**Jacob Hefele**

**Professor Villa**

**MSA680**

**12/18/14**

**Data Exploration Report**

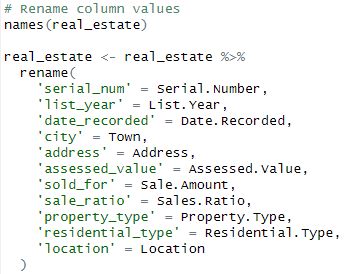
For my data exploration project, I decided to use a dataset provided by the US government via Data.gov. This dataset had information about real estate transactions in the state of Connecticut since 2000. It was massive, and difficult to work with, containing about 300,000 records and taking up 119 MB. For this reason, I did some initial data extraction in excel, taking only the most recent 100,000 records. This made the data easier to work with in R and PowerBI.

**Data Cleaning**

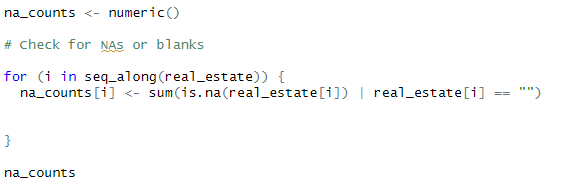
More cleaning needed to be done once the data was put into R. Immediately, the data was messy because it contained an “assessor’s remarks” section that took up a ton of space and contained many nulls. I decided that this wasn’t going to help in any exploratory analysis, though it was interesting to look through, so I got rid of the column from the dataset which I called real\_estate.



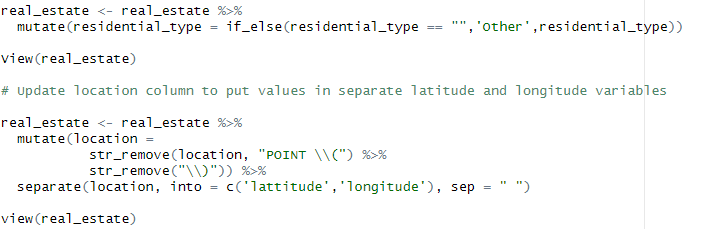
Secondly, I renamed the columns into names that were easier to work with and more closely matched the way I name variables in my programming projects.

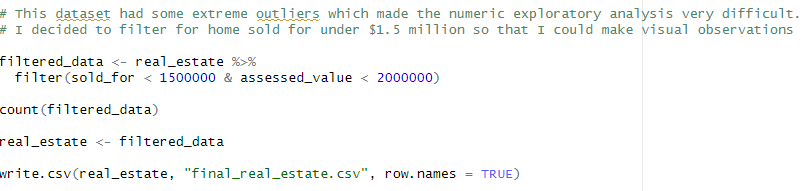


Even after removing the assessor’s notes, there were still NAs to check for.



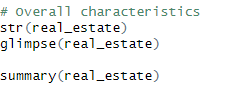
The output showed me that the residential\_type field had many NA values, so I decided to create a whole new type called ‘Other’ to account for these so that the exploratory analysis could go more smoothly.

The last part of the data cleaning process wasn’t added until I began doing my analysis. There were outliers that were significantly impacting my visualizations, indicating that some homes in Connecticut were being sold for prices approaching hundreds of millions of dollars. The majority of the houses were selling for under 1.5 million, so I decided to keep my analysis for this subsection of values. Because the dataset was so large, the number of extremely high selling prices was muddying any meaningful exploratory analysis.

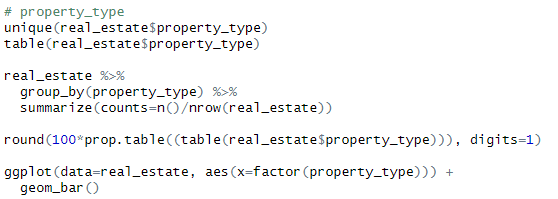
**Exploratory Analysis**

Much of the tests I ran didn’t give much insight into the data however there were a few that I found particularly interesting especially when comparing assed value home prices and actual selling prices.

My initial approach was to get some overall characteristics about the data.

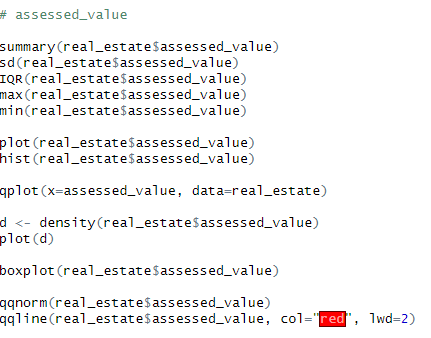


I then performed univariate analysis of the variables. For example, this is what I ran for the property\_type variable.



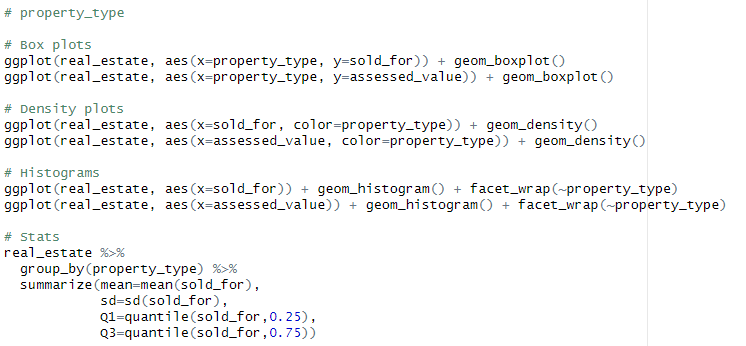
Much of the categorical data followed this structure in order to make insights.

My numeric data is where I spent much of my time and energy exploring.



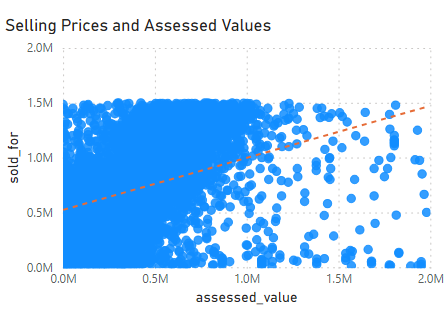
Most of my insight came from bivariate analysis. My aim was to see what cities had the highest selling prices and assed values, and to discover how the selling prices and assessed values interact with each other. This is more clearly shown, however, in the PowerBI visualizations.

Here is an example of the property\_type bivariate analysis with selling prices and assessed values.

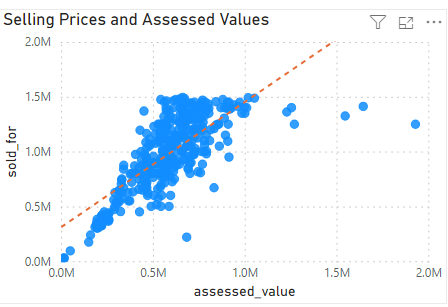
**PowerBI Visualization**

My first aim was to discover which cities had the highest average selling prices and assessed values.

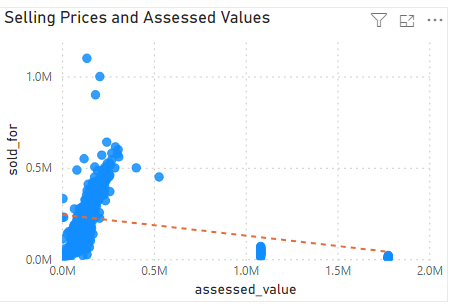
I then tried to understand the correlation between selling prices and assessed values.



Because of the size of the dataset, the visualization doesn’t effectively show the correlation between the two variables, though the trend line indicates that there is a correlation. I then decided to filter based on single a single city to see if a smaller sample size more adequately showed this correlation.



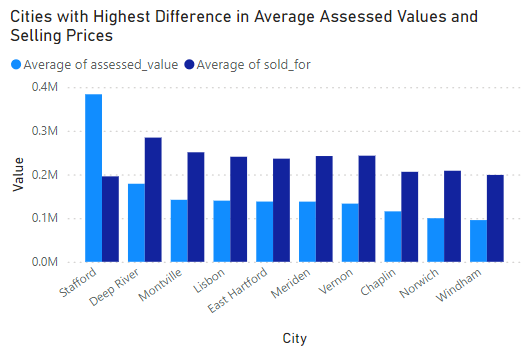
New Canaan was the first city I chose, and this allowed me to actually see the correlation between these two variables. This remained true for each city for the most part, however, I did find some cases where this wasn’t true—just as in the case of Stafford, CT.



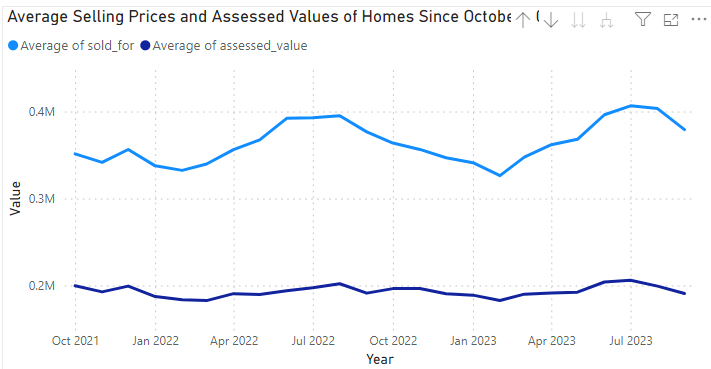
It appears that there are a few extreme values that are skewing the correlation in this instance. This is micro chasm of the much larger issue this dataset has with extreme house prices skewing data and making exploratory analysis difficult.

It make sense that selling price and assessed value are correlated—this isn’t much of a surprising discovery to make, but to actually see it demonstrated in data was really interesting, especially because this is a real-world dataset.

This did inspire me, however, to look into instances where cities had large disparities between assessed values and selling prices, so created another bar chart showing cities with the highest disparities.



Lastly, it was interesting to see how housing prices changed over time, and I found this relatively difficult to do in R. Luckily, PowerBI made this visualization easy to create and clearly demonstrates trends over the last 3 years.



Overall, this dataset, through unwieldy at times, was extremely fun and interesting to explore. As a Connecticut resident that may be looking to purchase a house in the next decade, these insights directly applied to my life and have allowed me to make my own educated decisions based on observations that I have made completely on my own.