MBAS903: BUSINESS ANALYTICS FOR ECONOMICS AND MARKET ENVIRONMENTS

Assessment 2

Abstract

Analytics report for grouping of housing characteristics for Ames Housing

Medha Luthra
Chinmay Datar
Shruti Jawale

Executive Summary

Like every industry these days, the real estate industry too collects data and can potentially expand and grow its business rapidly and innovatively by harnessing data analytics tools. Ames Housing, over the course of 4 years recorded the characteristics and the details of the sale of 300 houses. Our task in this report is to identify ways in which these houses can be categorised based on their similarities. K-means cluster analysis techniques have been applied to complete the analysis.

This report utilises the variables pertaining to size of the house, the age, quality of materials and finishes, and the number of bathrooms and identifies 5 relatively unique clusters of houses. These clusters can be seen to have a distinct relationship with the sale price of the home, implying their relevance and applicability for pricing in the future

Table of Contents

Introduction

The real estate industry plays a key role in every economy, the extent of the industry makes it very appealing to investors. In Australia, real estate has a market size of over 32 Billion dollars and has increased faster than the economy overall. The market size of this industry has grown 1.5% per year on an average in the last five years, in Australia. Reports suggest that the industry revenue is expected to grow consistently over the next 5 years, the steady rise in residential housing prices and housing transfers will presumably support the industry in this period. The key factor that affects the house pricing, purchases and overall industry revenue is positive consumer sentiment.

Property sale prices are impacted by a myriad of factors. Some of these factors are unrelated to the property itself, like supply and demand, conditions in the economy, interest rates on housing loans, the prominence of the location and demographics of the area, all of which vary over time. The macro-environmental fluctuations mean that the price at which the property sells in two consecutive years, even when there is no damage or improvements made during that time, may be vastly different. However, at the same point in time, and in the same area, the sale value of the property is also moderated by its size, the presence of architectural features like extra bathrooms, parking spaces, garages and outdoor spaces and their sizes, the quality of materials used and the functionality of the property's layout. The relationship of these features with the sale price of the house is more tangible and is controllable to some extent by property developers, architects, homeowners, estate agents, or asset managers.

Ames Housing group sold 300 houses over the period 2006-2010 and recorded the characteristics of these properties. The aim of this analysis is to identify distinct groups of houses based on their characteristics, through a cluster analysis. The use of clustering to divide a dataset based on similarities can help the organization not just to develop pricing strategies, but also to identify and reach market segments for different types of homes.

Methodology

The dataset provided includes data describing the house and its sale. The purpose of the analysis is to classify houses based on their inherent features. For this reason, variables that pertain to the sale have not been considered in the analysis.

Selecting Variables for the cluster analysis

In order to select variables that would be relevant for the cluster analysis, we made certain assumptions about the considerations of buyers and divided the variables into broad topics

1. The size of the house
2. The condition of the house
3. The amenities of the house
4. The aesthetic features of the house

For creating clusters, we decided to focus on the features that describe the size and condition of the house- as these considerations are more immediate when selling or buying a house. Buyers may have their own preferences for amenities and aesthetic features, within those seeking the same size.

The variables explored for the clustering are:

Variable Name	**Type**	
Heating Quality and condition	Category	Condition
Presence of central air conditioning	Category	Amenity
Style of Dwelling	Category	Aesthetic
Above Grade (Ground) Living area square feet	Measure	Size
Age of house when sold	Measure	Condition
Basement area in square feet	Measure	Size
Bedrooms above grade	Measure	Size
Lot Size in square feet	Measure	Size
Number of Full Bathrooms	Measure	Amenity
Number of Half Bathrooms	Measure	Amenity
Overall Condition of the House_final	Measure (calculated)	Condition
Overall Material and Finish of the house_final	Measure (calculated)	Condition
Size of Garage in Square feet	Measure	Size
Total Area of Decks and Porches	Measure	Amenity
Total number of Bathrooms (half bathrooms counted 10%)	Measure	Amenity

Table 1: Variables explored for clustering

Note: The dataset includes 2 columns representing a rating of “Overall condition of the house’. We have calculated an average of these as the ‘Overall Condtion_final’ variable. The dataset includes 2 columns representing a rating of “Overall material and finish of the house. We have calculated an average of these as the ‘Overall material and finish of the house_final’ variable.

Descriptive Analytics

We examined the variables using SAS visual analytics, and explored the mean, median, mode, central tendency, variation and the shape of the graph for every measure in the dataset. We also explored the distribution of the categorical variables.

To identify the variables that could be most effectively used for the cluster model, we created a correlation matrix using all the measures in the data set except those pertaining to sales, i.e. Sales Price in Dollars, Sales price> 175000, Year of Sale, Season when House sold, Month Sold(MM).

Cluster Analysis

The cluster analysis has been completed using SAS Visual Analytics Cluster Analysis programming. The algorithm uses K-means clustering with a default k-value of 5. In order to get a reasonable understanding of the types of houses we iterated the combinations of variables, checking the centroids and fit of the clusters in each attempt, with one variable belonging to each broad topic, for multiple numbers of clusters (k).

To derive an approximation of the ideal number of clusters for the variables in the model, we used the elbow method.

The derived clusters were ultimately examined with the variable of interest- 'Sales Price in dollars', to determine the range each type of house typically sold for in the period 2007-2010. A box plot was used to visualise this relationship as this graph allows us to view the price as a range and also view houses in the cluster that are outliers.

Results

Based on the explorations and iterations of clustering within the dataset, the variables finalised for the clusters were:

Above grade living area
Age of the house when sold
Number of Full Bathrooms
Overall Material and finish of the house

Descriptive Analysis of the variables

We investigated the characteristics of the measures selected for the clusters. This includes central tendency, standard deviation, variance, skewness and coefficient of variation. Central tendency of the measure is determined by the skewness of the histogram. For a symmetric distribution of a measure the mean is taken as the central tendency while for a positively skewed histogram median is considered to be a better estimate of central tendency. The characteristics of the chosen measures are in table 1.

	Above grade living area	Age of the house when sold	Number of full Bathrooms	Overall material and finish of the house
Mean	1130.74	45.89	1.62	5.39
Median	1135	45	2	5
Mode	864	4	2	6
Coefficient of Variation	50.57	59.8801	39.5737	17.3497
Standard Deviation	232.65	27.48	0.44	0.93
Variance	54,125.74	754.98	0.66	0.87
Skewness	-0.3905	0.1953	0.5397	-0.4837
Kurtosis	-0.3328	-0.5449	-0.3946	-0.1433

Table 2: Measure details for selected variables

Age of the House when sold

The distribution we observe that the average age of the house when sold is around 45 years. It also depicts that the measure 'age of the house sold (in years)' is right skewed.

Number of Full bathrooms

The average number of houses with full bathrooms is 5.39. Whereas the histogram for this measure illustrated that it is right skewed.

Overall material and finish of the house

The distribution of 'overall material and finish of the house' is left-tailed. Average rating for a house in the data set is about 5.5 with modal rating for the houses in the dataset being 6.

Above grade living area

The distribution of 'above grade living area' is skewed left. The average living area that is above grade in most houses is 1130.74 sq. ft.

Correlation Matrix

After looking at the characteristics of the measures we find the correlations between the measures. We take all the measures in the dataset except the sales price to eliminate the measures that have high correlation with other measures.This is done as using two highly

correlated variables creates confusion in the cluster algorithm, redundancy in the cluster and increases time to process the data to form the clusters.

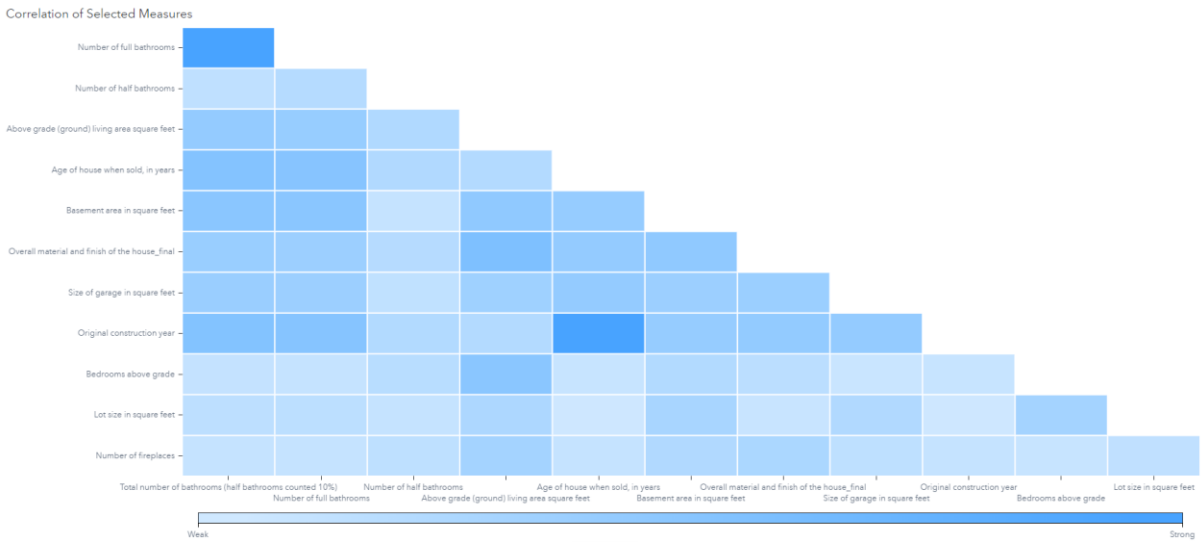

Figure 1: Correlation Matrix of Housing Characteristics Measures

As Total number of bathrooms is derived from Number of Full Bathrooms, and Age of the house is derived from construction year and Year sold, only 1 of these variables was selected so not to confuse the cluster model.

Cluster Model Iterations

We developed the model building upwards from 3 variables. The first iteration of the model utilised the variables: “Above Grade (ground) Living Area, Overall condition of the house and Total number of Bathrooms. These variables were selected based on the assumption that the size, functionality and livability of the house would be valuable descriptors of the product in the market. This attempt was built for examined with k= 3,4,5 and 6; and found to be in effective as the clusters’ centroids were not sufficiently variated and a substantial number of houses lay outside of SAS’s modelling of the clusters’ ellipses.

Next, we tried with 6 variables selected on similar criteria as before. We added “Basement area, lot size and garage size”. We repeated the process and examined with k=3,4,5,6 and 7 but found that clusters were not satisfactory as the centroids for more than 1 variable were similar.

Cluster Analysis Model

The 4 chosen measures are as follows:

- Above-grade living area
- Overall material and finish of house
- Age of the house when sold
- Number of full bathrooms.

After using the elbow method on these four measures and the SAS clustering algorithm we discovered that 4 or 5 clusters gave us satisfactory results. Based on the appearance and logic of how the clusters were segregated we chose to go ahead with 5 clusters, wherein each cluster has its combination of unique features.

Table 3: Sums of Squares of clusters for K

K	SS
2	799.427
3	659.733
4	593.626
5	410.625
6	368.632
7	334.265
8	315.041

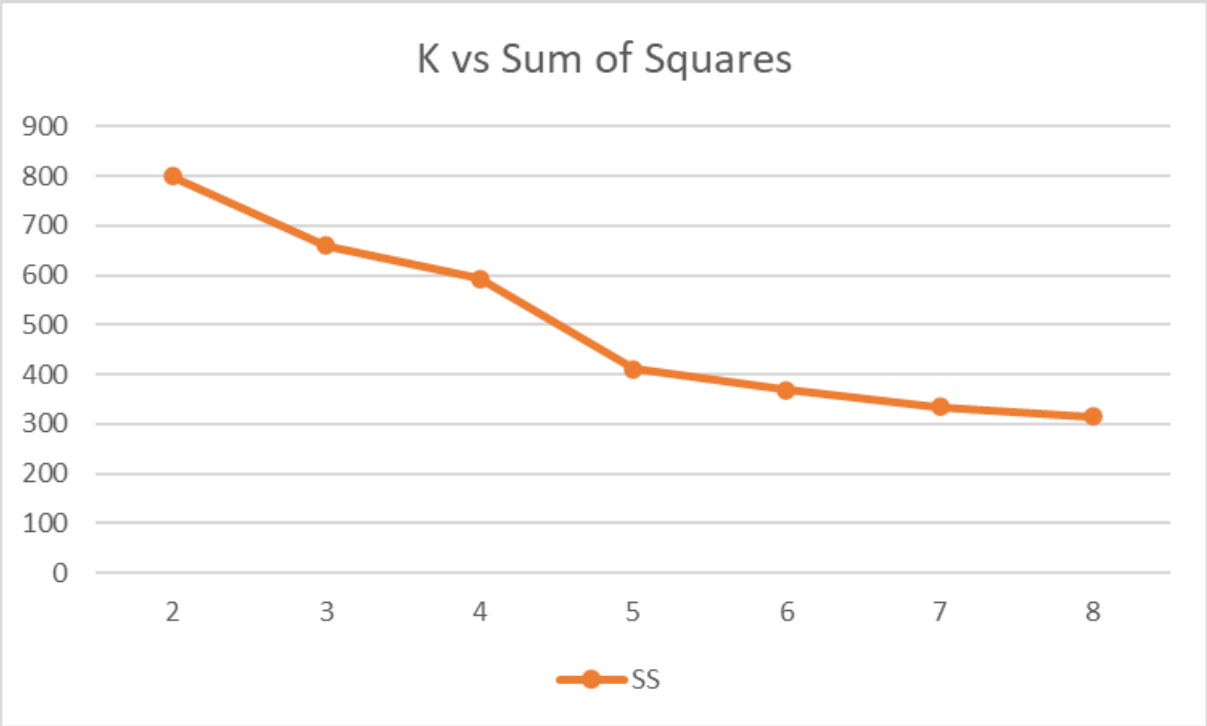

Figure 2: Line graph to identify ideal value of K (Elbow method)

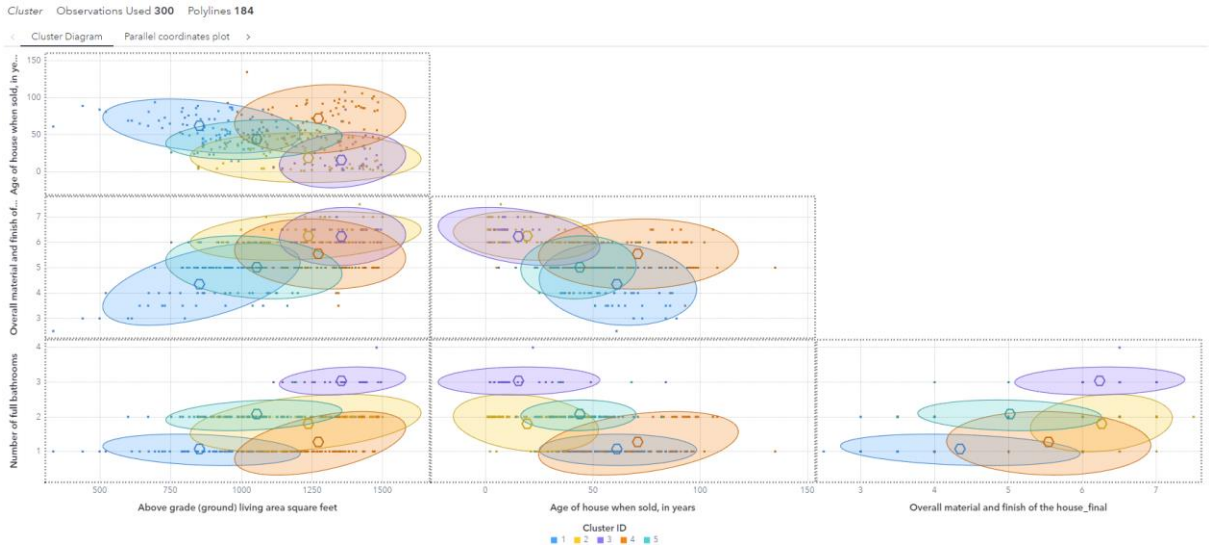

Figure 3: Cluster Diagram

Although all clusters have their unique characteristics, we observe that there are also similarities to be found in them.

Cluster ID	Above grade (ground) living area square feet	Age of house when sold, in years	Overall material and finish of the house_final	Number of full bathrooms
1	858.96923077	61.630769231	4.35	1
2	1227.7761194	18.71641791	6.25	2
3	1358.2142857	15.5	6.23	3
4	1278.8611111	71.263888889	5.54	1
5	1044.4117647	43.25	5.01	2

Table 4: Centroids of Clusters

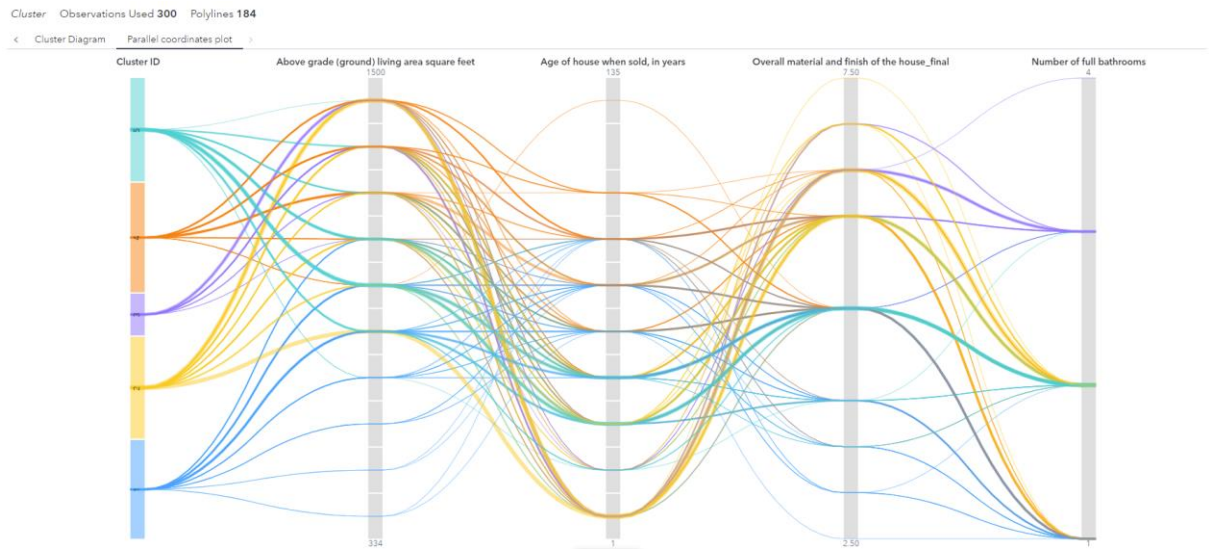

Figure 4: Polyline Co-ordinates Plot

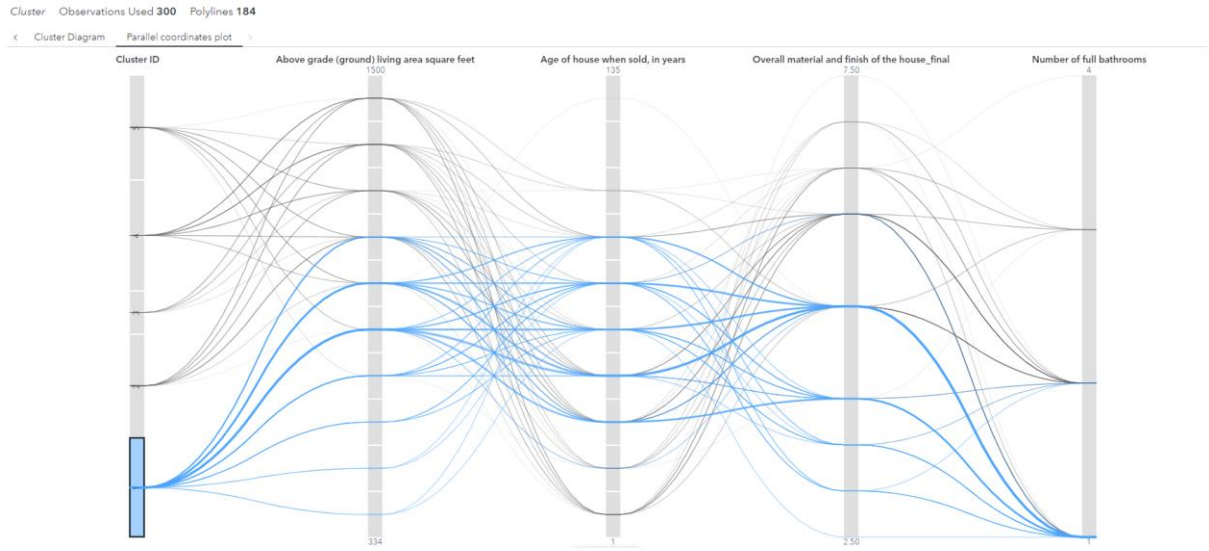

Figure 5: Cluster 1 Polyline coordinates plot

The cluster ID 1 consists of 65 houses with an average above grade living area of 859 Square feet, one bathroom, an average age of 61 years and an average rating for material and finishes of 4.3.
This cluster defines homes that are small, older, with fewer amenities and are not rated very well. This is understandable considering the age of the house that it will have old and have out-of-date elements. It can be assumed that the sale prices of these houses will be lower.

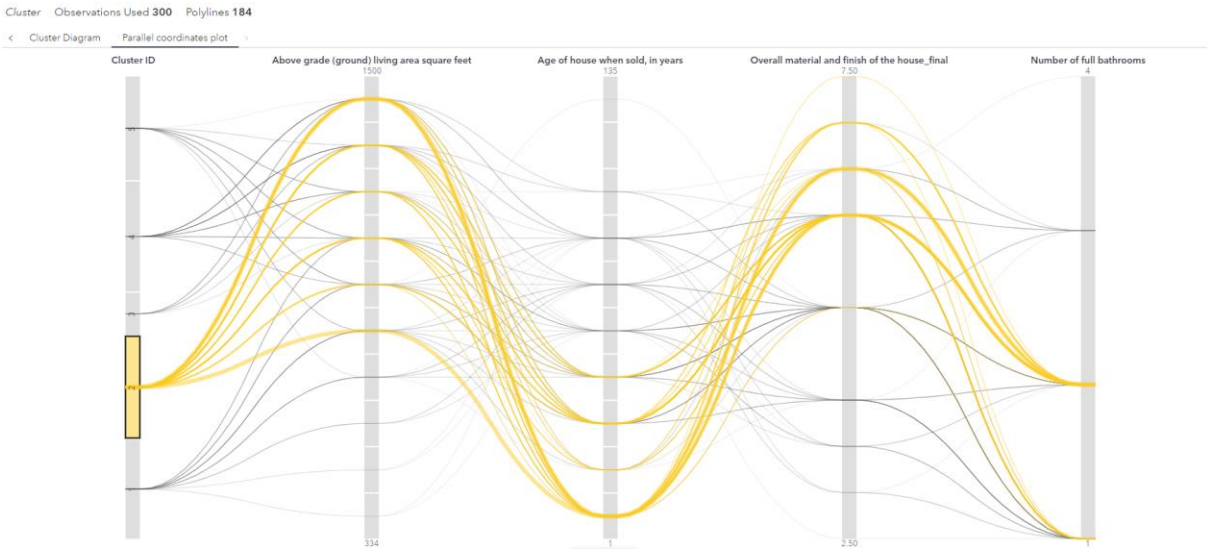

Figure 6: Cluster 2 Polyline coordinates plot

In cluster ID 2 consists of 67 houses, with a larger living area that cluster 1. Most houses in this cluster have a living area of around 1228 square feet, generally with 2 bathrooms. These homes are relatively new, with an average age of 19 years. The houses in this cluster are rated high in material and finish. As the house is comparatively modern its fitting that it has a good rating as well as a higher number of bathrooms compared to older houses.

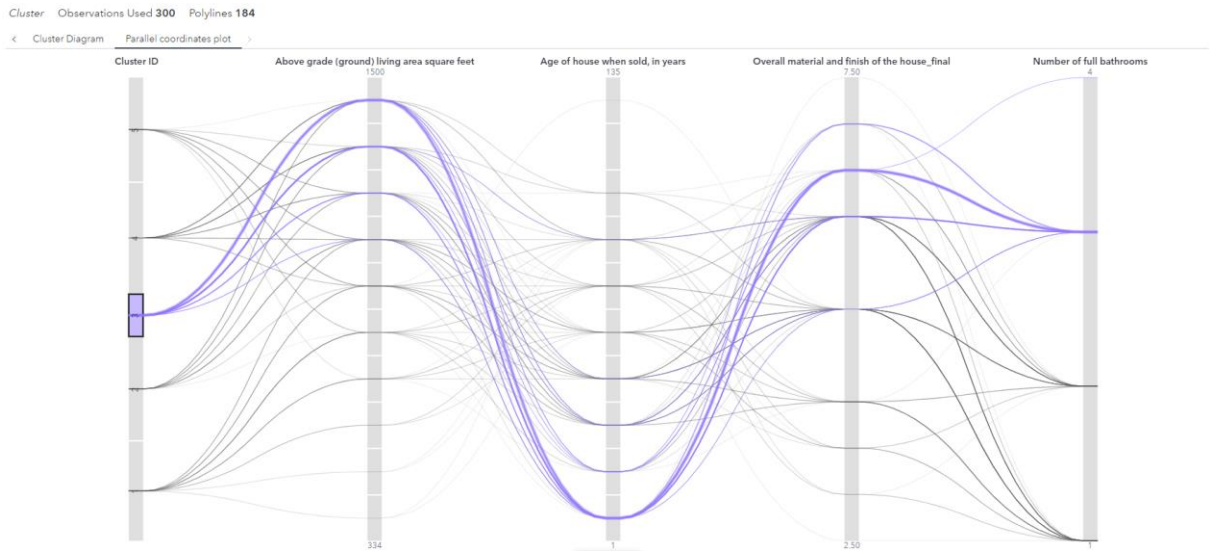

Figure 7 Cluster ID 3 Polyline Coordinates plot

From Cluster ID 3 consists of 28 houses. We observe that the age of the houses in this cluster is the lowest at 15 years and generally has the highest number of full bathrooms. The size of houses in this cluster are similar to those in cluster 2, and are rated well in material and finish.

The houses in cluster 3 are differentiated from cluster 2 by the number of bathrooms- 3.

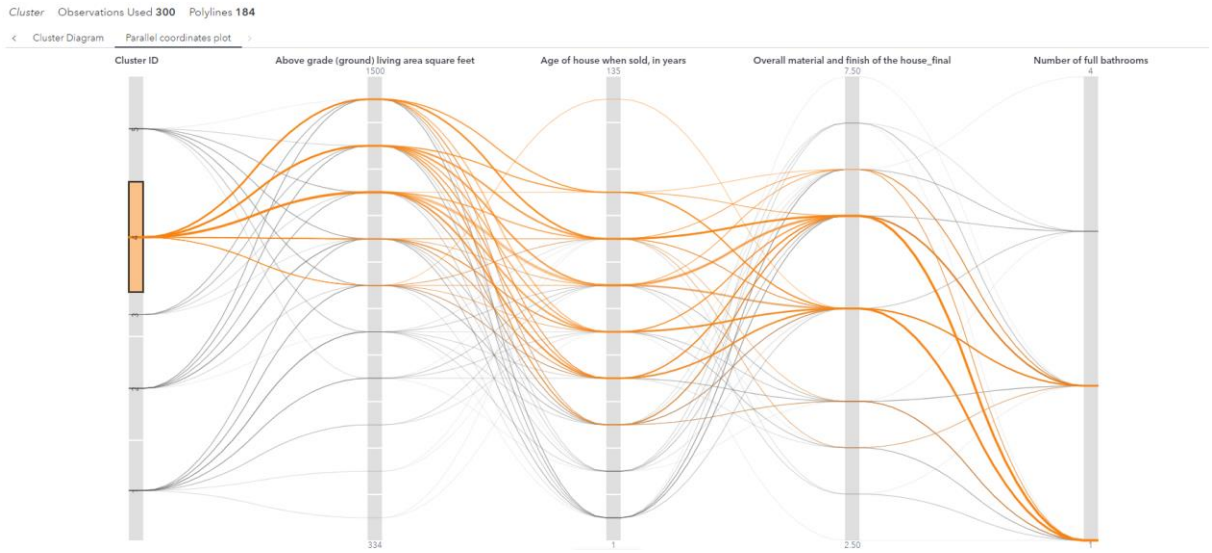

Figure 8 Cluster ID 4 Polyline Coordinates Plot

Cluster ID 4 consists of 72 houses with a median above grade area in the upper mid-range among the data-set. These houses are older, average age is 71 years, These houses have 1 or 2 bathrooms, which may be considered less for the size of the house. The material and finish rating is around 5.5.

These houses are likely to be in the middle of the market range for sales price as the relative size is moderated by the lack of amenities and the condiiton of the house.

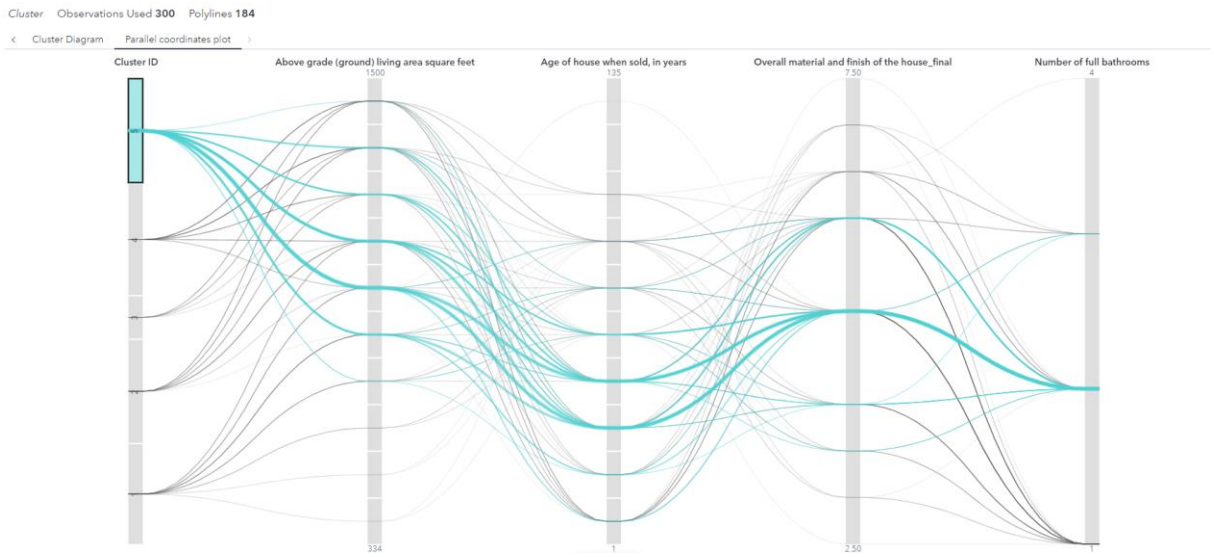

Figure 9 Cluster ID 5 Polyline Coordinates Plot

Cluster 5 consists of 68 mid-sized houses that are not new but not old, and has material and finishes that are rated average. These older houses are seen to have 2 bathrooms.

Relationship between Cluster IDs and Sales Price

By reviewing the clusters we can draw certain hypotheses regarding their sales prices. For example, as cluster 1 has homes of smaller area, lower amenities and lower relative quality of materials, it can be assumed that these homes would sell for the lowest price. However, when comparing the sales data with the clustered houses, we find that the houses in cluster 1 sold for relatively higher than those in cluster 4, which despite a larger livable area, and reasonable rating of material and finish, generally sell for less. This can be attributed to the age of the houses which is 71 years on average.

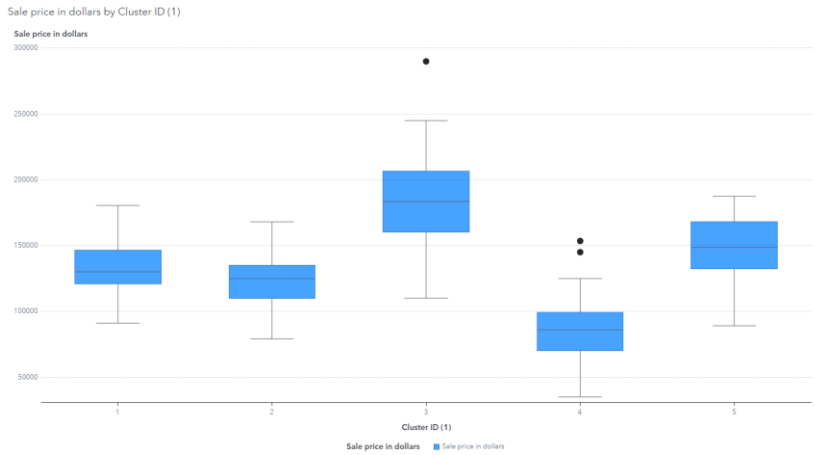

Figure 10 Box Plot - Clusters and Sales Price

Cluster 3 which appears to contain the most recently constructed, large size houses has the highest sales price.

Conclusion

The dataset provided for this assessment consisted of data on the various features of houses that play a role in its pricing and sale. Each buyer has their own set of preferences, therefore based on our assumptions we chose features that we believe play the key role in the real estate business market. We used the correlation matrix to identify variables with low correlation to each other to be used for clustering. Following that we analyzed those variables using SAS Analytics, by running them through the cluster analysis algorithm.

Through the 5 clusters we were able to identify groups of houses that were similar in nature as well as their unique characteristics. Through this analysis we deduced the relationship between the various features of a house and its market value. The variable that majorly affects the pricing of a property is the age of the house. Houses with a higher age are less likely to have modern amenities, but their pricing differs on the basis of the living area. Modern houses with large liveable areas have the highest selling price.

Another benefit of cluster analysis is that not only can it help determine the market value of properties, it can also be used for detecting the target audience for particular properties. Based on this information realtors can direct prospective buyers towards houses within their price range with their preferred amenities.

Appendix

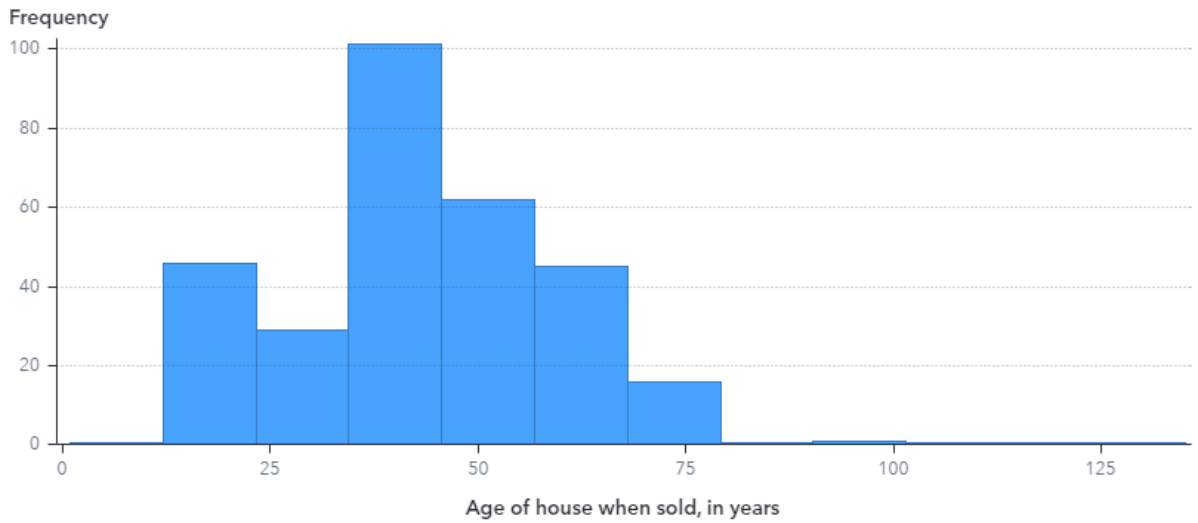

Figure 11 Distribution of Age of House

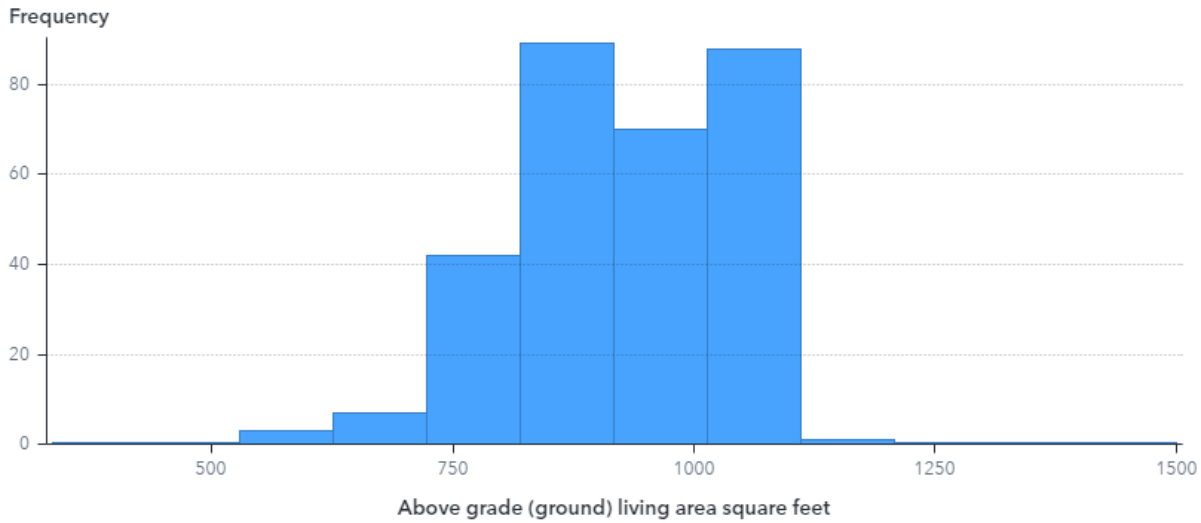

Figure 12 Distribution of Above Grade Area

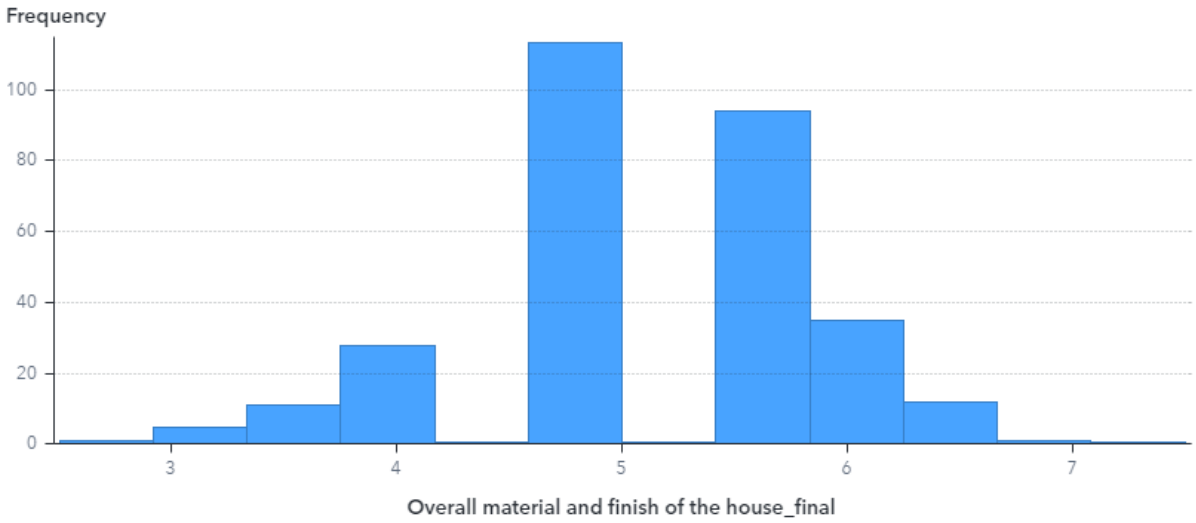

Figure 13 Distribution of Rating of material and finishes

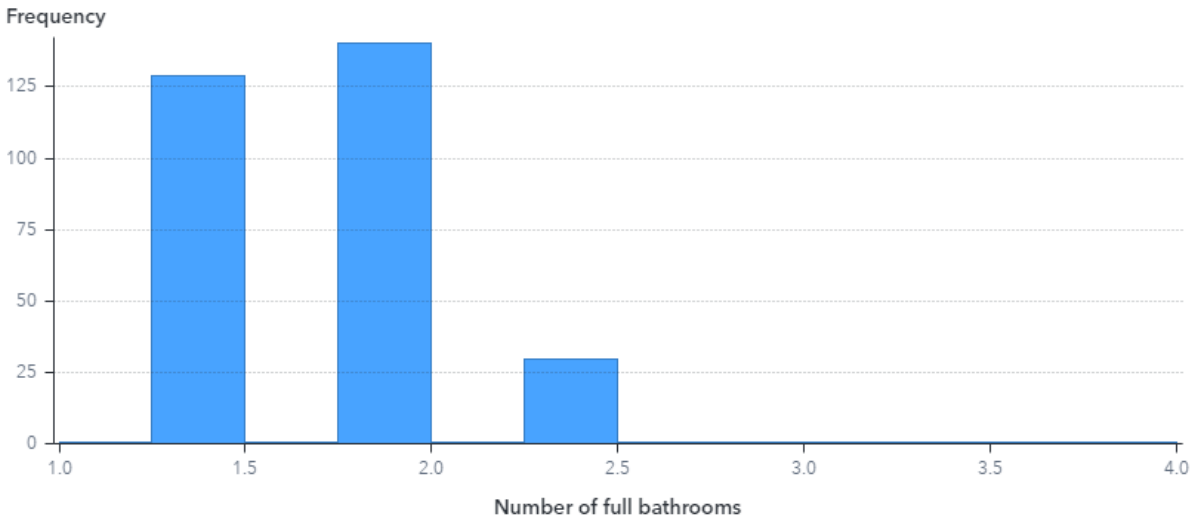

Figure 14 Distribution of Number of full bathrooms

Cluster ID	Observations	RMS of STD	Within cluster SS	Min centroid-to-observation	Max centroid-to-observation	Nearest Cluster	Centroid Distance	Average Distance
1	65	1.2673165191	104.39592537	0.3839410266	2.9998990504	5	1.9638639016	1.1749591788
2	67	1.2424782387	103.43139563	0.589383797	1.9541428008	5	1.8136463189	1.1901785198
3	28	0.9913955178	27.520222037	0.3839513848	2.5063484385	2	1.9185106595	0.8575632172
4	72	1.362202288	133.60284528	0.6142495103	2.6610561179	5	1.963638728	1.2928044347
5	68	1.0318460102	72.400020836	0.0714911123	2.160845361	2	1.8136463189	0.9118956921

Figure 15 Cluster Summary