough, the differences between Owned Homes and Rented Homes were evident in the household income. While both were not particularly separable, it is interesting that rented homes are more difficult to distinguish groups in household income than owned homes. Perhaps it is because owning homes in NYC are much more expensive than in other places so perhaps people that can own homes in NYC would have a higher bracket of pay than people who rent (which can have a much larger spectrum of incomes). Broken Windows and monthly rent are the two most difficult to separate between Owned Homes and Rented Homes respectively. It's difficult to say what the broken windows index score means, but the monthly rent makes sense as listed in the explanation for the household income. Perhaps with more time, the score on "Value" for owned homes would be worth exploring further to see if it gives insights on the question referring to the improvement of housing conditions in NYC over the past years.

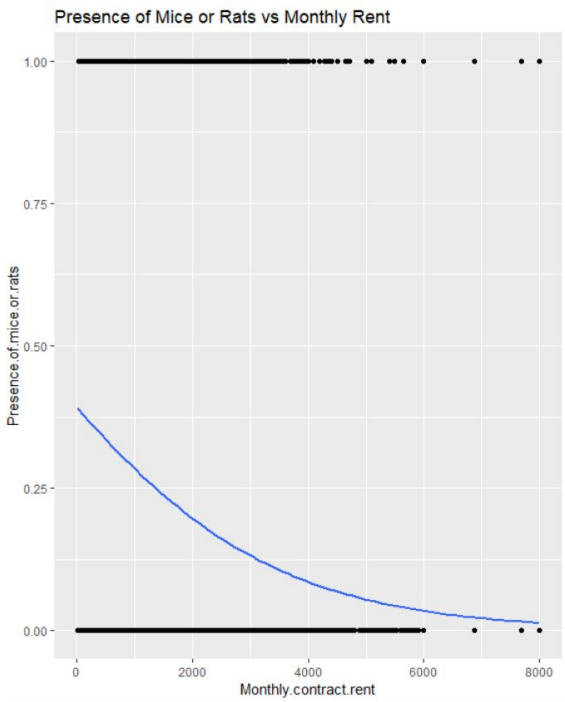
Gini Index

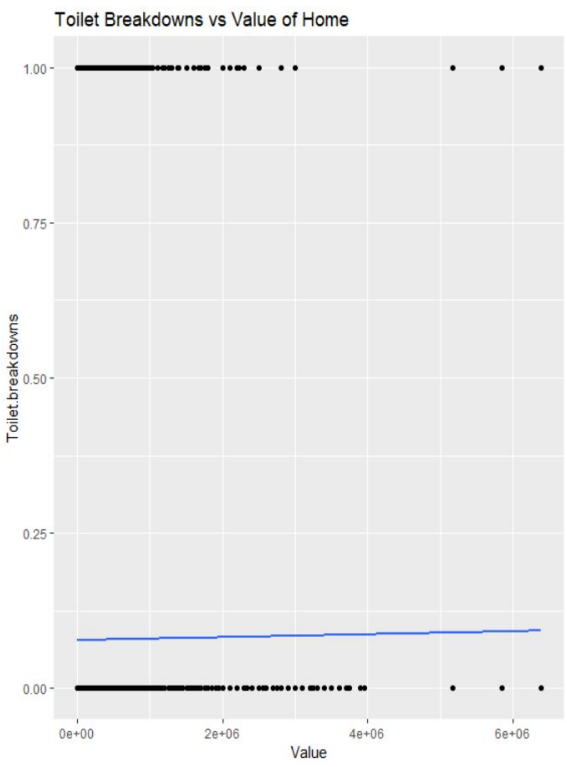
Owned Homes	Rented Homes
Broken Windows	monthly rent
0.95236	0.80992
Value	Out of Pocket Rent
0.44563	0.81604
number of units	number of units
0.45363	0.23161
number of stories	number of stories
0.40452	0.2812
number of rooms	number of rooms
0.19089	0.19499
household income	household income
0.53058	0.79677
number under 6	number under 6
0.10966	0.13158
number under 18	number under 18
0.25873	0.26486

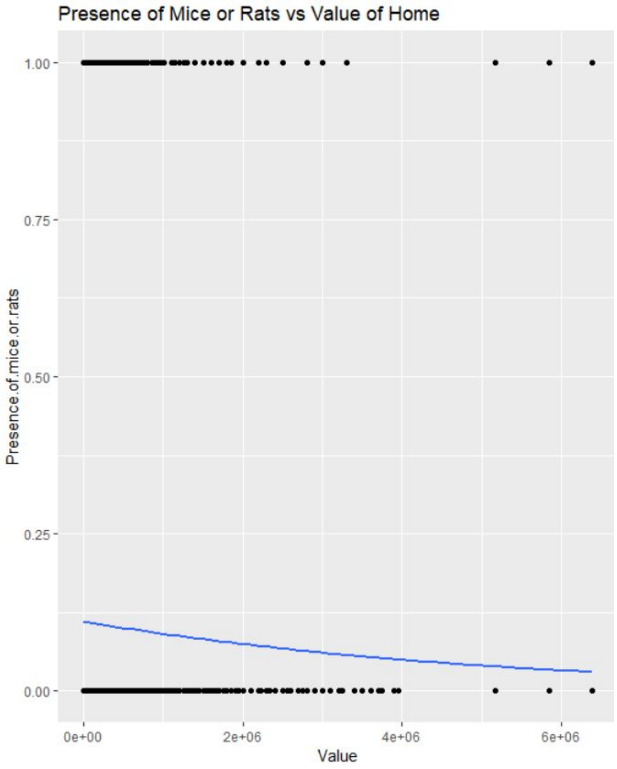
Our next analyses used linear regression. Starting with units which were own by the resident, we attempted to predict household income using home value, number of rooms, number of bedrooms, and occupants under 18. Although we surmised this would be a meaning prediction, the model performed poorly, with an adjusted R^2 value of 0.1462. Upon further exploration of the owned-unit dataset the highest performing model we were able to create predicted the number of bedrooms using number of rooms, number of stories, and occupants under 18. Although this model did have an adjusted R^2 of 0.7311, we felt the model had little use in terms of a meaningful or usable prediction. The rented dataset yielded similar findings. A meaningful model, predicting monthly rent using occupants under 18 and household income yielded an adjusted R^2 of 0.009571 while the highest performing, yet less meaningful, model predicting number of bedrooms using number of rooms, occupants under 18, monthly rent, and occupants under 6 yielded an R^2 of 0.7884. Unfortunately we were unable to find a model that was both high performing and meaningful. After analysing both datasets using a variety of linear regression models we have concluded that these models are not well suited to make predictions using our chosen data. In future models, however, we would like to retry linear regressions using smaller, and potentially more homogenous, datasets, such as by further grouping the data by suboroughs and analysing each individually.

We next used logistic regression to predict specific issues within each rented and owned units. The particular issues we attempted to predict were toilet breakdowns and presence of rats or mice. In each dataset we used home value or monthly rent for owned and rental data respectively, as our explanatory variables. One unexpected finding was that an increase in home value of owned units were actually more likely to have had a toilet breakdown during the surveyed year. Two initial hypotheses are that more expensive homes are historic buildings and have dated plumbing, or that more valuable homes have a greater number of toilets, and therefore a higher likelihood of breakdowns; though these hypotheses still need to be further explored. The second conclusion we found from our logistic regression was as expected, units with lower rent were much more likely to have had rats or mice during the surveyed year.









Our final analysis used a random forest classifier, and was built on the original dataset from the first part of this project. This data was focused on categorical features which contributed to an overall index of housing quality based on the number of deficiencies; the index we created was also included in the analysis. We attempted to predict the borough in which a unit is located based on information about its condition, the year surveyed, and other unit features. The table below is the confusion matrix of our classifier. The accuracy of this model turned out to be 41.16%. Though this is not a particularly effective model in terms of pure prediction, it is twice as effective as pure chance of being correct which would be 20% (or 1 out of five). Again, while we found a model that is somewhat effective, it is not particularly insightful. Rarely would there be a real-world solution in which one would have lots of information reg

	Bronx	Brooklyn	Manhattan	Queens	Staten Island
Bronx	4853	588	640	408	36
Brooklyn	1920	12024	1682	3589	848
Manhattan	2980	3532	14228	3344	222
Queens	1870	4674	875	11164	2229
Staten Island	101	394	8	459	1127

Unfortunately, it appears that the data we cleaned for this project yielded either poorly performing or not particularly insightful models. Moving forward we would like to add more categorical columns to assist in our model building as well as adjust for inflation so that we may include all of the datasets. Additionally we would like to consider breaking the data down further, for example by suborough, or including external data to see if there are either any other explanatory variables not already included in the datasets, or if the given data can help us draw any other conclusions. While the data we found is interesting, we believe we have significantly more information to mine that can help us come up with more meaningful and interesting findings.

Part III Grouping and Clustering

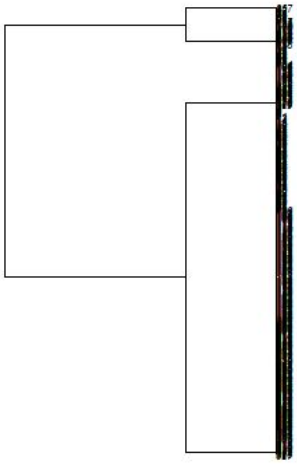
The final portion of the project focused on grouping and clustering portions of the data. Since the data we were using had over thousands upon thousands of instances of data, we had to scale down the sample to about 1% so that the data could be clustered and grouped more easily.

We implemented a hierarchical clustering model on the data. In total, there were 10 features considered in the unsupervised learning model, thus creating a 10 dimensional euclidean distance from one point to the next. Among the features considered were the following: value, number of units, number of stories, number of rooms, number of bedrooms, household income, etc.

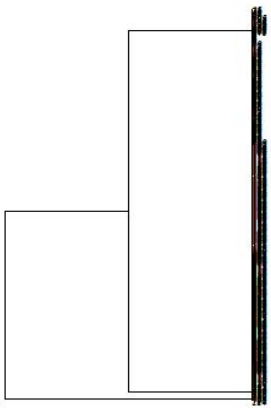
We started by running a hierarchical model that grouped the data into various distinct subsections. Because this was an exercise in grouping data together, the Gini Index we found of the data was useful in determining how we would go about grouping the data. We looked at the Gini Indices for both the rented and the owned homes and decided to see how the hierarchical model would group the data based on the highest, lowest, and middle quartile Gini Scores.

We started by grouping the lowest Gini Index data to see how the model did it. Below are the results.

Number of Individuals Under 6 (rented)

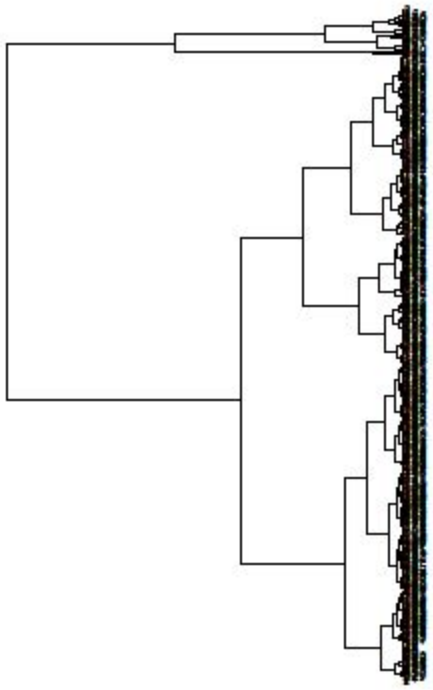


Number of Individuals Under 6 (owned)



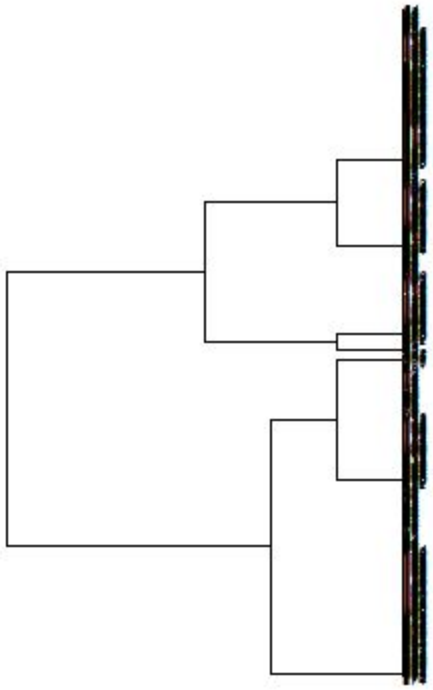
As you can see, the difficulty is in the identities of the values. We can't really gather too much info on this since we don't really know what each individual value represents. The label for each input is the index (row number) that was placed there (ex: Joe Schmo's building on 17th st that's rented and has 9 other features associated with it). Because each individual value has no real signalling distinction between it other than its individual index row, this model doesn't tell us much more than how many groups were able to be made from the data. Out of the 330 individual homes looked at in this sample of the total data, the lower Gini Index correlated to fewer groupings.

Out of Pocket Rent (rented)

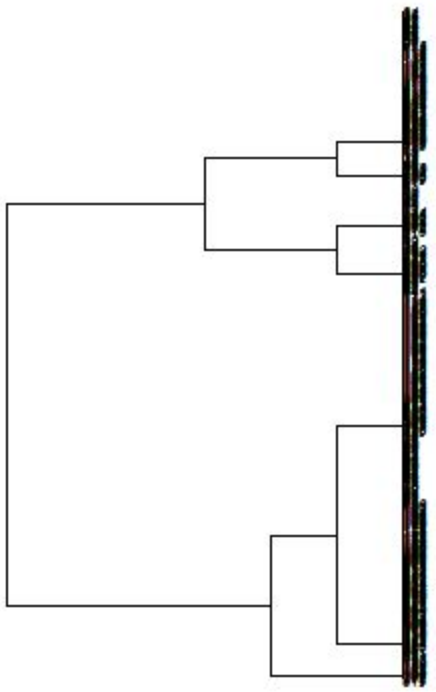


While Broken Windows was the other group that had the highest Gini Index, we didn't decide to do anything with the group data since it's categorical and doesn't even give us any information that's useful.

Num of Stories (owned)



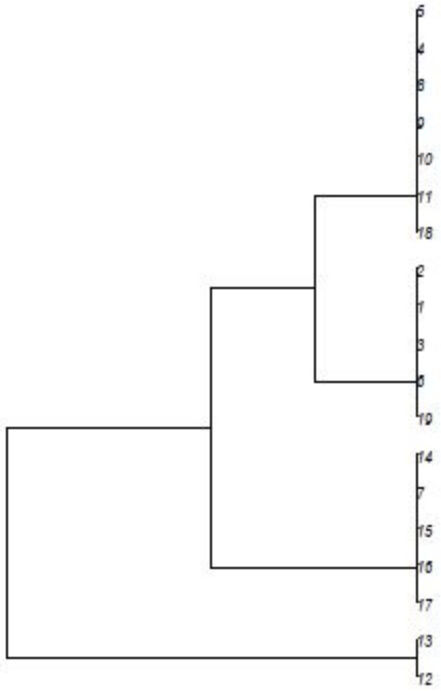
Num of Stories (rented)



Number of Stories was the final group we applied the hierarchical model on and it was able to separate the data into more groups than the number of individuals under 6 grouping. This seems to hold consistent with the idea that the more pure a data column is, the fewer distinctions in groups the model will find. The fundamental take away from these models is that it shows us how many groups we might be able to break down a subset of a column into and how populous that group would be if we did so. While we don't know the specifics of how the model split the data into groups, we can see that if the data was split into more groups then there are more factors concerning how differentiable the data is. If we could get more data on how the model split the data itself, then we might have more to go off of on seeing what significance the different groups carried.

Lastly, so that we could at least see some implementation of the hierarchical model and how it distinguished data, we took a micro sample of the data (20 values) and applied the same grouping model to it to see how the model split the groups. Below is the grouping we found.

Number of Stories (rented)

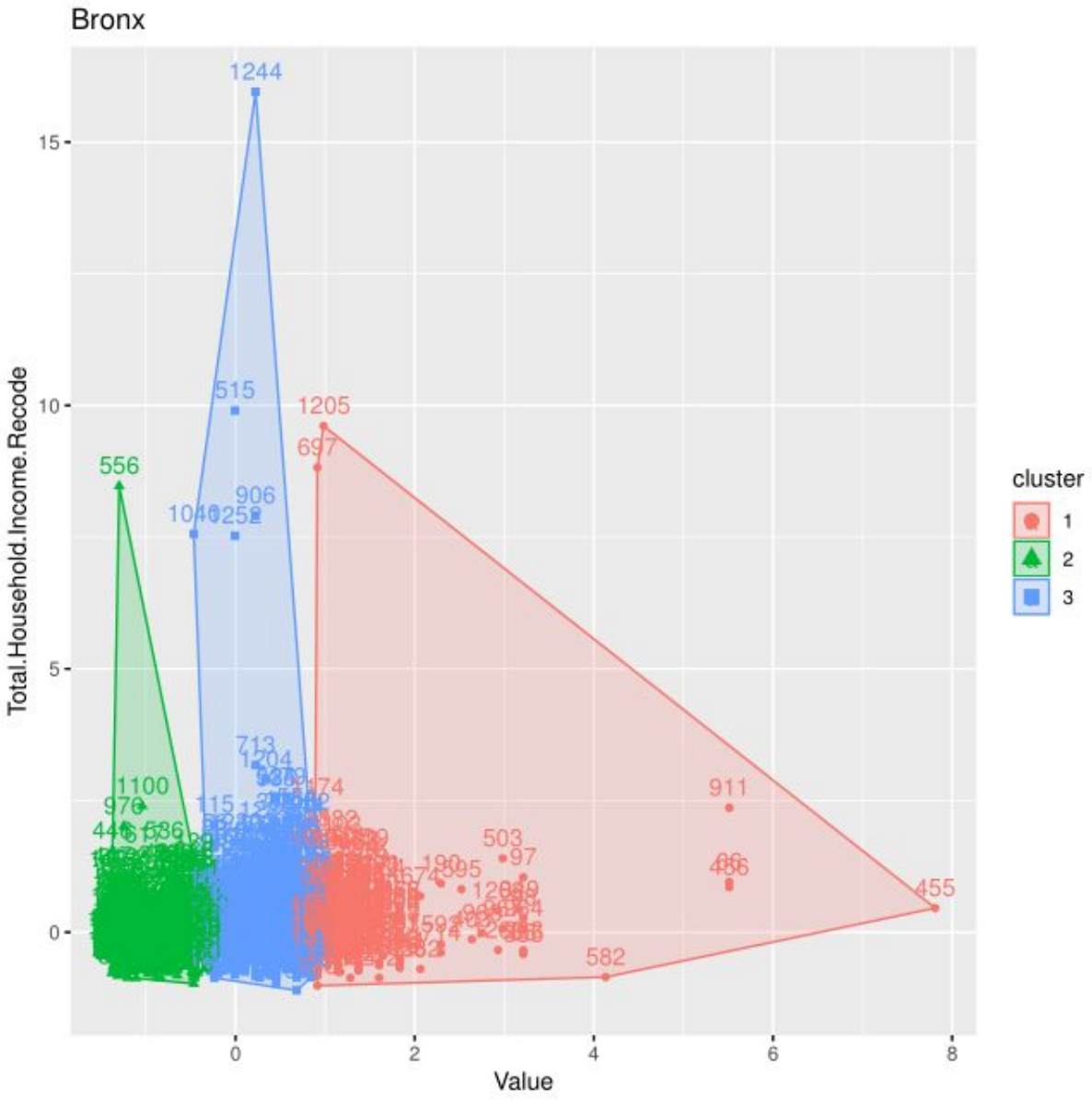


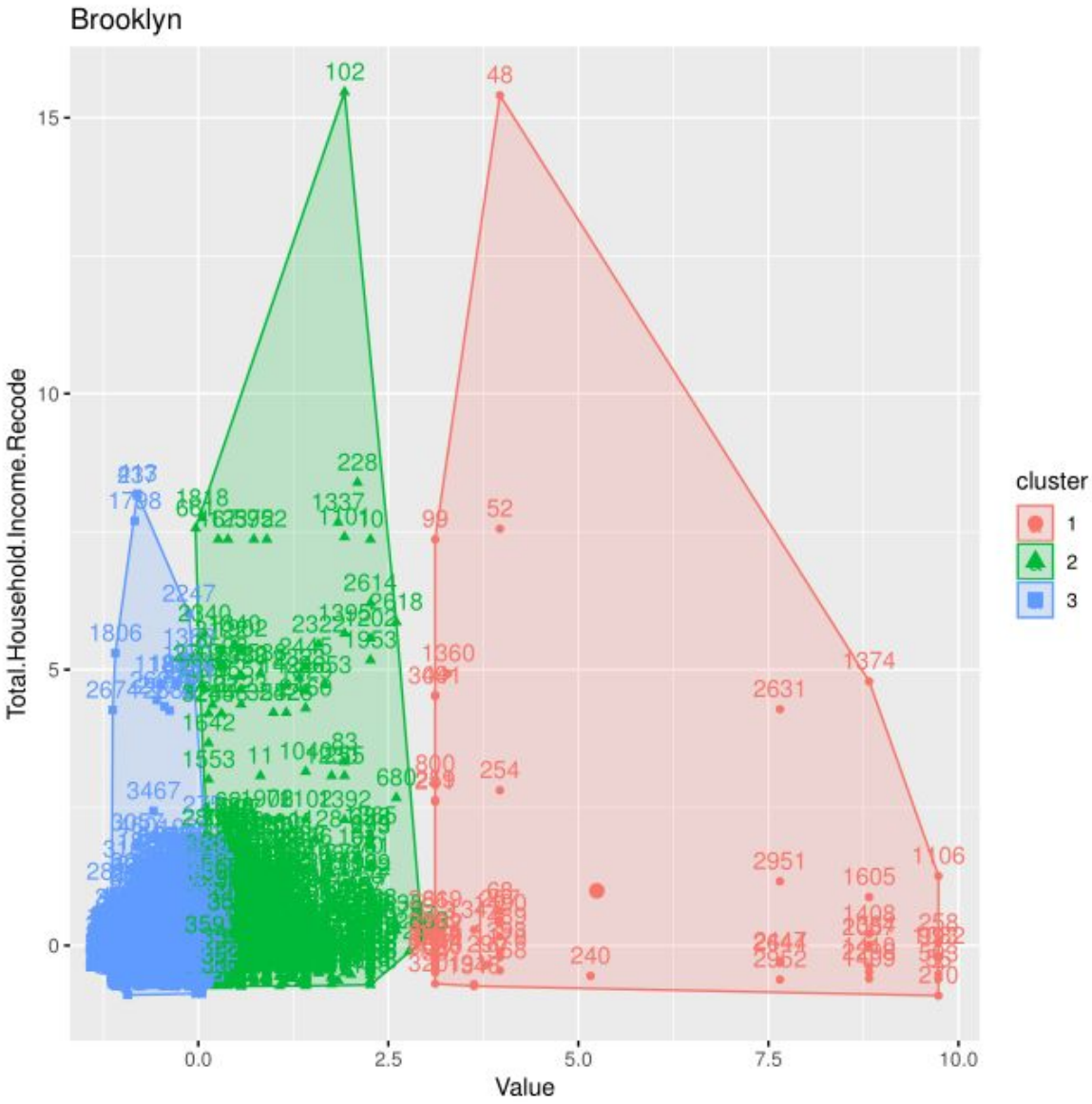
From the model above, we can see that with a data set the size of 20 rows, the model distinguished 4 groups whereas when there were 330 rows, the model distinguished 7 distinct groups. This, unsurprisingly, shows us that the more instances that there are, the more groups that can be created from the data. What is interesting, though, is that the increase in the number of groups is not linear.

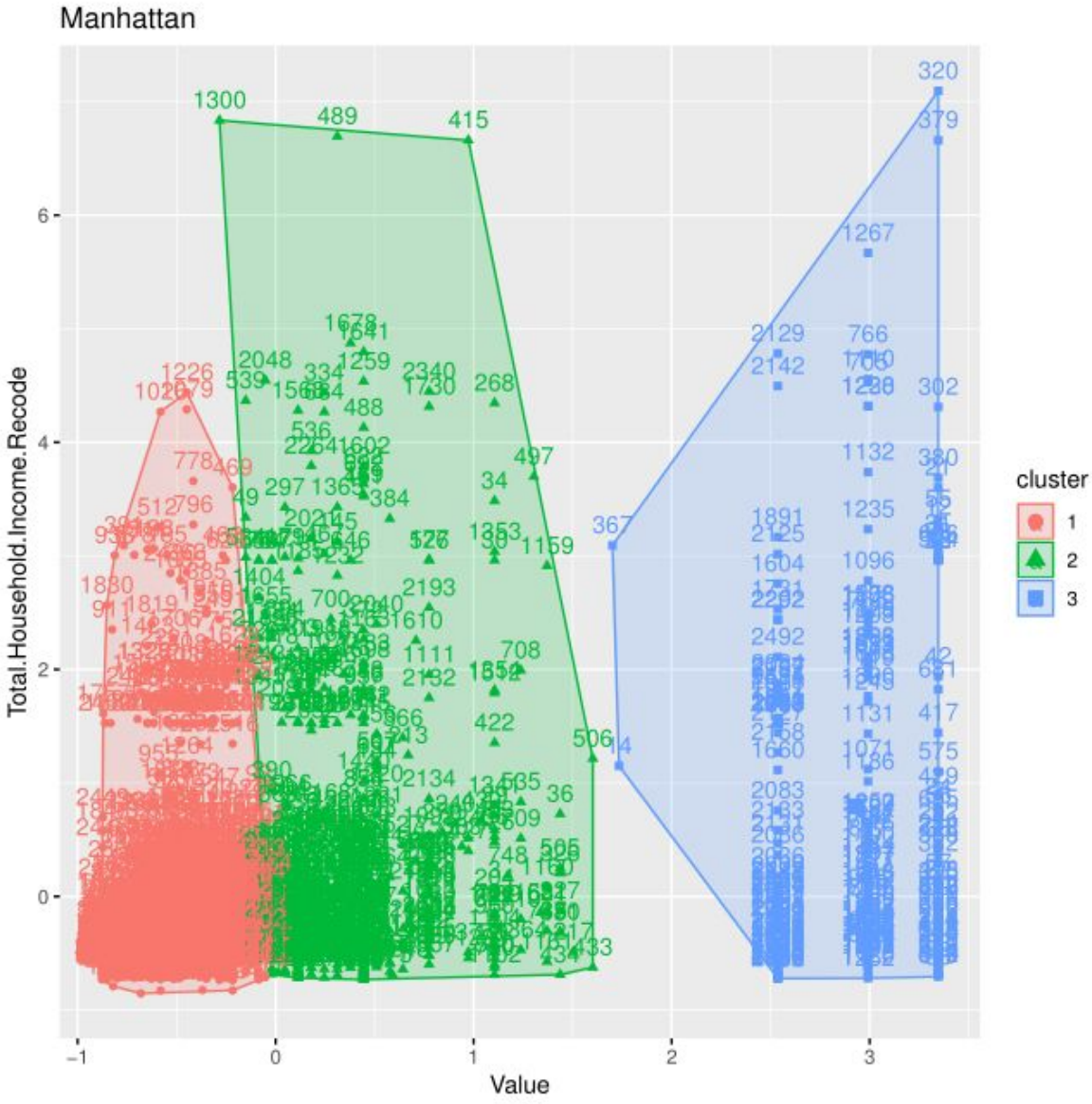
Perhaps this data would be a bit more understandable if the individual row ID's meant more (perhaps we could focus on a specific characteristic that is particularly unusual about any individual row) and that might give us more data on how the differences were noted between the

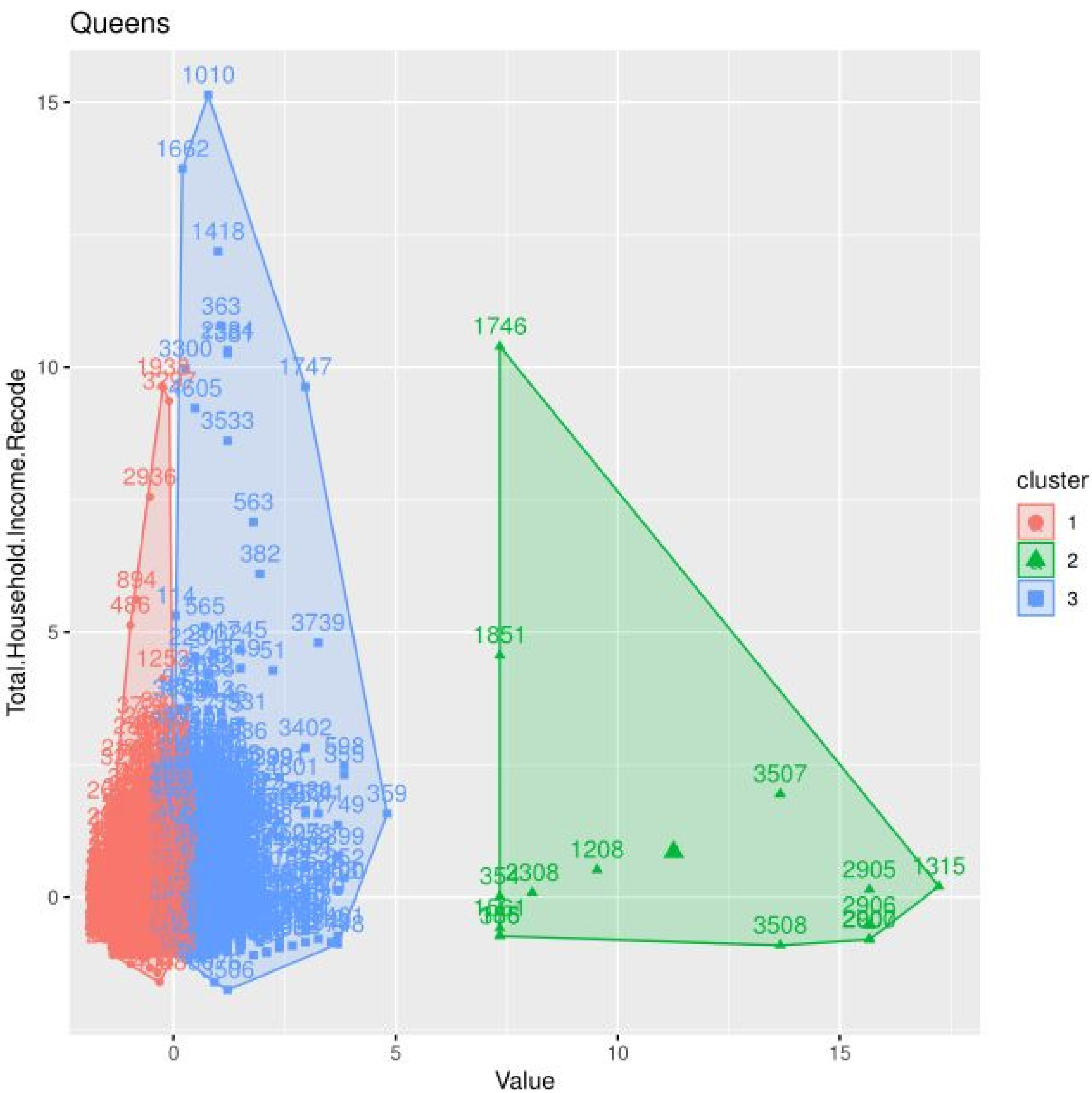
super small sample. Alternatively, if we reduced the number of features to 2-3 instead, we might have had more useful information to consider when looking at how the data was grouped. For future note, we will also try scaling all the data so that every value is within the same range. Since Euclidean distance is being used to understand group the data, scaling everything the same way would be useful in seeing differences in the data itself. However, as it stands now, the inferences that can be made from the hierarchical grouping is very limited.

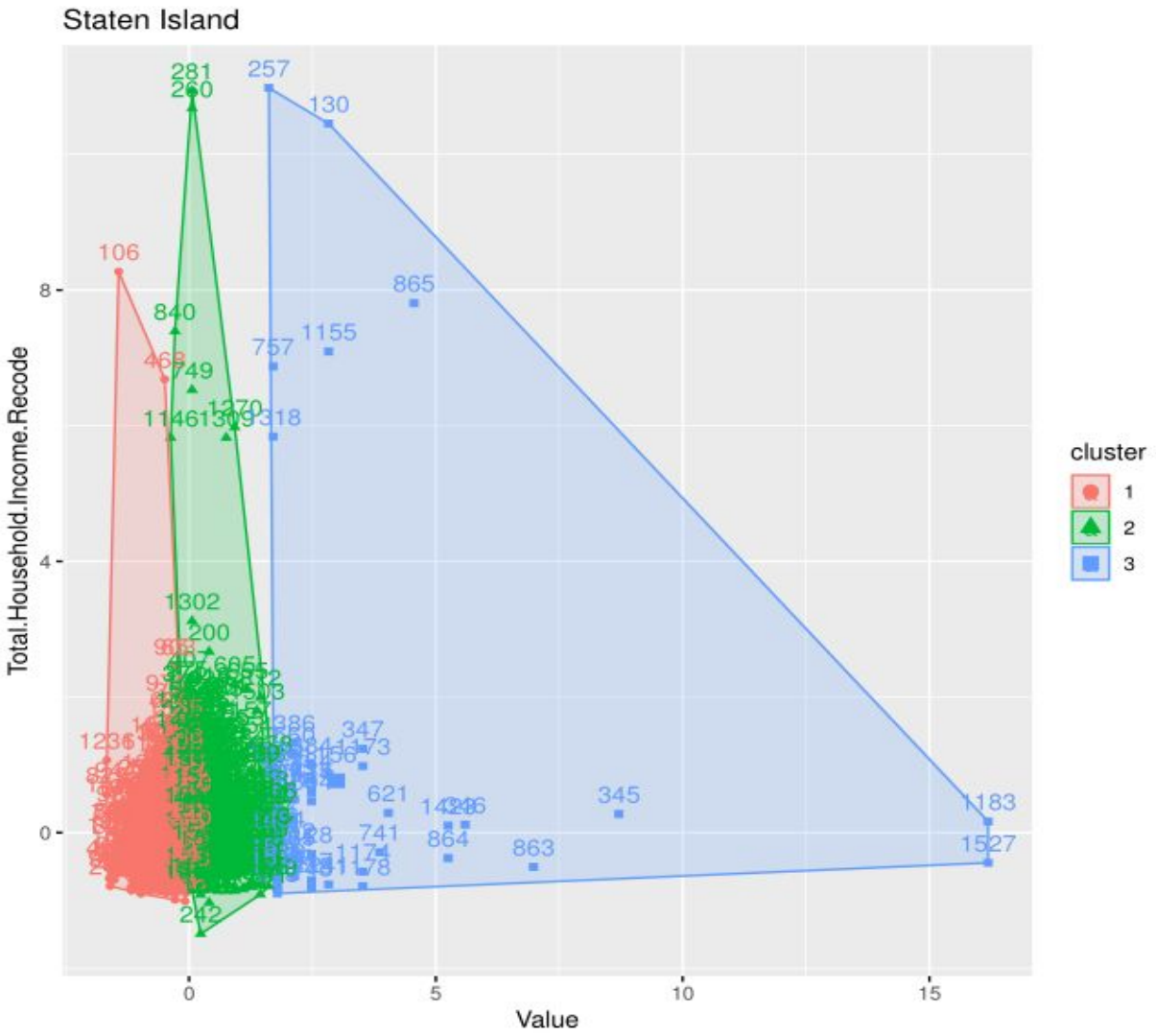
For our K-means analyses we were able to get a visual projection of three tiers of economic classes for each borough. In a sense showing a breakdown of upper, middle and lower classes for each area by looking at both the value and household income for owned residences.




Brooklyn
Total.Household.Income.Recode
Value
cluster
1
2
3
0
5
10
15
0.0
2.5
5.0
7.5
10.0


Manhattan
Total.Household.Income.Recode
Value
cluster
1
2
3


Queens
Total.Household.Income.Recode
Value
cluster
1
2
3
15
10
5
0
0
5
10
15
1010
1662
1418
363
3300
1747
3533
2936
563
382
894
486
114
565
3739
1253
3402
359
1746
1851
3507
1208
2308
1315
2905
3508



In many ways these graphs appeared much as one would expect based on general knowledge of the New York City region -- Manhattan is relatively wealthy, Queens has quickly gentrified with little middle ground. But perhaps what is most surprising is the Staten Island and Bronx have such similar shapes with their clusters, even though the Boroughs are quite different in a lot of ways.

Through the course of this project we were able to use many different tools to analyze data regarding housing in the New York area. We were able to glean a lot of different information, everything from the impact of hurricanes on the more coastal regions of the city to Staten Island rat infestations and their relation to home value. Unfortunately, one of the clearest conclusions that we ran into time and time again was the necessity for good data. Far too often we were faced with unremarkable projections and analyses which seemed to be no more insightful than general knowledge. It's almost obtuse, but data is clearly so critically important to data analyses.