Unsupervised Learning Final Project

June 2021

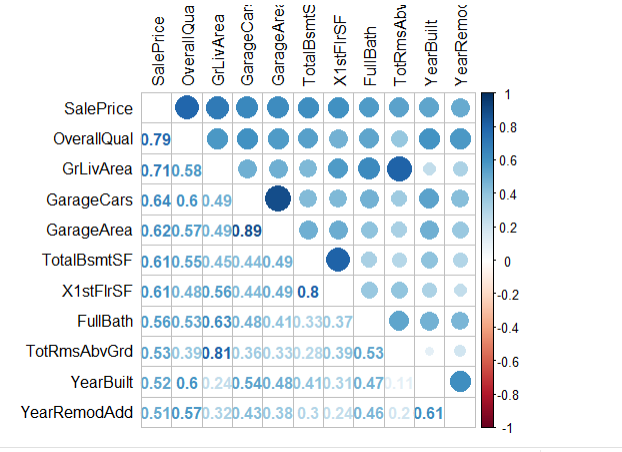
**Introduction**

This project regards the Ames, Iowa dataset and the 1,500 housing properties, and their sales price between 2006 to 2012. The objective of this project is to see whether, using various models, we can bring more clarity for potential homeowners and homebuyers of predicted sales price based upon historical data. Additionally, the data set has both continuous and non-continuous variables with 82 explanatory variables. The dataset contains 23 ordinal, 23 nominal, 22 continuous, and 14 discrete variables. The numerous continuous variables are data describing the metrics surrounding the description of the housing properties: including square feet, living area, number of rooms, baths, and kitchens. The non-continuous variables discuss more details about the housing and its area including neighborhoods, garage type, and sale type.

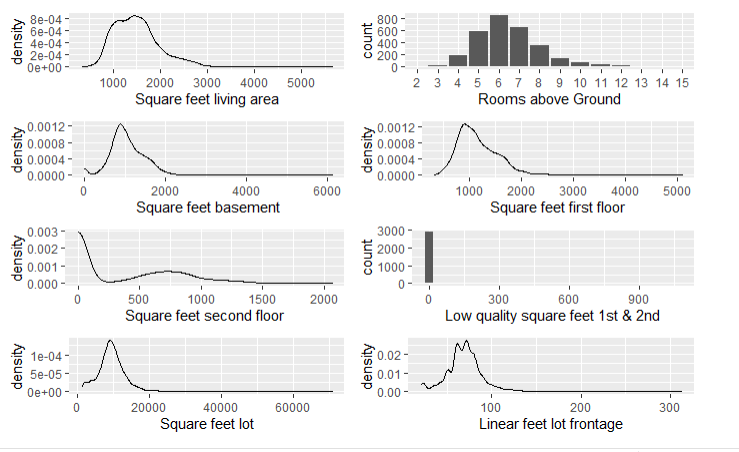
Given this large amount of historical data, this project aimed to create models to help predict sales price of properties based on similar descriptive data. The models we created were the K-means cluster on pca, k-means cluster on original data, and k-means cluster on pca data. Two of our models required Principal Components Analysis (PCA) which saw our dimensions being reduced to try to remove ineffective or redundant variables. The goal was to produce better performing models and we tested both using the pcas changed data, and the original data to see which model performed the best. We found that the k-means cluster model on the original data performed the best with a nearly 86% accuracy rate.

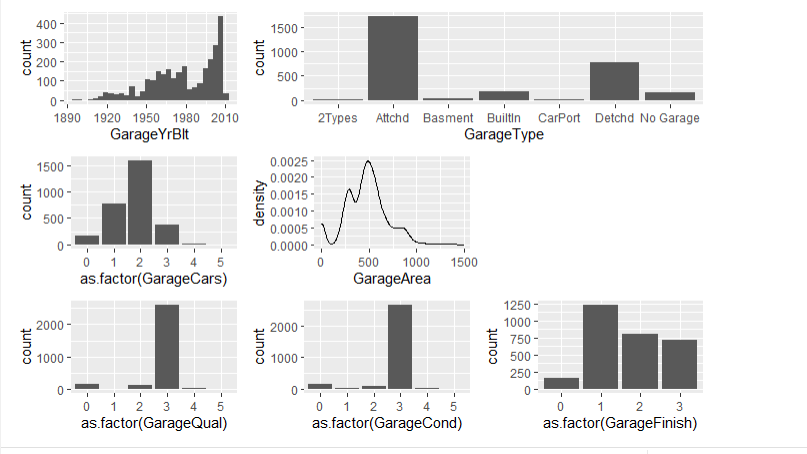
**Methods and Findings of Exploratory Data Analysis (EDA)**

In performing exploratory data analysis (EDA) for this project, we created histograms, and counts for viewing relevant and missing data. We created a correlation matrix to view the relationships of the other variables with our response variable sales price:



What was interesting to note, was that the variables with high correlation to our response variable were all continuous variables, which would have good bearing for our model. I also reviewed which variables had missing or low amounts of data, and the data which only had zeroes instead of actual values. These variables could possibly be removed as they are redundant in comparison with other variables. Here are some of the histograms we created for EDA:





The EDA shows that there are quite a few variables with null or low number of values which we will discuss how we treated in the next section.

**Data Preparation Description**

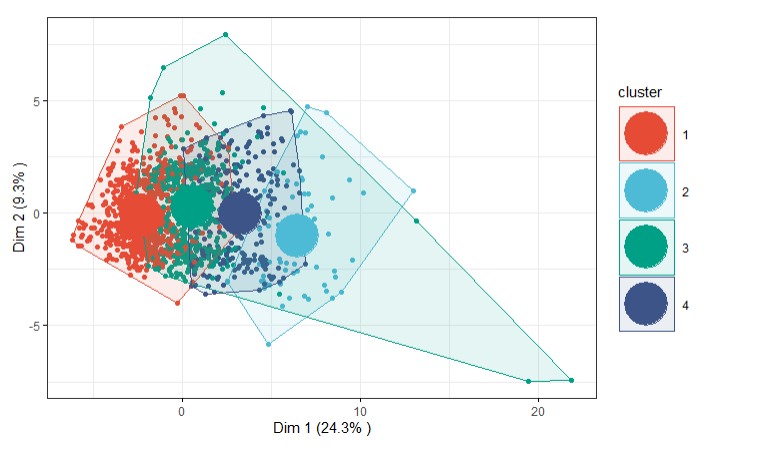
After viewing the results from the EDA, we can see that there are quite a few variables with low correlation to our response variable and quite a few with null or missing values. We decided, for this project, to use k-means clustering, pca, and hierarchical clustering models to try to answer our research questions. These variables require continuous numerical data variables and thus we will remove the categorical or noncontinuous variables. We could have created dummy coding and transform the variables with one hot encoding to numerical variables, but instead decided to go with just the continuous variables for this project, largely because these variables had a high correlation with the response variable. This left about 34 variables for us to create our models on, but of these, nearly half had null data which needed to be fixed. We chose to impute the missing values for these variables by selecting the mean of each variable. We could have chosen other forms like median or mode, but felt due to the large amount of present data, that mean would help create the best model. There were not a large amount of outliers, and the outliers that were there, we felt needed to be included in order to create a more accurate model.

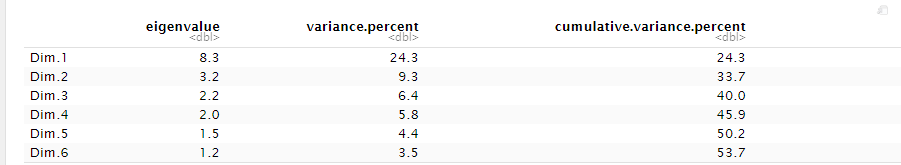
**Analysis Methods**

The first model we decided to create was the PCA. This was done to reduce dimensions on our data and try to weed out redundant variables. We then tested our models against this pca data and our original data. We chose to create a k means clustering model to cluster our data together to try to find underlying patterns within the data. Our goal with using k-means was to find clusters in the data with which we can use to predict future sales price. We first tested to find the ideal number of clusters for our data. For both the pca data and original data we found that 4 clusters was the ideal amount. We then tested our data creating a model to test the pca and original data using 4 clusters. For the pca data, we run the data through a pca model which reduces dimensions and shows us the amount of variance in the model. We selected the two highest dimensions of variance to graph, but noted that there were four dimensions with our k-means analysis which accounted for nearly 50% of the variance. For the hierarchical clustering method, we again found that 2 clusters were the optimal choice for the model. Similar to the k-means clustering, the hierarchical clustering transforms the data into clusters based upon similar characteristics.

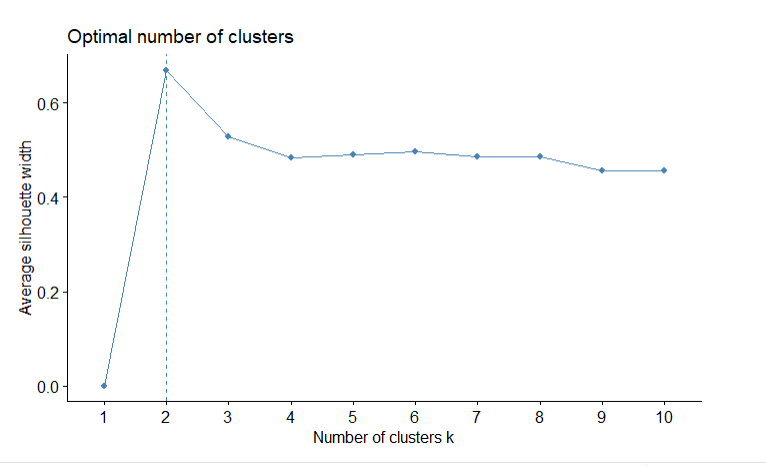
**Final Model Presentation**

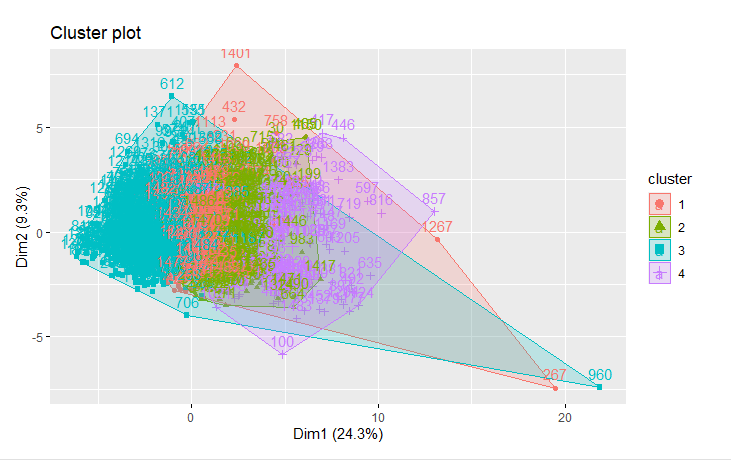
As noted in our introduction, the k-means clustering on original data performed the best. This was followed by the hierarchical clustering model on pca data, with k-means clustering on pca data finishing last. The pca was successfully able to reduce the dimensions, and we noticed that the data had been transformed, but it ended up having a worse effect on our accuracy of our models. The k-means with pca performed at only 48.7% percent with only 53.7% of the variance explained when reduced to six dimensions. Both of these are rather low, and suggest that this pca model would need more adjustments to be trusted.



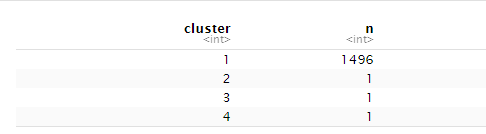


The hierarchical model on the other hand, was difficult to ascertain for accuracy. However, I was able to assess the cluster results against our original response variable of sales price and was pleasantly surprised at how much of the data corresponded.

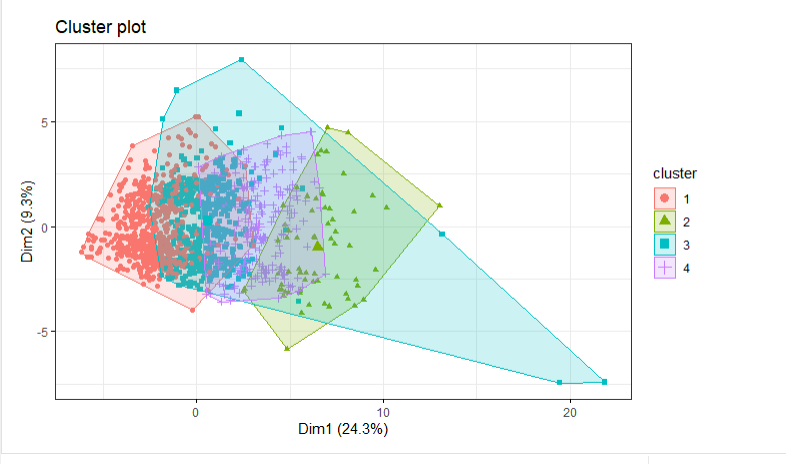


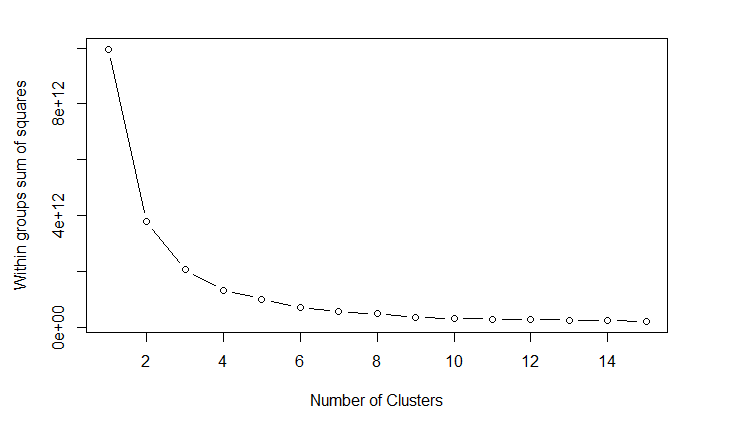


* The Hierarchical clustering model showed interesting results, especially when I cut the tree. Most of the results for k=4 showed up in the first cluster. However, when I cut the tree with k=300, the clusters were more spread out evenly. Obviously, there shouldn’t be 300 clusters but it was interesting to note. This model also suggest that further work can be done to improve upon the accuracy.



The k-means on the original data was the highest performing model with nearly 86% accuracy – which was a successful first round for modeling. This model shows that homeowners and homebuyers can expect that they can get a better picture of valuation of homes in the Ames, Iowa market based upon specific data provided for the homes. Although not 100%, the high figure shows great potential in reducing uncertainty!





**Conclusion and Reflection**

This was a very successful first attempt on using different models for the Ames, Iowa Dataset. One of the most exciting things is how much future considerations there are for improved accuracy and modeling. First, we can add the non-continuous dataset by applying one hot encoding or dummy variables to change the categorical values to numeric. This will allow us to increase the dataset even further, adding more depth. We can also use these other variables to weed out those without high correlation or are redundant. Next, we can improve upon our models by changing the missing data we imputed. For this project, we used the means of the variables, instead we can try to find other statistical calculations or even find missing data points to impute the data with. Lastly, we can build other models and improve upon our current ones. The hierarchical clustering model can be improved with further experimenting on cutting the trees, and manipulation of the data!