Karan Bhargava

Travis

Lance

Rachel

Alex

Gil

**Midterm Write-Up**

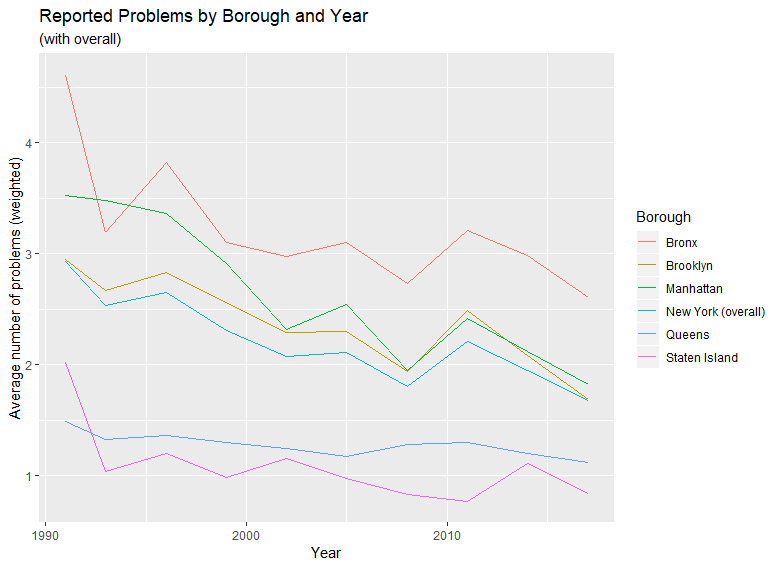
Research Question:

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 question. Our group looked into the research topic focused on the quality of housing improvement over the years from 1991 to 2017.

In order to address this research problem, we took the 5 datasets and concatenated them into a single data frame. We then combined similar variables to reduce redundancy, to compress data into more manageable sets of independent variables, and to make the data consistent across the data sets. 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 2017, they consolidated X\_d1 and X\_d2 into a new variable-X\_d12- which combines the issues listed in \_d1 and \_d2 into a single variable. 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. Among other things, we also renamed variables or inputs to have more meaningful information for the reader to understand instead of integer classifications. Once the data frames are refactored, cleaned, etc, we joined them together to make everything cohesive for the data manipulation and graphing. We removed the unnecessary columns and outputted the new data frame into a CSV file entitled, “housingAggregation.csv”.

We created two different graphs from the data after wrangling it. The first graph we made plotted the average reported problems about the building conditions over time. The graph displays the average problems reported broken down by borough and shows an overall descending number of problems being reported.



This seems to hold 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 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 in the building in the entire city of New York. The index was weighted based on weighting information provided by the similarly run American Housing Survey. We weighted external structural issues higher than we weighted other miscellaneous issues (presence of pests, broken handrails, etc). In doing so, we were able to take the average reported issues to incorporate the severity of a problem for each borough and make a more descriptive graph of the data verifying a distinct disparity within the boroughs and the quality of the housing in each. (I need info on what line 391 is doing because I’m not following it). I also don’t have much info on an index itself since I just see how the weighting of the problems effected the “avg problems in building”.

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. You can see below what the unweighted graphs look like (need Gil to send an image of his unweighted graph-ideally without the smoothing lines).

In conclusion, the data we were able to represent from the dataframes seem to agree with the premise that, while housing in NYC has improved over the years, there is still a disparity from one borough to the next in the quality of the buildings as exemplified by the average reported problems in each of the boroughs. An interesting consideration we are looking into is the relationship between GDP and the problem reporting over the years. We also want to see how significant events that hit NYC in this time period might have effected the reporting of problems (ex: Hurrican Sandy, the recession, etc). If we can graph and find some patterns in this, it might shed light on which borough tends to be more negatively effected by serious events. 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. Conclusion is incomplete because I need to see how Gil’s data and graph compares with Travis’s findings, and I also need to understand Travis’s index a bit more.