Proposed Index of Housing Quality for New York City in Selected Years

Vincient Arnold, Paul Franklin

February 2019

Cavalytics Club, Spring Data Challenge

University of Virginia

1 Research Question

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.

2 Data Summary

2.1 Data Description

From the website of the ASA: "To facilitate ASA entries, the NYC Department of Housing Preservation and Development has provided microdata files from interviews with occupied households from the 1991 through 2017 NYCHVS cycles. These files have been prepared to provide consistent variable names over time. As such, they differ slightly from the files available at census.gov. ASA data challenge submissions may use either these data files or any of the materials available on the NYCHVS homepage on census.gov." Specific definition of each variable can be found in the accompanying PDF codebooks on the ASA website.

2.2 Advantages and Disadvantages

The primary advantage of this data is its robustness. This data provides a highly detailed picture of housing in New York City, including variables as typical as average monthly rent to as scrupulously detailed as the number of cockroaches observed in the building during a visit. The evident disadvantages largely concern the data structuring. Specifically, the data for variables like income or rent, while expectantly quantitative and continuous, were given in the data class of levels, instead of integer or float values. This created significant issues in the data cleaning phase of the work flow, as well as presented further difficulties with NA values, since the data entry analysts chose to signify NA values with a set of numbers, like 99999. Further, there was a notable change in column names and level indicators in the 2017 file compared to all previous files. This created special difficulty in the scoring index, as a new function and scoring system had to be created specifically for 2017. These discrepancies were unexpected, and thus fixing the bugs presented involved a significant amount of manual parsing through the files, which was inefficient.

3 Methodology

The methodology of this project was the source of a great amount of deliberation and debate. Ultimately, we set out to be as consistent, honest, and accurate as possible in creating an index for housing quality. The primary difficulty we had to overcome was the ever-present conundrum of statistics: correlation versus causation. Initially, we sought to include variables like income, internet availability, monthly gross rent and other variables of their

ilk in our scoring system. However, we soon realized that while a house being expensive and having full access to WiFi is almost *correlated* with high quality, it is the not the *cause* of high quality. We arrived at the conclusion that only variables directly related to quality of a house should be included in our scoring system.
A full list of these variables can be found in the Python code but to give a few examples, we included things like presence of cracks in outer walls, chipping or pealing paint in interior walls, presence of leaking water inside, missing or cracked windows, etc. The scores were calculated by starting with a perfect score of ten on a scale of 0-10, and then subtracting a certain number of points for every instance that a negative occurrence appeared. More technically, we used a number of if-statements, simple Boolean filters to verify whether or not a particular condition was met, and then used the 'then' portion of the 'if-then' statement to subtract a given number of points or fraction of a point for each 'True' result. There was weighting involved, though we tried to keep it to a minimum. Variables that were weighted were particularly detracting of quality, such as being listed as 'dilapidated' or having no plumbing. Below we show the scoring regime. If each item is perfect, the housing quality score will be a 10. In the below table we show the amount of points taken off for an issue in any category.

Demerit	Possible Point Deduction
missing materials	0.5
major cracks in outside wall	0.5
loose or hanging cornice or roofing	0.25
broken or missing windows	0.25
rotten or loose window frames	0.25
boarded up windows	0.25
missing stair railings	0.25
loose, broken, or missing steps	0.25
sagging or sloping floors	0.25
depressions in floors	0.25
holes or missing flooring	0.5
dilapidated condition	1.5
deteriorating condition	0.75
few plumbing facilities	0.25
no plumbing facilities	0.75
toilet breakdowns	0.25
no full kitchen facilities	0.25
no kitchen facilities	0.75
dysfunctional kitchen facilities	0.75
between 1 and 2 heating breakdowns	0.25
between 3 and 4 heating breakdowns	0.75
presence of mice or rats	0.5
cracks or holes in interior walls/ceiling	0.5
holes in floors	0.25
broken plaster or peeling paint	0.25
leaking water inside	0.75

Table 1: Table of demerit point reductions.

It should also be noted that in the code, if statements were used so that some of these demerits will not occur simultaneously. For example, a building

is either listed as dilapidated or deteriorating, but not both. Thus, 'double jeopardy' does not occur. (Thus a simple summation of the above demerit points will result in a score greater than 10, but this does not occur in the program due to the boolean filters).

4 Exploratory Data Analysis

Though the code is given in a separate section, here we include graphics of histograms created for each year. These will illustrate the (minimal) changes in housing quality over the years in New York City. It should be noted that this overview does not group by borough or sub-borough unit, as this would involve far too many graphs to include in any deliverable report. However, in the later data visualization section, a more meaningful detailed graphic of housing quality scores by sub-borough unit will be given in the form of a colored map, which will be far easier to read and interpret. The purpose of the following graphics is to illustrate the distribution of housing quality per year and how these distributions have changed throughout the years.

Figure 1: Year = 1991

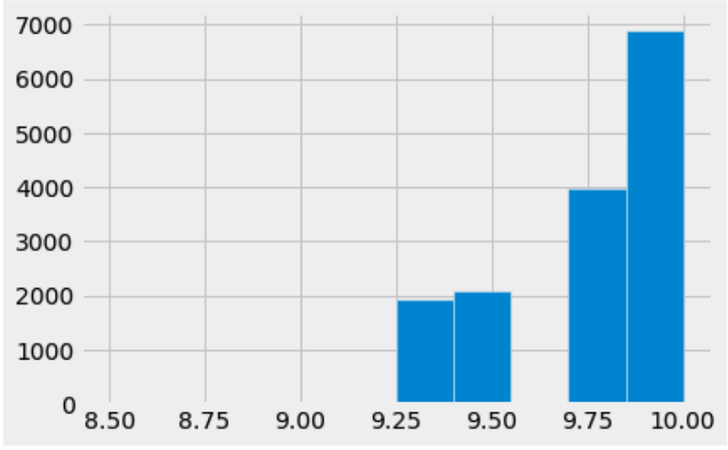


Figure 2: Year = 1993

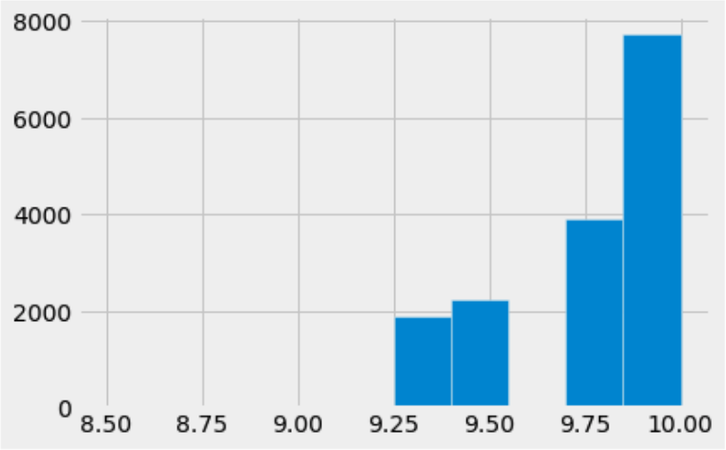


Figure 3: Year = 1996

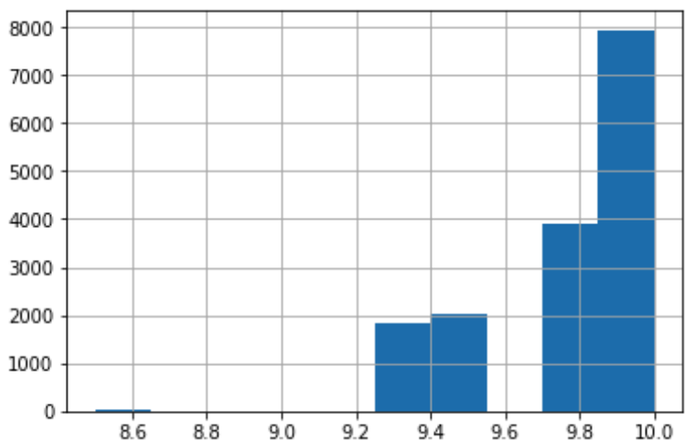


Figure 4: Year = 1999

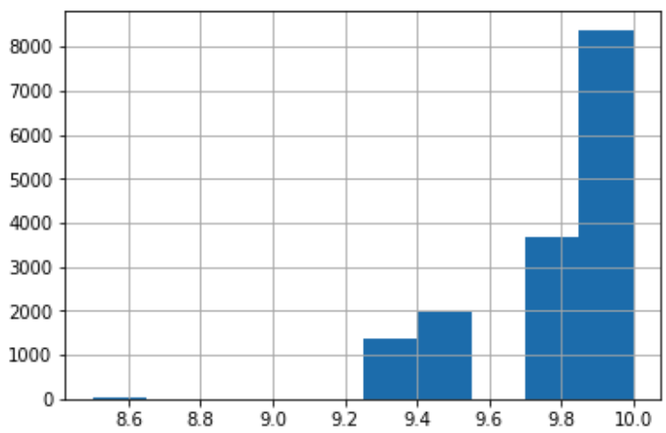


Figure 5: Year = 2002

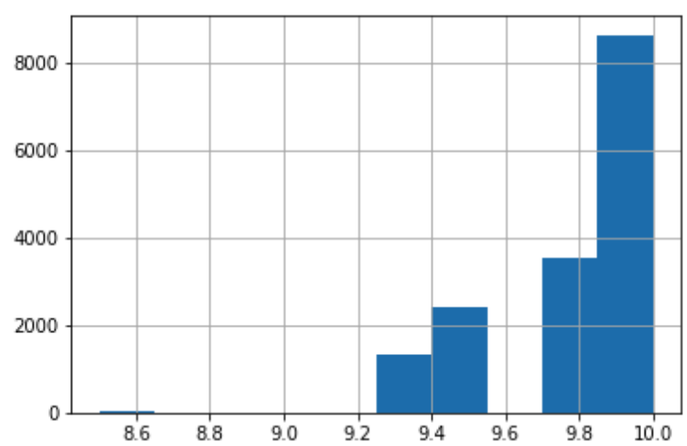


Figure 6: Year = 2005

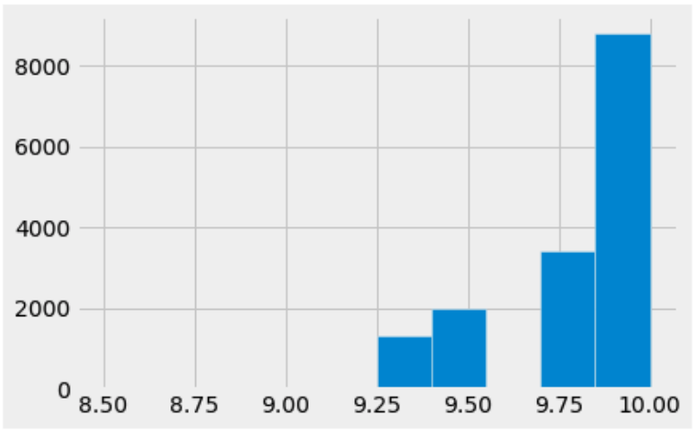


Figure 7: Year = 2008

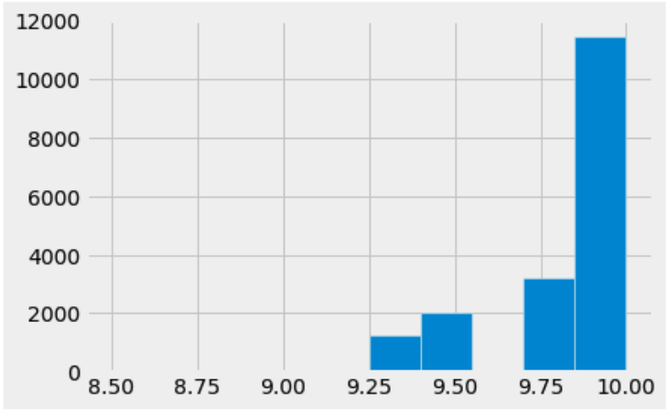


Figure 8: Year = 2011

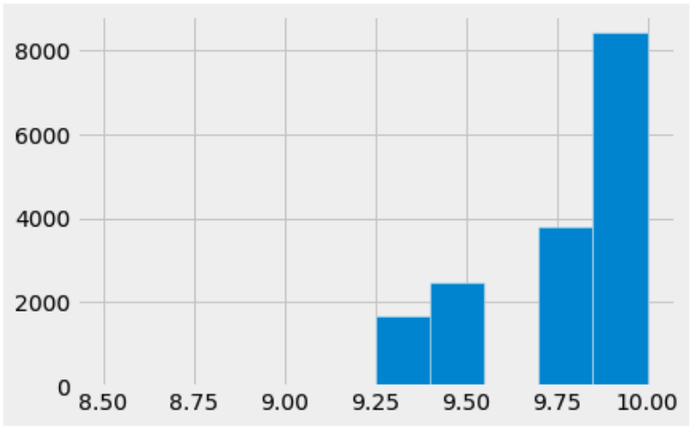


Figure 9: Year = 2014

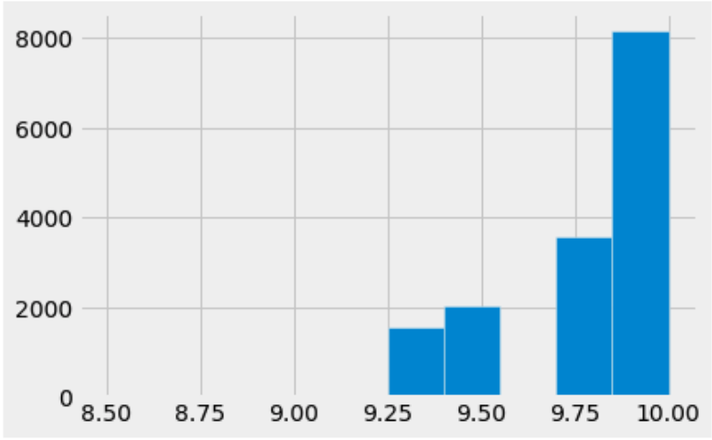
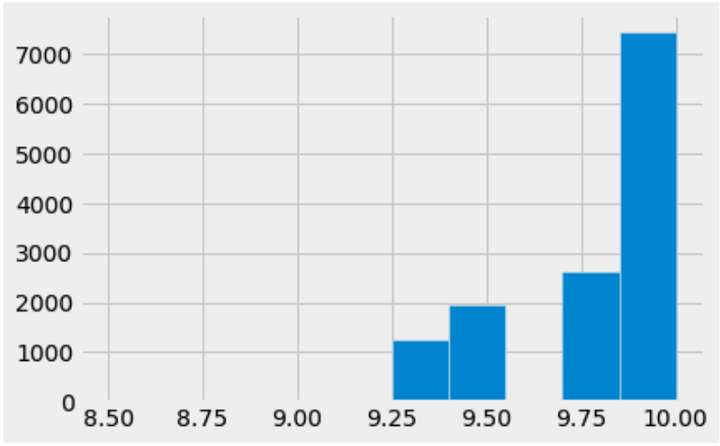


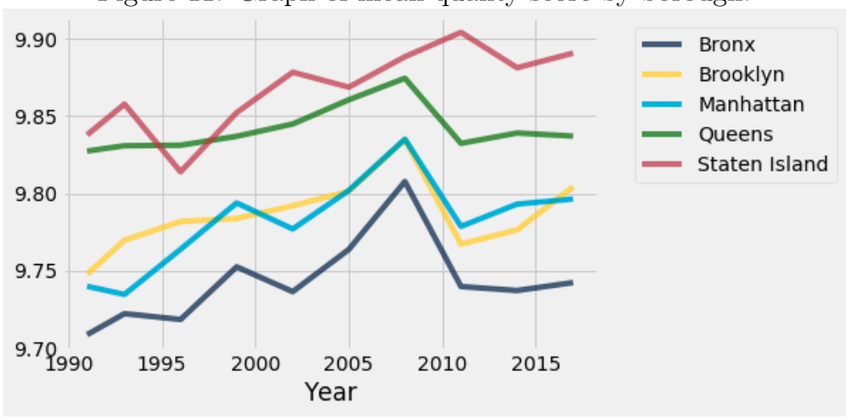
Figure 10: Year = 2017



5 Conclusion

Notably, the distributions of the housing quality scores are relatively consistent across years. Also perhaps the most important takeaway here is that most buildings in New York City between 1991 and 2017 are of remarkably high quality. Also 2008 seems to represent an outlier year. In this case, it appears based on the histogram that there are actually more observations. However, it is likely that this is simply the result of a number of low scores that aren't large enough to show up on the y-axis for count. Finally, as we show below, the average housing quality index score changes for each year and is notably different for each borough.

Figure 11: Graph of mean quality score by borough.



Some key takeaways from the results as illustrated in the above graphic are as follows. First, it's noteworthy, if not expected, that the quality of housing in NYC shows a general upward trend over time. The city should be proud of these efforts. However, secondly, the average quality score fluctuates over time, and moreover seems to be affected by the economic state of the nation (viz the dip around and following the 2008 housing crisis). Further, of course we can discern some more at-risk boroughs of New York. Specifically, the Bronx is consistently at the lower end of the spectrum while Staten Island seems always to have very high average scores. Finally, housing inequality seems to have gotten worse. Take 1991. The difference between Staten Island and Queens was quite small. However, in 2017, despite marked improvement in both boroughs compared to 1991, the gap between Staten Island's and Queens' average quality score has widened dramatically. Similar divergence is noted between the Bronx and Manhattan.
Finally we have below a table of the most common demerits. This table illustrates the most common housing issues in NYC for the given time period.

Figure 12: Percentages of demerits by year.

% of Houses with Condition in 1991	% of Houses with Condition in 1993	% of Houses with Condition in 1996	% of Houses with Condition in 1999	% of Houses with Condition in 2002	% of Houses with Condition in 2005	% of Houses with Condition in 2008	% of Houses with Condition in 2011	% of Houses with Condition in 2014	Condition name
0.892677	0.924051	1.04749	0.985925	0.314584	0.7011	0.523531	0.324	0.4693	lacking plumbing facilities
11.6652	9.55696	9.40198	9.26899	8.14773	9.01138	8.15372	8.93141	9.86834	toilet breakdowns
0.765152	0.778481	0.939563	0.758903	0.396376	0.791149	0.456697	0.501284	0.606179	lacking full kitchen facilities
3.41634	2.67722	2.86313	2.36103	2.25242	1.6016	1.19744	2.0357	2.08578	kitchen facilities not functional
18.1287	15.4241	14.3791	11.3446	10.7588	11.9123	7.70259	11.9697	12.7363	Heating breakdowns
23.8808	22.7342	20.9307	18.3045	20.5801	18.5309	14.8705	20.1675	17.162	Mice and rats
15.4843	13.6709	12.9063	11.0657	11.1866	10.465	7.68588	11.9575	11.198	cracks or holes in interior walls/ceiling
7.20183	7.02532	6.29126	5.50042	5.46747	5.16498	3.98775	5.34906	4.77773	Holes in floors
18.3569	17.0506	16.2456	13.7446	13.546	13.218	8.10359	14.1154	13.551	broken plaster or peeling paint
21.2699	19.3165	19.3309	16.3132	15.811	15.1926	11.3283	17.9851	15.715	leaking water inside
0.530237	0.240506	0.406298	0.382694	0.257959	0.463112	0.21721	0.158944	0.215096	missing material
1.28197	0.64557	0.958608	0.577285	0.67321	0.61105	0.662768	0.51351	0.580107	major cracks exterior
0.946372	0.677215	1.00305	0.836739	0.78646	0.791149	0.729602	0.605208	0.801721	roofing
2.8324	1.77215	2.03149	1.54375	1.3653	0.913359	0.75188	1.10038	0.795203	windows missing
2.22834	2.22152	2.29177	1.73834	1.00038	1.16421	1.06934	1.08815	0.997262	window frames
0.738305	0.626582	0.641188	0.609717	0.55996	0.45668	0.467836	0.782492	0.619215	boarded windows*, *stair railings
1.88603	1.91772	2.13941	1.50483	1.23317	1.24783	1.09719	1.3877	0.827793	stairs missing
4.28217	3.56329	4.30422	3.85938	3.88197	3.1646	2.40602	3.69238	3.08956	floors sagging
1.23498	1.10127	0.774505	0.875657	0.446709	0.861903	0.445558	0.580756	0.391083	depressions in floors
2.12766	1.61392	1.95531	1.75131	1.5855	1.74953	1.54832	1.65057	1.20584	depressions in floors OR slated or shifted door frames

Notably, mice and rats remained alarmingly high, as did cracks/holes in walls. Overall, most demerits were highly uncommon and, with the exception

of the mice and rats, the ones that were common were not very serious. Another notable exception however, was leaking water and heating breakdowns. Common in low-income neighborhoods, these types of issues represent a large obstacle to high quality housing, yet are relatively easy to fix. Our suggestion for the city is to isolate the at-risk neighborhoods and then focus in on the most common, and generally also the easiest to fix, issues.