Data Wrangling Report

This project was started with multiple datasets from Zillow regarding median rent and median home values by city. One of the first difficulties in the cleaning process was that cities were not consistently classified in the same way across all datasets. For example, in the median rental price dataset the state, metropolitan area, county name and city were all in separate columns. In the median home value dataset there was one column named “RegionName”, a RegionID column that had a series of numbers and no columns for state or county. This meant that in order to pull out the areas I wanted I had to make a separate object for each dataset containing the cities that I wanted to examine and then had to check the dataset afterwards to see if I had the results I wanted. This process involved some degree of trial and error but the data was small enough to do a visual inspection and correct the data as necessary.

There are multiple datasets and each set contained at least one action that was unique to that dataset but there are similarities in my approach to all of them. After filtering for the cities I then removed columns that were irrelevant. At this point the data was in a very “wide” format. From here the rows are the cities(variables) and most of the columns are dates(observations) which is the opposite of what a tidy structure is. From here I gathered all of the date columns into a single column called “date”, the value of the rent or home in another column, and the name of the city in the first column. This is the because the cities can be considered subjects and the observations for them all follow the exact same structure, even if some of the data is incomplete for some of the cities.

After this I convert the “date” column using the lubridate package and then from here the data is ready to graph and model. For the rent dataset due to the format of the titles it required a little more wrangling with changing the names of the columns to remove certain characters and substitute different ones to make the dates more friendly to the lubridate package. Also for the median rent dataset the dates all started with an “X” character. I have no idea why this is the case but it was something I fixed in the wrangling process. The data was also arranged in alphabetical order by city.