Assignment #2

Jonah Muniz

Abstract

The purpose of this research is to use unsupervised learning methods to categorize houses in Melbourne Australia into submarkets. The goal of categorizing the houses into submarkets is to then be able to assign relators to relevant territories, appraise houses correctly and identify under or over-valued houses for their submarket. Clustering analysis will be the primary unsupervised learning method that will be used in this research analysis. Both Hierarchical and K-Means clustering techniques will be used on the Melbourne Housing dataset. The research will conclude with defined categorizations for each submarket and how a real estate firm can leverage these submarkets to optimize their business.

Keywords: Unsupervised Learning, Hierarchical Clustering, K-Means Clustering, Australia Housing Market Segmentation

Introduction

When a real estate firm is assigning relators to a territory or appraising a houses value, it is important to understand all of the factors that lead to success. For example, if a realtor is often selling similar type homes, they will start to develop a specialized skillset and strategy to be successful. It is advantageous for the real estate firm to assign realtors to housing types that they are familiar to selling. A real estate firm having the ability to appropriately appraise houses and find the outliers in the market will also lead to potential unrealized revenue. The Melbourne Housing dataset from Kaggle[1] will be used to see if housing submarkets within Melbourne can be uncovered.

Literature Review

Utilizing unsupervised learning methods to analyze and segment a housing market is not a new analysis. Similar analysis has been preformed in other countries outside of Australia to find geographic areas where the price of the houses are constant and individual housing characteristics are available for

purchase[2]. The first research that will be covered used hierarchical models and single-family property transactions over the 1/1995 through 1/1997 period to examine housing market segmentation within metropolitan Dallas. The research was able to identify separate submarkets and decompose the difference among the submarkets into price and quantity/quality effects. Another research paper looked to analyze housing submarkets in Sydney and Melbourne Australia[3]. The goal of the research was to develop a way to define submarkets in large metropolitan areas that are internally homogeneous and externally heterogeneous. The research used principal components, cluster analysis and hedonic regressions to define their submarkets. The team was able to define 5 submarkets within the Sydney and Melbourne metropolitan areas.

Methods

In order to uncover submarkets in the Melbourne housing market the Melbourne housing data will be acquired from the Kaggle.com. The dataset is then ingested into R for further data cleansing and analysis. The Melbourne dataset contains 34857 observations and 21 variables. Once the data is within R a new dataset is created to only include rows that have values for all columns. This new dataset is 8887 observations and 21 variables. A new dataset is then created to only include the numeric variables for cluster analysis. This dataset contains 8887 observations and 9 variables. These 9 variables will be the variables used to perform cluster analysis. Prior to any cluster analysis exploratory data analysis is performed. All duplicated rows of data are removed from the dataset. This removed 16 rows of data that was duplicated in the dataset.

Once the data has been fully cleaned an elbow plot can be created to understand what the ideal number of clusters are for the dataset. The ideal number of clusters will be represented on the plot as the point where increasing the number of clusters does not lead to any additional value, the elbow of the plot. Once the number of clusters is determined the two clustering methods, hierarchical and k-means can be conducted. Using the dist function the distance between points within the cleaned dataset can be

determined. Now that you have a distance matrix for all the observations the hclust function can be used to create a hierarchical clustering. The complete method was used to conduct the hierarchical clustering. Once the hierarchical clustering has been completed, the cutree function can be used to determine the number of clusters to obtain from the analysis. The hierarchical clustering was cut so 3 clusters are produced. The cluster assignments can then be combined with the original dataset for labeling visualizations. A tSNE plot can be constructed to visualize the observations on two principal component axis and label and color each observation based on their cluster assignment.

The similar process can be followed for a cluster analysis using the k-mean clustering function. The same cleaned dataset can be used as the input to the kmeans function. The number of clusters needs to be identified in the kmeans function unlike hierarchical where the number of desired clusters is determined after the analysis is conducted. Once the kmeans clustering analysis is performed the clusters can be visualized by leveraging tSNE plots and labeling data points with their cluster assignment. The centers for each cluster can also be evaluated to describe the clusters and their unique qualities. The above process can repeat for different cluster amounts until an outcome is produced that can be easily communicated to leadership. The ideal clusters will create distinct submarkets within the Melbourne housing market that the real estate firm can leverage to assign agents as well as evaluate if a property within a cluster is under or over-valued compared to similar properties.

Results

Following the above methodology, the Melbourne Housing Full file was imported into R and saved as a data frame called mydata. The necessary R packages were loaded into the session and a data frame consisting of all observations with data across all 21 variables was created. This act of data cleansing reduced the data set from 34,857 to 8,887 observations. In order to perform the clustering analysis only numeric variables should be used as clustering variables. Due to this requirement a separate data frame called workdata was created consisting of only 9 variables: Rooms, Price, Distance, Bedroom2,

Bathroom, Car, Landsize, BuildingArea and YearBuilt. One of the last steps prior to running any clustering analysis is to ensure none of the observations were duplicated in the workdata data set. Removing duplicated observations reduced the total observations from 8,887 to 8,871. The final data set prior to clustering consisted of 8,871 observations and 9 variables.

Now that the data has been cleansed and prepared for analysis an elbow plot can be constructed to help determine what the ideal number of clusters are. The elbow plot produced from the workdata dataset can be found in the appendix as Figure 1. As can be seen in the elbow plot it seems that the ideal number of clusters is 3 based on the lack of increased value for increasing the number of clusters past 3. Now that there is some understanding of the ideal number of clusters, the two different type of cluster analysis can be performed.

First complete hierarchical clustering was performed on workdata. The hierarchical clustering tree diagram can be found in the appendix as Figure 2. As can be seen in the tree diagram, it is hard to see where each property lies on the diagram but there seems to be a lot of small clusters that roll into 3 main clusters. These 3 clusters were then determined by using the cutree function. The clustering assignments were then combined with workdata to create a new data set labeling each observation with its cluster assignment. TSNE plots were then created to visualize the observations and the cluster labels. 2 PCA dimensions were used to plot each observations and each observation was labeled and colored based on their cluster label. An example of the tSNE plot from the hierarchical clustering can be found in the appendix as Figure 3. Perplexity values were experimented with to determine the best visualization of the clusters across the two tSNE dimensions. Comparing the tree diagram to the tSNE plot it is not surprising to see that majority of the observations fall under one cluster label.

Now that one of the clustering analysis methods was performed and visualized the k-mean clustering analysis can be performed. For the k-mean clustering the same data set, workdata, can be leveraged.

Unlike hierarchical clustering, the desired number of clusters must be identified within the kmeans function. The number of clusters was selected and the kmeans function was run and saved. The kmeans function produces a couple of outputs, the most of import of which are the clusters and the centers of each cluster. A tSNE plot was then created to visualize the observations across 2 PCA components with each observation labeled and colored by their cluster assignment. An example of the tSNE plot can be found in the appendix as Figure 4.

Now that both cluster analysis was performed the clearest clustering analysis for management should be selected and analyzed further. Reviewing the tSNE plots of both the hierarchical and k-means clustering analysis it was determined that the k-means clustering analysis was the best. The centers for each of the three clusters were then analyzed to understand the qualities of each cluster. A table can be found in the appendix as Table 1 of the center of the k-means clusters. Reviewing Table 1 it can be seen that there are three submarkets within the Melbourne housing market. Reviewing the clusters across the variables you can see that cluster 3 has the highest mean number of rooms, price, bedrooms, bathrooms, car, landsize and building area. This submarket can be regarded as the large family homes cluster. These properties also tend to be located closer to the shore. Cluster 2 seems to consistently have the lowest mean when looking across the same variables as cluster 3 and also is located the furthest on average from the coast. This cluster can be regarded as small family homes. Lastly, it seems that cluster one is consistently in-between cluster 2 and 3 in terms of their center across each variable. Cluster 1 seems to be the middle tier homes.

Conclusions

Now that the Melbourne housing data has been used to determine submarkets within the Melbourne housing data a real estate company can leverage these submarket insights to allocate their realtors to the appropriate submarket for their skill set. For example, the most senior and experiences realtors can be assigned to cluster 3 properties due to their high sale price and mean bedroom and bathroom

characteristics. Having an experience realtor sell houses in this cluster will be key to ensure potential clients feel that they are in experienced hands when looking to make such a large investment. If a brand-new realtor was asked to sell houses in cluster 3, they may feel overwhelmed and over their head. I would suggest assigning newer realtors to cluster 2 properties. These properties are on the lower end in terms of size and price. This is a great type of housing to build skillsets and grow a realtor's expertise.

I would also suggest that the real estate firm leverage the characteristics of the clusters to evaluate if a potential property is above or below market value. For example, if a property is identified within cluster 3 but is priced lower then the mean price of cluster 3 the property is undervalued. This could be a great opportunity for the firm to flip a house and make a profit.

Using unsupervised learning techniques submarkets within the Melbourne housing markets have been uncovered in efforts to help a potential real estate firm better allocate their realtors and evaluate potential properties. Using k-means clustering each cluster and its characteristics can be clearly communicated to leadership and help drive business decisions on where the firm should invest in properties and where they should assign their realtors.

Appendix

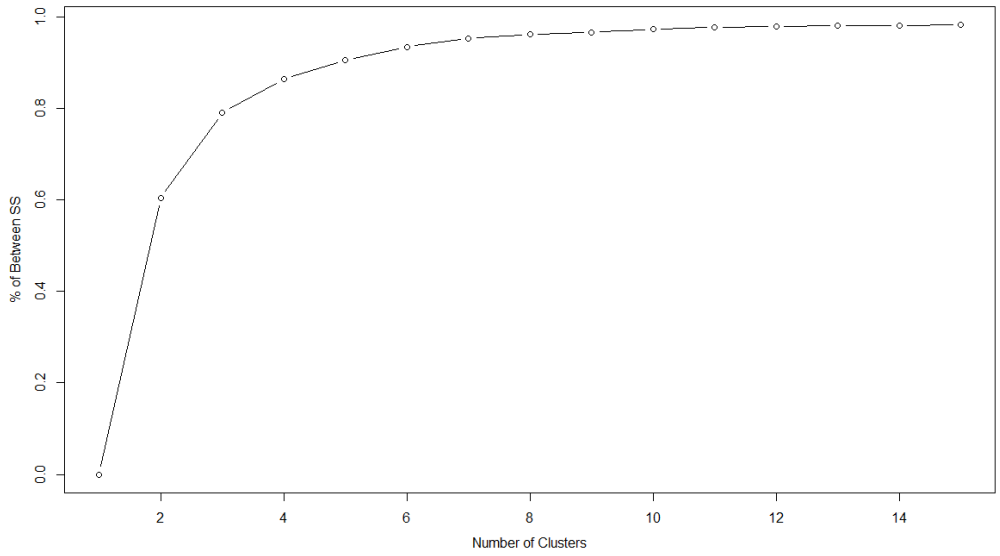

Figure 1: *Elbow plot for the workdata data set*

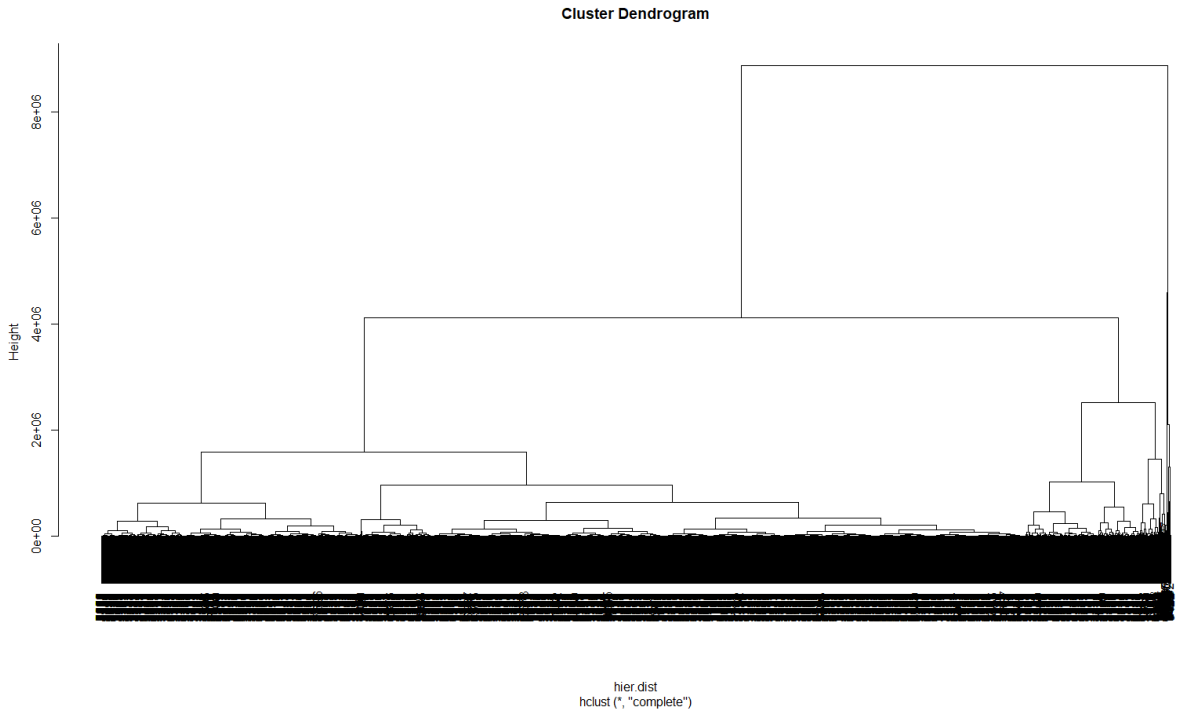

Figure 2: *Complete Hierarchical Clustering Tree Diagram for all Observations*

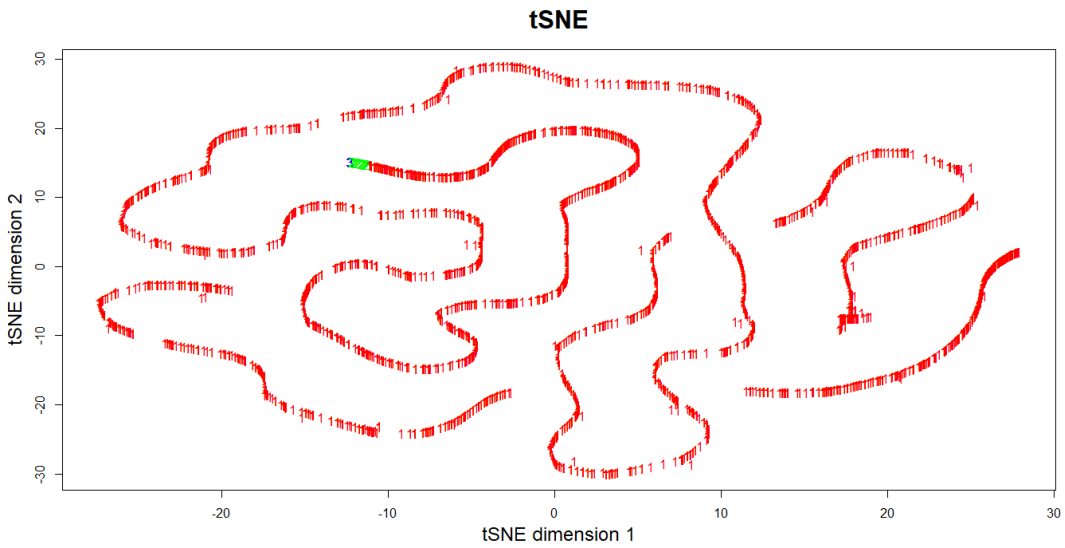

Figure 3: *tSNE plot of the hierarchical clustering analysis with cluster assignment labeled and colored*

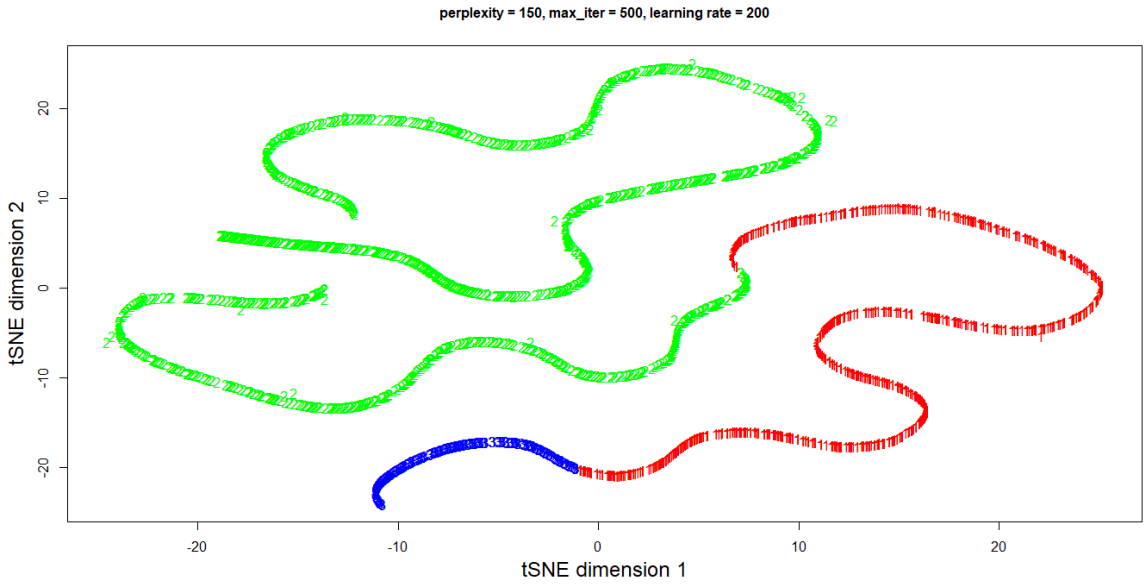

Figure 4: *tSNE plot of the k-means clustering analysis with cluster assignment labeled and colored*

	Rooms	Price	Distance	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt
1	3.443879	1479496.9	9.251253	3.407192	1.831457	1.778060	528.6059	174.9323	1950.344
2	2.812511	702213.5	12.525150	2.801813	1.455304	1.594742	499.3293	123.3680	1975.629
3	4.154229	2907354.7	8.013101	4.114428	2.552239	2.195688	721.7728	270.3051	1945.900

Table 1: *K-means clustering analysis cluster centers for each of the 9 clustering variables*

References

1. Pino, Tony. “Melbourne Housing Market.” Kaggle, October 14, 2018. https://www.kaggle.com/anthonypino/melbourne-housing-market.
2. Goodman, Allen C., and Thomas G. Thibodeau. “Housing Market Segmentation.” Journal of Housing Economics. Academic Press, May 25, 2002. https://www.sciencedirect.com/science/article/abs/pii/S1051137798902297.
3. Bourassa, S. & Hoesli, Martin & MacGregor, Bryan. (1997). Defining Residential Submarkets: Evidence from Sydney and Melbourne. Ecole des Hautes Etudes Commerciales, Universite de Geneve-, Papers.