Module Assignment 8: Portfolio Assignment Option 1

Farhan Faseehuddin

MIS 470-1

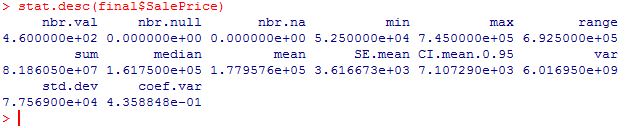
Colorado State University

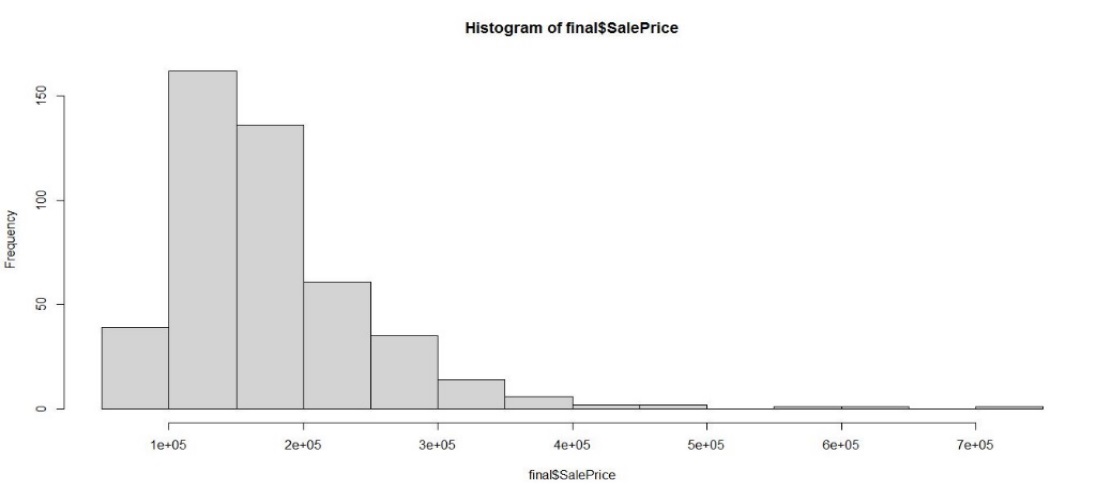
Professor Osama Morad

July 26, 2021

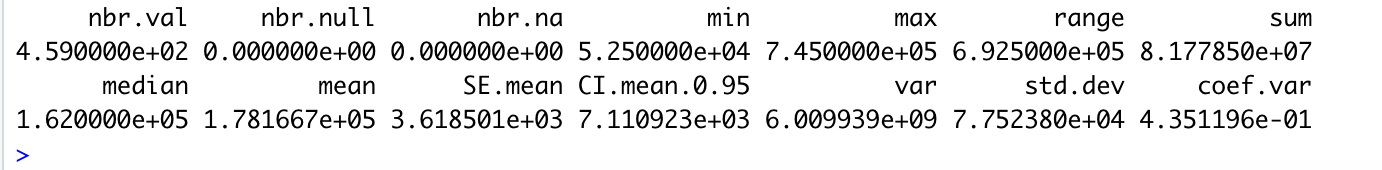
***Part I-Housing Testing Summary Statistics and Histogram***

***House Testing Summary Statistics of Sale Price***



***House Testing Histogram of Sale Price ***

***House Training Summary Statistics of Sale Price***



***House Training Histogram of Sale Price***

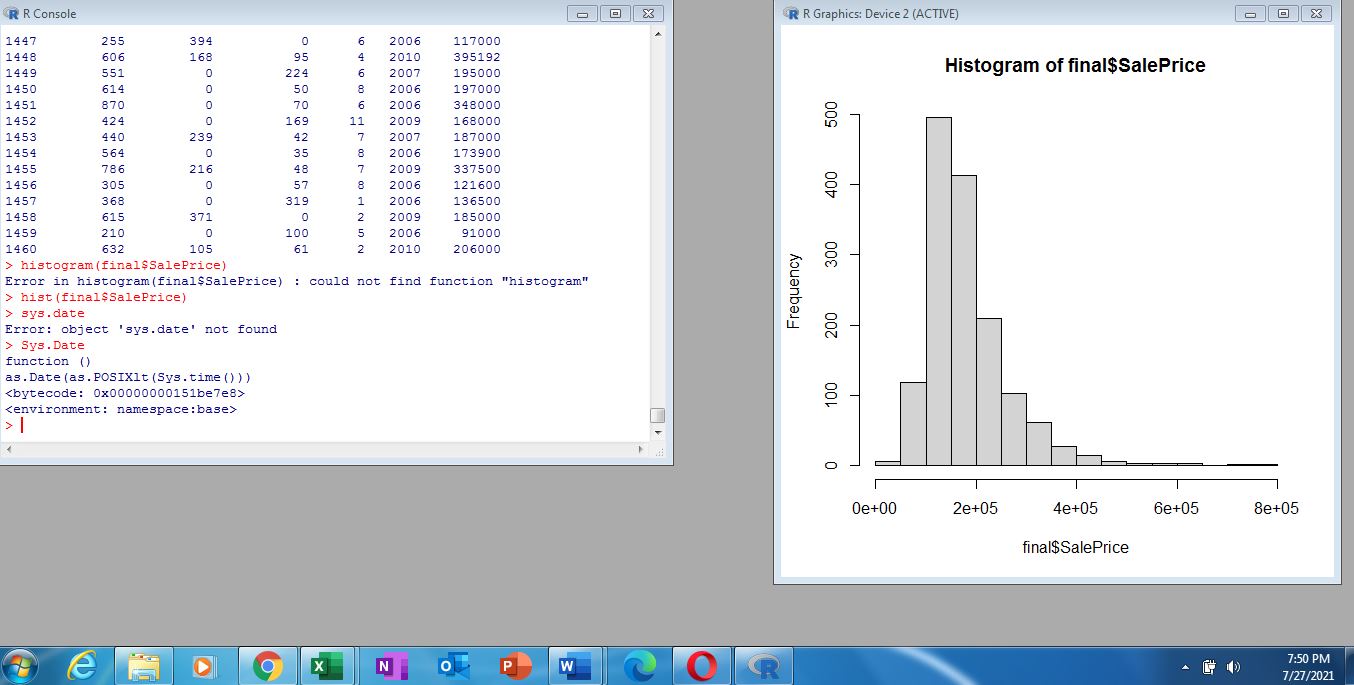
Chart, histogram

Description automatically generated

***Compare and Contrast the two datasets***

In the house-training dataset, the mean of the dataset is 1.781667E05, or 178,166.7 and the median of the dataset is 1.620000E05 or 162,000. In the house testing dataset, the mean of the dataset is 1.779576E05 or, 177,957.6 and the median of the dataset is 161,750.0. In the testing dataset, both the mean and median are lower. To further understand these two numbers, I wanted to find the min and maximum ranges. The minimum and maximum values of the house-training dataset is 52,500 and 745,000.00, respectively. While in the house testing dataset the minimum and maximum values are the same as the house-training dataset. Since, both the minimum and maximum values are the same in both datasets, the standard deviation becomes even more critical. Standard deviation measures the differences of every observation or datapoint from the mean (BMJ, 2021). The standard deviation in the house-training dataset is 7.752380E04, or 77,523.80. In the house-testing dataset, the standard deviation is, 7.756900E04 or 77,569.00. In both cases the datapoints are fairly spread out. Conversely, the difference of the two standard deviations in the datasets is quite minimal, or 45.2. Since the minimum and maximum ranges are the same, it may not be too surprising to see the small difference of the standard deviation between the two datasets.

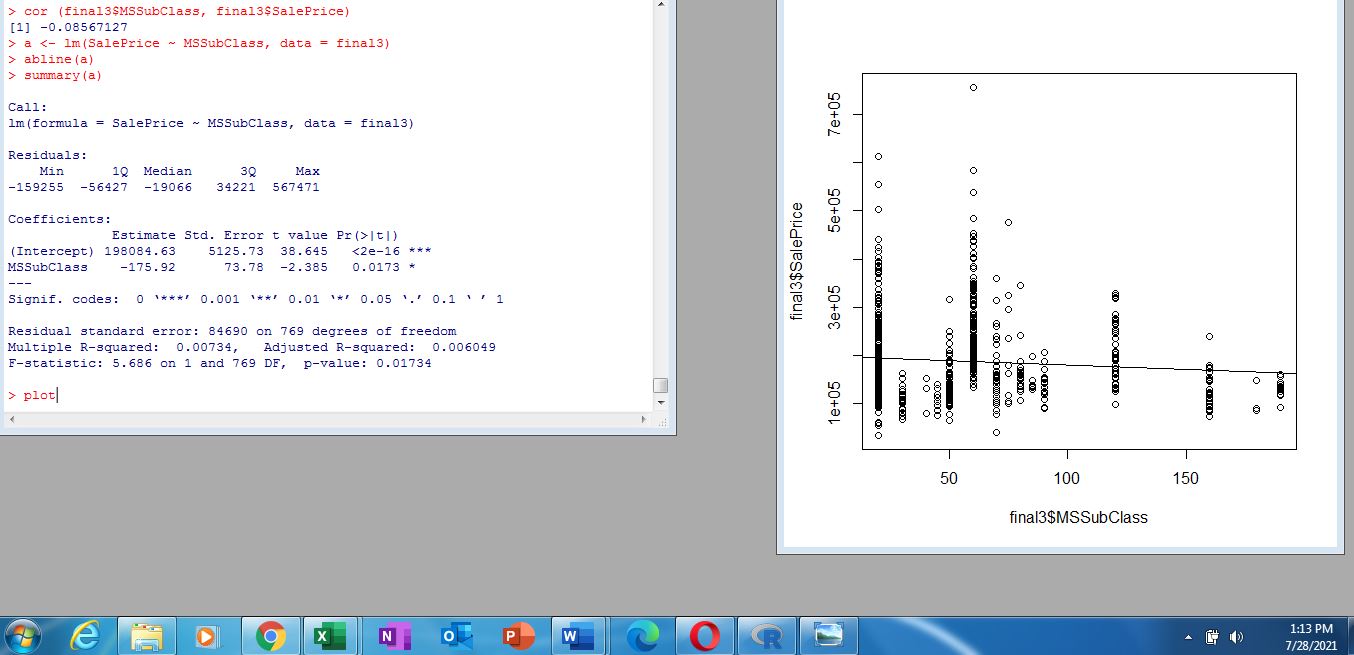
***Part II Combining Datasets***

******

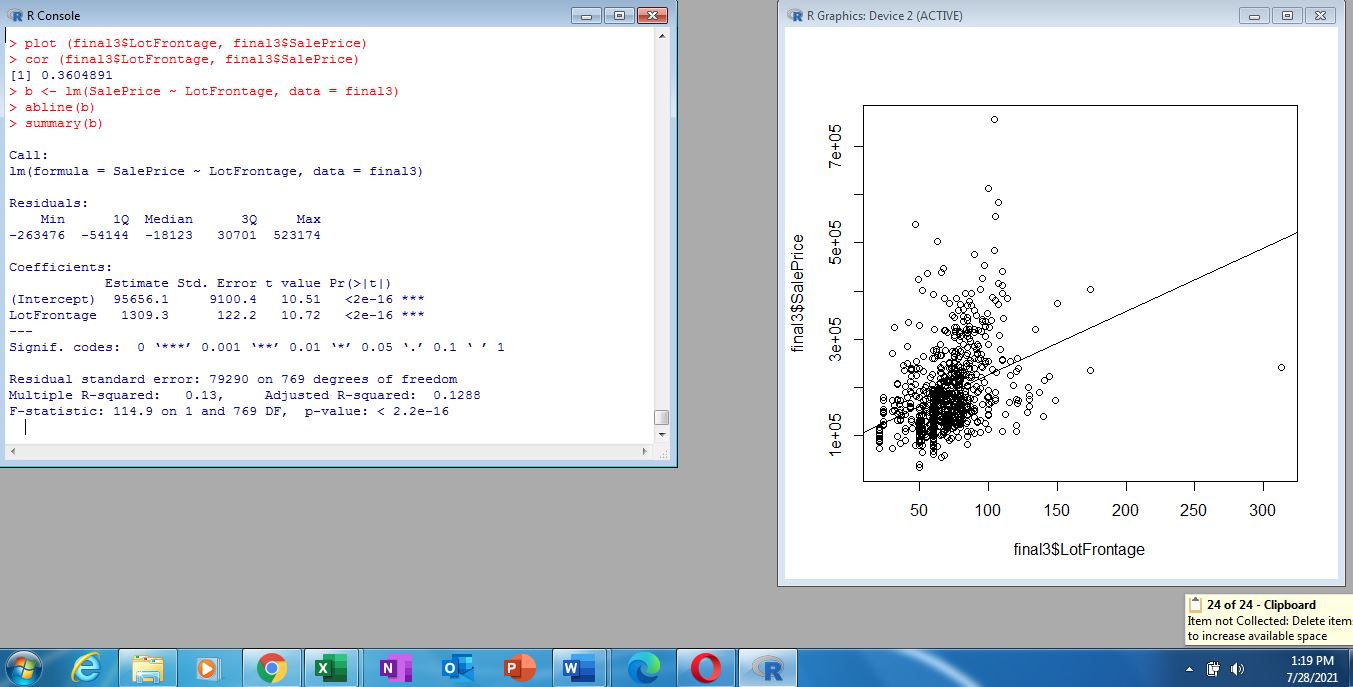
The above illustration is the histogram for the combined datasets, and in Part I, there is the histogram for sales price in the testing and training datasets. In all three cases, the histogram is skewed to the right. To understand why it is skewed the way it is, it is important to understand if it is positively or negatively skewed. Since the mean is larger than the median and the mode, this means that it is positively skewed, which we see in the histograms (Doane and Seward, 2011, 3).

***Part III***

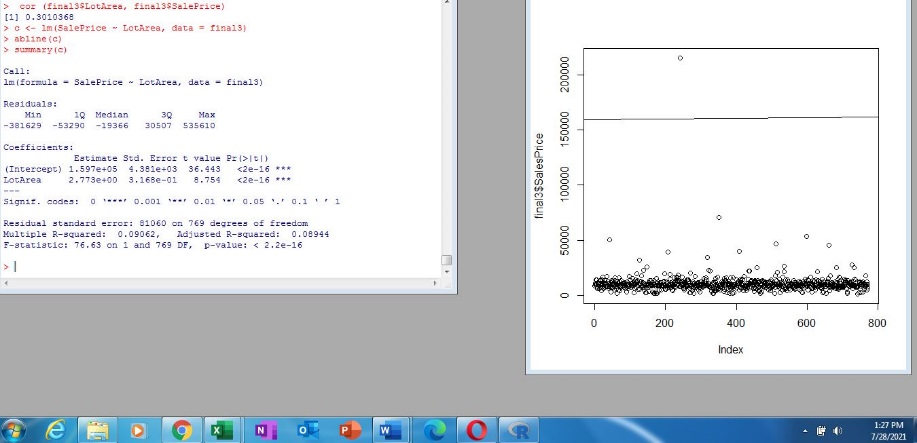
***Sales Price with MSSubClass Regression Model***

******

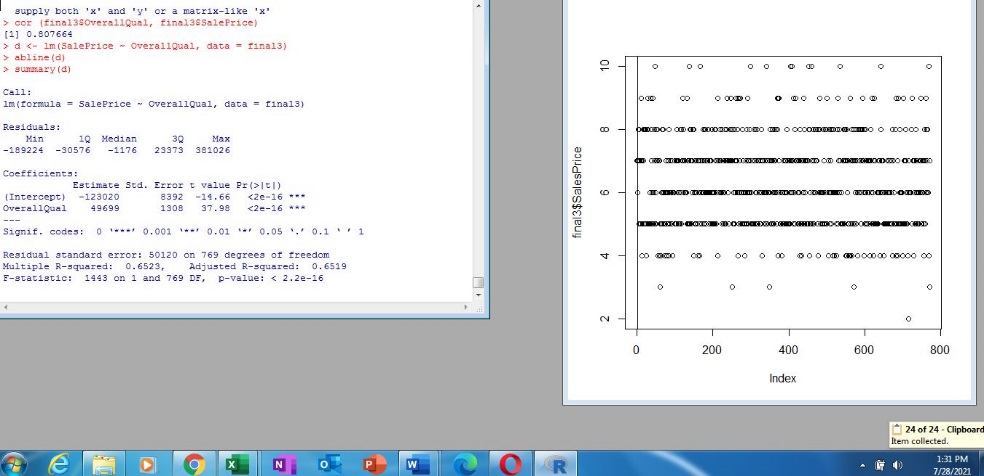
***Sales Price with Lot Frontage Model***

******

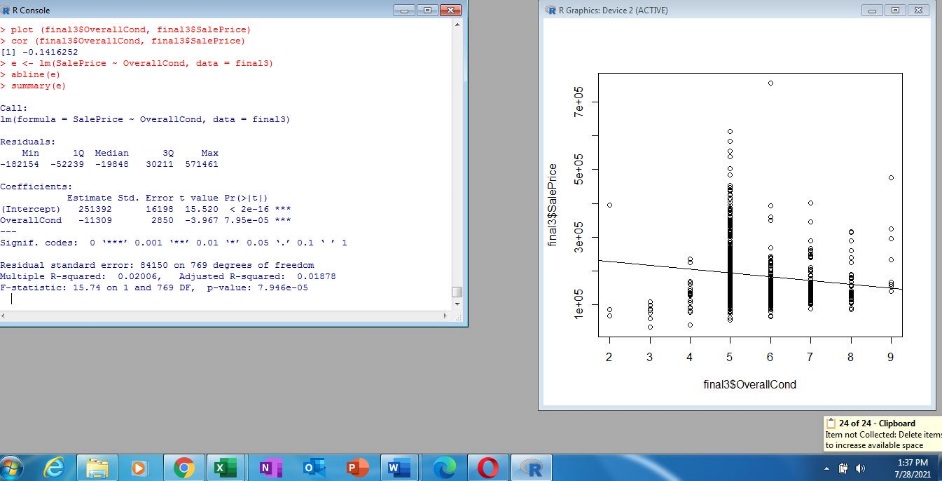
***Sales Price with Lot Area***

******

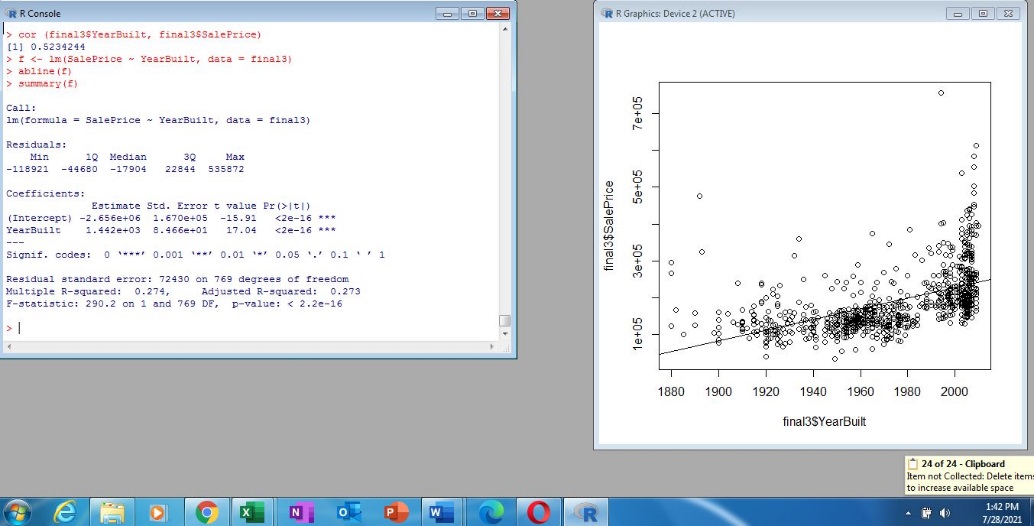
***Sales Price with Overall Quality***

******

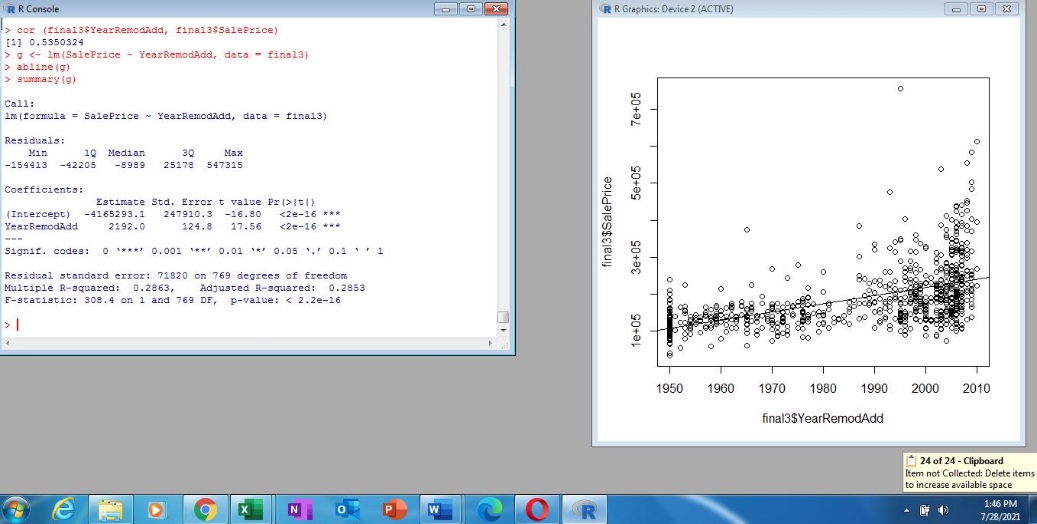
***Sales Price with Overall Condition***

******

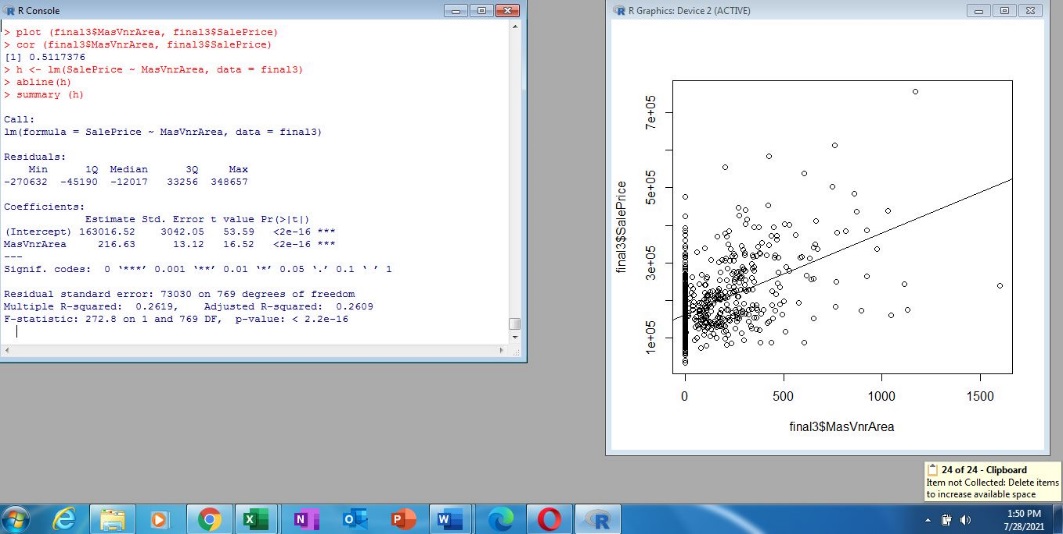
***Sales Price with Year Built***

******

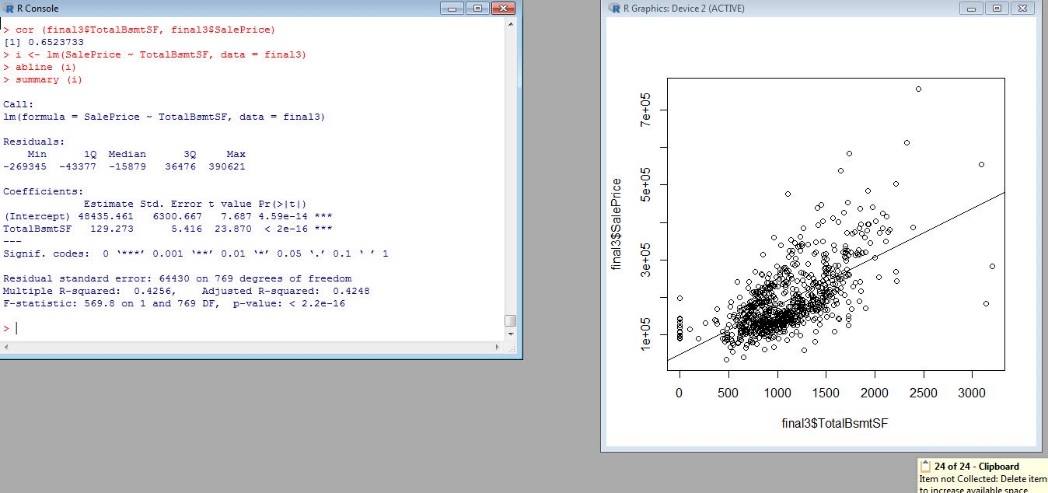
***Sales Price with Year RemodAdd***

******

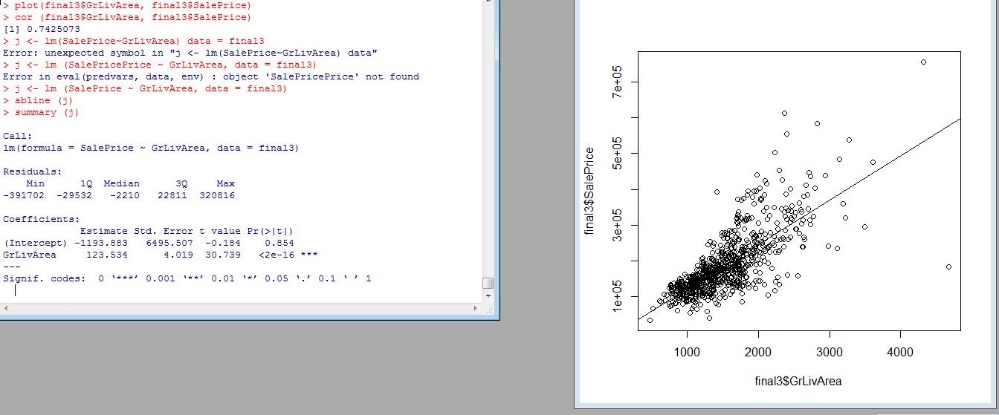
***Sales Price with MasVnrArea***

******

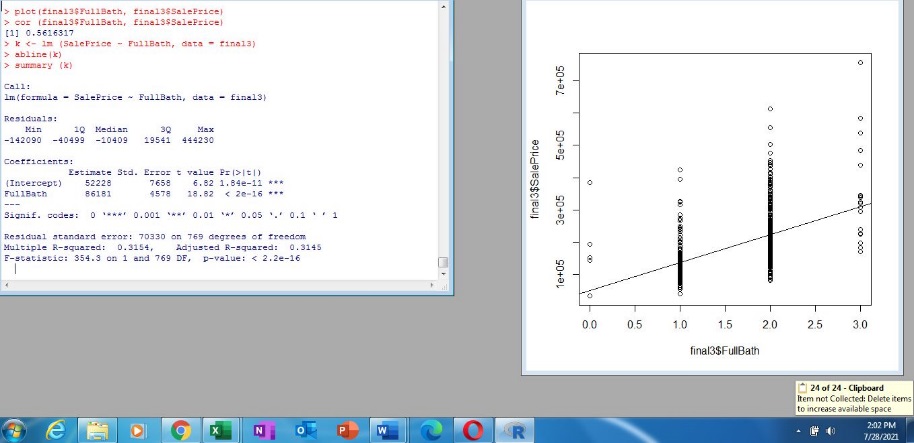
***Sales Price with Total BsmtSF***

******

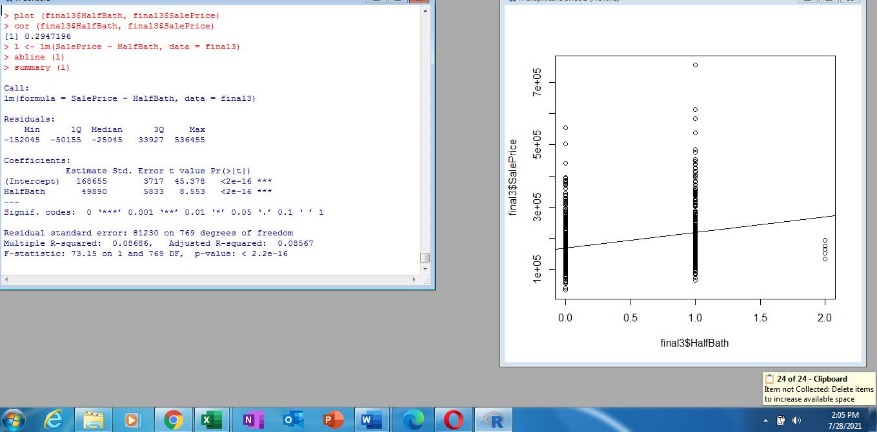
***Sales Price with GrLivArea***

******

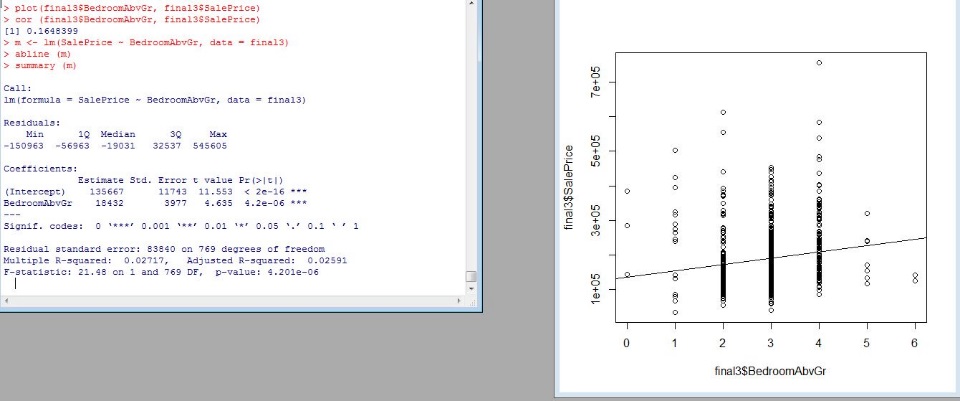
***Sales Price with Full Bath***

******

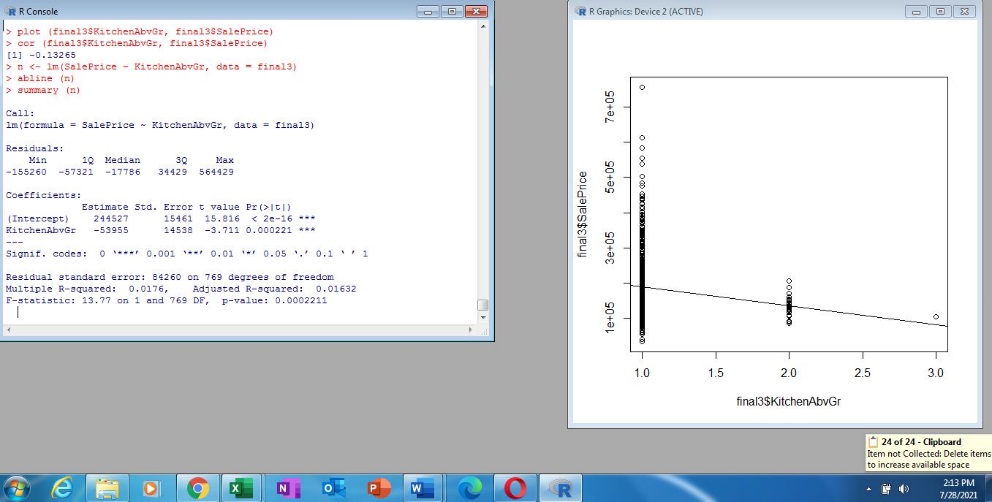
***Sales Price with Half Bath***

******

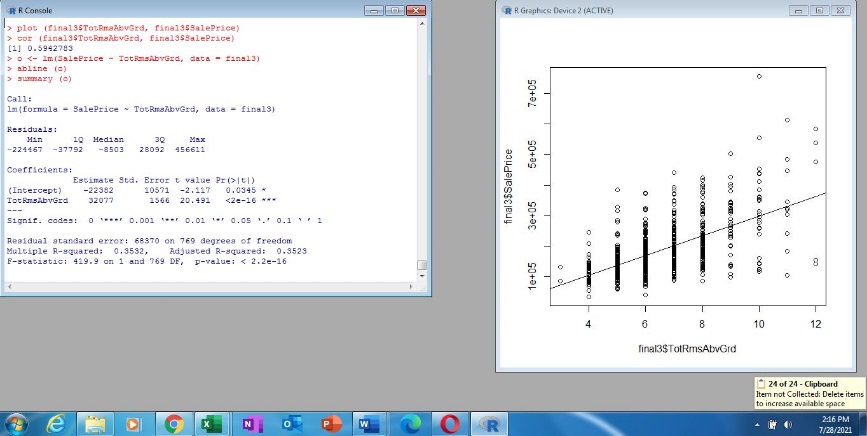
***Sales Price with Bedroom Above Ground***

******

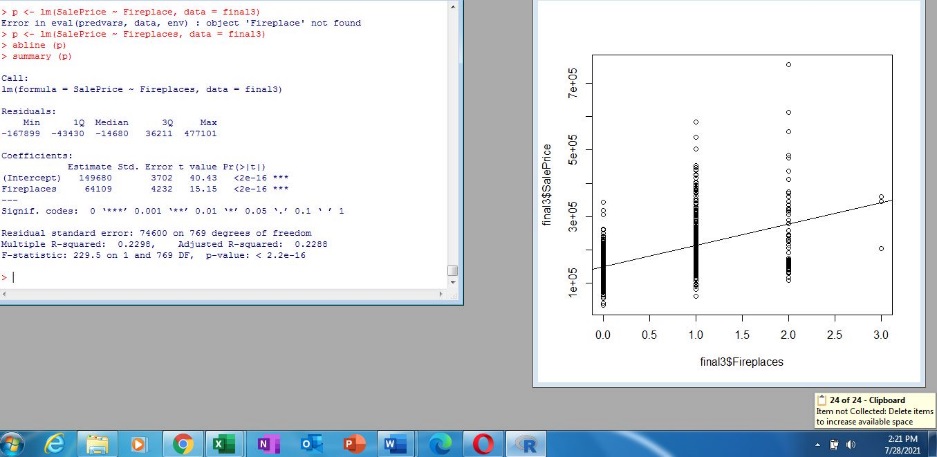
***Sales Price with Kitchen Above Ground***

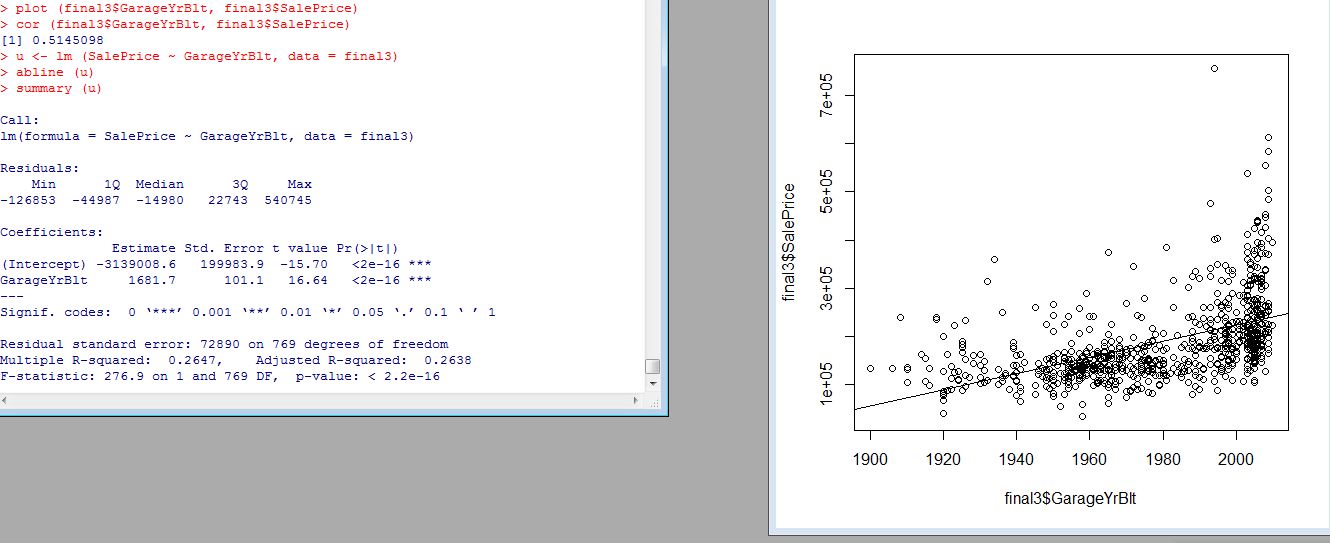
******

***Sales Price with Total Rooms Above Ground***

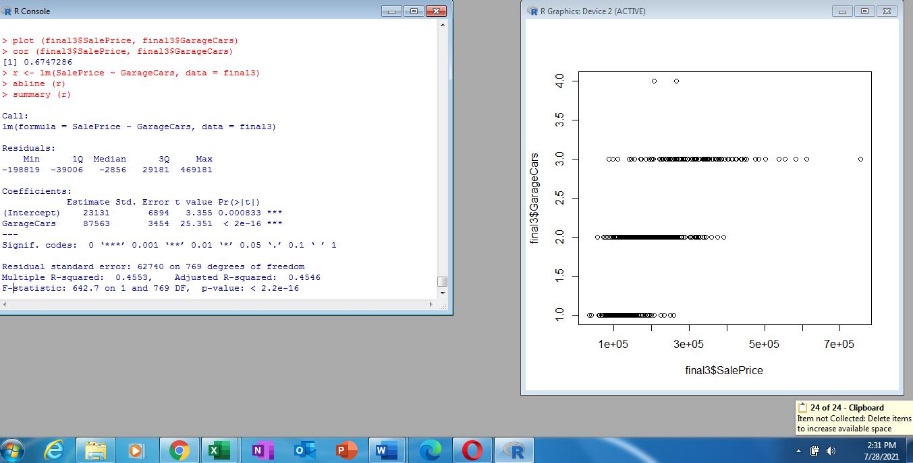
******

***Sales Price with Fireplaces***

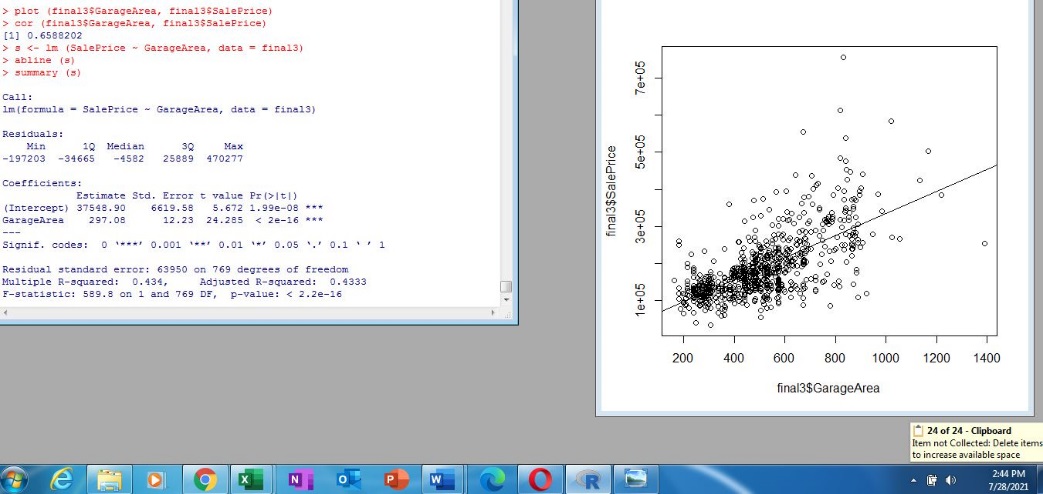
******

***Sales Price with Garage Year Built***

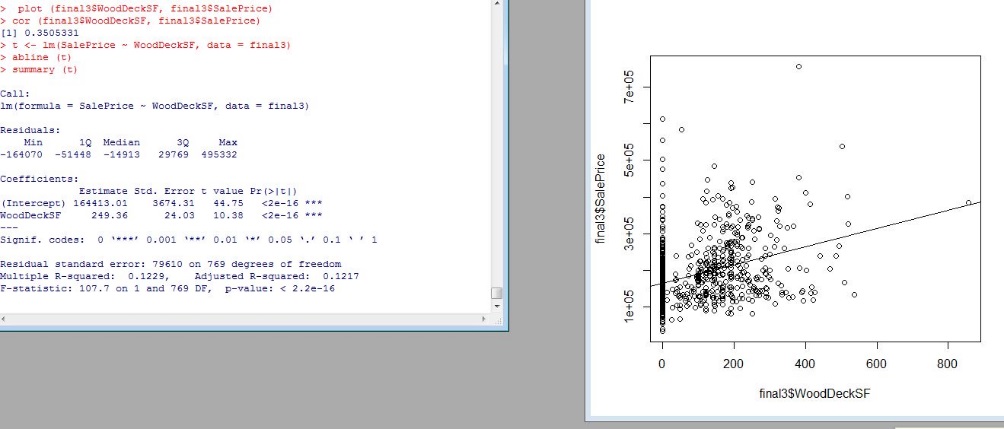
***Sales Price with Garage Cars***

******

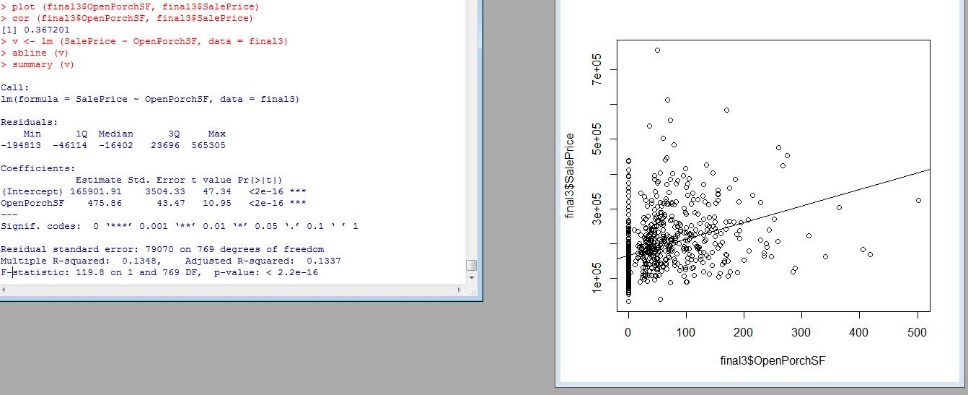
***Sales Price with Garage Area***

******

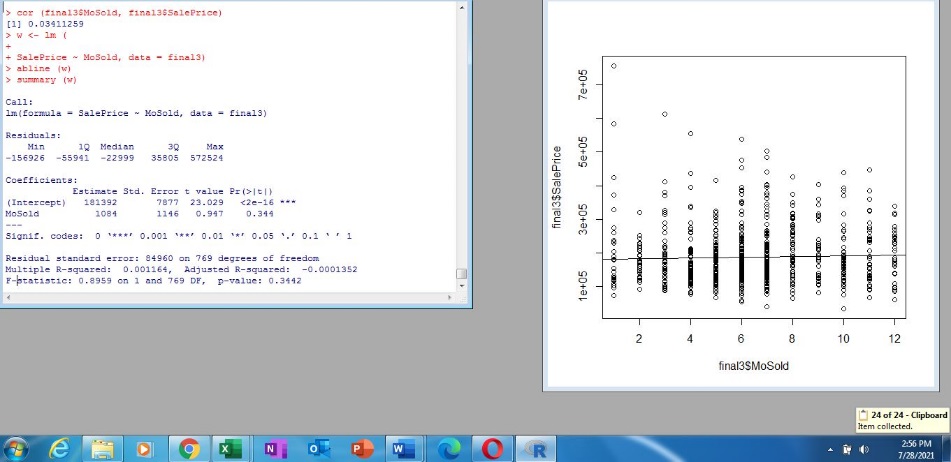
***Sales Price with Wood Deck Square Footage***

******

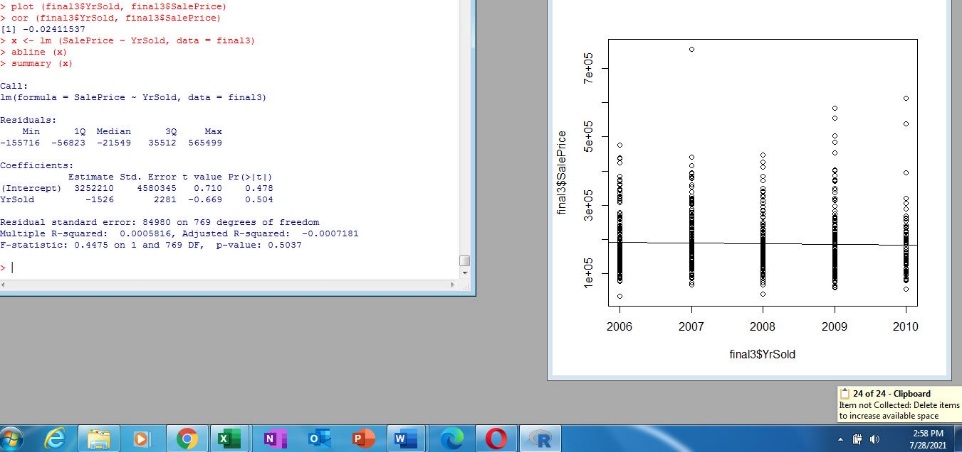
***Sales Price with Open Porch Square Footage***

******

***Sales Price with Month Sold***

******

***Sales Price with Year Sold***



***Part IV Significant Factors***

The following are statistically significant in relation to sales price: wood deck square feet, garage area, open porch square footage, garage cars, garage year built, fireplaces, total rooms above ground, kitchen above ground, bedroom above ground, half bath, full bath, grlivearea, total basement square footage, masvnrarea, remodadd, year built, overall condition, overall quality, lot area and lot frontage model.

Sales price with garage area, sales price with garage cars, sales price with garage year built, sales price with fireplaces, sales price with total rooms above ground, sales price with full bath, sales price with grlivarea, sales price with masvnrarea, sales price with year remodadd, sales price with year built, sales price with overall condition, sales price with overall quality all had the same p value of 2.2 e-16. All these values are highly significant, and it would make sense that as the more rooms, more total area of the house, more space with car garages, newer houses, and newer remodeled houses all have a direct correlation with price or the sale of the house. These features tend to make the house bigger. However, what is surprising is all these factors have the same p value.

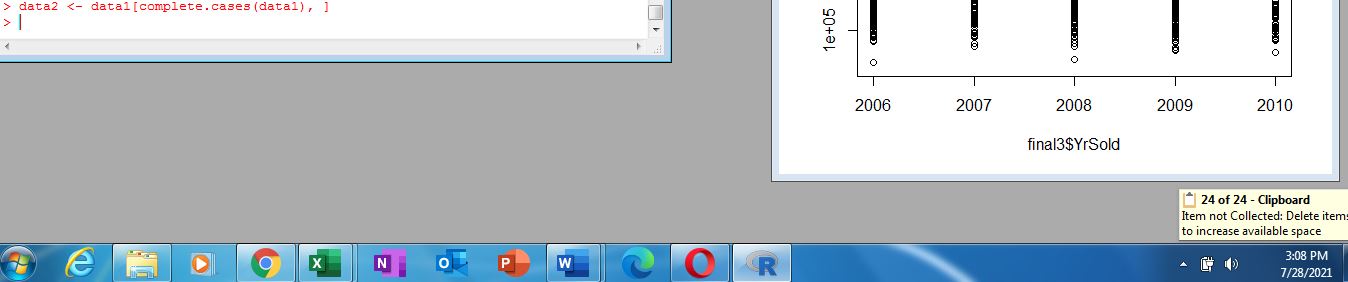
There are other statistically significant factors in respect to sales price. Wood deck and open porch have a significant factor of 1.556 e-12, and 2.07 e-11 respectively. This is logical since a bigger deck and a bigger porch would mean that the house has space in the backyard to build something like this. This would lead to the lot of the house. The sales price with lot area has a significant value of 4.691 e-11, and the sales price with lot frontage model has a significant number of 3.161 e-10. This would also make sense since the lot of the house should, and in this case does have a positive effect on the price of the house. The bigger the lot of the house means that there is a bigger space where the house can be built on.

Kitchen above ground, bedroom above ground, and half bath(s) all are statistically significant in terms of sales price. Kitchen above ground has a p value of .001918, bedroom above ground has a p value of .0006224, and half bath(s) has a p value of 1.837 e-07. The bigger the kitchen, the more bedrooms and bathrooms should all positively affect the sales price of the house, and in this case they do. However, what is interesting in this case, is the above ground portion. It seems that the bedrooms and bathrooms, below ground, or in this case the basement, is omitted from the research. I would imagine they do have a positive impact on sales price, but from this dataset, it is not stated.

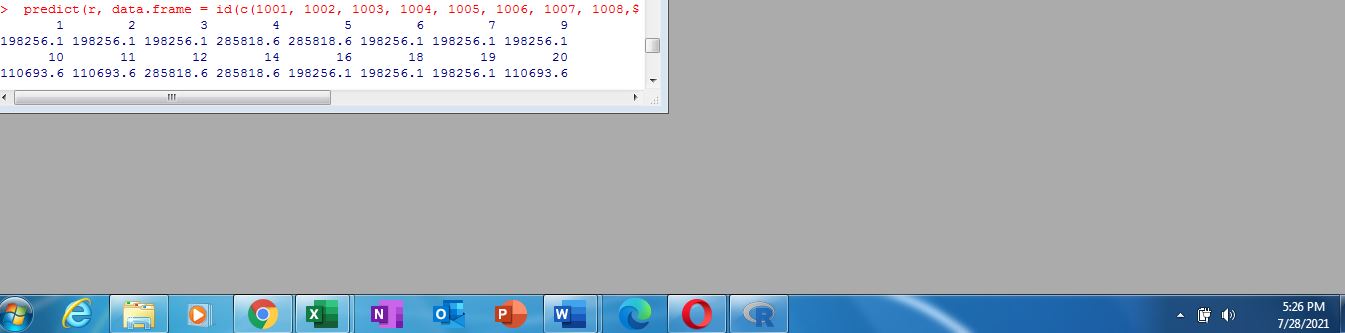
The final factor that is significant in terms of sales price is the sub class regression. The p value for this variable is .05, which is right at the cutoff for a significant value. I would imagine this is the neighborhood, or suburbs of where the house is located. I am surprised that this p value is not more statistically significant. I would think that more people would also value the neighborhood, school district of where the house is, as an important influence of their decision of buying a house, which would increase the sales price of the houses in the neighborhood.

***Part V***

Dropping rows using complete cases screenshot:



***Part VI Predict Sale Price***

******

The above is a screenshot of the predict sale price using the house testing dataset.

***Part VII Compare and Contrast the Datasets***

The values for the first twenty rows of the house are as follows

|  |  |
| --- | --- |
| SalePrice | Mine |
| 82000 | 198256.1 |
| 86000 | 198256.1 |
| 232000 | 198256.1 |
| 136905 | 285818.6 |
| 181000 | 285818.6 |
| 149900 | 198256.1 |
| 163500 | 198256.1 |
| 88000 |  |
| 240000 | 110693.6 |
| 102000 | 110693.6 |
| 135000 | 285818.6 |
| 100000 |  |
| 165000 | 285818.6 |
| 85000 |  |
| 119200 | 198256.1 |
| 227000 |  |
| 203000 | 198256.1 |
| 187500 | 198256.1 |
| 160000 | 110693.6 |
|  |  |

My numbers are not close to the actual sale price. My model is off by thousands of dollars. Another problem that happened is the sales price in my model keeps repeating, so there is some sort of issue that is going on here.

Resources

BMJ. (2021). *Mean and Standard Deviation.* <https://www.bmj.com/about-bmj/resources->

readers/publications/statistics-square-one/2-mean-and-standard-deviation

Doane, David & Seward, Lori. (2011). Measuring Skewness: A Forgotten Statistic?.

J. Stat. Educ.. 19. 10.1080/10691898.2011.11889611.

***Appendix***

Summary Statistics of entire housing testing dataset

