An Analysis of Housing Quality in New York City

Damien Chambon[1], Jacob Gerszten[2]
[1] University of Virginia, Charlottesville, VA, USA (dlc8mt@virginia.edu)
[2] University of Virginia, Charlottesville, VA, USA (jeg6bk@virginia.edu)

Abstract
The physical condition of a person's home plays a large role in that person's overall quality of life. This paper attempts to measure housing quality through a standardized index and determine factors, such as social groups and economic characteristics, that impact living conditions. Specifically, we focus on disparities based on ownership status, comparing renters and owners. Using data collected by the New York City Housing and Vacancy Survey, we analyze relevant factors, such as location, house value, and demographic information, and how the impact of these factors on the housing index changes over time. We first conclude that renters and owners do in fact face different levels of housing quality. We also point out that some factors become more detrimental to housing quality over time while others become beneficial. We recommend further investigation into these housing quality disparities to inform policies that improve quality of life in New York City.

Key Words: housing quality index, "broken window" theory, ownership status, factors explaining housing disparities, principal component analysis

1. Introduction

To some observers, houses with cracked walls and broken windows are simply eyesores for the neighborhood, but the impact on residents goes beyond just outward cosmetic looks. It has been shown that housing conditions impact the quality of life for residents through different factors such as health, mood, and self-esteem[1]. Indeed, living in cold housing can cause health issues[2] as well as living in damp housing can be the reason for respiratory diseases[3]. Furthermore, it has been reported that poor housing conditions can produce depression[4] and can have a negative social impact[5]. Therefore, identifying differences in housing quality across citizens of a large city such as New York can provide beneficial insight into finding solutions to issues that communities face. To do so, we created a housing quality index which we then used to identify drivers of differences in housing quality across communities using data from the New York City Housing and Vacancy Survey.

We first defined several questions which we then attempted to answer. Our city of interest, New York City, faces many challenges related to high levels of poverty, crime, and homelessness. Despite its decrease, the share of New York City population living at or near the poverty rate is 43.1%[6]. On top of that, in 2016, more than 59,000 property crimes took place in New York City, making it one of the cities with the highest property crime rate[7]. Finally, homelessness levels have recently been on the rise: it has been reported that more than 63,000 people have slept in New York City shelters in February 2019[8].

The city is known for its "broken windows" policing policy, where the city police inferred that little issues such as broken windows provide an environment prone to crime. This policy suggests that lower levels of housing quality are directly linked to higher levels of

crime[9]. Thus, our main goal was to understand what factored into housing quality so that we could give recommendations to improve living conditions in New York City.

Our first task was to create an index that characterizes an individual building's overall condition. Due to the large number of variables related to the building conditions in our data set, we had to summarize and combine multiple building variables into one value for each observation, while maintaining as much original detail as possible. We then used principal component analysis to create our index.

Second, we focused on the differences in building conditions depending on the ownership status using nonparametric statistical tests. We hypothesized that ownership status does in fact have an impact on housing quality.

Finally, we tried to understand other relevant variables which impact housing quality. Most importantly, we tested to see if certain factors affect renters and owners in different ways. We also tracked the changes in the importance of certain predictors over time.

2. Data

The data that we use comes from the New York City Housing and Vacancy Survey (NYCHVS)[10] which contains survey results of the New York City housing stock and population from 1991 to 2017, collected every three years. The surveys, sponsored by the New York City Department of Housing Preservation and Development (HPD), are representative of housing across all boroughs within New York City. The dataset contains roughly 200 variables and 10,000 unique buildings per year. The data reflects New York's size and diversity and encompasses a wide range of responses designed to capture detailed information about each location. The dataset contains surveys for 10 years between 1991 to 2017. The exact number of variables and observations differ between years, but they are mostly similar over the data collection period.

Before analyzing the dataset, we were forced to remove several unusable variables and recode other variables. In the end, we removed 34.57% of our data going from 156,230 to 102,218 observations, while keeping 31 variables. The subsequent analysis was conducted in RStudio.

3. Methods

3.1 Creating a Housing Quality Index

To create our Housing Quality Index (HQI), we first identified variables related to housing quality which we wanted to include in our index.

Variables included in the HQI: *Wall Severity, Window Severity, Stairway Severity, Floor Severity, Building Condition, Toilet Breakdowns, Kitchen Functioning, Mice and Rats, Cracks in Walls, Holes in Floors, Broken Plaster, Water Leakage*

We created a subset of the data with all variables related to the housing conditions with binary variables for each condition. We wanted to keep the differences between the commonality of certain issues, i.e. we did not want some conditions to be weighted the same if they were not as common as others. In other words, for two households with the same number of issues, the one that faces fewer common issues would have a worse HQI.

We chose to use principal component analysis (PCA) to reduce the dimensionality of the housing conditions, described by the aforementioned variables, into one index. To get the value, the features are weighted depending on how much new information they add: if they bring little additional information, or in other words, they have low variance, they will have a low impact on the final value. While the method is an efficient way to create an index, it is not easy to interpret. Therefore, in our analysis, for a given household, we cannot determine whether a high index value is due to having many issues or only heavily weighted issues. We can only use this measure as a comparison between houses. Our initial values ranged from -0.403 to 2.935. To standardize these values, we scaled them from 0 to 10, with a value of 0 representing no issues and a value of 10 corresponding to a household where all possible issues have been reported.

3.2 Comparing housing quality distributions between owners and renters

To test whether or not the HQI distributions for owners and renters are the same, we used a nonparametric test, the Mann-Whitney Test. This allowed us to keep the logarithmic distributions of the HQI. In this test, the test statistic will be equal to the number of pairs (x_i, y_j) where $x_i < y_j$ and where x_i belongs to the set of the renter HQIs and y_j belongs to the set of owner HQIs. For this test, our null hypothesis was the following:

$$H_0: F_{renters} = F_{owners}$$

where $F_{renters}$ is the distribution of the housing quality index for renters and F_{owners} is that of the owners. Our alternative hypothesis was the following:

$$H_0: F_{renters} \leq F_{owners} \text{ for all } x$$

with a strict inequality for at least one. This implies that the values of the index for the renters are larger than for the owners.

3.3 Determining other relevant factors

We wanted to determine relevant factors that impacted housing quality besides ownership status. We first made several visualizations to try and identify potential factors. We then created a linear regression model with HQI as the response variable to test whether or not these other factors were significantly associated with housing quality.

Thus, the linear regression model was the following:

$$HQI = \beta_0 + \beta_1 HouseholderSex + \beta_2 HouseholderAge + \beta_3 HouseholderHispanicOrigin + \beta_4 DurationOfStay + \beta_5 NumberOfUnits + \beta_6 OwnerInBuilding + \beta_7 NumberOfStories + \beta_8 NumberOfRooms + \beta_9 PlumbingFacilities + \beta_{10} KitchenFacilities + \beta_{11} LengthOfLease + \beta_{12} ResidentRating + \beta_{13} HouseholderIncome + \beta_{14} Year + \beta_{15} Borough + \beta_{16} NumberOfPeople + \beta_{17} Status$$

3.4 Testing the impact of ownership status on housing quality

To test the significance of ownership status on different factors, we created a linear regression model with the HQI as the response variable for each status, i.e. one linear regression for owners and one for renters. With those two models, we were able to compare the coefficient of the different factors between the two ownership statuses. In order to determine if the differences were significant, we also used a linear regression with the entire dataset where the housing condition index was the response variable, but this time, we added the status as an interaction term.

We computed the difference between the coefficients of the regression for the renters and the regression for the owners so that we could understand for what type of ownership certain predictors were detrimental or beneficial. We also calculated the percentage that the difference represents compared to the value of the coefficient for the renters.

3.5 Understanding changing importance of factors over time

One of our main goals was to track how the importance of the predictors evolved over the years. Since the number of predictors is quite large, we decided to compare the predictors that were significant in all the years. To do so, we first conducted a linear regression for each year in the dataset, where the HQI was the response variable. After this initial step, we conducted a linear regression similar to the first one we did, but this time, we added the year variable as an interaction term on all other variables.

4. Results and Discussion

4.1 Creating the Housing Quality Index

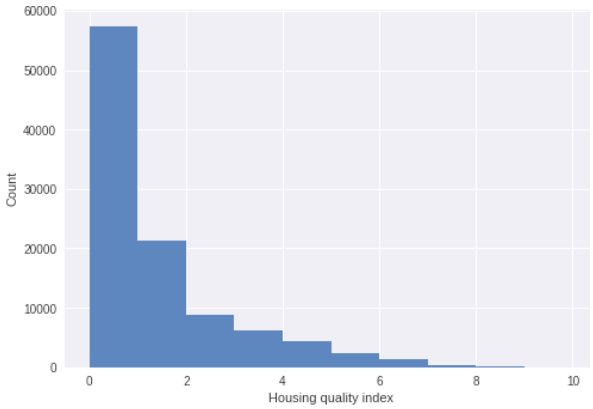
The HQI generated meaningful insight into the distributions of housing quality across certain subgroups. Summary statistics for the index are provided in Table 1.

Table 1: Some statistics of the housing condition index

Statistics	Value
Mean	1.207
Standard Deviation	1.639
Minimum Value	0
Q1	0
Q2	0.28
Q3	1.7
Maximum Value	10

With this data, we have an overview of the housing quality in New York City. First, we notice that there are people that have no issues at all with their household, whereas some people have all possible issues. However, the latter does not appear to be common since the third quantile is 1.7, meaning that 75% of the residents have an HQI lower than 1.7. Indeed, with a mean of 1.207 out of a scale from 0 to 10, people on average have few issues. Figure 1 contains a histogram of the overall HQI. The distribution is extremely right skewed and appears to be logarithmic.

Figure 1: HQI Distribution



4.2 Comparing housing quality distributions between owners and renters

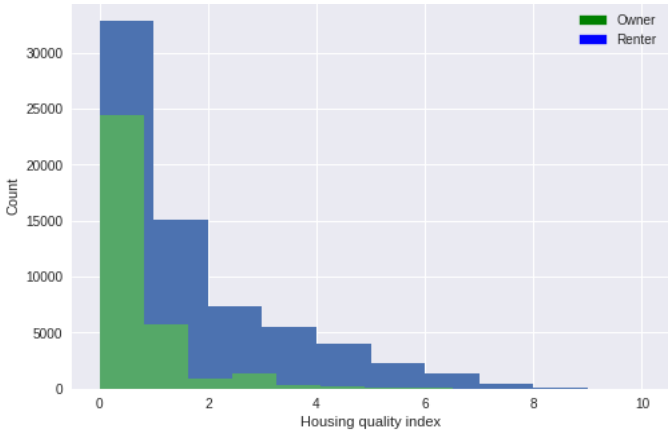
There are almost twice as many renters than there are owners, showing that New York City is a city where most people rent. Table 2 contains the means of the two groups:

Table 2: Mean value of the housing quality index per status

Status of the residents	Mean value of the housing quality index
Owner	0.547
Renter	1.523

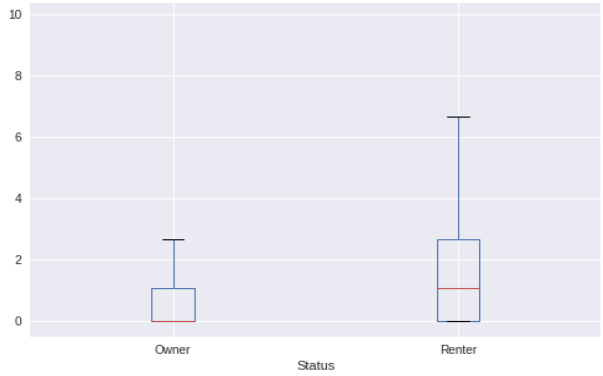
We note that the index is higher on average for renters than owners, which suggests that renters face more issues with their dwellings. The distributions of both groups are presented in Figure 2.

Figure 2: Housing quality indices grouped by status



The distributions have the same general shape as the overall HQI distribution and appear to be logarithmic. However, a major difference is that the distribution for renters is more right-skewed, which suggests that renters tend to experience higher indices.

Figure 3: Boxplot of the HQI grouped by status



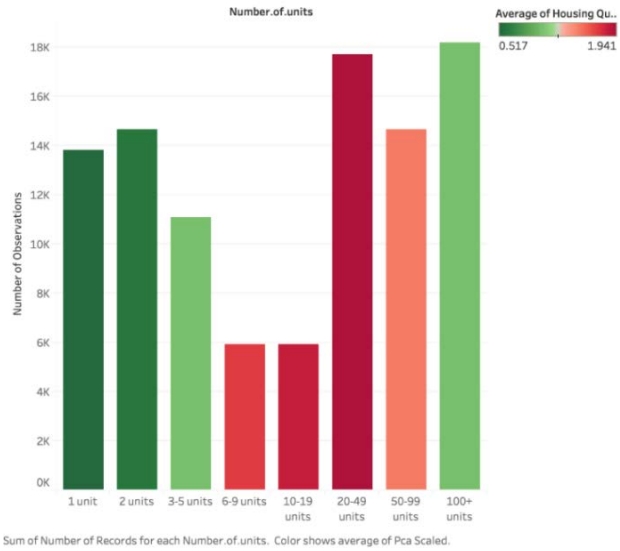
In Figure 3, we find the values for renters are more spread out. For the owners, 75% of the observations have an index lower than 1. We can even see that the median is 0: 50% of the renters have an index equal to 0 or, in other words, have no issue with their dwelling. In contrast, 25% of the renters have a value between 2.5 and 7 for the index, while the range is from 1 to 3 for owners. It is worth noting that the median for the value of the index for renters is equal to 1, which is still not high on the scale, despite the long right tail.

The Mann-Whitney test statistic was equal to 770984750 and the p-value had a value very close to 0. Since the p-value is very close to 0, we can conclude that the difference between the values of the housing quality indices is significant. In other words, renters do tend to have more issues with their dwellings than owners.

4.3 Determining other relevant factors

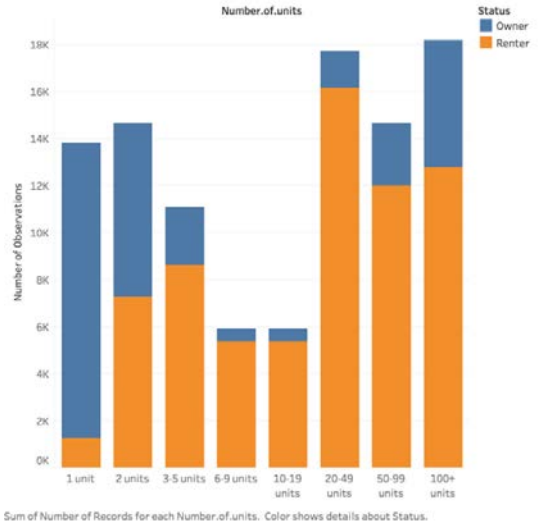
To identify other important factors, we created a linear regression model with HQI as the response. Before fitting the linear regression, we wanted to get some overview of our data to get an idea of how several variables were related to the housing condition index.

Figure 4: HQI of dwellings grouped by number of units



In Figure 4, we see what kinds of buildings are most common in New York City. Most of the residents live in small-size buildings (1 to 5 units) and in medium-to-large-size buildings (more than 20 units). We notice that the smallest count is for houses with 6 to 9 units with only 6,000 observations. That means that the results from the linear regression will be significant even for those values of the variable number of units as the number is large enough. When considering the average HQI for each type of building, we see a trend: for small-size buildings, residents tend to face fewer issues as well as buildings comprising more than 100 units. This implies that housing conditions are the worst for medium-size buildings.

Figure 5: Status of owners grouped by number of units



Owners are more likely to live in buildings of 1 to 2 units, as shown in Figure 5. Since owners are potentially more likely to fix the issues since they directly suffer from the consequences, that could explain why those buildings have a lower HQI. Buildings with more than 100 units are mostly rented. Since such dwellings are large and potentially better managed managers may be quicker to fix those problems compared to smaller places.

Figure 6: HQI against household value, by status

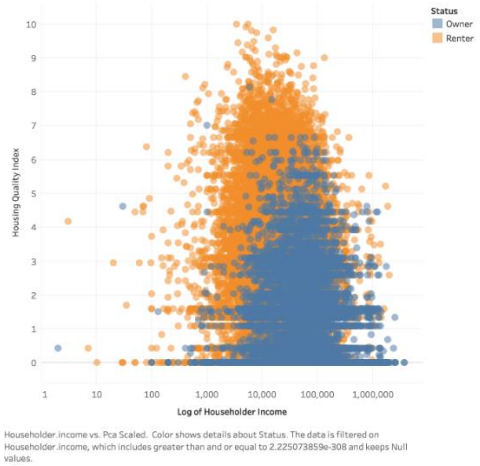
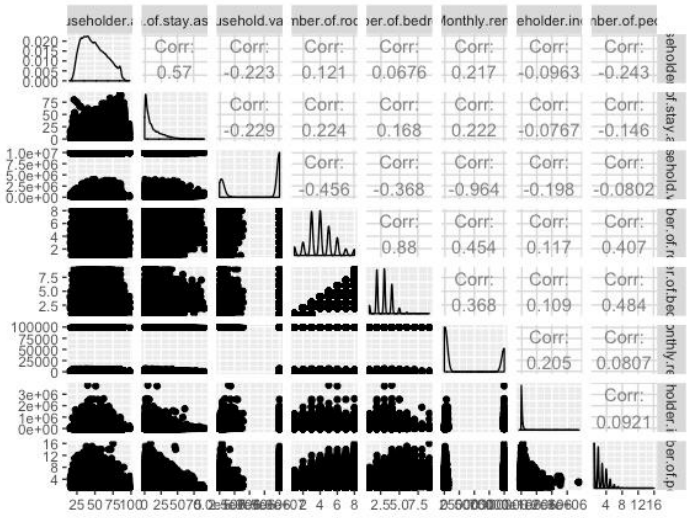


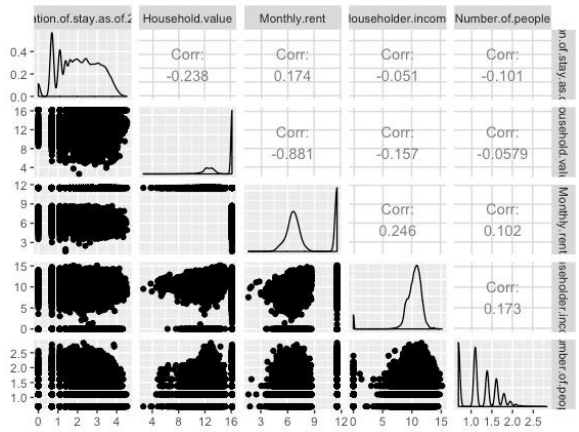
Figure 6 shows us more about the relations the householder income variable has with the housing quality index. We also had to use the logarithmic function on the values of the householder income to make it easier to understand. First, there does not appear to be a strong relation between the log of the householder income and the HQI. In fact, the data looks quite normally distributed, suggesting that there may not be a linear relationship between the two variables. We will see the significance of the income to predict the HQI after fitting the linear regression. We see that owners tend to have higher incomes. We can also see that owners do tend to have a lower HQI, confirming our earlier results.

Figure 7: Scatterplot matrix of the numerical variables



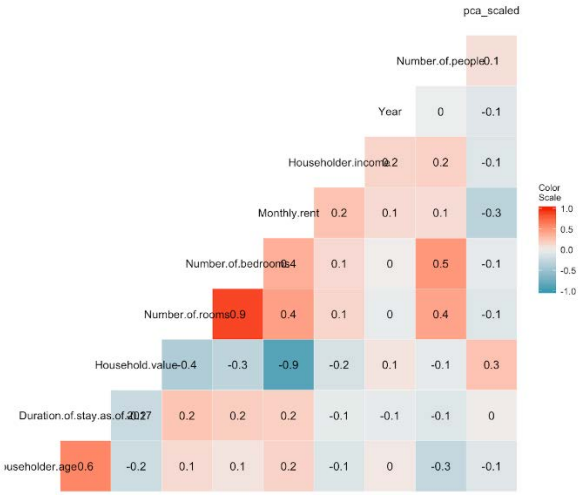
Before fitting the regression, we created a scatter plot matrix in Figure 7 for all the numerical variables to check for normality. The first column (householder age) as well as the fourth (number of rooms) and fifth one (number of bedrooms) appear quite normal, which validates the assumption. The other variables all look logarithmic. Furthermore, the third column (household value) and the sixth one (monthly rent) look somewhat different: they have two peaks. For those two variables, the peak on the right of each distribution can be explained by the placeholder value used when the value was not known, e.g. for renters, the household value is not known so the value assigned is 9999999, vice versa for the owners with the monthly rent. Therefore, for those two variables, if we look only at the first peak, the actual values, we can see that they also look logarithmic.

Figure 8: Scatterplot matrix of the numerical variables, after a log transformation



Other than the peak for the household value and the monthly rent explained above, all variables look more normally distributed. However, the first column (duration of stay as of 2017) as well as the number of people variable still do not look normal.

Figure 9: Correlation matrix of the numerical variables



Another important assumption to check is for multicollinearity using Figure 9 which is a correlation matrix of the predictors. We observe strong evidence of multicollinearity within 2 pairs: the household value with the monthly rent, and the number of rooms with the number of bedrooms. For the first pair, this issue will be quite easy to deal with. In fact, as previously stated, there is no observation that has two actual values for those variables at the same time since one of the two variables will have a placeholder variable based on the status of the resident, e.g. 999999 for household value for renters and 999999 for monthly rent for renters. Therefore, in our analysis, we will never include the two variables at the same time. When we will study our entire dataset, we will not take those variables into account. When we will study renters, we will include monthly rent in our analysis but not household value, and vice versa for owners. Thus, we will not have to deal with the placeholder values. Regarding the second pair, number of rooms vs number of bedrooms, the multicollinearity makes sense. It is highly likely that the more rooms a dwelling has,

the more bedrooms there are. Since the correlation is quite high (around 0.9), we have to remove one of the two variables. Given that the variable for the number of bedrooms was less normally distributed than the number of rooms, we should remove the number of bedrooms. Finally, there is a somewhat positive correlation between the age of the householder and the length of time for which they stayed in a unit. This shows that as people grow older, they usually stay in the same place.

We fitted a linear regression for all observations with all the predictors remaining against the HQI. As mentioned before, we did not include 2 predictors: the monthly rent and the household value.

Figure 10: Linear Regression Output

```
Call:
lm(formula = pca_scaled ~ ., data = inputdata_noHval_noRent)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1936 -0.8954 -0.2985  0.6858  8.2982

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                20.3101315  1.0895147  18.641  < 2e-16 ***
Householder.sexMale                        -0.1277379  0.0090120 -14.174  < 2e-16 ***
Householder.age                            -0.0050864  0.0003390 -15.006  < 2e-16 ***
Householder.hispanic.originNot hispanic    -0.0916637  0.0110311  -8.310  < 2e-16 ***
Duration.of.stay.as.of.2017                 0.1596086  0.0056452  28.274  < 2e-16 ***
Number.of.units10-19 units                  0.5431686  0.0294700  18.431  < 2e-16 ***
Number.of.units100+ units                   0.4667773  0.0324735  14.374  < 2e-16 ***
Number.of.units2 units                     -0.0510832  0.0181619  -2.813  0.00491 **
Number.of.units20-49 units                  0.6186338  0.0273400  22.627  < 2e-16 ***
Number.of.units3-5 units                    0.0683456  0.0231488   2.952  0.00315 **
Number.of.units50-99 units                  0.5126514  0.0299255  17.131  < 2e-16 ***
Number.of.units6-9 units                    0.4707507  0.0285895  16.466  < 2e-16 ***
Owner.in.buildingYes                       -0.3871926  0.0174503 -22.188  < 2e-16 ***
Number.of.stories11-20 stories             -0.4391045  0.0298607 -14.705  < 2e-16 ***
Number.of.stories21+ stories               -0.6000316  0.0339000 -17.700  < 2e-16 ***
Number.of.stories3-5 stories                0.1302123  0.0167146   7.790 6.75e-15 ***
Number.of.stories6-10 stories              -0.0618238  0.0234286  -2.639  0.00832 **
Number.of.rooms                             0.0281653  0.0040049   7.033 2.04e-12 ***
Plumbing.facilitiesYes                     -0.7832277  0.0712394 -10.994  < 2e-16 ***
Kitchen.facilitiesYes                      -0.5747839  0.0716446  -8.023 1.05e-15 ***
Length.of.lease2 years                      0.1303658  0.0129213  10.089  < 2e-16 ***
Length.of.leaseBetween 1 and 2 years       -0.0937677  0.0538346  -1.742  0.08155 .
Length.of.leaseLess than 1 year             0.2008169  0.0394413   5.092 3.56e-07 ***
Length.of.leaseMore than 2 years            0.1350409  0.0283609   4.762 1.92e-06 ***
Length.of.leaseNo lease                     0.1137432  0.0164083   6.932 4.17e-12 ***
Length.of.leaseOwner-occupied              -0.3380522  0.0365013  -9.261  < 2e-16 ***
Resident.ratingFair                         0.8727032  0.0145510  59.975  < 2e-16 ***
Resident.ratingGood                         0.2101445  0.0117985  17.811  < 2e-16 ***
Resident.ratingPoor                         2.0014028  0.0233058  85.876  < 2e-16 ***
Householder.income                         -0.0047918  0.0025533  -1.877  0.06056 .
Year                                       -0.0091412  0.0005423 -16.856  < 2e-16 ***
BoroughBrooklyn                            -0.1140209  0.0144690  -7.880 3.30e-15 ***
BoroughManhattan                            0.0464058  0.0153852   3.016  0.00256 **
BoroughQueens                              -0.4040945  0.0150505 -26.849  < 2e-16 ***
BoroughStaten Island                       -0.3193648  0.0231097 -13.820  < 2e-16 ***
Number.of.people                            0.3737017  0.0133392  28.015  < 2e-16 ***
StatusRenter                               -0.2762324  0.0381730  -7.236 4.64e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.397 on 101746 degrees of freedom
Multiple R-squared:  0.2728,    Adjusted R-squared:  0.2726
F-statistic:  1060 on 36 and 101746 DF,  p-value: < 2.2e-16
```

From the results in Figure 10, we can spot several things. A lot of, if not all, the predictors are considered significant such as the age of the householder, the number of units in the

building, the length of the lease, the area of the dwelling and so on. We will only consider variables that are significant at the 0.001 threshold. With this threshold taken into consideration, that means that the householder income does not play in role in determining whether a house experiences more issues than the others. Interestingly, we can notice that the status of the residents (owners vs renters) is significant but has a negative coefficient. That means that, for dwellings with the exact same characteristics, a renter will have on average a HQI lower than the one of an owner, by a value of 0.27 out of 10. This contradicts our previous finding that showed that renters have higher housing condition index, meaning that they have more issues. This result can be explained by the somewhat low R squared value we obtained: 0.2728. This value tells us that our model explains 27.28% of the variance in the HQI. Therefore, other factors not present in our model impact the index.

4.4 Testing the impact of ownership status on housing quality

After seeing how variables impact the HQI, we wanted to see if the impact of certain variables on the HQI changes depending on the status owner/renter. To do so, we conducted two linear regressions with the HQI. In order to determine if the differences were significant, we also used a linear regression with the entire dataset, this time adding status as an interaction term. With this model, we performed an ANOVA to determine whether adding the interaction term made the predictor significant or not. Figure 11 contains the output.

Figure 11: Changes in coefficients over time results

```
Analysis of Variance Table

Response: pca_scaled
                                    Df Sum Sq Mean Sq   F value    Pr(>F)
Householder.sex                      1   2738  2738.0 1423.8416 < 2.2e-16 ***
Householder.age                      1   2661  2661.3 1383.9640 < 2.2e-16 ***
Householder.hispanic.origin          1   5540  5540.3 2881.1512 < 2.2e-16 ***
Duration.of.stay.as.of.2017          1   1995  1995.2 1037.5657 < 2.2e-16 ***
Number.of.units                      7  22445  3206.4 1667.4466 < 2.2e-16 ***
Owner.in.building                    1   7414  7414.3 3855.6930 < 2.2e-16 ***
Number.of.stories                    4   1799   449.8  233.8979 < 2.2e-16 ***
Number.of.rooms                      1   1041  1041.0  541.3328 < 2.2e-16 ***
Plumbing.facilities                  1    525   524.7  272.8425 < 2.2e-16 ***
Kitchen.facilities                   1    194   193.7  100.7183 < 2.2e-16 ***
Length.of.lease                      6   1023   170.5   88.6737 < 2.2e-16 ***
Resident.rating                      3  22819  7606.2 3955.4677 < 2.2e-16 ***
Householder.income                   1      1     1.4    0.7351  0.391231
Year                                 1    529   528.9  275.0378 < 2.2e-16 ***
Borough                              4   2127   531.6  276.4734 < 2.2e-16 ***
Number.of.people                     1   1535  1534.5  798.0001 < 2.2e-16 ***
Status                               1    102   102.2   53.1387 3.131e-13 ***
Householder.sex:Status               1      6     5.6    2.8899  0.089137 .
Householder.age:Status               1     50    49.9   25.9246 3.557e-07 ***
Householder.hispanic.origin:Status   1     39    38.6   20.0870 7.408e-06 ***
Duration.of.stay.as.of.2017:Status   1    429   428.9  223.0596 < 2.2e-16 ***
Number.of.units:Status               7    349    49.9   25.9580 < 2.2e-16 ***
Number.of.stories:Status             4    114    28.5   14.8210 4.135e-12 ***
Number.of.rooms:Status               1    113   113.5   59.0057 1.586e-14 ***
Plumbing.facilities:Status           1     20    19.7   10.2544  0.001364 **
Kitchen.facilities:Status            1      3     3.1    1.6351  0.201001
Resident.rating:Status               3   1093   364.3  189.4739 < 2.2e-16 ***
Householder.income:Status            1      1     1.2    0.6375  0.424604
Year:Status                          1     98    97.9   50.9345 9.613e-13 ***
Borough:Status                       4    255    63.8   33.1544 < 2.2e-16 ***
Number.of.people:Status              1    376   376.5  195.7802 < 2.2e-16 ***
Residuals                       101718 195599     1.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

An important finding that we can get based on this output is that the ownership status does not change the impact of the income on the HQI. Indeed, the p-value is quite large (more than 0.42). That means that for two people, one owner and one renter, with the same income, they will have, on average, the same issues. Conversely, for two people, one owner and one renter, that stayed for the same time in a dwelling, they will have, on average, a very different number of issues.

Since the number of variables significant with the interaction term "Status" is quite large, we decided to focus on those with a p-value lower than 0.001. We computed the difference between the coefficients of the regression for the renters and the regression for the owners. We also calculated the percentage that the difference represents compared to the value of the coefficient for the renters so that we could compare the differences across the variables.

Figure 12: Understanding the changing importance of factors over time

Variable	Renters: coeff.	Owners: coeff.	Difference	Percentage of diff. Compared to Renters
Householder.age	-0.00610	-0.00258	-0.00353	-57.74477
Householder.hispanic.originNot hispanic	-0.07733	-0.07256	-0.00476	-6.15702
Duration.of.stay.as.of.2017	0.21772	0.03780	0.17992	82.63874
Number.of.units10-19 units	0.62793	0.24168	0.38624	61.51085
Number.of.units100+ units	0.54479	0.24777	0.29701	54.51919
Number.of.units2 units	0.06707	-0.03958	0.10665	159.01470
Number.of.units20-49 units	0.68286	0.30327	0.37959	55.58772
Number.of.units3-5 units	0.22059	-0.08938	0.30997	140.51625
Number.of.units50-99 units	0.58491	0.28745	0.29747	50.85662
Number.of.units6-9 units	0.56609	0.18799	0.37810	66.79182
Owner.in.buildingYes	-0.39499	-0.17035	-0.22464	-56.87272
Number.of.stories11-20 stories	-0.46151	-0.26057	-0.20095	-43.54084
Number.of.stories21+ stories	-0.67852	-0.32281	-0.35571	-52.42447
Number.of.stories3-5 stories	0.18635	0.10556	0.08078	43.35147
Number.of.stories6-10 stories	-0.00076	-0.05345	0.05268	6890.50866
Number.of.rooms	0.02807	0.01324	0.01483	52.83946
Resident.ratingFair	1.00086	0.51532	0.48554	48.51238
Resident.ratingGood	0.27164	0.16762	0.10402	38.29441
Resident.ratingPoor	2.13002	1.06364	1.06637	50.06413
Year	-0.01310	-0.00378	-0.00932	-71.15060
BoroughBrooklyn	-0.13334	-0.01485	-0.11849	-88.86474
BoroughManhattan	0.03456	0.12943	-0.09487	-274.47889
BoroughQueens	-0.50908	-0.17303	-0.33605	-66.01050
BoroughStaten Island	-0.34148	-0.20125	-0.14023	-41.06591
Number.of.people	0.50397	0.07938	0.42459	84.24874

In this figure, we can see different information regarding the coefficients from the regressions previously mentioned. Only the significant predictors are featured in this table. We will interpret only the predictors where the difference between renters and owners is large compared to the values of the coefficients. Let us first look at the largest percentage, which is 6,890%. This percentage is much larger than the others suggesting that owners are much more advantaged than renters on this specific aspect, and more precisely when the number of stories is between 6 and 10. This percentage means that the impact of living in such a building on the housing condition index is greater for renters compared to owners. Therefore, it would not be wise for renters to buy dwellings in such buildings. If we look more closely at the coefficients, we can see that they are very small, suggesting a small impact on the HQI. After looking at the differences of coefficients for the other values of the "Number of Stories" variable, it turns out that living in a building with more than 21 stories is 56% more beneficial for renters.

If we look at the two next largest positive percentages. Those deal with the number of units equal to 2 or between 3 and 5. For those specific values of the "Number of Units" variable,

the percentage of the difference are above 140%. Those number show us that the housing condition index of renters is 140% more impacted when they live in a building with 2 units compared to owners. That impact is 159% more important for renters compared to owners when they live in a building with 3-5 units. Those percentages show that this is the type of buildings where the difference between owners and renters is the most significant. It would be interesting to understand why that difference is especially large for those buildings. We notice that for all the values of the "Number of Units" variable, the differences are positive, which means that there is no type of buildings where it is more advantageous to be a renter given the housing condition index.

Lastly, if we look at the smallest percentage difference, we see that it corresponds to the variable Borough being equal to Manhattan. It has a value of -274%, meaning that living in Manhattan is 274% more beneficial for renters than owners. In other words, if you live in Manhattan, the average increase of the housing condition index of your dwelling will be 274% smaller than that of owners. This finding is quite intriguing. We could posit that if you are an owner, living in Manhattan could end up being detrimental for the housing condition index because the life is quite expensive in this area, and you would not spend as much money on your own dwelling as if you were living in a cheaper area such as Brooklyn, where the difference is smaller. It is interesting to note that it looks like it is more beneficial to be a renter on all areas of New York City, though it contradicts our previous finding that being an owner was related to a lower housing condition index. We could infer that living in the Bronx (the baseline) may be much better for owners, or that other variables may compensate for the differences in coefficients detrimental to owners. Indeed, the coefficients are quite small. After seeing the differences of coefficients between owners and renters, we decided to focus on their changes of values over the years.

4.5 Understanding changing importance of factors over time

One of our main research questions was also understanding how the coefficients of the predictors of the previous linear regression evolved over the years. To tackle this topic, analyzing the entire dataset as a whole might not be quite relevant. Since the number of predictors is quite large, for our analysis, we decided to compare the predictors that were significant in all the years. To do so, we first conducted a linear regression for each year in the dataset, where the HQI was the response variable and the other variables (except the year) were predictors. By doing so, we were able to have the coefficients for each predictor across the years. After this initial step, we conducted a linear regression similar to the first one we did, where the HQI was the response variable and the other variables were the predictors, but this time, we added the year variable as an interaction term on all the variable. Once this was done, we decided to only take into account the variables whose p-value were smaller than 0.001 once the interaction term was taken into account. In other words, the variables that we selected had significant changes in terms of impact on the HQI over time.

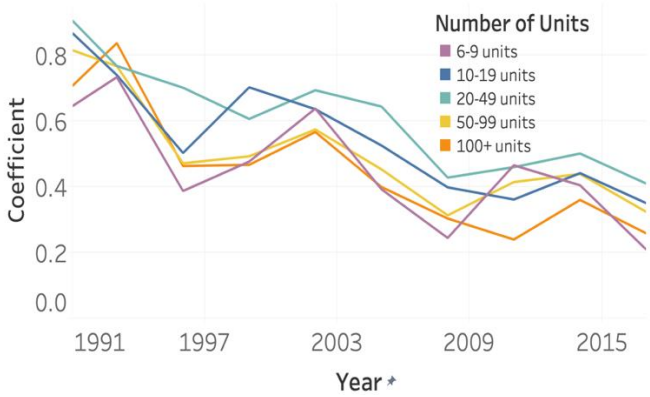
In the end, we had 11 predictors out of the 16 used in our models. Regarding categorical variables, we used an ANOVA table to know if the categorical variable was significant as a whole. For such variables, we only kept the values for which the p-value was really close to 0 in our analysis. The predictors that we kept were: the sex and age of the householder, whether the householder has a Hispanic origin, the number of units, whether the owner was in the building, the number of stories, whether there are kitchen and/or plumbing facilities, the length of the lease, the rating of the resident about their neighborhood and the borough.

Figure 13: RStudio output of the ANOVA, with year as an interaction term

```
Analysis of Variance Table

Response: pca_scaled
                                Df Sum Sq Mean Sq   F value    Pr(>F)
Householder.sex                  1   2738  2738.0 1413.6075 < 2.2e-16 ***
Householder.age                  1   2661  2661.3 1374.0165 < 2.2e-16 ***
Householder.hispanic.origin      1   5540  5540.3 2860.4424 < 2.2e-16 ***
Duration.of.stay.as.of.2017      1   1995  1995.2 1030.1080 < 2.2e-16 ***
Number.of.units                  7  22445  3206.4 1655.4615 < 2.2e-16 ***
Owner.in.building                1   7414  7414.3 3827.9795 < 2.2e-16 ***
Number.of.stories                4   1799   449.8  232.2168 < 2.2e-16 ***
Number.of.rooms                  1   1041  1041.0  537.4419 < 2.2e-16 ***
Plumbing.facilities              1    525   524.7  270.8813 < 2.2e-16 ***
Kitchen.facilities               1    194   193.7   99.9944 < 2.2e-16 ***
Length.of.lease                  6   1023   170.5   88.0364 < 2.2e-16 ***
Resident.rating                  3  22819  7606.2 3927.0371 < 2.2e-16 ***
Householder.income               1      1     1.4    0.7298 0.3929387
Year                             1    529   528.9  273.0610 < 2.2e-16 ***
Borough                          4   2127   531.6  274.4862 < 2.2e-16 ***
Number.of.people                 1   1535  1534.5  792.2643 < 2.2e-16 ***
Status                           1    102   102.2   52.7568 3.802e-13 ***
Householder.sex:Year             1     36    36.2   18.6898 1.539e-05 ***
Householder.age:Year             1    429   429.0  221.5152 < 2.2e-16 ***
Householder.hispanic.origin:Year 1     31    31.3   16.1712 5.791e-05 ***
Duration.of.stay.as.of.2017:Year 1      1     1.2    0.6064 0.4361338
Number.of.units:Year             7    396    56.5   29.1820 < 2.2e-16 ***
Owner.in.building:Year           1     29    28.5   14.7192 0.0001248 ***
Number.of.stories:Year           4    137    34.3   17.7037 1.545e-14 ***
Number.of.rooms:Year             1      1     0.7    0.3581 0.5495449
Plumbing.facilities:Year         1     48    47.5   24.5465 7.266e-07 ***
Kitchen.facilities:Year          1     24    24.0   12.3740 0.0004356 ***
Length.of.lease:Year             6     80    13.3    6.8551 2.738e-07 ***
Resident.rating:Year             3     87    29.0   14.9894 9.442e-10 ***
Householder.income:Year          1      3     3.1    1.5855 0.2079736
Year:Borough                     4    227    56.7   29.2964 < 2.2e-16 ***
Year:Number.of.people            1      9     8.6    4.4634 0.0346314 *
Year:Status                      1      7     7.2    3.7059 0.0542230 .
Residuals                   101711 197001     1.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 14: Impact of "Number of Units" on the HQI over time



In Figure 14, we see the changes in coefficients for the significant values of the variable “Number of Units”. Interestingly, we see that they all have similar changes over the years: they all increase and decrease at the same time. The overall trend is a decrease, meaning that they are representative of a lower HQI, compared to the baseline which is “1 unit”. In other words, living in a single-unit dwelling is less and less related to a lower HQI. We can also see here a decrease in coefficients in 2008, especially for the number of units being between to 6 and 9. We can interpret this value by saying that the buildings comprising between 6 and 9 units were less impacted by the 2008 financial crisis compared to single-

unit dwellings, eventually bridging the gap in housing conditions of those two types of houses. While this graph shows that larger numbers of units are related to fewer issues with a dwelling, those findings contradicts with the previous graph, where we showed that living in taller buildings meant more issues. Those differences in findings could be interpreted by the balance that each variable contributes to for the final value of the HQI.

Figure 15: Impact of "Number of Stories" on the HQI over time

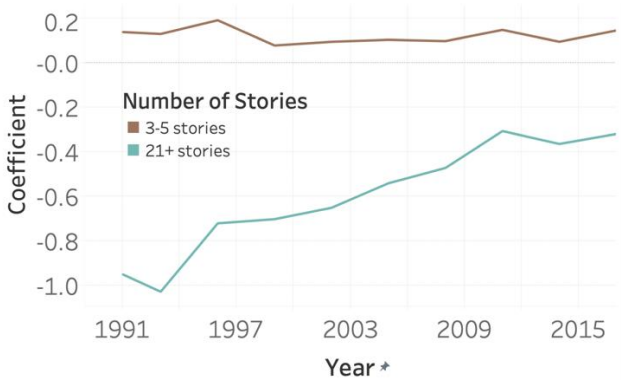
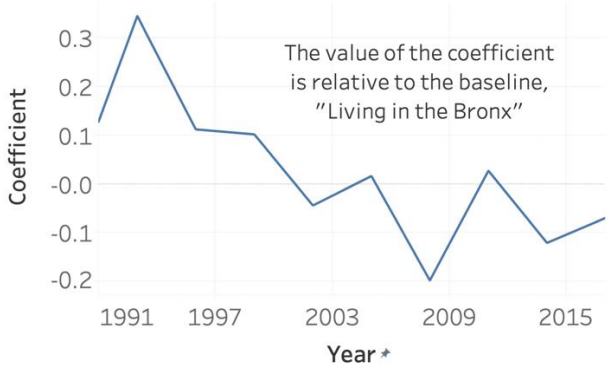


Figure 15 helps us see how the number of stories impact the HQI in different ways over time. For buildings with a number of stories between 3 and 5, we see that it has the same impact on the HQI over the years. When we compare that trend to that of buildings with more than 21 stories, we note that there is quite a big difference. Indeed, dwellings in such buildings encounter more issues over time as the coefficient has been steadily increasing since 1994. It is worth noting that its increase appears to slow down. In any case, the conditions in buildings with more than 21 stories is deteriorating.

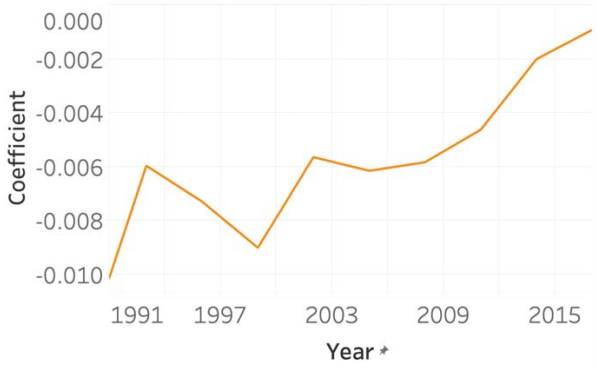
Figure 16: Impact of "Living in Manhattan" on the HQI over time



It is interesting to see in Figure 16 that for previous years, living in Manhattan meant more issues with your dwelling compared to the Bronx but it became less true as time passed. Indeed, the value of the coefficient was around 0.3 and then steadily decreased. We notice a sudden decrease in 2008 for that coefficient, which is at the same time of the economic crisis that affected the world. Since the coefficient got smaller quite fast and eventually decreased lower than 0, that shows that people living in the Bronx were more affected by that economic crisis than the people living in Manhattan. Indeed, the populations in the Bronx are poorer and the economic crisis accentuated those financial issues. We can point out that, from 2008 onward, the coefficient stayed very low and even negative, which

shows that the Bronx has never fully recovered from the 2008 crisis: from that point on, it is worse to live in the Bronx than in Manhattan in terms of housing conditions.

Figure 17: Impact of "Age of Householder" on the HQI over time



Finally, in Figure 17, we see the changes in the coefficients for the variable age over the years. We can see that after successive increases and decreases, since 2005, the coefficient keeps increasing. That can be interpreted by the fact that, as a householder gets older, the HQI will increase, and that average increase in the index gets larger each year. It shows that older people are more and more likely to experience issues with their dwellings compared than before. A reason for that could be that, as the residents get older, so does their dwelling, which makes them more prone to issues. Indeed, we saw in the correlation matrix in a previous section that the correlation factor between the age of a householder and the duration of their stay in a dwelling is equal to 0.6, meaning that as people get older, they tend to stay in the same place. Therefore, the condition of old buildings in New York City should be examined by New York City in order to help the elders live in better conditions. It is worth noting that the coefficient is really low so that means that the impact is quite small, but it is still significant.

4. Conclusion

Over the course of this analysis, we focused on several research questions to have a better understanding of the housing market in New York City. Using our HQI as a measure of housing quality for owners and renters, we saw that owners had significantly fewer issues than renters. To determine other factors driving this difference, we fitted a linear regression and found out that several variables we assumed had an impact on the index such as the income were not as linearly related to the index as others. Some variables like the borough or the length of the lease have a strong impact on our index. However, it is worth noting that our model could only explain about 28% of the variance in the index, meaning there may be other omitted factors.

We then investigated if the impact of certain coefficients on the index differed depending on ownership status. We found out that it was much more beneficial to be a renter in Manhattan than an owner, but much more detrimental to be a renter in a building with 6-10 stories than an owner, regarding the HQI. Those differences may need their own study to understand why certain types of buildings are more beneficial for owners and vice-versa. Finally, we focused on the changes in coefficients over the years. We saw that a lot of predictors of the HQI experienced significant changes over the years. For example, it was interesting to see that during the 2008 crisis, living in Manhattan suddenly became related

to lower HQI compared to living in the Bronx, though it was not true for other years. We also noticed that living in buildings with 11-20 stories has rapidly become a source of numerous housing issues, as well as the age of residents. Observing, and potentially predicting those trends, could help curb the impact of those detrimental factors.

Acknowledgements

We acknowledge the help of the UVA Statistics Department, specifically Tianxi Li and Jordan Rodu, for their help and guidance. We want to thank the American Statistical Association for organizing the Data Challenge Expo. Finally, we would like to thank the city of New York City for providing the data used for this project.

References

1. Krieger, J., and Higgins, D. L. (2002), "Housing and Health: Time Again for Public Health Action," American Journal of Public Health, 92, 758–768. https://doi.org/10.2105/AJPH.92.5.758.

2. Evans, J., Hyndman, S., Stewart-Brown, S., Smith, D., and Petersen, S. (2000), "An epidemiological study of the relative importance of damp housing in relation to adult health," Journal of Epidemiology & Community Health, 54, 677–686. https://doi.org/10.1136/jech.54.9.677.

3. Billings, C. G., and Howard, P. (1998), "Damp housing and asthma.," Monaldi archives for chest disease = Archivio Monaldi per le malattie del torace, 53, 43–49.

4. Hopton, J. L., and Hunt, S. M. (1996), "Housing conditions and mental health in a disadvantaged area in Scotland.," Journal of Epidemiology & Community Health, 50, 56–61. https://doi.org/10.1136/jech.50.1.56.

5. Dunn, J. R., and Hayes, M. V. (2000), "Social inequality, population health, and housing: a study of two Vancouver neighborhoods," Social Science & Medicine, 51, 563–587. https://doi.org/10.1016/S0277-9536(99)00496-7.

6. "Poverty Measure - NYC Opportunity" (n.d.). Available at https://www1.nyc.gov/site/opportunity/poverty-in-nyc/poverty-measure.page.

7. "Montana through Ohio" (n.d.). FBI, Available at https://ucr.fbi.gov/crime-in-the-u.s/2016/preliminary-semiannual-uniform-crime-report-januaryjune-2016/tables/table-4/state-cuts/table_4_january_to_june_2016_offenses_reported_to_law_enforcement_by_state_montana_through_ohio.xls.

8. "Study finds homelessness at record highs in NYC" (2019), *FOX 5 New York*, Text.Article, , Available at https://www.fox5ny.com/news/study-finds-homelessness-at-record-highs-in-nyc.

9. Wilson, J. Q., and Kelling, G. L. (1982), "Broken windows," Atlantic monthly, 249, 29–38.

10. “NYCHVS in the ASA Data Challenge Expo” (n.d.). Available at https://www1.nyc.gov/site/hpd/about/nychvs-asa-data-challenge-expo.page?#.