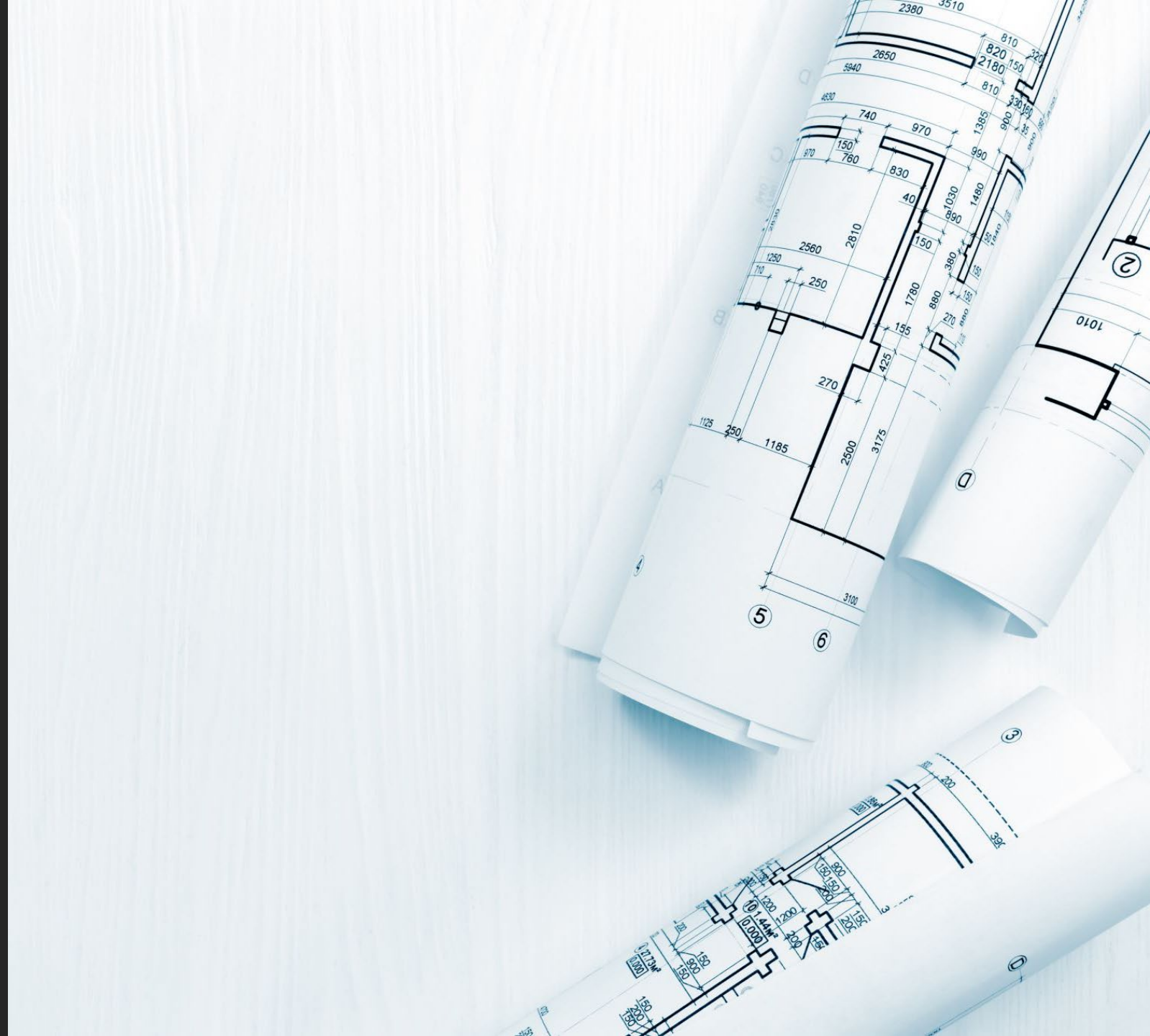
AMES Housing Survey

MSDS 411 – UNSUPERVISED LEARNING

BRANDON MORETZ &
SEAN PRENTIS



Executive Summary

The goal of this presentation is to summarize our findings in the AMES Iowa housing data set in order to answer the following questions:

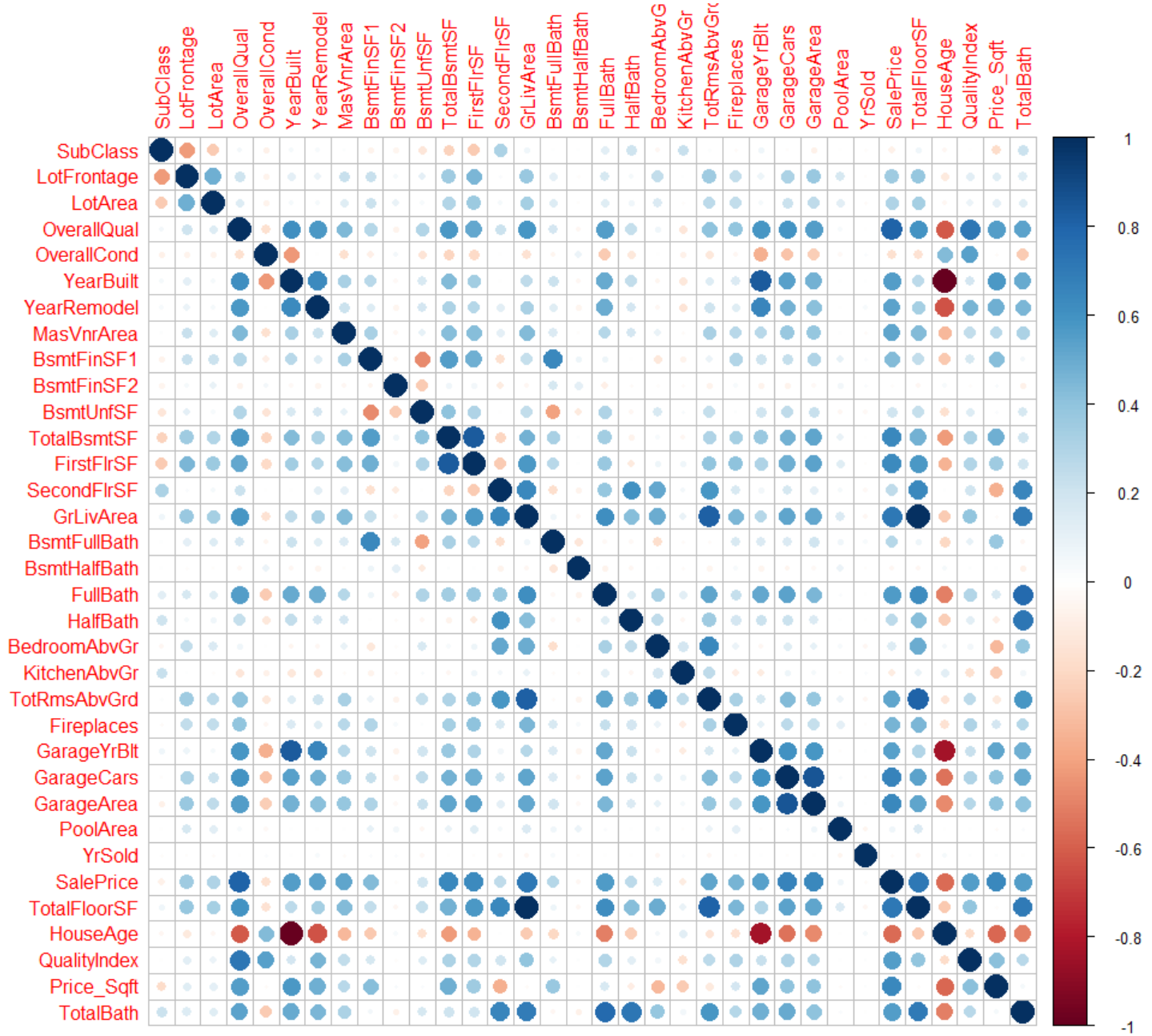
1. What are the features of these homes that have the most distinguishing characteristics?
2. How can we form a concise set of descriptors that accurately reflects the variation in the homes, minimizing the number of individual variables?

To help answer these questions, we derived two bespoke variables do give a holistic representation of these qualities. They are two categorical variables that define quality and value and will be used extensively throughout this analysis.

Variable Correlations

We started our analysis by surveying the numerical variables in the data set by computing the correlations between them and using a standard correlation plot.

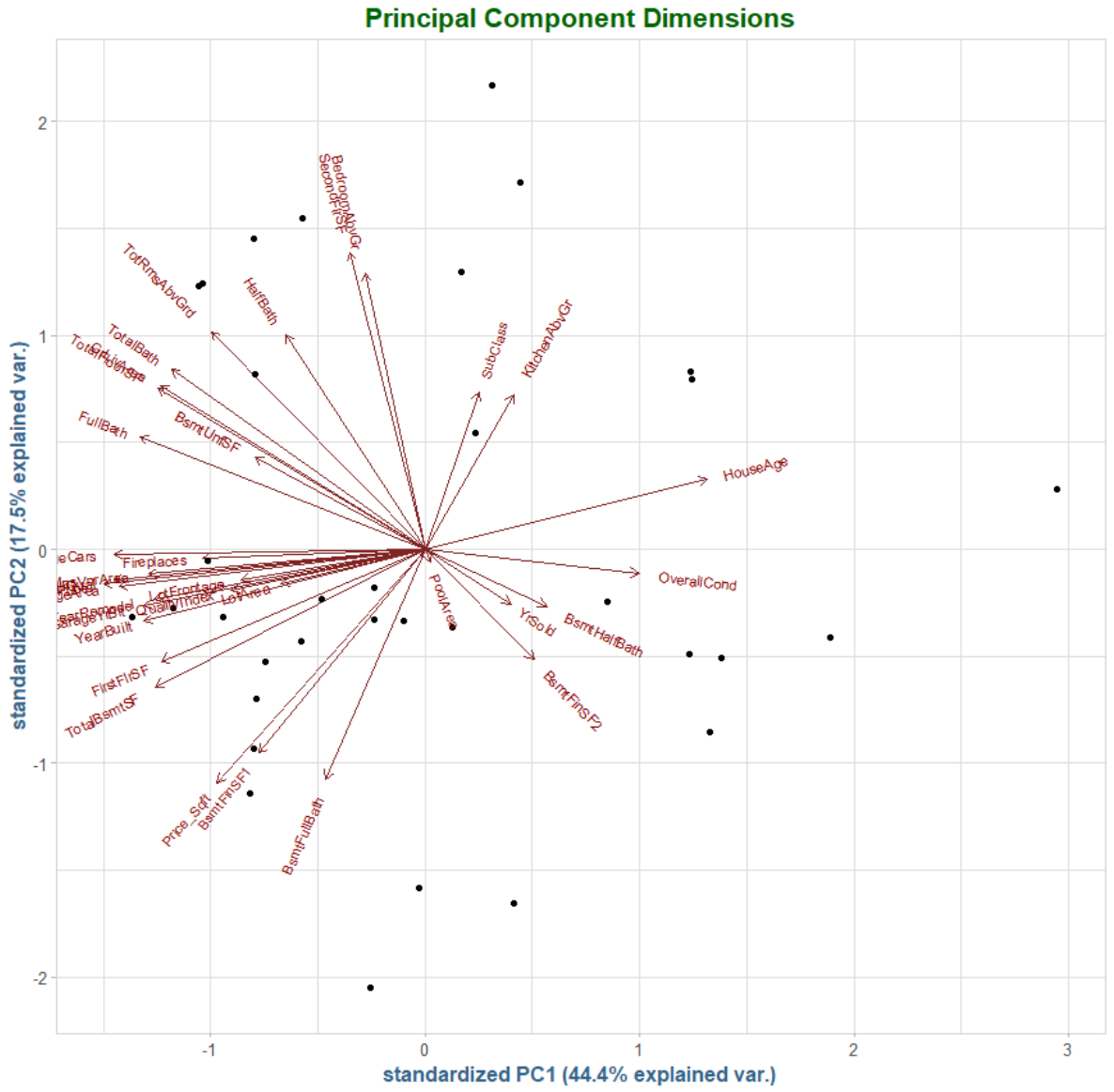
The dimensionality of this data set adds an additional level of complexity, as the variables here are reflective of only 43% of the variables in the data.



Principal Components Analysis

Due to the high-dimensionality of the data set, we employed a principal component analysis as a dimensionality reduction technique.

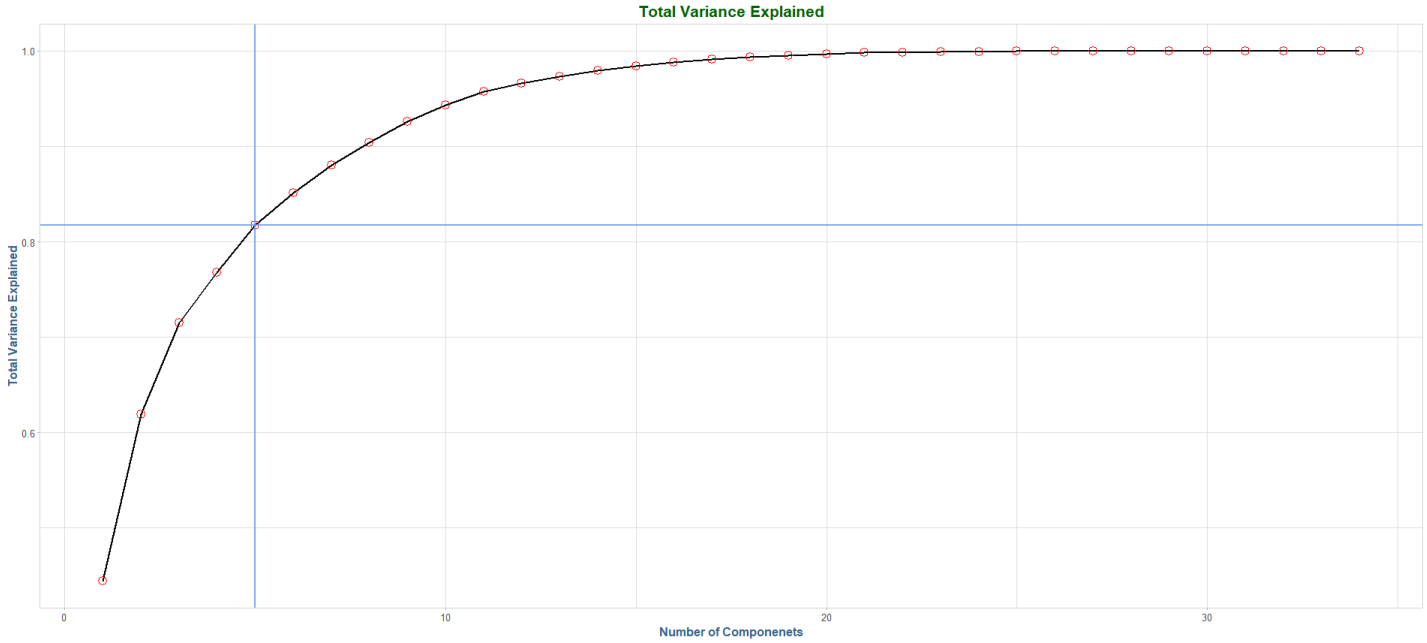
We notice a great deal of attribute clustering with the cosmetic attributes (Fireplaces, Masonry Veneer, Garage, Living Area, etc.)



Principal Component Analysis, Continued.

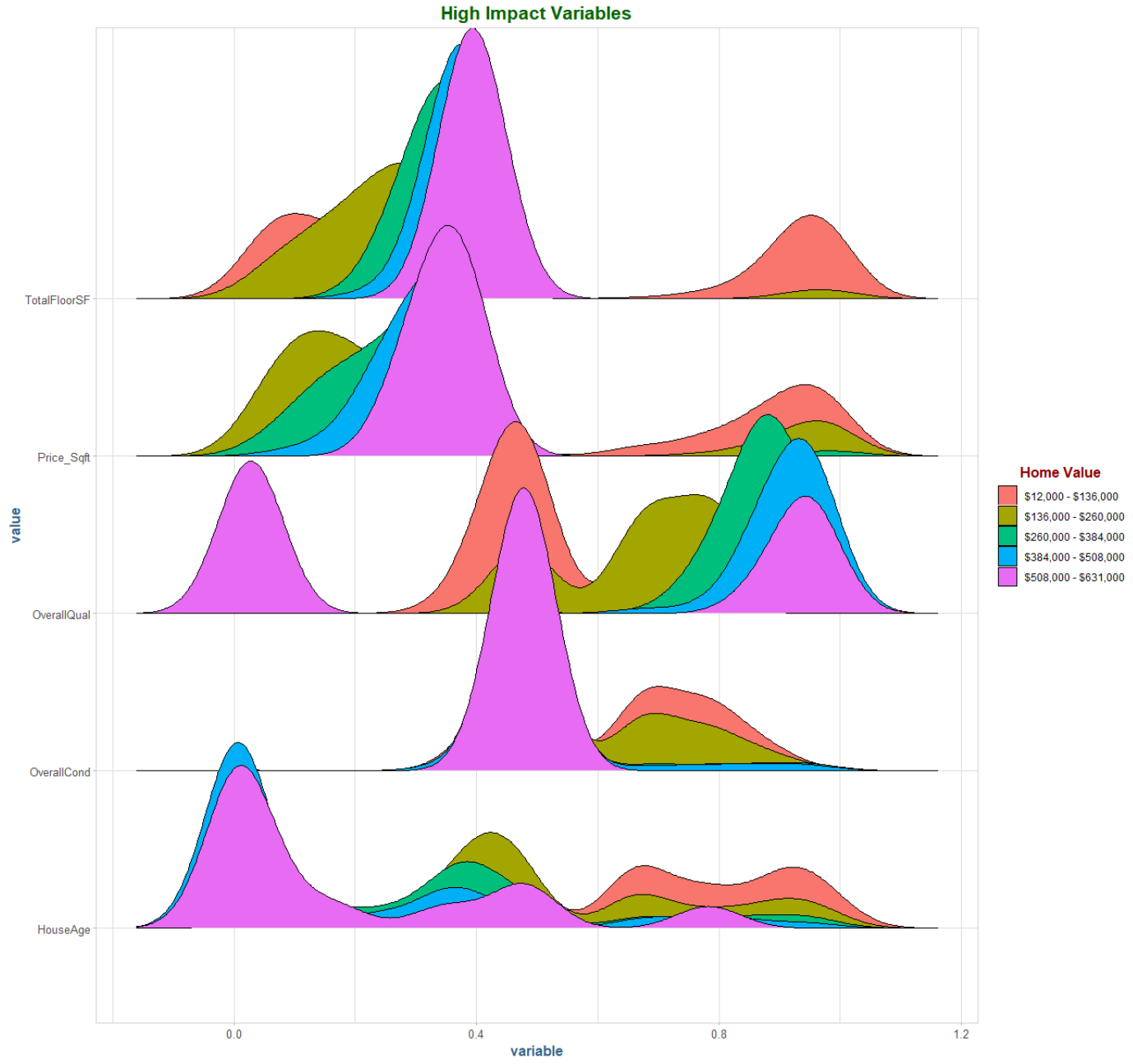
Above, we see the attribute clustering in a slightly different angle, and color-coded by type (cosmetic, temporal, lot/land, house size, quality).

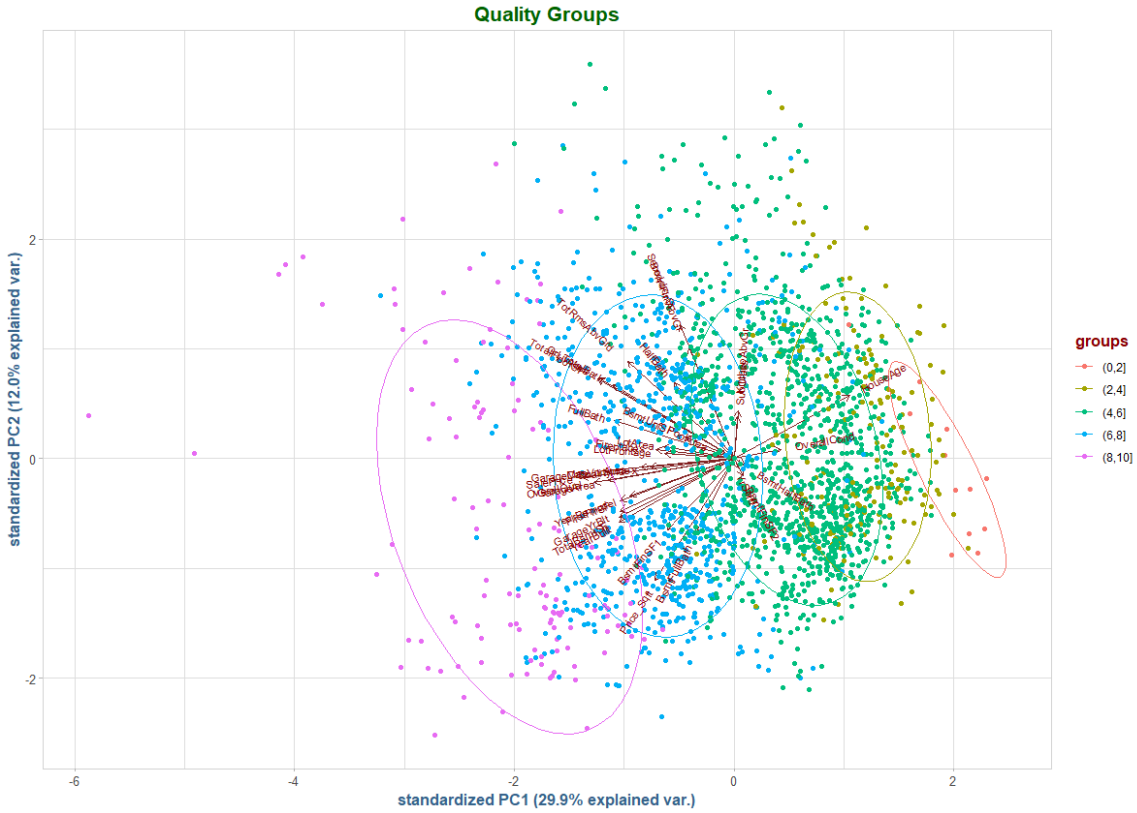
In using these components for further analysis, we set the cut-off at 80% of the explained variance, which is contained within the first 5 components.



HIGH IMPACT VARIABLE DISTRIBUTIONS

These variables are orthogonal in the principal component space. Combined, they span most of the area in the reduced space, so they exhibit a great deal of explained variance in the data set.





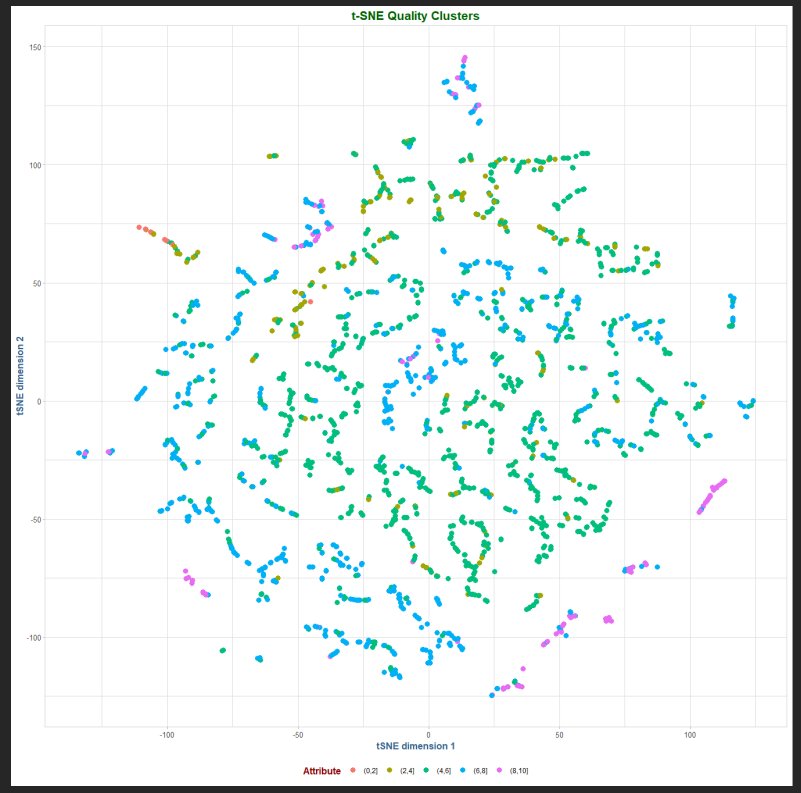
Clusters of homes by value and quality attributes.

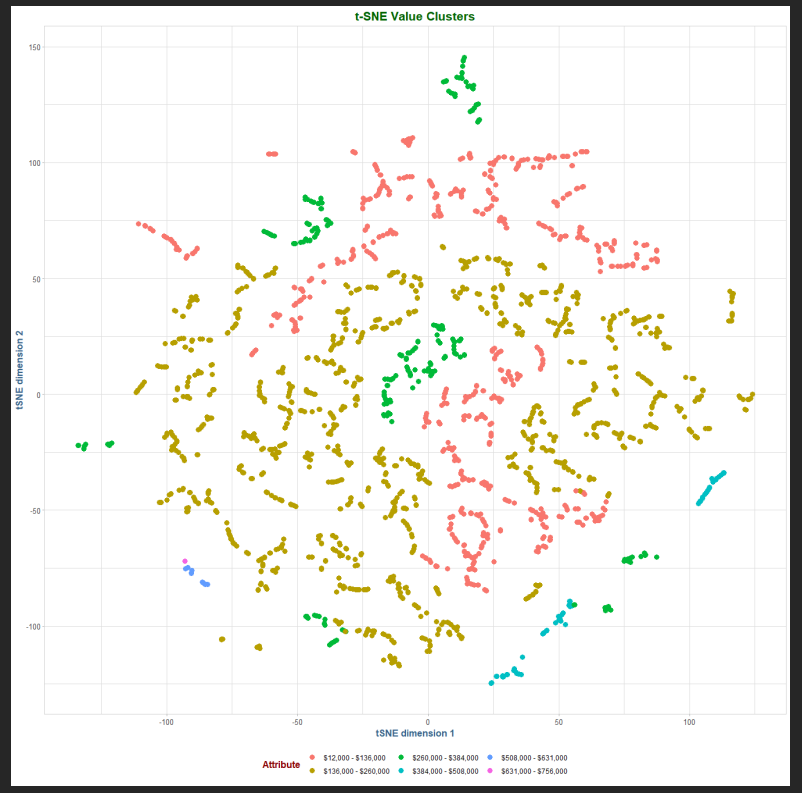
t-SNE Analysis

Using t-SNE analysis, we can see clear separation of attributes into similar clusters. In our principal component analysis, we observed similar clustering by attribute categories. Here we see more disperse groupings, however, the similarities in the attributes remains strong.

t-SNE Analysis, Continued.

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 similar quality and value characteristics.

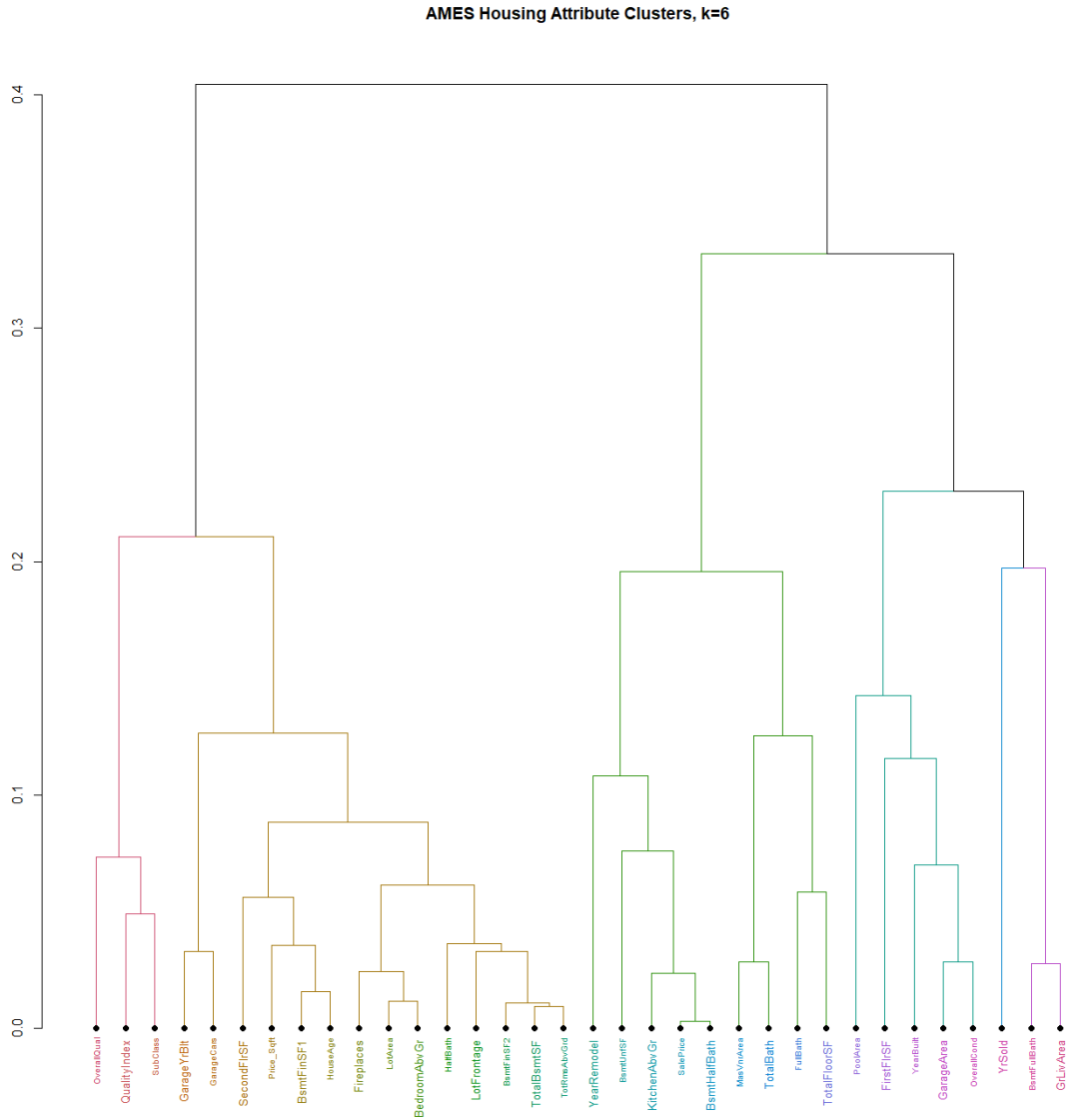




Hierarchical Clustering Analysis

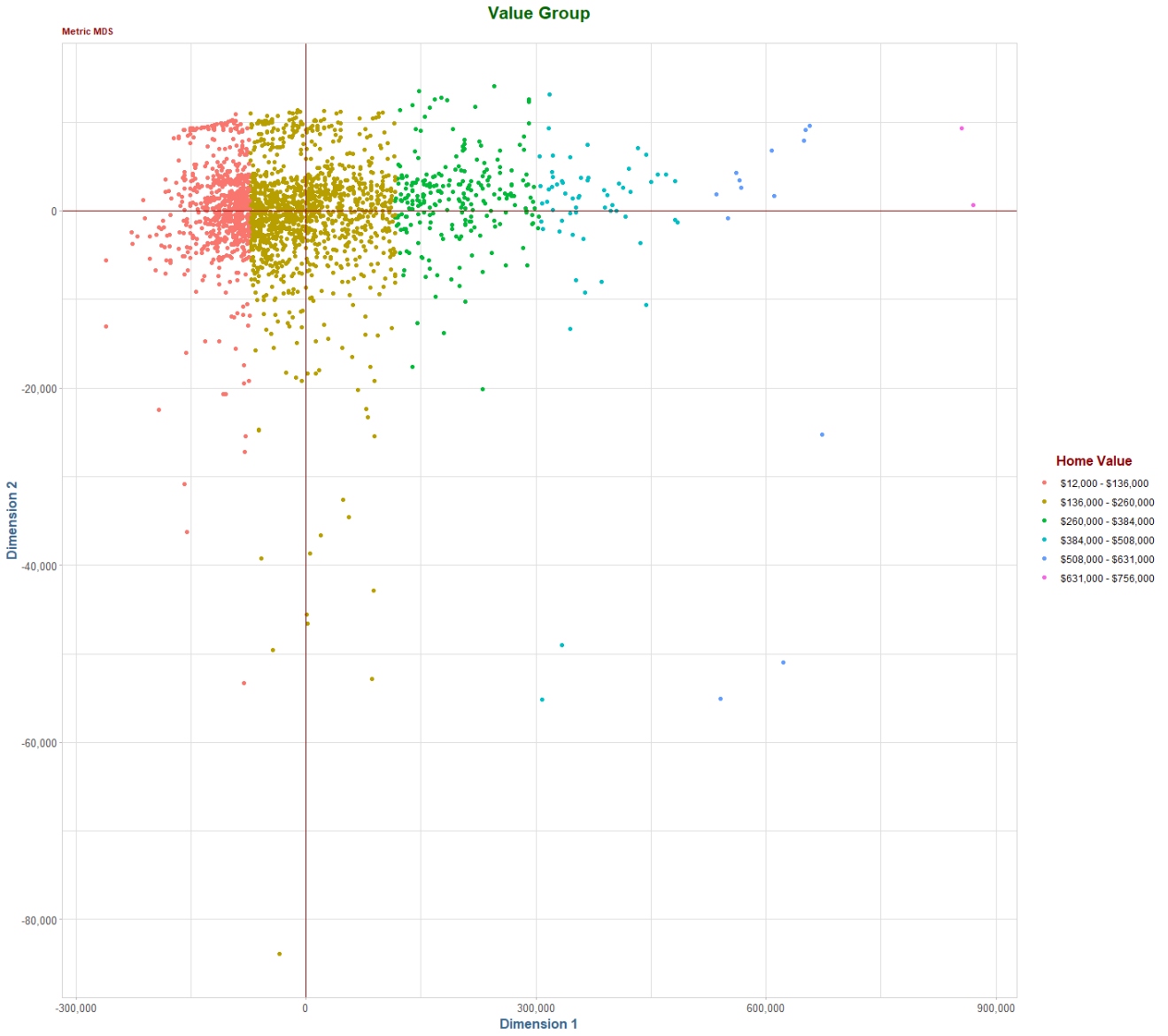
Using hierarchal cluster analysis, we can see with six clusters that the home attributes fall into similar groupings that we saw with both principal components and t-SNE.

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



Multidimensional Scaling

Using multidimensional scaling, 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.



Multidimensional Scaling, Continued.

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.

Conclusions & Further Analysis

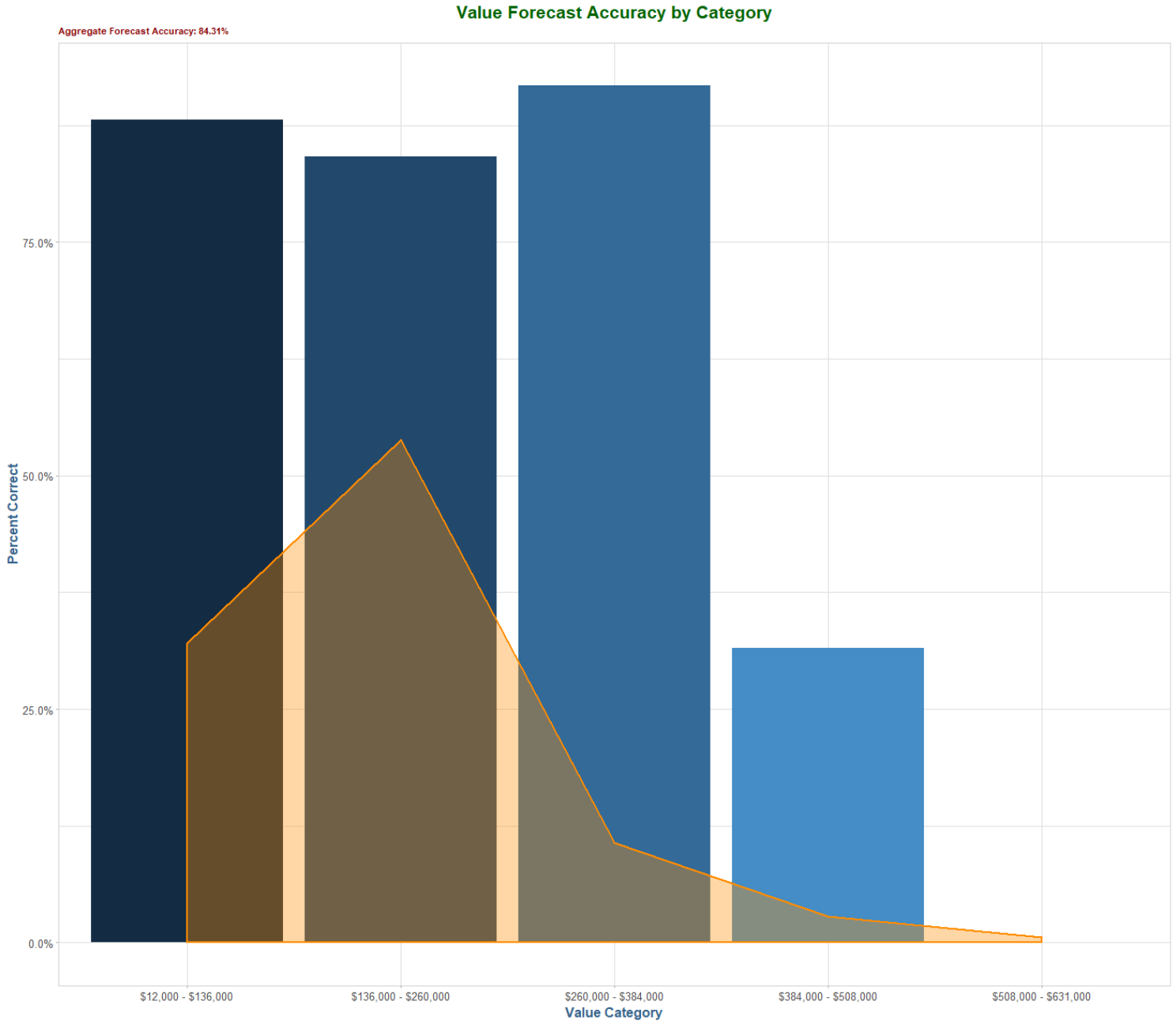
We started this analysis with two basic questions:

1. What are the characteristics of these homes that have the most distinguishing characteristics?
2. How can we form a concise set of descriptors that accurately reflects the variation in the homes, minimizing the number of individual variables?

We saw several techniques that visualized both the clustering behavior of the individual attributes based upon the underlying attribute type of the variable, such as temporal, cosmetic, lot/area and size.

We also observed the clustering of homes that have similar value and quality metrics. The value attribute was more pronounced, especially in t-SNE and MDS. However, the quality metric was quite persistent in its clustering behavior.

Can we use this information to simplify further analysis and categorization?



Predicting Value

We started this analysis by using principal components to reduce the dimensionally of our data set. We noted that using only eight components we could explain over 80% of the variance in our data.

We used these eight components to derive a regression model to predict the value of a home based upon these components.

In our test sample (70/30 split), we were over 84% accurate in our predictions using this reduced space, which is quite impressive.

Predicting Quality

Further, we used the same components to derive a predictive model for the quality attribute.

As we saw in many of the previous analysis, the separation of quality was not quite as strong as value. Thus, while the model is overall quite successful, we were not able to achieve the level of accuracy seen in the value category.