MTH390Q Data Science  
Larry Breeden  
Elizabeth Foster

Gemma Hoeppner

Kelsey Smith

Nathanael Whitney

For Dr. Nels Grevstad

Midterm Project

Introduction

The focus of our project is to understand the housing conditions faced by residents of New York City. As a necessary first step, we created a housing quality index from the NYCHVS data set. This will help us examine potential quality of life issues immigrants might face in New York.

Methods

Our first task was to create a housing quality index, which involved a lot of data wrangling. We selected variables that related to the physical condition of housing units, being careful to choose only values present in every data set from 1991 – 2017. In total we selected 20 variables for our index, which we used to create a subset of these categories for each year the housing census was conducted. In preparation for compiling all years into one data set, we used mutate() to add a column for the year.

After creating a new subset of variables for each year, all the subsets were combined into one data frame using the rbind() function. Subsequently, we used the names() function on the data frame to change the names of the columns to be more succinct and useful.

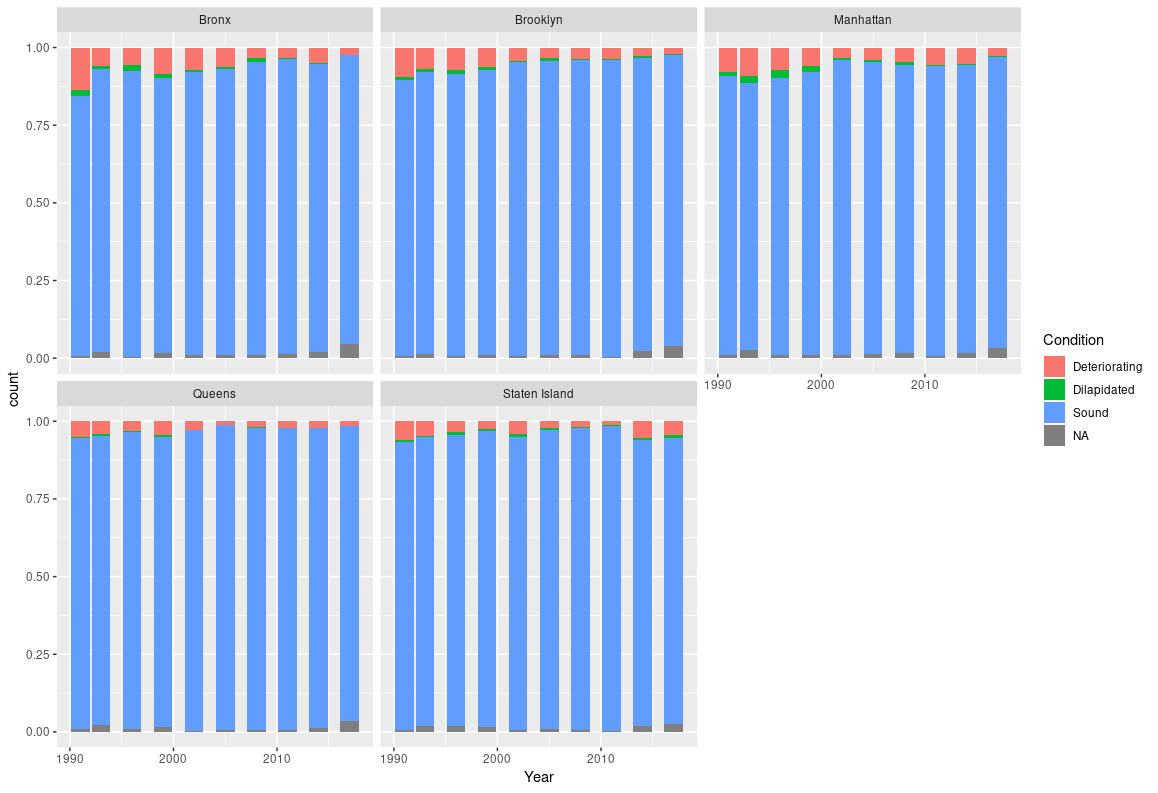
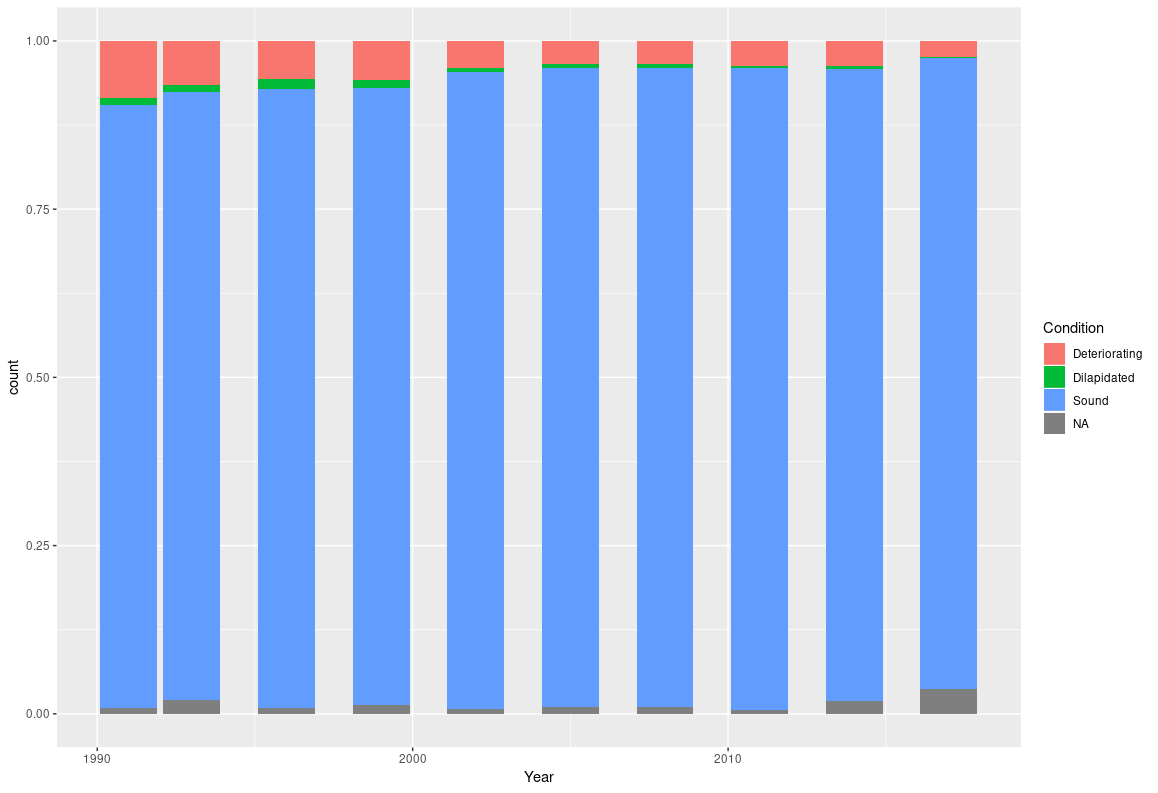
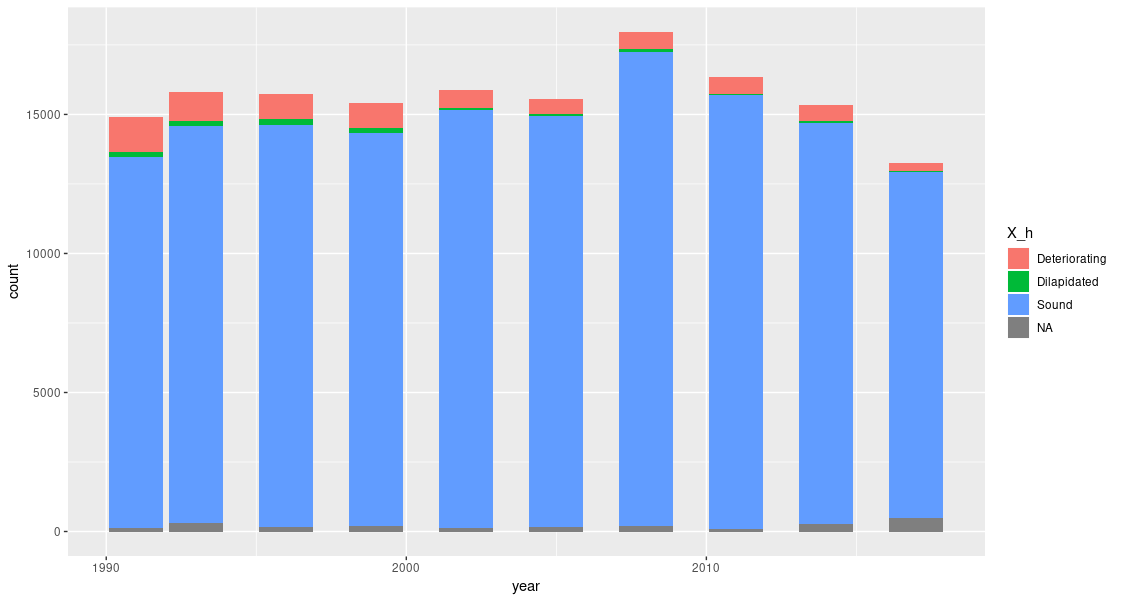
Our data was tidied further by changing the values representing the five boroughs to respective borough names and using NAs to replace relevant numerical values representing non-reported data.

To create a weight for each item, we subjectively determined which data points were more important to the overall condition of the housing unit using a scale from one to ten. While this might seem odd, we were following the original methodology of the NYCHVS. For example, we rated “No Plumbing Facilities” as more severe than “Broken Plaster or Peeling Paint on Interior Walls.” We then changed the original values in the data sets to our assigned values. This was done in R by selecting the columns (via $ and single brackets) we needed to modify and using the assign function to insert the new value. An example is (for column 2):  
nych\_all$Walls\_cracks[nych\_all$Walls\_cracks == 1] <- 8

Results

The minimum index value a housing unit could obtain was 17, and the maximum value was 122. The average index value (mean(nych\_all\_cleaned$index\_value))) was 25.4, implying that most housing in New York is fairly sound. Without a deeper dive, this does mimic the *Condition of Building* variable across the provided data sets (“Dilapidated”, “Sound, or “Deteriorating”).

Below is a plot of the proportions for each year that were “Deteriorating”, “Dilapidated”, or “Sound”. We feel that it was less useful than we were hoping, given the large proportion of buildings rated “Sound.” However, this aligns with our index value distribution, and improvements from the early 90s in the proportion of deteriorating housing can be seen.



Graphical displays

|  |  |
| --- | --- |
|  | These five plots display by borough a plotted point per house in the dataset, arranged by year and index value.  ggplot(nych\_all) + geom\_point(mapping = aes(x = Year, y = index\_value, color = index\_value)) + facet\_wrap(facets = ~Borough) + ylab("Index Value") |
|  | Plot of mean index value over time, colored by borough  ggplot(nych\_all) + geom\_smooth(mapping = aes(x = Year, y = index\_value, color = Borough)) + ylim(0, 50) |
|  | An overlay of the smooth line on top of the dotplot of all houses in the dataset, colored by borough  ggplot(nych\_all, mapping = aes(x = Year, y = index\_value, color = Borough)) + geom\_point() + geom\_smooth() |

Note: The above graphs include index values with NA values included, which may skew our comparison.

Conclusions

We ran into challenges because of inconsistencies in the data set, the large size of the data set, and our attempts to account for such. For example, one of the challenges was what to do with any housing units that contained NAs in one or more of the variables. In attempting to do analyses using the index we devised, we realized that any units containing NAs would skew our index by virtue of having a lower number by default, regardless of whether the unit was affected by the housing quality variable in question. That meant that a unit with no plumbing should receive our highest point value possible (10). However, if that variable happened to be not reported, then its index value could be misleading.

Due to this, we should have either looked at only rows that were complete (which would at least ensure accurate comparison, despite removing a portion of our data set), or developed a different method that takes into account non-reported data (such as normalizing our index, perhaps by dividing by the number of variables.)

Additionally, while we chose the twenty variables for simplicity (since they were consistent across all the data and spoke to the condition of the housing), we are interested in expanding the variables addressed and making other refinements. An added challenge would be determining how to factor variables in that aren’t consistent across all years without skewing the overall index value.

The above is still probably scratching the surface. We sampled all the relevant data from all eight spreadsheets. This may have ended up skewing our distribution towards one or more years, since the spreadsheets do not have an equal number of rows (nor was presence of unreported values uniform across the spreadsheets.) It may have been better to take the same size sample from each spreadsheet, e.g. 2000 rows each. We also did not do any histograms, which would not have been as susceptible to skew from unreported values and might have helped us better determine what the shape of the distribution was, where important features were, and where the outliers were. It is also unknown how values in our index correlate. It could be that some values in our index are tightly correlated with each other and are thus redundant because they are measuring the same thing, despite appearing to be unrelated. We do not know enough about the data yet. Despite room for growth, we think we have done a great start to the project. We have a clear direction to move forward on, and our hope is to make improvements as we do so.