Project Description

MSDS 411

Introduction:

Using the Ames data set, I will develop a model to help predict the selling price. The original data contains individual residential properties sold in Ames, IA, from 2006 to 2010. The data includes 82 columns, 23 nominal, 23 ordinals, 14 discrete, and 20 continuous variables. It contains 2930 total observations.

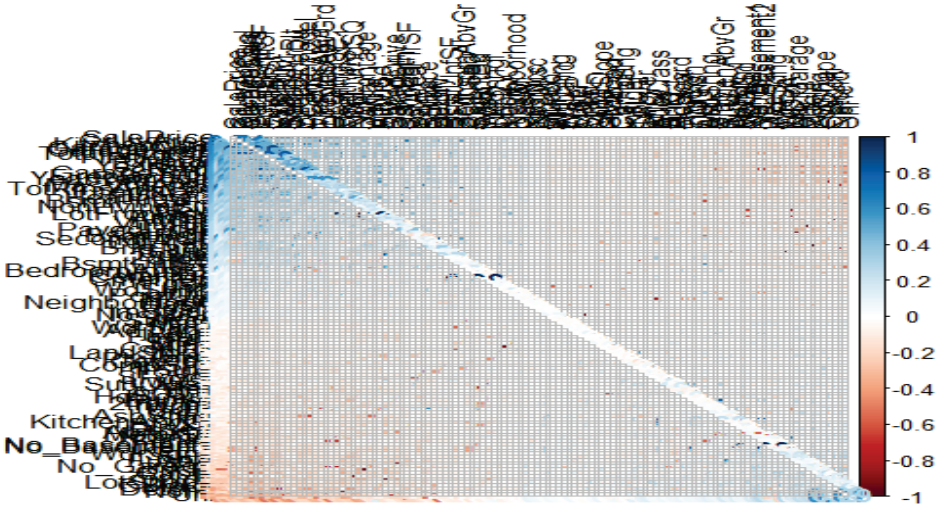
Data Preparation Description:

Looking at the N/A's in the data set, there are 13960. Not all N/As are wrong but will have to be changed to help the analysis. Also, to use more of the data, I changed a lot of character columns into numbers. For most columns “EX” = 5 and “Po” = 1. For the LotFrontage, MasVnrArea/Type, and GarageYrBuilt, I changed the NA's to the means of their neighborhoods(by_grouping). Then similar variables were together to make one variable, i.e., mydf$Exterior <- (mydf$ExterQual + mydf$ExterCond)/2. Other variables added together were: mydf$Bsmt, mydf$Garage, mydf$NonLivingSQ,

Then the columns that were combined were removed—changing the variables from 82 to 69. Once all the character columns were converted and spread out, there are 186 numeric columns

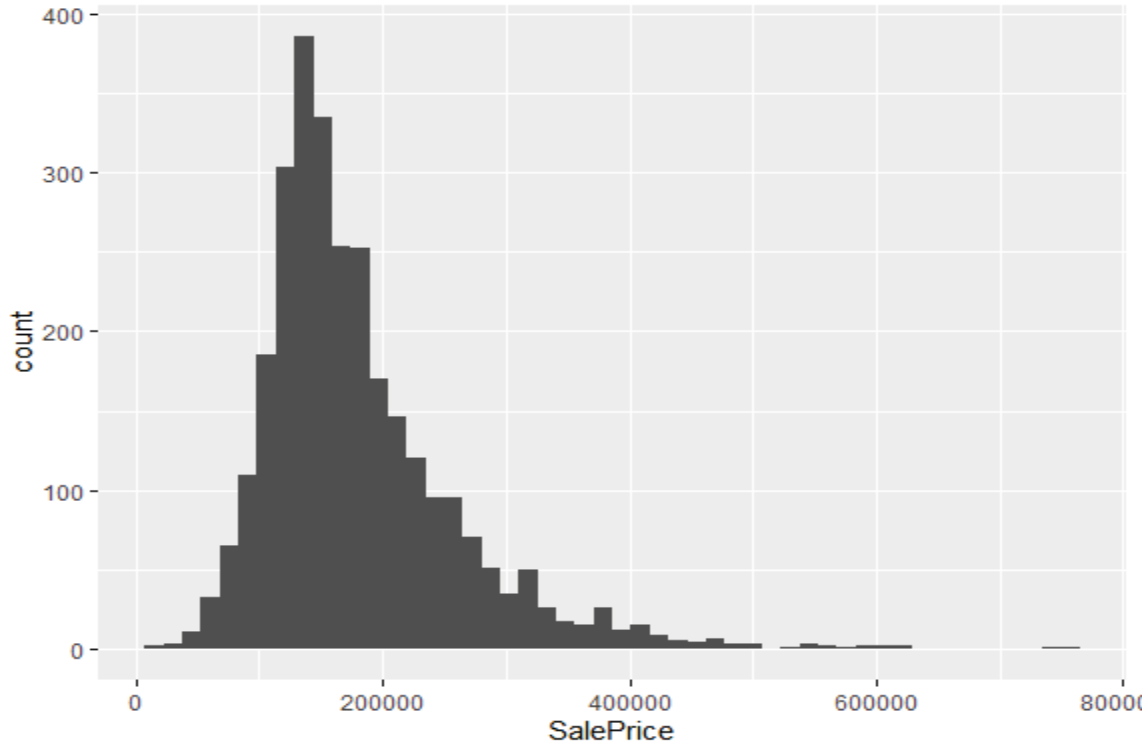
Exploratory Data Analysis (EDA)

Looking at the N/A's in the data set, there are 13960. Not all N/As are wrong but will have to be changed to help the analysis. Once the data was all converted and spread out, the N/A's are now zero. Looking at the Correlation plot of the 186 variables is ugly. The graph is sorted based on its correlation to SalePrice.

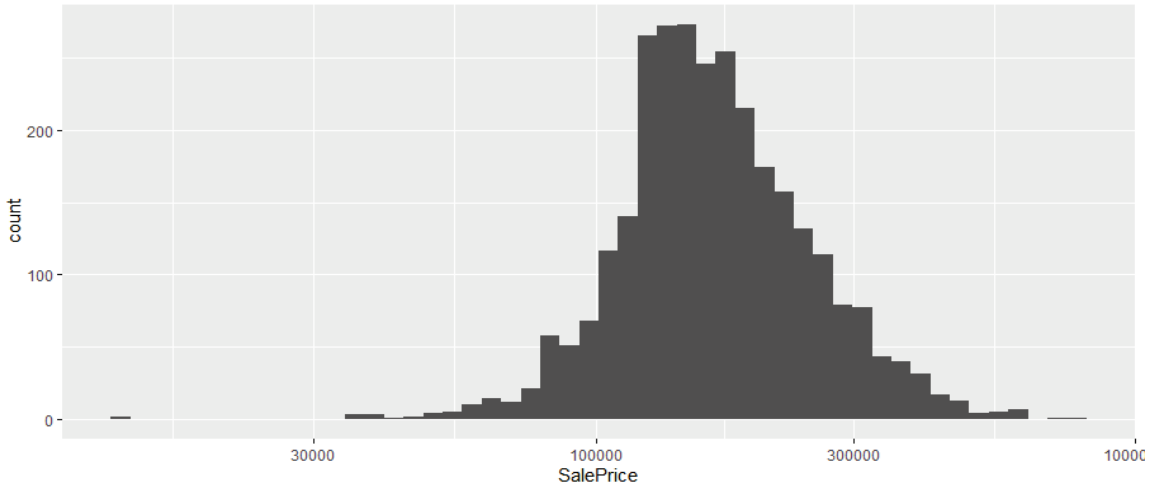


From the above correlation graph, you can see that red seems to cover over 50%. Plus, the dark blue is just a few on top.

Looking at SalePrice price, it seems to be a little under 200k and skewed to the left. Let's check it by logging it.

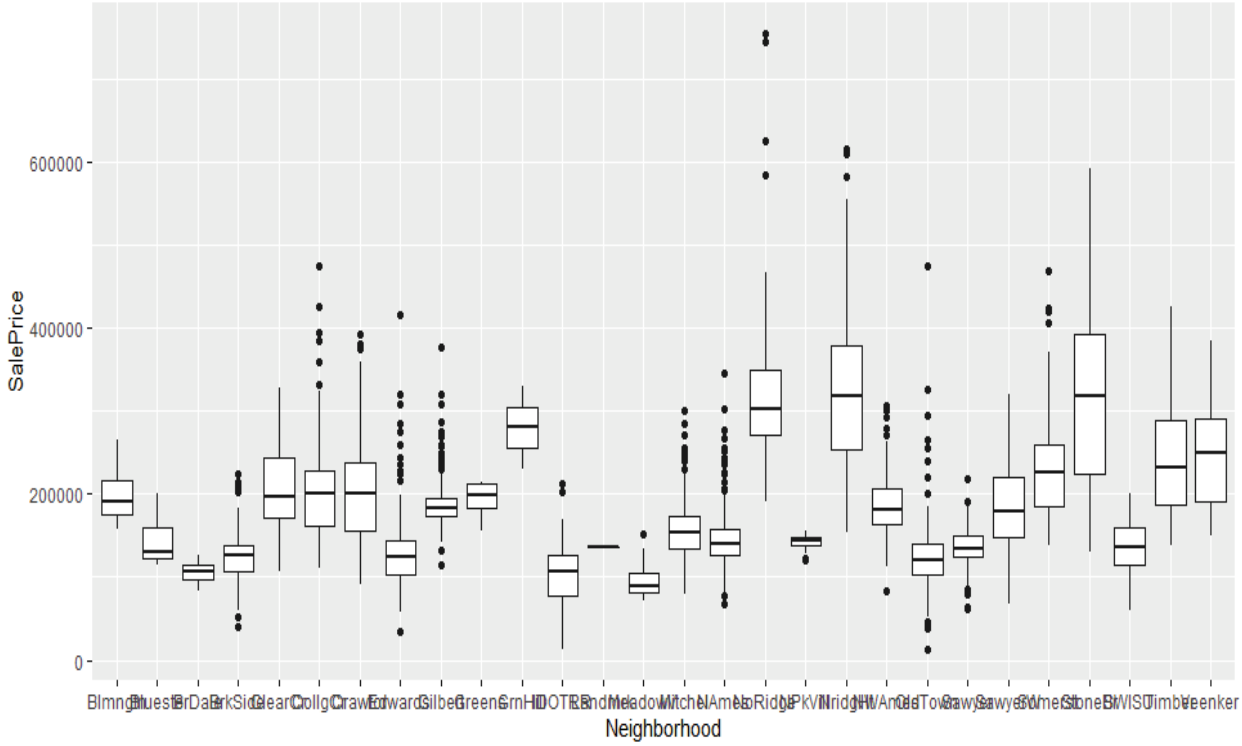


Logging the prices does change things a bit. The middle now looks more even and with more bars.

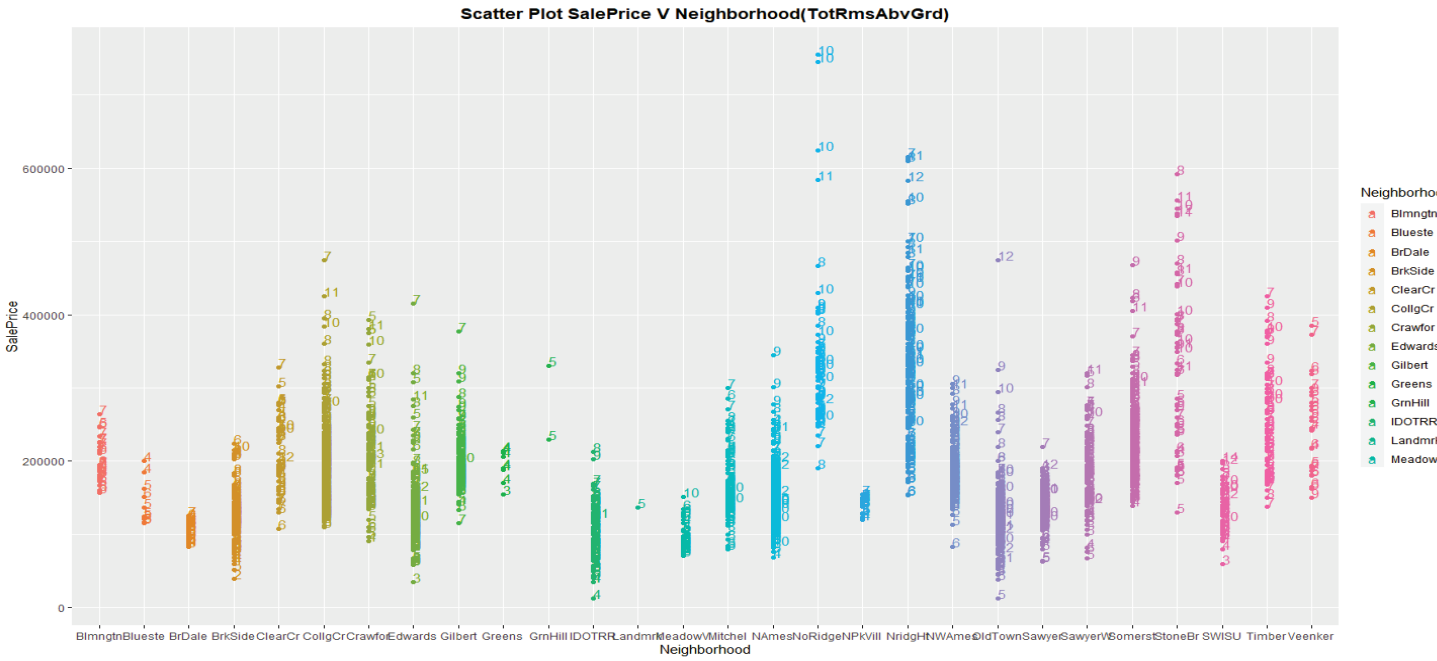


Analysis

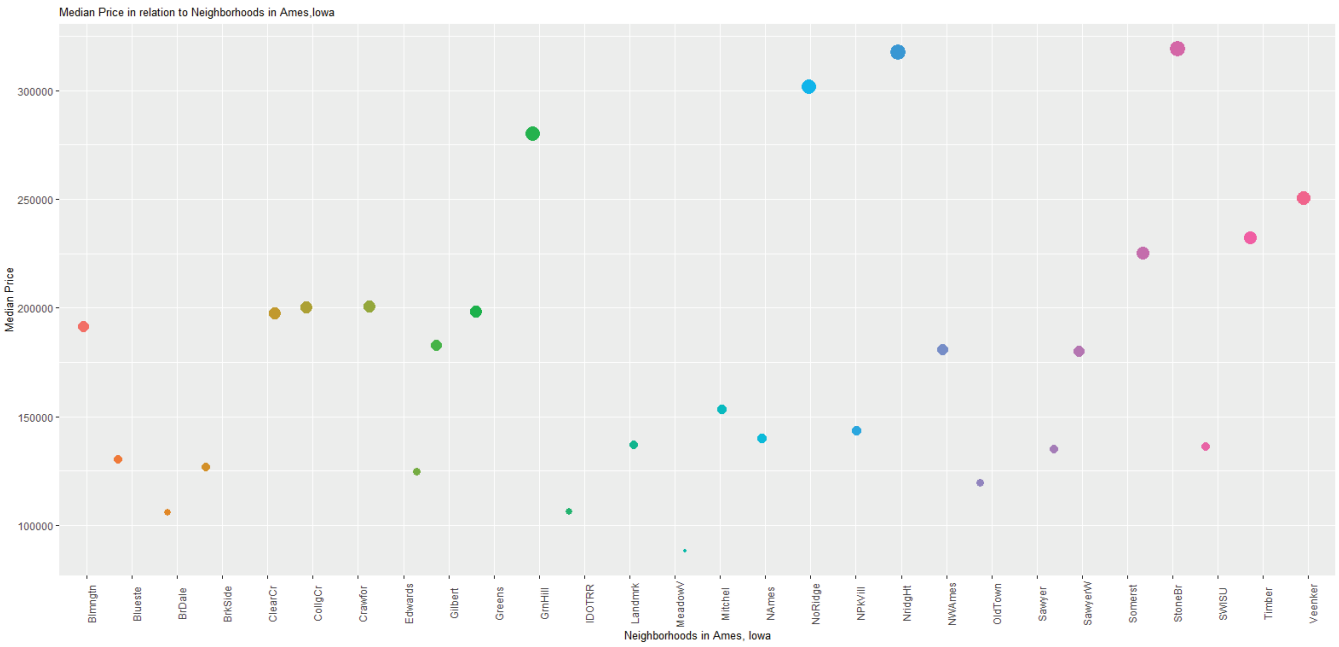
Since Location is considered one of the essential factors looking at the SalePrice by Neighborhood, it seems to confirm that. However, to many outliers to make that case yet.



Taking a further look at the saleprice/neighborhood thought it would be interesting to see how many rooms the houses have, which should also give us a hint on the size.

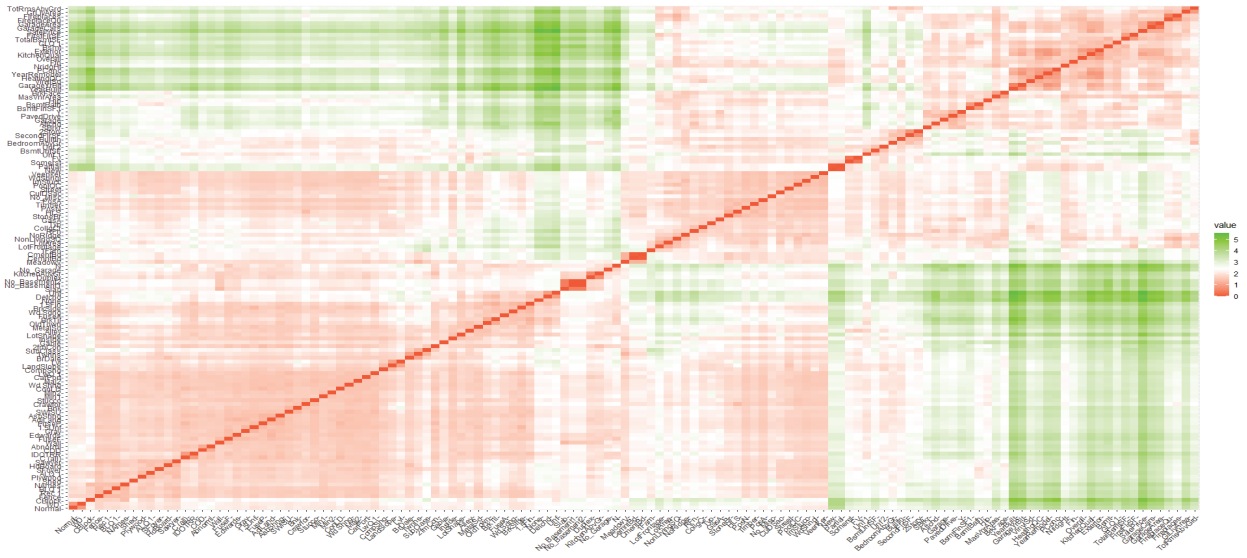


It looks to be that size and location help.



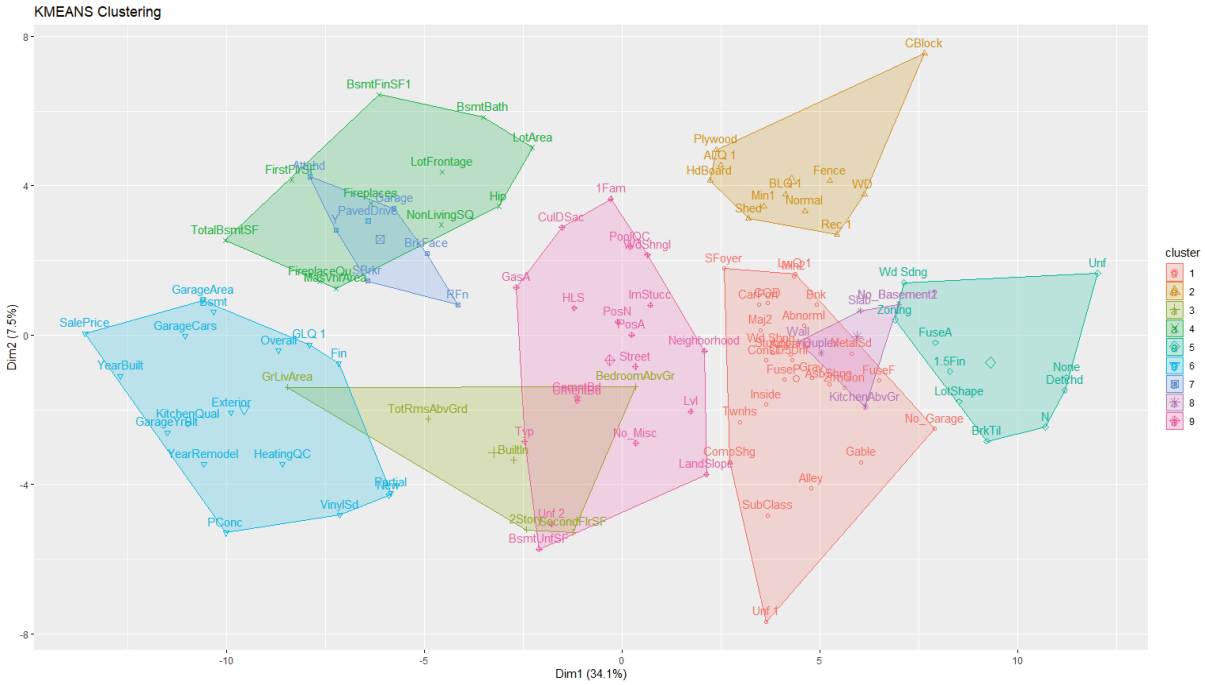
Find the two graphs above interesting. If you put one directly on top of the other, you can see all the prices and the big long lines from the top chart and then on the following plot just the averages. I thought the two blue in the middle would be reversed due to the outliers.

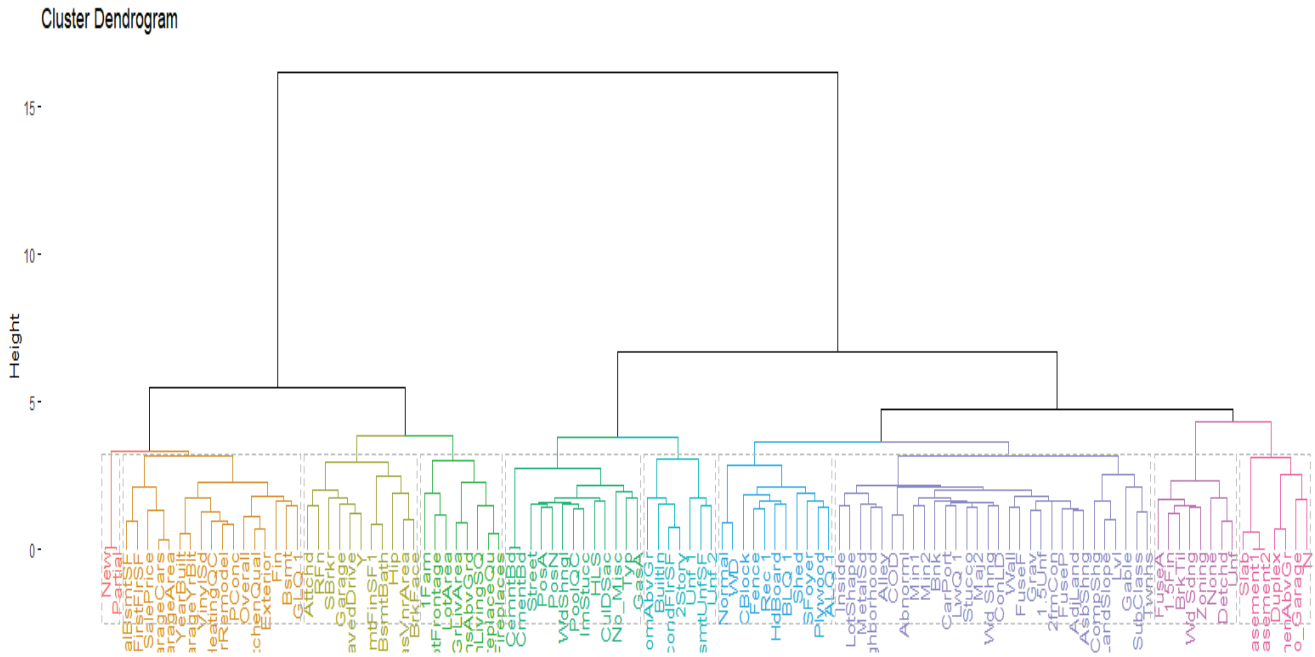
We have already looked at the correlation above. Below is the distance map.



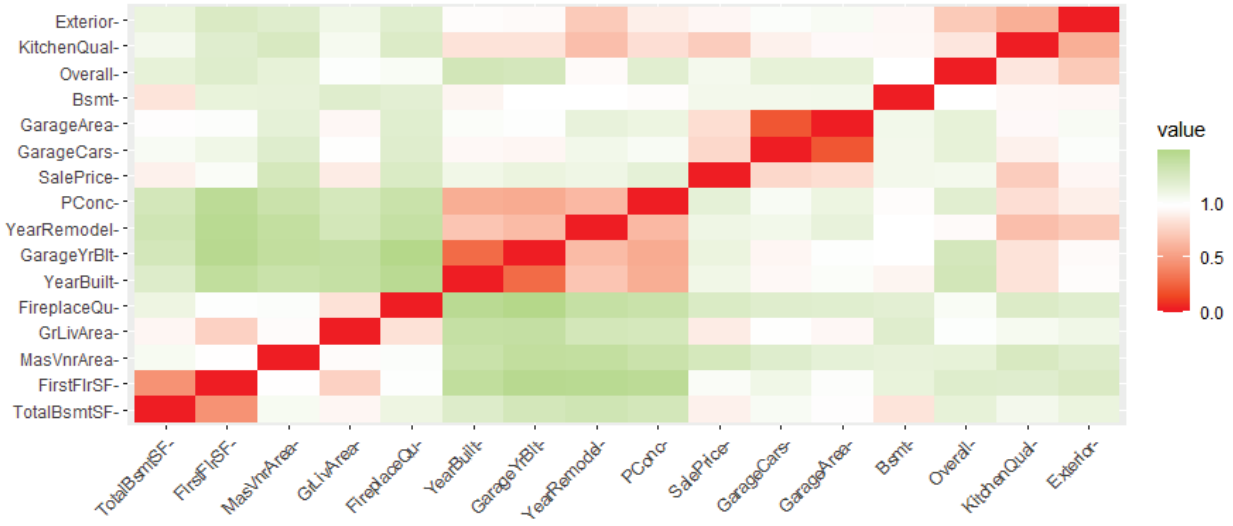
The above distance graph is in the same order as the correlation plot. It does look like the correlation plot above. With only a couple of areas of green and the dark green is barely there.

So we have our correlation and distance information. There are a few good variables that we can use, but there is also a lot of noise, so we still need to cut it down a bit more but still capture most of the data/patterns. PCA should help here. Using Caret with the 114 variables tells us that to capture 95% of the variance, PCA needs 46 components. Using the k -means clustering, we get nine groups from the variables. We will cut down on the noise by taking out the data with a < 0.5% correlation. It leaves us with 16 variables. This doesn't seem like a lot, but it seems the majority of the good data was one top from the past graphs.

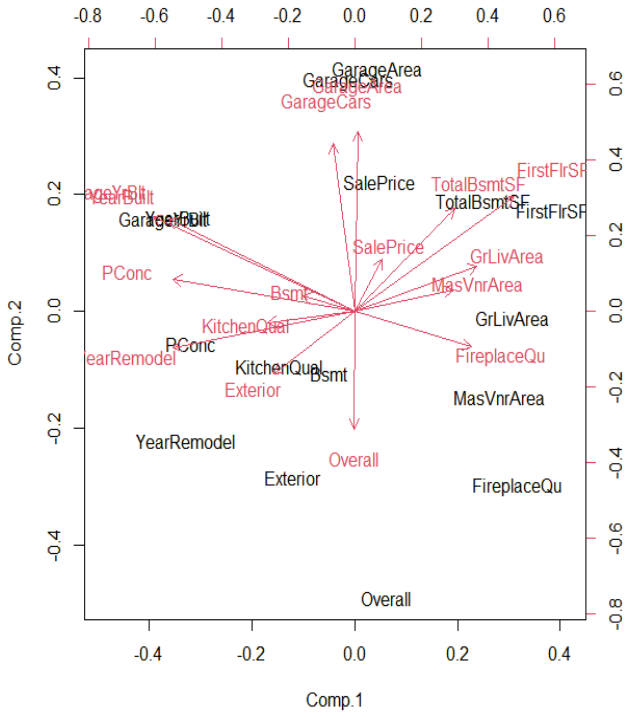




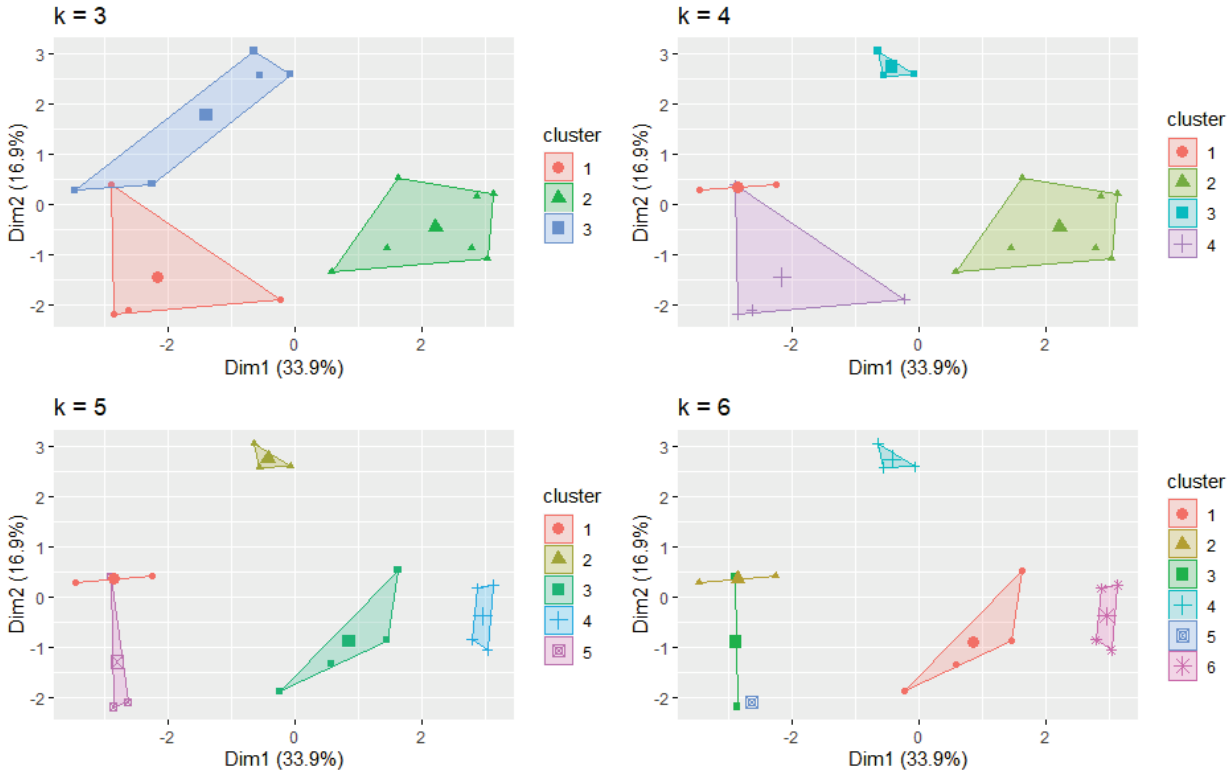
With just the 16 the distance graph is more green but still contains a lot of red.



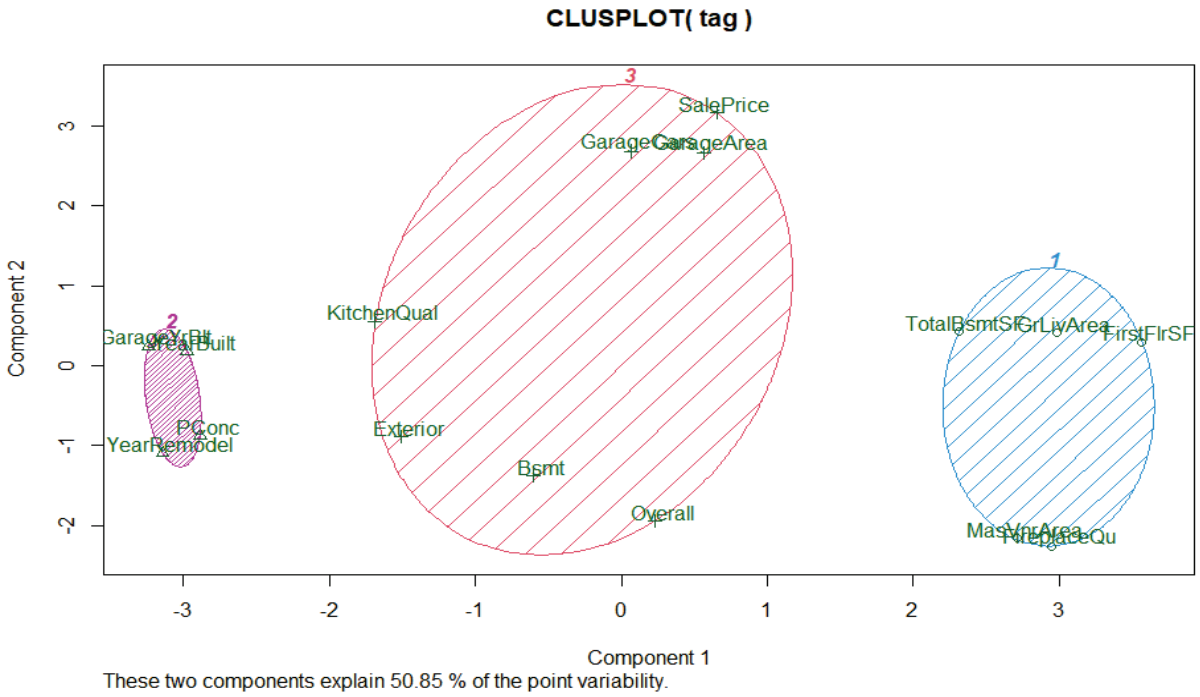
The biplot shows some nice groupings separated nicely



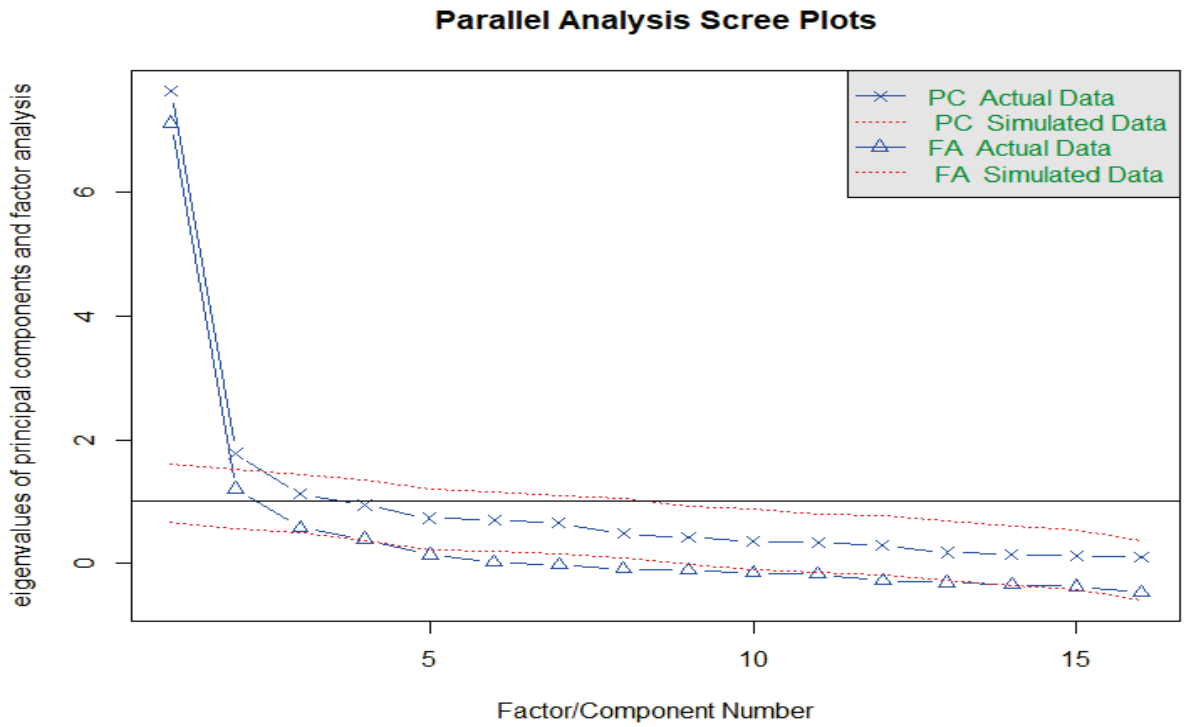
Using the k-means clustering with cuts 3-5, K =3 contains all points within its clusters.



Clustering the data into three groups is interesting when you see the labels. Like there are garage sections in two and three.

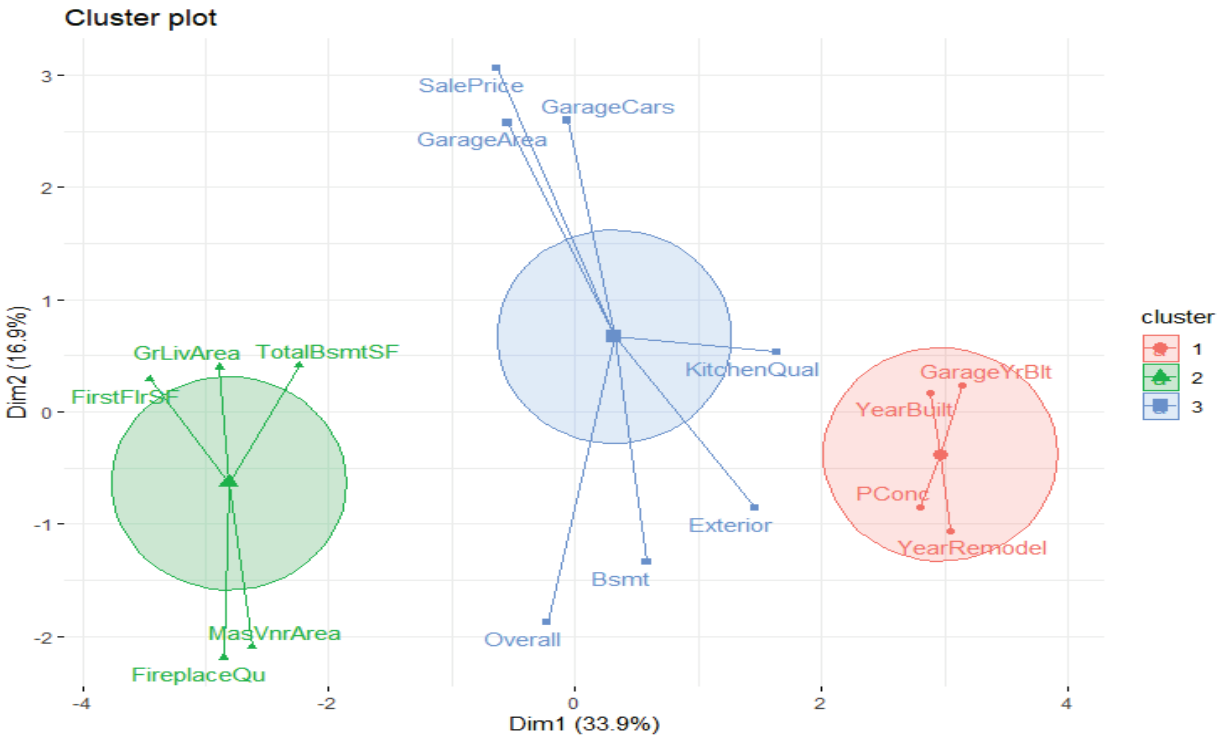


The Parallel analysis requires two components and four factors.



and four factors

I like how this cluster plot shows the distance from its node. Cluster 1 is all within its circle, with most of them dealing with dates/age of the houses.



The main question that I kept asking myself going back to 410 was if all that data was not numeric was not used. This time I was able to find some examples on how to mutate the character columns into numeric. Was it worth it to convert everything? I don't think so. Using PCA, I was able to get the same type of predictions as I did when I used them all

"The predictions are" Full model						
1	2	3	4	5	6	
163295.3	108315.1	137492.8	246274.6	163295.3	193171.9	
"Actual price"						
215000	105000	172000	244000	189900	195500	
"The predictions are" Top16						
1	2	3	4	5	6	
163295.3	129389.6	172150.6	246274.6	163295.3	193171.9	
"Actual price"						
215000	105000	172000	244000	189900	195500	
	1	2	3	4	5	6
Dif Full	-51705	3315	-34508	2274	-26605	-2329
	-51705	24389	150	2274	-26605	-2329

Conclusion/Reflection:

These never seem to go as fast or turn out as I had planned them. For this assignment, I chose the Ames dataset, which was covered pretty thoroughly in 410. I prefer it because I kept wondering why we seem to leave so much data out. Or not always converting the string data to numeric and just kicking out the rows and columns with N/A. Even though I had some excellent examples of converting all the data, there are always issues. I do not have too much experience with R, I like R, but I have spent a lot more time in python. For me, visually, as I worked through the examples or tried my own. I started to understand what was going on visually better—the three charts with the SalePrice and Neighborhood. The first two graphs are very informative, then the third using one dot relays the same information.

Being in investments, the topic of distance is very appealing to me. Whether it's price, volume, time, or any other factors that you can think of. I was looking at graphs or information. Many repeatable patterns are seen in the stocks that are outperforming. Of course, these patterns are seen in the average stocks too. It would be interesting to look at the past performers and see what variables contributed to its success. For example, stocks trading in a range for over three months and then take off. Is there something that started to change? Was volume picking up, was internal or large institutions starting to get on board.

Again, sorry I have been so late these past few weeks. With the sale of Wells Fargo Asset Management, my days are much longer now in preparation for moving out. I would much prefer to have my time spent on this.