Brandon Moretz & Sean Prentis
5th December 2019
MSDS 411: Unsupervised Learning
Final Assignment Ames Iowa Housing data

Introduction

This report will look at the feasibility of applying an unsupervised method in the selection of features for our chosen data set. The focus will be in signifying potential features of the instances which distinguish the most from one another. In doing so, we aim to form a concise set of descriptors that accurately reflects the variation in the homes, minimizing the number of individual variables. The two core questions we look to answer as follows:

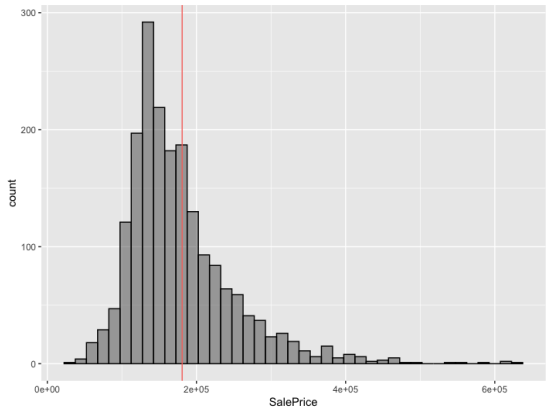
- What are the attributes of these homes that have the most distinguishing characteristics?
- How ca we for a concise set of descriptors that accurately reflects the variation in the homes, minimizing the number of individual variables?

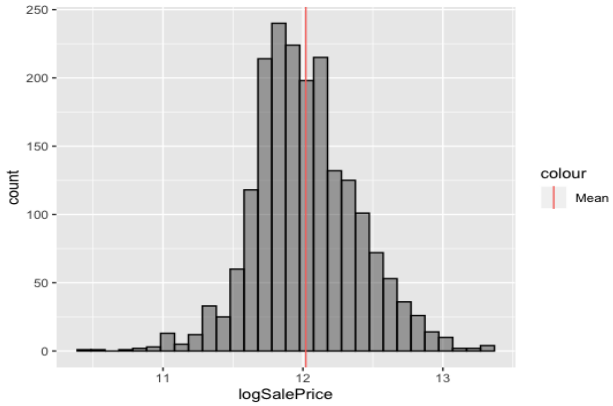
The unsupervised methods applied to the data set are as follows, Principal Component Analysis (PCA), Cluster Analysis, t-SNE, Hierarchical Clustering, and Multidimensional Scaling (MDS). Further details for each method used will be discussed in the Analysis Method section. The final step will assess the overall performance of the variable selection process through applying them to a predictive linear regression model. This will allow for a better understanding of the methodologies business practicality.

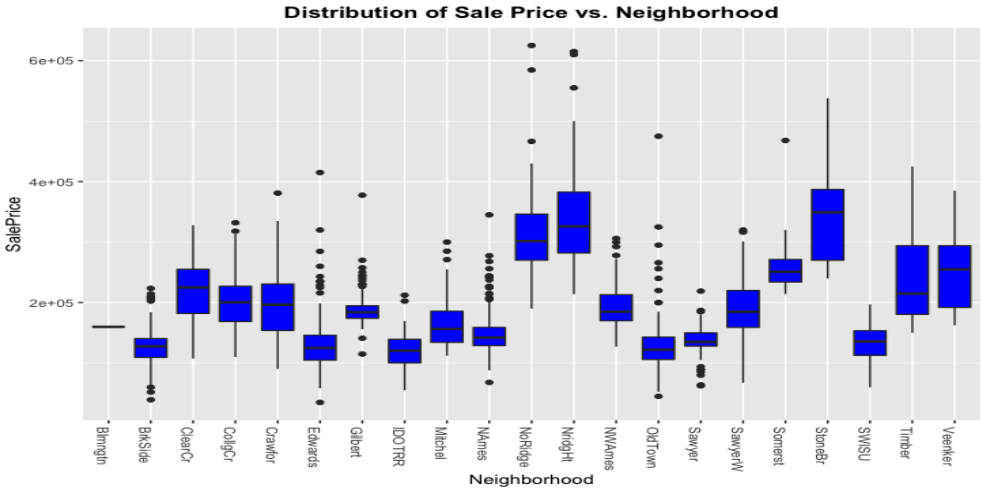
EDA Methods & Findings

The Ames Iowa Housing data set contains 82 features and 2930 rows/instances. Each row describes one home in the set area, encompassing both qualitative and quantitative measures of a home. These measures range from the Sales Price, Square Footage, number of rooms, type of façade, and sales condition to name some of them; the extensive list is included in the appendix.

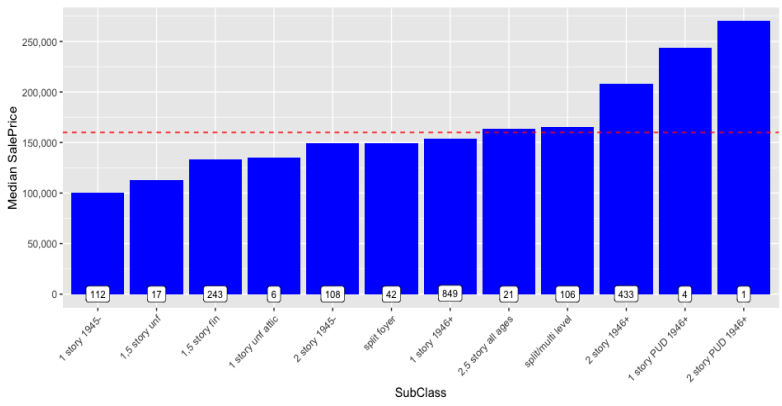
The first part of our analysis focuses on the distribution of sales prices, in its original form and in using a log transformation. As we can see below the original sales price is rightly skewed, which could be the results of the number of homes located in each neighborhood. This can be seen in the following boxplot, which compares the sales price for each neighborhood to one another. Between the original sales price chart and neighborhood breakout we can see a number of outliers that could impact the performance of our models.

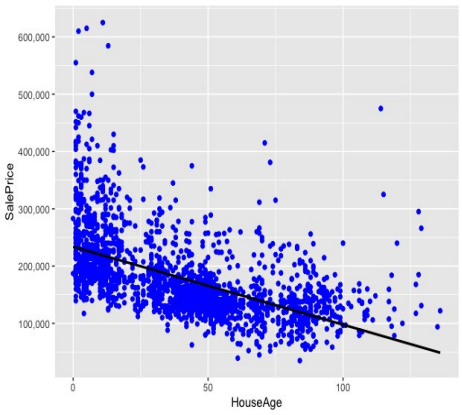




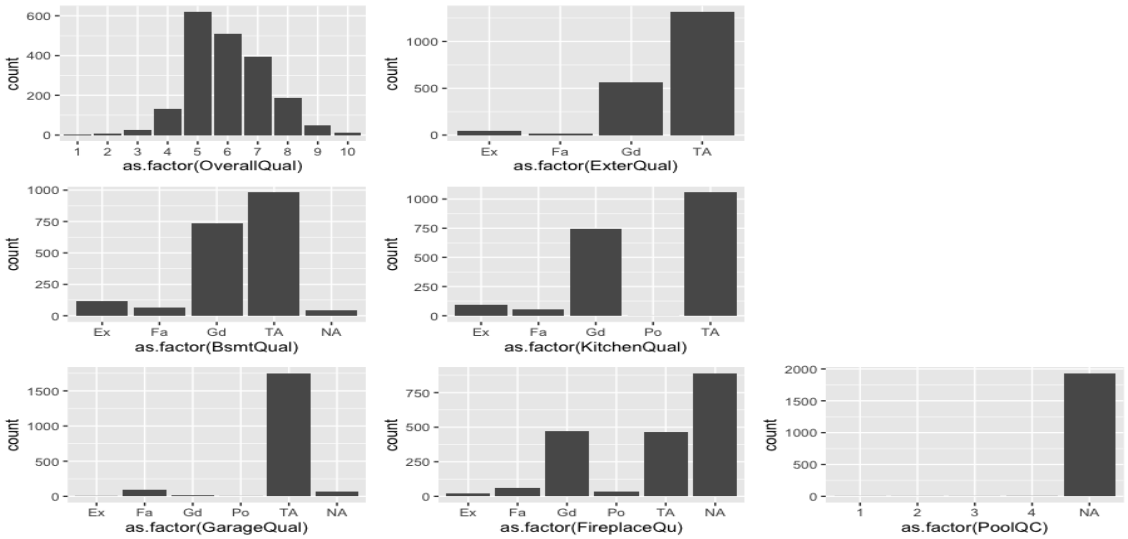


The next assessment takes place on the subclass of homes and the house age vs. sales price. The redline in the first chart indicates the median sales price for all of the homes includes in our analysis. As we can see there are three subclasses with higher median sales prices than the rest that are pulling up the overall median price. Transitioning to the next chart, we can see a downward sloping trend towards sales price and a houses age, the older the house the lower the house price.

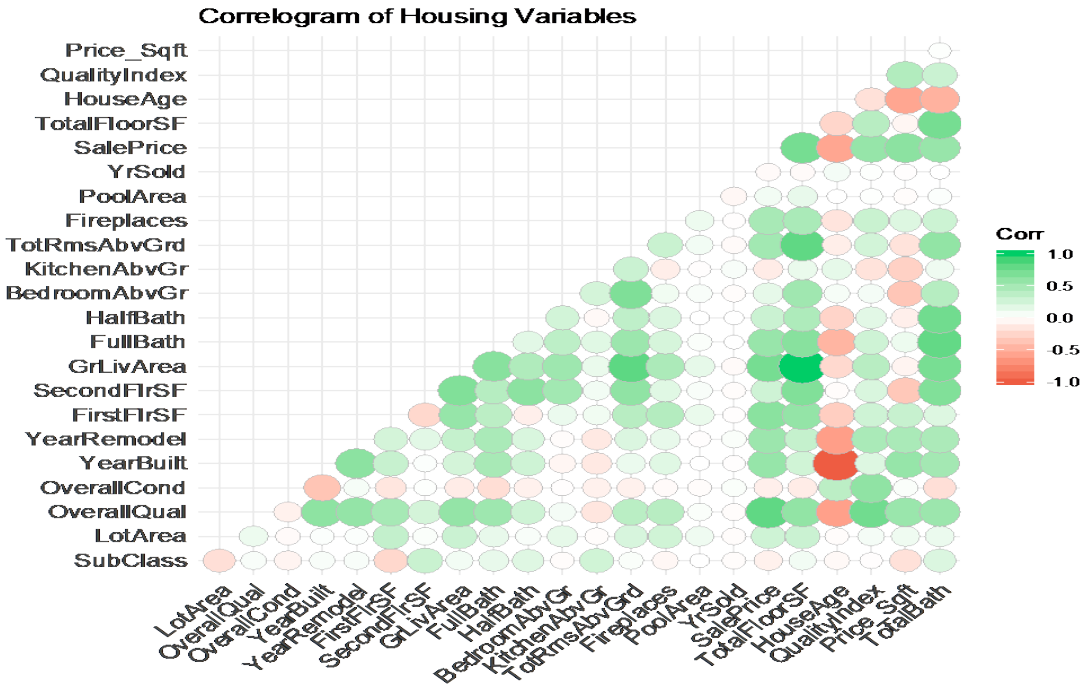




Below we look at selected group of factor variables. We can see there is not an evenly distribution across the factors



After removing numerical variables with significant overlap in explained variance a correlation plot was generated. High overlap was found in the variables describing the size of different types of porches, log sales price, month sold, and PID & SID. The chart shows a strong positive correlation for variables describing the size and overall quality of homes. There is one obvious negative correlation with year built and house age, as the older the home the higher in age it is. The dimensionality of this data set adds an additional level of complexity, as the variables here are reflective of only 43% of the variables in the data.



Data Preparation

To get a clear understanding of the potential limitation in taking a near "hands off" approach, anomalies and outliers were not assessed or removed and instead kept in the raw data set. There are missing values in this dataset. No backfill, use of the average or other forms of replacing the missing values for this iteration, instead, their impact will be assessed as the variables are incorporated into the model. The step of handling missing values for this process is of great importance for this assessment for packages used are unable to work with data that contains NA's or blank. An example of this can be seen in many of the basement variables which have observations that are either blank or NA's. For this assignment we interpreted blanks and or NA's as homes that do not have a basement as a home for sale does not always have one. We did first look to remove instances which appeared to be duplicates and were not complete. This adjustment resulted in the reduction of 658 instances and a net reduction of five features, bringing the structure of the data to 2272 rows and 77 variables.

Six new features were created for this process, they are as follows:

New Feature	Feature 1	Calculation	Feature 2
TotalFloorSF	FirstFlrSF	+	SecondFlrSF
HouseAge	YrSold	-	YearBuilt
QualityIndex	OverallQual	*	OverallCond
LogSalePrice	SalePrice	Log()	
Price_Sqft	SalePrice	/	TotalFloorSF
TotalBath	FullBath	+	HalfBath

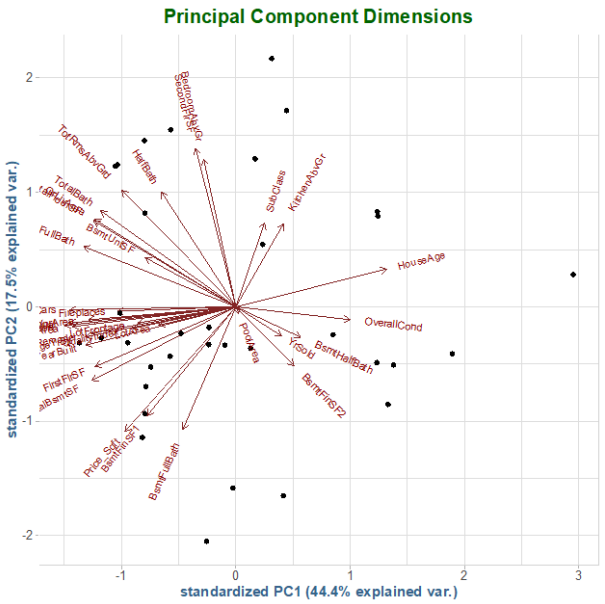
These variables were created with the goal of bringing better performance to the models. In addition to the new features above, groupings of the housing Sales Price and Quality Index were developed. The reduction in values should provide a more robust model in the manner of providing a performance measure for not only the overall model but in our designated groups. Does the model have better predictive capabilities for a specific subset or are the results consists throughout?

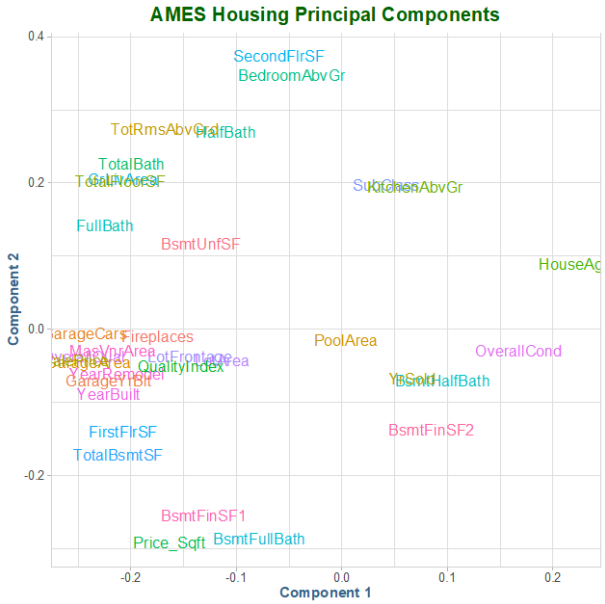
Analysis Method & Preliminary Results

The following portion of the report will be broken out by the unsupervised method used.

PCA

With a large number of numerical features, the use of PCA allows for the test in the ability to reduce the dimensionality of data and assist in the discovery of potential patterns. In allowing for the algorithm to complete the process, we allow for discovery of similarities that would not have been immediately apparent through a manual application. Doing so showcased a great deal of attribute clustering within the cosmetic attributes. This can be seen below in both formats of the PCA charts.



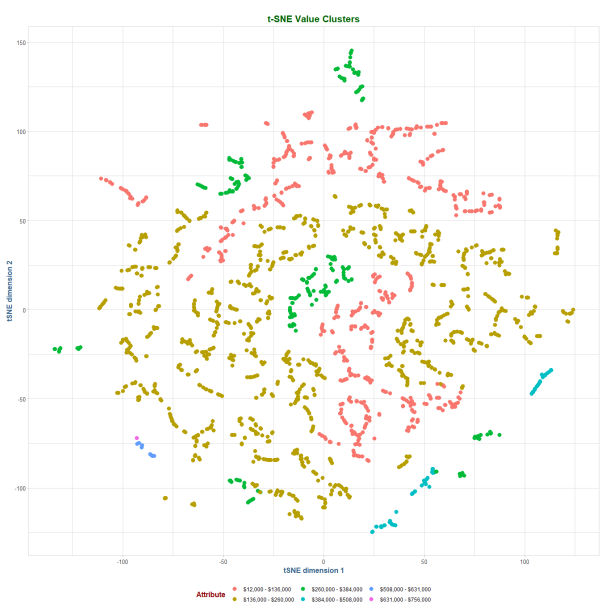


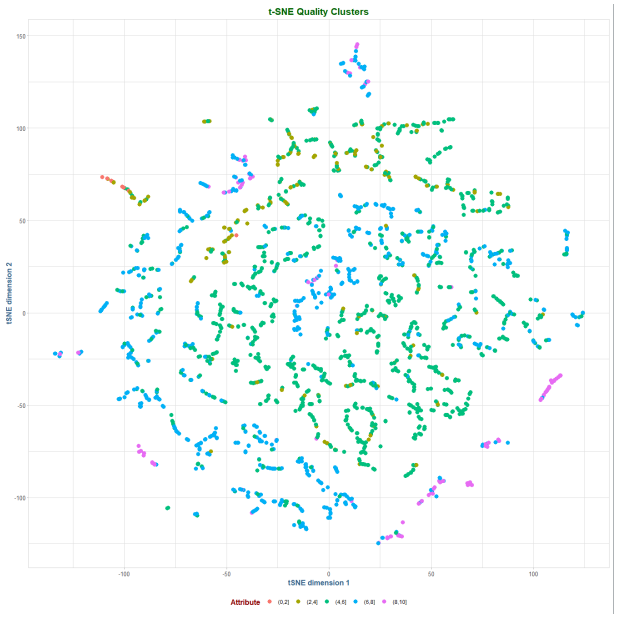
Combining the above charts with standard deviation for each component we can see that in order to explain 80% of the variation within the data eight components should be included. This determination comes from looking at components with values greater than one which are considered significant.

t-SNE

Using t-SNE analysis, we can see clear separation of attributes into similar clusters. In our principal component analysis, we also observed similar clustering by attribute categories. While PCA designated eight potential clusters, it appears t-SNE creates five to six. The below chart was created with the following dimensions, perplexity = 3.5, learning rate = 35, and iterations 5,000. Here we see more disperse groupings, however, the similarities in the attributes remains strong.

Continuing our t-SNE analysis, the homes exhibit some clustering characteristics when looking at the quality and value attributes. Homes with similar quality and value can be found in distinct areas of the plots, indicating there are similarities in homes with comparable quality and value characteristics.



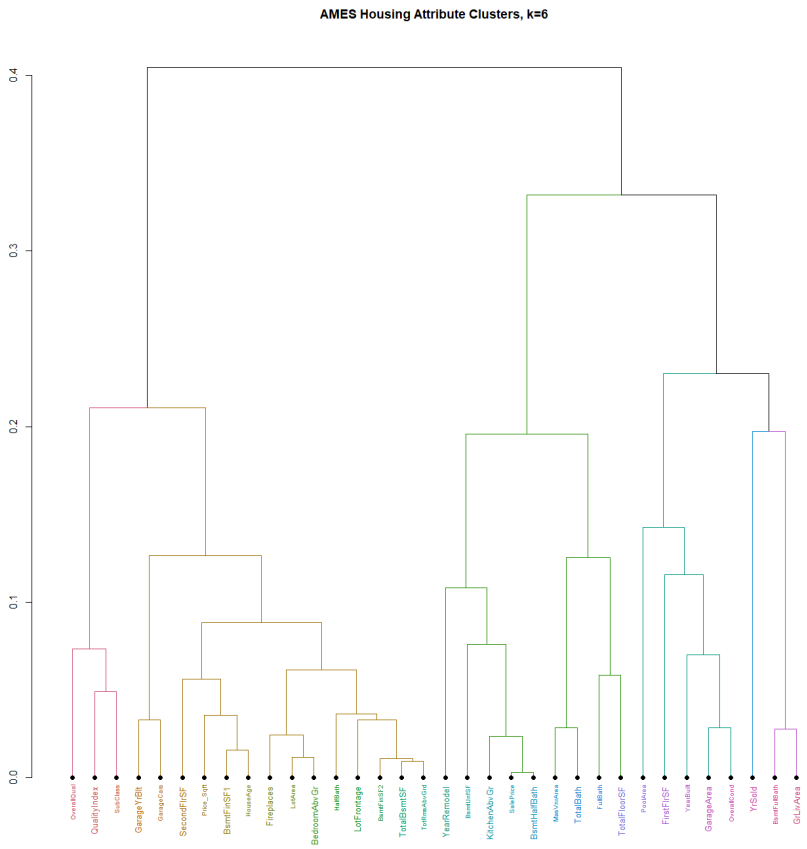


The chart above offers a visually appealing representation of the data and variation, but a clear distinguishable clustering is lacking as the attributes overlay each other.

Hierarchical Clustering Analysis

Using hierarchal cluster analysis, we can see with six clusters that the home attributes fall into similar groupings that we saw with both principal components and t-SNE.

There are clear similarities in variables with cosmetic, temporal, lot/land and size specifications.

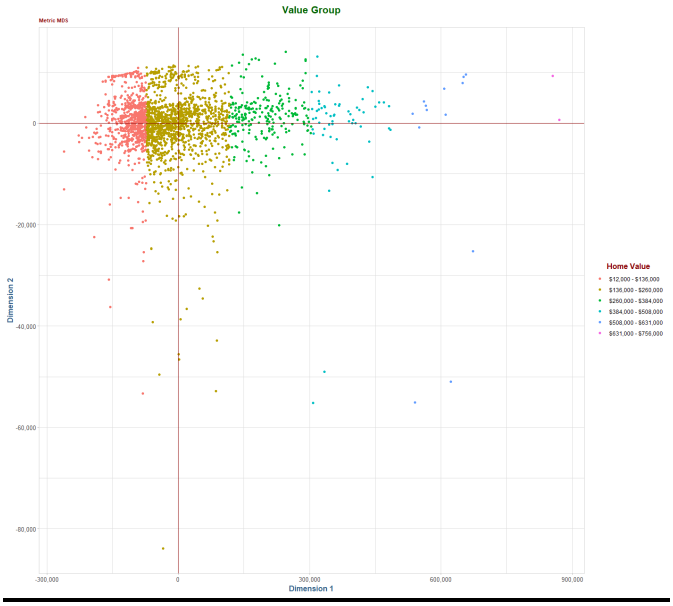


This hierarchical cluster analysis was performed using the average method, which seemed to yield the most accurate clusters based upon the temporal/cosmetic/size/cost variable groupings we have seen in previous techniques. We also tried ward.d, complete and McQuitty methods before setting upon the average method.

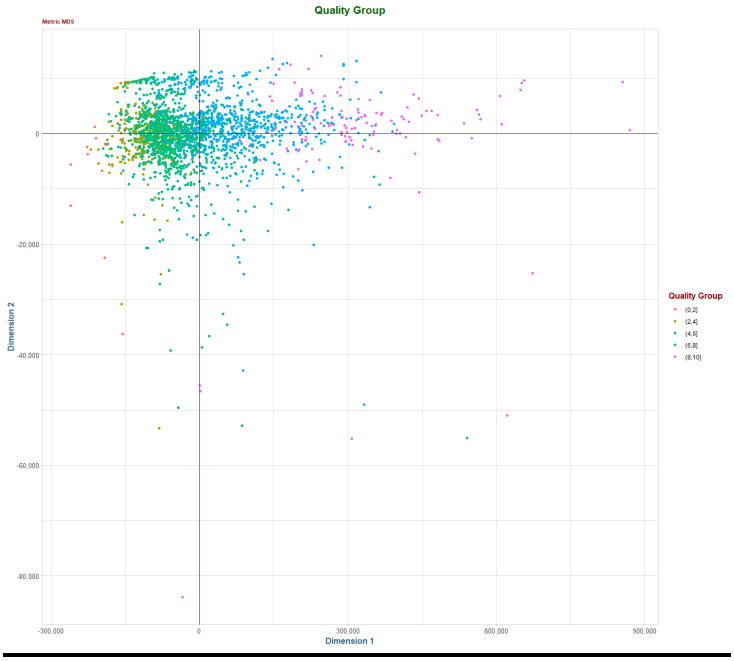
The results of the hierarchical cluster analysis were the most disappointing in all our modeling techniques. We noted the results of the analysis, and moved on to other techniques, ultimately choosing to discard the results of this analysis in favor of more robust methods that provided further insight.

Multidimensional Scaling

The final step looks at the use of Multidimensional scaling. With this method we see relatively clear lines of distinction between homes that fall into a given value grouping. There are some outliers, however, this is the most distinct and unambiguous separation we have seen in the home value categories.

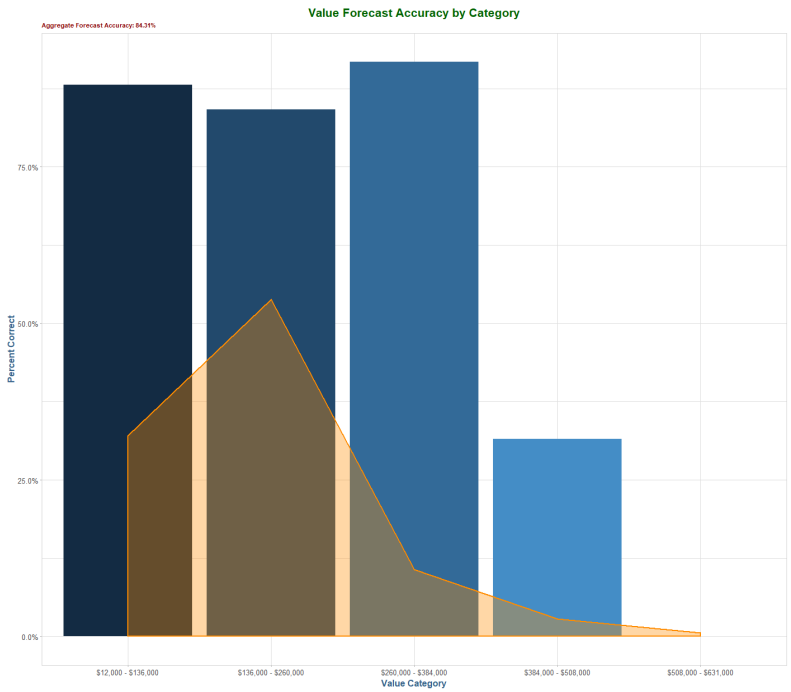


Applying the same technique to the quality groups, we continue to see visible lines of clustering, however, they are far less distinct and obvious than with the value group we saw previously. The quality group is quite persistent in the middle range of homes, in the 4-8 range.



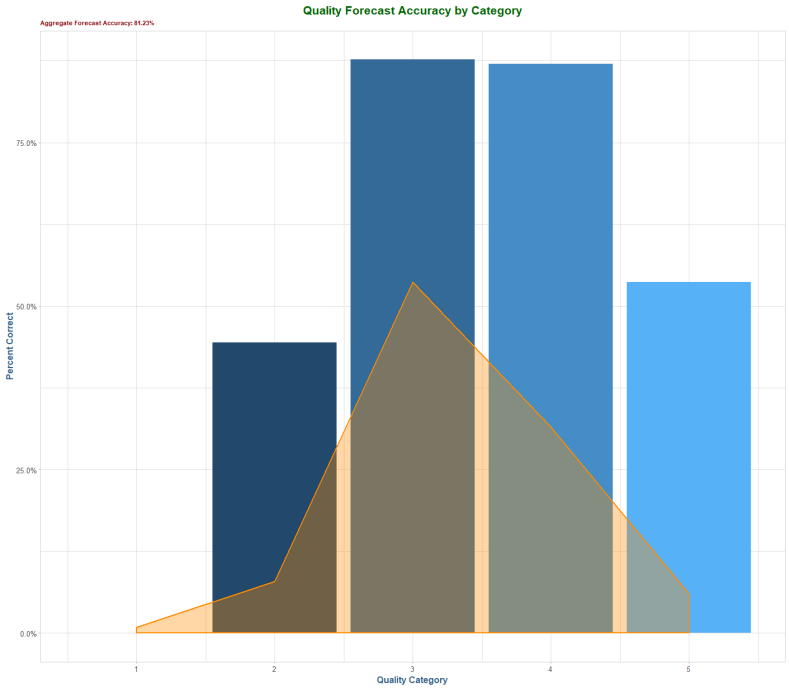
Final Model Presentation

For our final steps we applied all findings and learnings from the previous asessments to test the accuracy of creating a predictive linear model for the Ames Housing data set. Due to the ease in interpretation and a clear distinguishabilty between groupings, we used the results for the PCA model to generate the inputs. We noted that using eight components could explain approximately 80% of the variance in the data. In our test sample, using a 70/30 split, we were over 84% accurate in our predictions using this reduced space. These are impressive results. The data set contained extensive variations in such areas as how the home was sold, residential/commerical zoning, and density and without separting these out the model generated promising results.



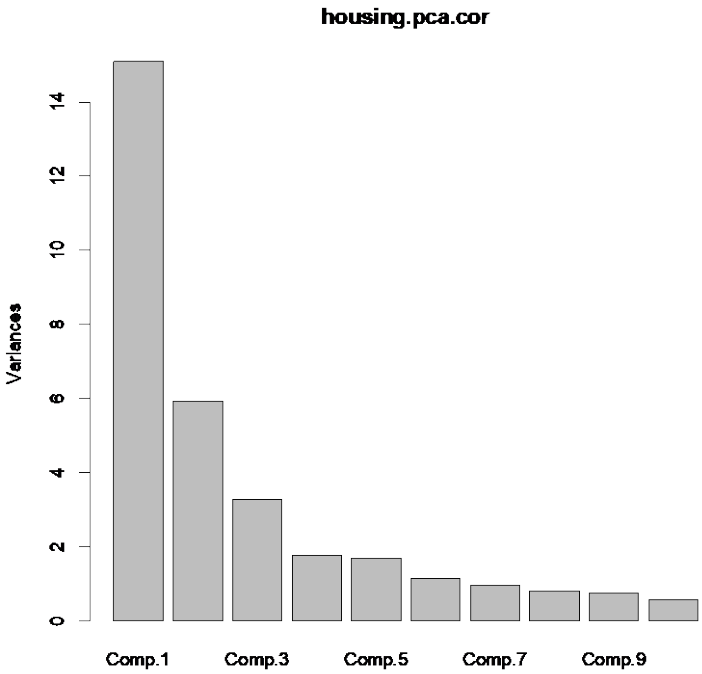
The predictiability of the model shows favorable results with the first three value categories. This could be a result of the high frequency in the data set and or within these groups there are a number of outliers, which were highlighted during the MDS method.

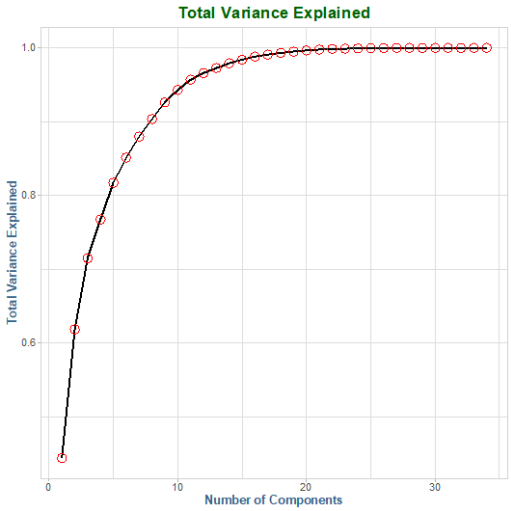
For a final round of analysis we derived a model using the quality attributes. Based on accuracy score this model provided slightly lesser results than the value groupings, however the distinguishabilty between the categories is not as strong.

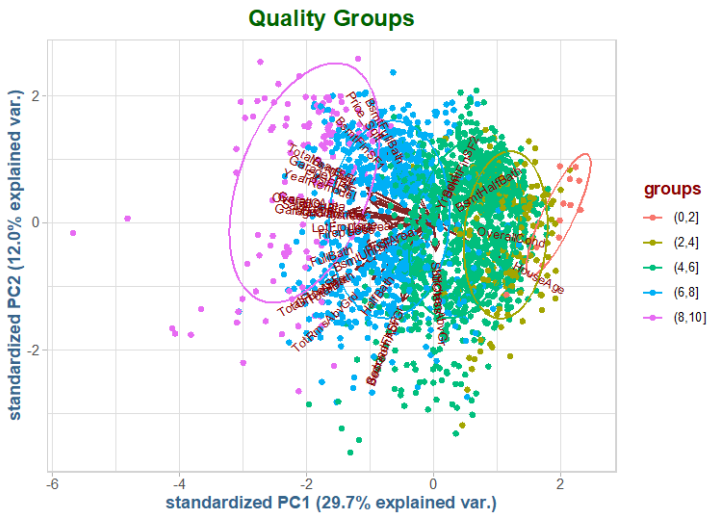


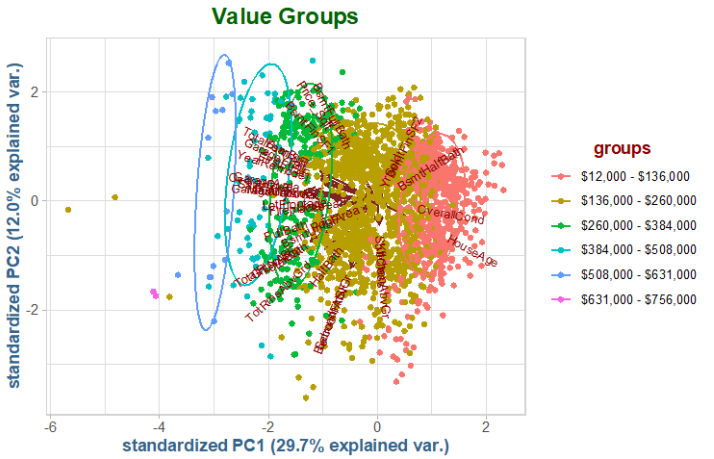
Conclusion and Reflection

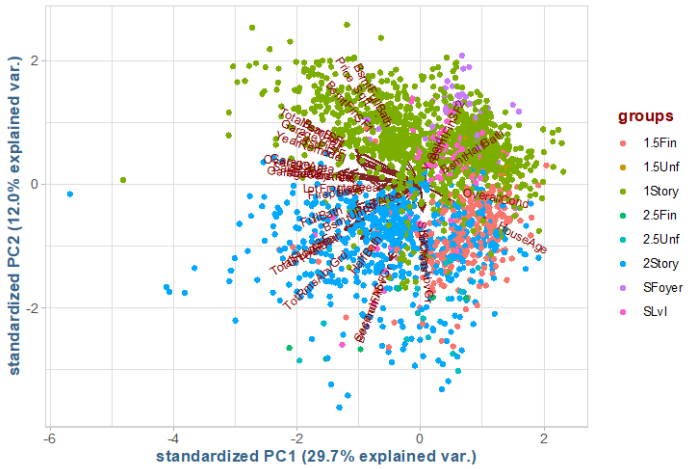
APPENDIX


housing.pca.cor
Variances
0
2
4
6
8
10
12
14
Comp.1
Comp.3
Comp.5
Comp.7
Comp.9


Total Variance Explained
Total Variance Explained
1.0
0.8
0.6
0
10
20
30
Number of Components


Quality Groups
standardized PC2 (12.0% explained var.)
standardized PC1 (29.7% explained var.)
groups
(0,2]
(2,4]
(4,6]
(6,8]
(8,10]


Value Groups
standardized PC2 (12.0% explained var.)
standardized PC1 (29.7% explained var.)
groups
$12,000 - $136,000
$136,000 - $260,000
$260,000 - $384,000
$384,000 - $508,000
$508,000 - $631,000
$631,000 - $756,000


standardized PC2 (12.0% explained var.)
standardized PC1 (29.7% explained var.)
2
0
-2
-6
-4
-2
0
2
groups
1.5Fin
1.5Unf
1Story
2.5Fin
2.5Unf
2Story
SFoyer
SLvl