Group 6 – Sean Atkinson, Sameer Khan, Parker Davis

Assignment #2

Abstract

The purpose of this work is to use cluster analysis to identify types of houses within the Melbourne real estate market. These classifications will enable us to find the most valuable investment opportunities for potential future acquisition as well as compare properties which have similar types. Two clustering methods are used and their results are compared to identify consistencies or lack thereof. Applying the clustering methods will enable the organization to make appropriate decisions with respect to future investment funding strategies and the realization of those investments against both current and future real estate data.

Keywords: K-means clustering, Hierarchical clustering, Scaling, t-SNE

Introduction

This research is being conducted to identify types of homes and whether they can be valuable investments. By utilizing cluster analysis with carefully executed feature selection clusters can be developed to determine a fair value for any given property. Cluster analysis can also identify undervalued and overvalued properties based on the price of similar types of homes. The development of the model will enable future potential analysis of investment opportunities. A current real estate dataset is used to build and validate the model's effectiveness in identifying potential real estate opportunities.

Literature Review

Real estate has long been an industry for making large amounts of money. People that can identify the best investment opportunities can see huge profits. At any given time thousands of properties can be on the market so cluster analysis has been used to better assess where great

opportunities may lie. The group data4help approached real estate clustering with the k-means methodology. The reason for this was its simplicity, adaptability, and guaranteed convergence (data4help, 2020). They used an euclidean distance method to determine how the properties related to one another. Choosing the amount of clusters to be used was determined by silhouette scores. The silhouette score gives insight into how far spaced a point in another cluster is when compared to points lying within the cluster. A higher score indicates that a point within a cluster has smaller euclidean distances to points within the same cluster than points in another cluster (data4help, 2020).

The utility of hierarchical clustering is described in the research by Hepsen & Vatansever, (2012), their research shows inference in terms of rental investment opportunities within the Turkish real estate market. Using hierarchical clustering the research provides indicators for investment portfolios based on rental incomes from multiple residential areas in Turkey. The utility of the model provides methods to diversify rental income and manage the risk of those investments over time.

Methods

The application of two clustering algorithms will be used to define the most appropriate model for real estate data and build a capability to discover investment opportunities. The approach for building the model includes steps to address the formatting of the data set. Any data set provided to the analytical team will need to go through preprocessing. This will identify specific columns of data for augmentation, address missing values and also build a scaled dataset that will allow for better inference and more accurate modeling results.

Preprocessing/Scaling

In order to properly analyze the data it was first scaled. The data was scaled by using the scale function in R. With this function the mean of all the data will be equal to 0 with a standard deviation of 1. In Figure 1 the utility of pairs() is used as a starting point to view relationships between 2 variables. The following matrix shows some patterns within the data for example, Price and Building Area have string alignment indicating alignment of the data and structure within the dataset.

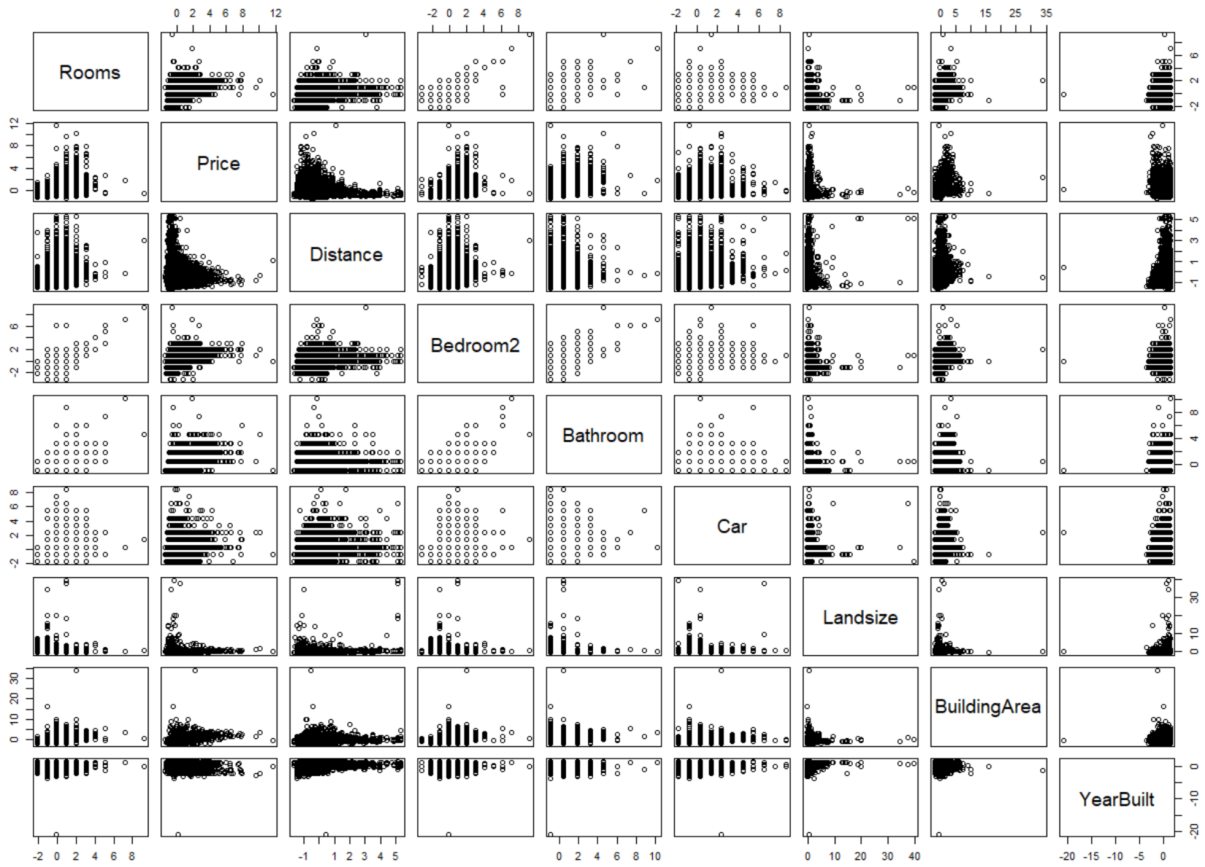


Figure 1: Pair plot for comparing each variable in the real estate dataframe

Next, the correlations between the column variables were examined to examine the strength of the relationship between them. This could give some insight into possible dimensionality reduction opportunities to make the dataset easier to understand and cluster. The

correlation matrix below shows a strong relationship between the variables describing different types of rooms in a given home. We also see moderate positive correlation between room variables, home price, and building area.

	Rooms	Price	Distance	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt
Rooms	1.00	0.48	0.28	0.96	0.62	0.40	0.10	0.61	0.03
Price	0.48	1.00	-0.23	0.46	0.46	0.21	0.06	0.51	-0.31
Distance	0.28	-0.23	1.00	0.28	0.12	0.26	0.14	0.14	0.31
Bedroom2	0.96	0.46	0.28	1.00	0.63	0.41	0.10	0.60	0.02
Bathroom	0.62	0.46	0.12	0.63	1.00	0.31	0.08	0.55	0.19
Car	0.40	0.21	0.26	0.41	0.31	1.00	0.12	0.32	0.14
Landsize	0.10	0.06	0.14	0.10	0.08	0.12	1.00	0.08	0.04
BuildingArea	0.61	0.51	0.14	0.60	0.55	0.32	0.08	1.00	0.06
YearBuilt	0.03	-0.31	0.31	0.02	0.19	0.14	0.04	0.06	1.00

Table 1: Correlation Matrix to measure variable relationship strength

Once the data is in a format and data frame ready for modeling a stage is added called T-Distributed Stochastic Neighbor Embedding (tSNE). The use of this technique is to provide dimensionality reduction or simply taking multidimensional data and producing a 2D plot of the information. See Figure 2.

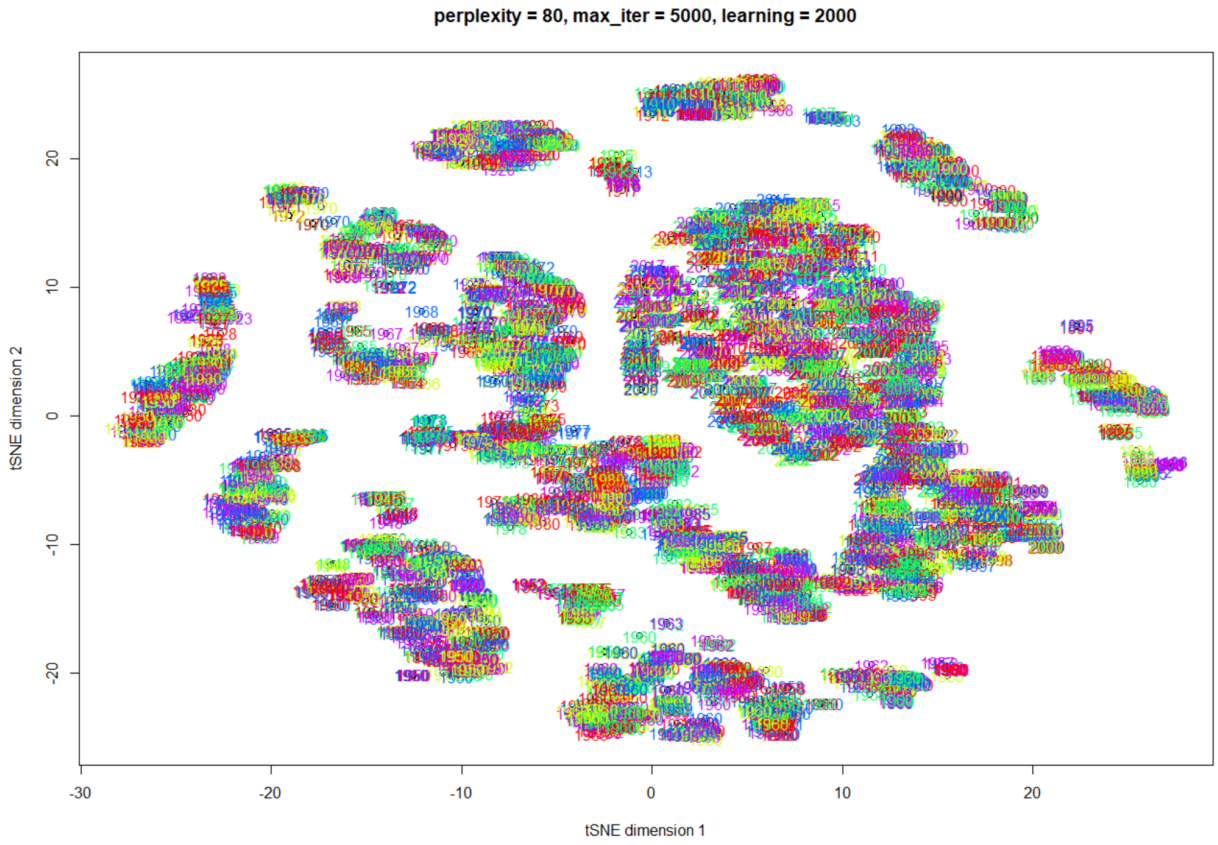


Figure 2: tSNE plot to provide cluster inference

The inference of the tSNE preprocessing identifies some common groups that may help determine the number of clusters within the clustering algorithms. The adjustment of hyperparameters will assist in determining structure of the data and act as an inference for clustering techniques described below. Given the alignment of multiple data elements to specific groupings shows with coloration that patterns exist that should be further explored with clustering techniques.. A perplexity score of 80 is used as the smoothness used for neighbor interpretation provides better clustering elements. The learning value of 2000 was used. This is relatively high but the higher learning rate did provide a faster method of step rate and building the visualization in Figure 2. An assessment of the hyperparameters is provided in Appendix A.

Clustering Algorithms

The input of the preprocess and scaling has been utilized as an input into the t-SNE dimensionality reduction algorithm. With this information the team will utilize two methods of clustering: hierarchical and K-means. These methods will identify the different housing types within our data set and address any outliers for the purposes of valuation and investment opportunities. The analysis starts with Hierarchical Clustering. We use two parameters for building the visualizations, “Complete” and “Average”.

Complete hierarchical clustering

Figure 3 below shows a visualization issue with building a dendrogram with the number of high dimensional features in our data set. Each feature is building a grouping with other similar points in the data and given the data set size of 8887 elements these complicate both the analysis and inference of this visualization.

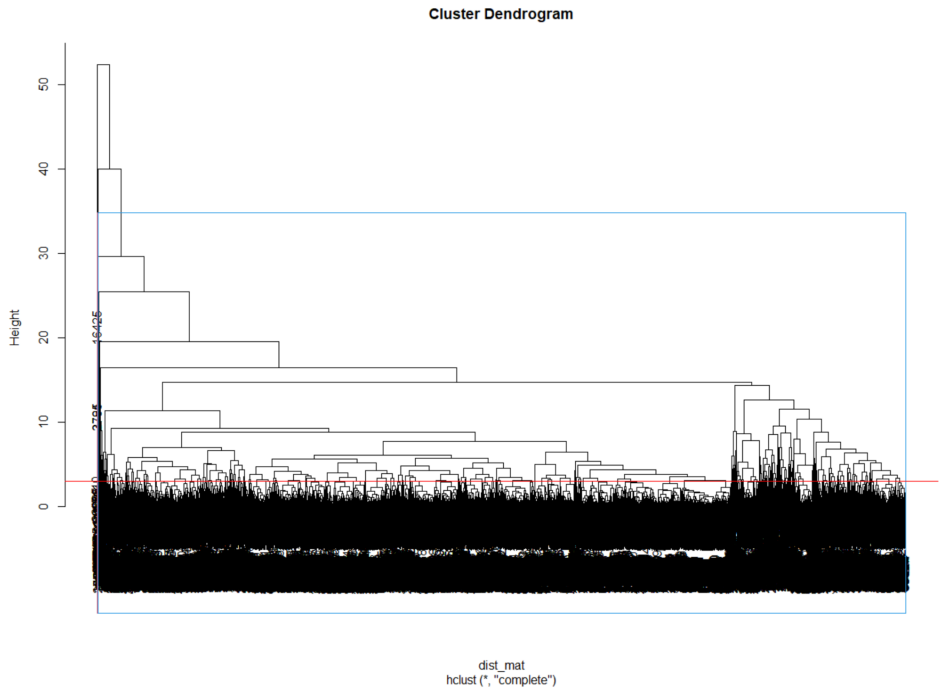


Figure 3 ‘Complete’ dendrogram for hierarchical clustering

In much the same manner as the “Complete” clustering method the utility of “Average” has the same issues as or previous hierarchical analysis. Figure 5 retains elements of complexity and building a specific measure of clusters to define housing type is not as apparent with the clustering method.

Average hierarchical clustering

This method employs the use of an average inner cluster distance where the number of pairs or divides by the sum of each pair of observations to form the average calculation.

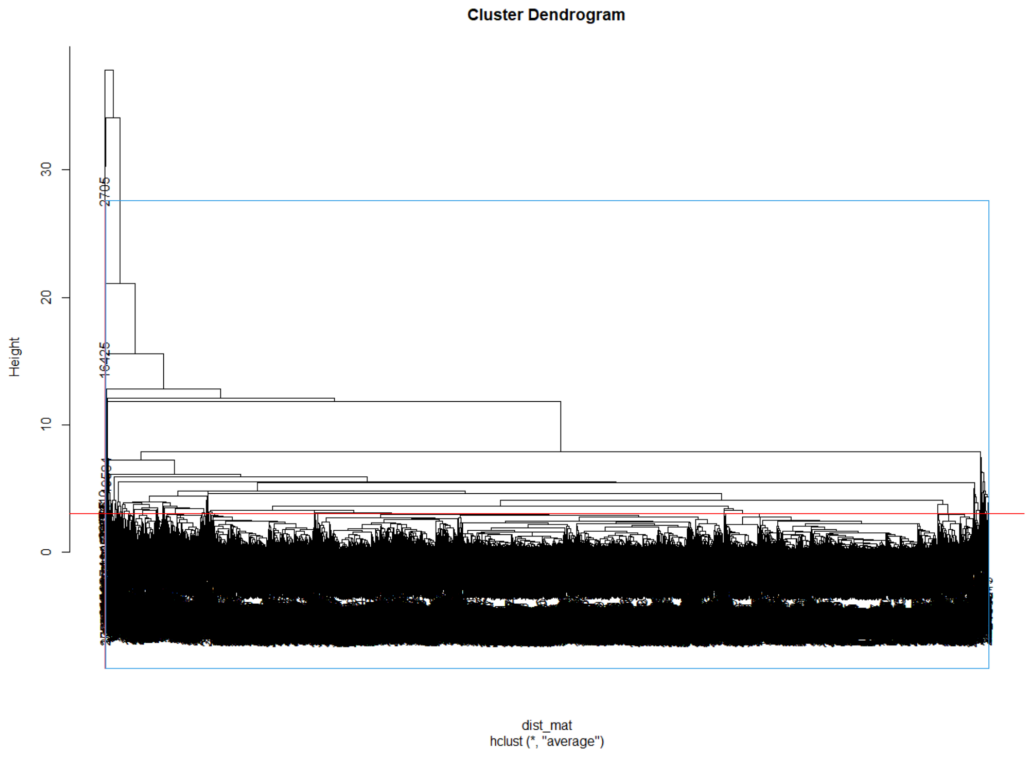


Figure 4 ‘Average’ dendrogram for hierarchical clustering

The interpretability of using hierarchical clustering shows that this form of clustering would not be of much use in the case of building identifiers for real estate investment opportunities. Figures 3 and 4 conclude that dendrogram visualization does make apparent the

requirements for examining specific identified investment opportunities or provide a means to communicate the models recommendations. Some methods were tried to make the dendrogram more readable, but they turned out to be fruitless. Reducing the number of columns in the dataset still formed a crowded dendrogram because the number of observations remained constant. Performing a random sample of the dataset reduces the number of observations and provides a clearer picture of the dendrogram, but it is still not a reliable method for identifying the number of clusters. A random sample might not be representative of the entire dataset, especially when it is so large. The figures of these two strategies are located in Appendix A.

The next clustering method used is k-means clustering. K-means clustering takes a user defined amount of centroids that will become the clusters. The distances to the data points are calculated from these centroids and the centroids are moved through an iterative process. Once the centroids have converged this will yield the optimal clusters for that amount of specified centroids. An additional calculation of a silhouette score can assist in finding the optimal number of centroids to set. The silhouette score calculation examines the distance between points in a cluster compared to the distance the clusters are separated. It is a metric used to calculate the "goodness" of a clustering technique. This calculation was performed in order to find the optimal amount of clusters for sorting the data points. This calculation is highly important because selecting too few clusters can result in dissimilar data points being in the same cluster. While too many clusters will result in clusters that should remain one cluster being potentially divided into multiple clusters.

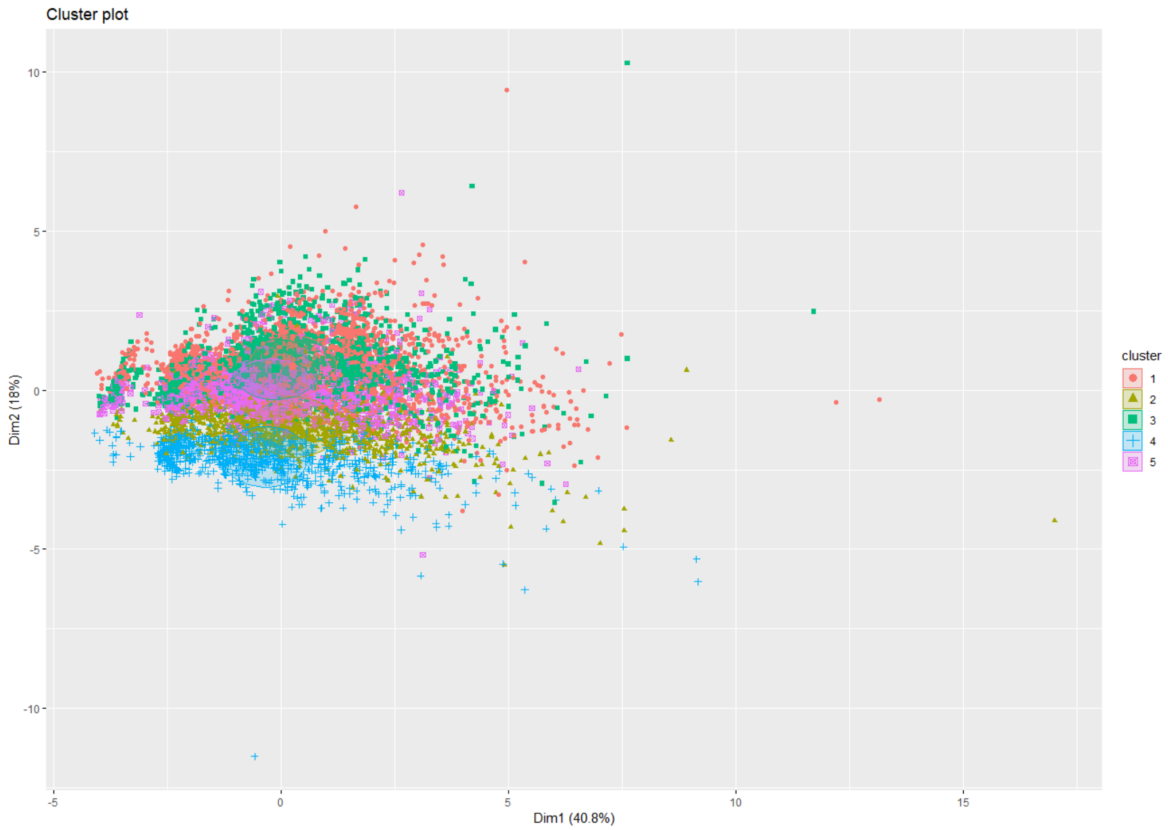


Figure 5 k-means plot to identify real estate investment opportunities.

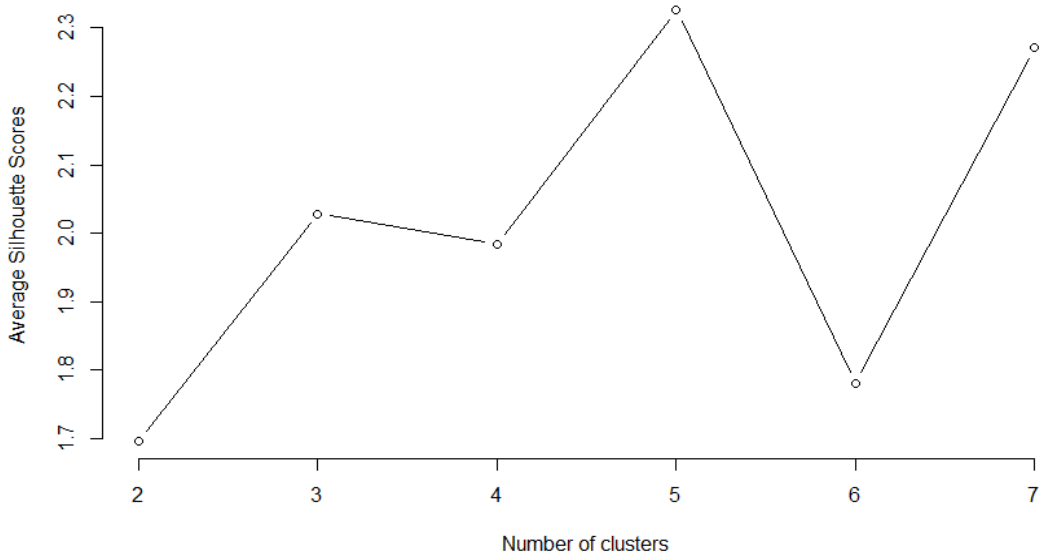


Figure 6 Silhouette Scores Generate to Find Optimal Number of Clusters = 5

```
Cluster means:
        Rooms      Price    Distance    Bedroom2   Bathroom        Car    Landsize BuildingArea   YearBuilt
1  0.01519623 -0.2218733  0.01241838  0.02128828 -0.1669798  0.1636056 -0.03088214   -0.1690171  0.27430494
2 -1.31715471 -0.7149040 -0.50112960 -1.30803149 -0.7277769 -0.5936566 -0.11786015   -0.7958823  0.35803182
3  1.28720814  1.4073773 -0.16397888  1.27144415  1.3145134  0.6854246  0.15222176    1.3737980 -0.09018513
4  0.61887713 -0.4748333  1.59029859  0.63355768  0.3883538  0.4723678  0.22635262    0.2503748  0.62598050
5 -0.19740343  0.4297252 -0.76431417 -0.21953786 -0.4018698 -0.6758630 -0.15696582   -0.2356557 -1.49381811
```

Figure 7 Cluster means for Respective Scaled Real Estate Categories

Results

Figure 5 identifies the following opportunities from cluster # and points to real estate opportunities at the axis (X, Y). The model now informs our decision making with respect to real estate investment. Using the k-means approach identifies the following opportunities as ones that require further investigation. In order to avoid purely guessing the number clusters in the k-means methodology silhouette scores were generated. Silhouette scores indicate how good the clusters are by calculating distance from one cluster to another. Larger silhouette scores indicate better clusters. From the silhouette chat in Figure 3 it can be seen that 5 clusters is considered optimal for this dataset. Looking at the type column from the original data set there are also 5 generically categorized types of houses: house, unit, townhouse, development site, and other residential. Therefore, the silhouette score and amount of house types align quite well. Figure 7 allows us to see from the top level scaled contents of each cluster. In the scaling the mean of the values are set to 0. In Cluster 1 many of the values are near the mean which would point to these homes being of the townhouse type. Cluster 2 falls on the low end of rooms, size, and price which would indicate properties of type unit. Cluster 3 has the largest amount of rooms and building size these could fall into the houses type. Cluster 4 has above average in rooms, lower price, and most recent year built which can indicate development sites or new homes. Cluster 5 has high prices with low rooms. This cluster represents the other residences that are similar to the other types but slightly different. This can be verified in Figure 5 by visualization; it can be seen that Cluster 5 is situated in the middle of all the clusters.

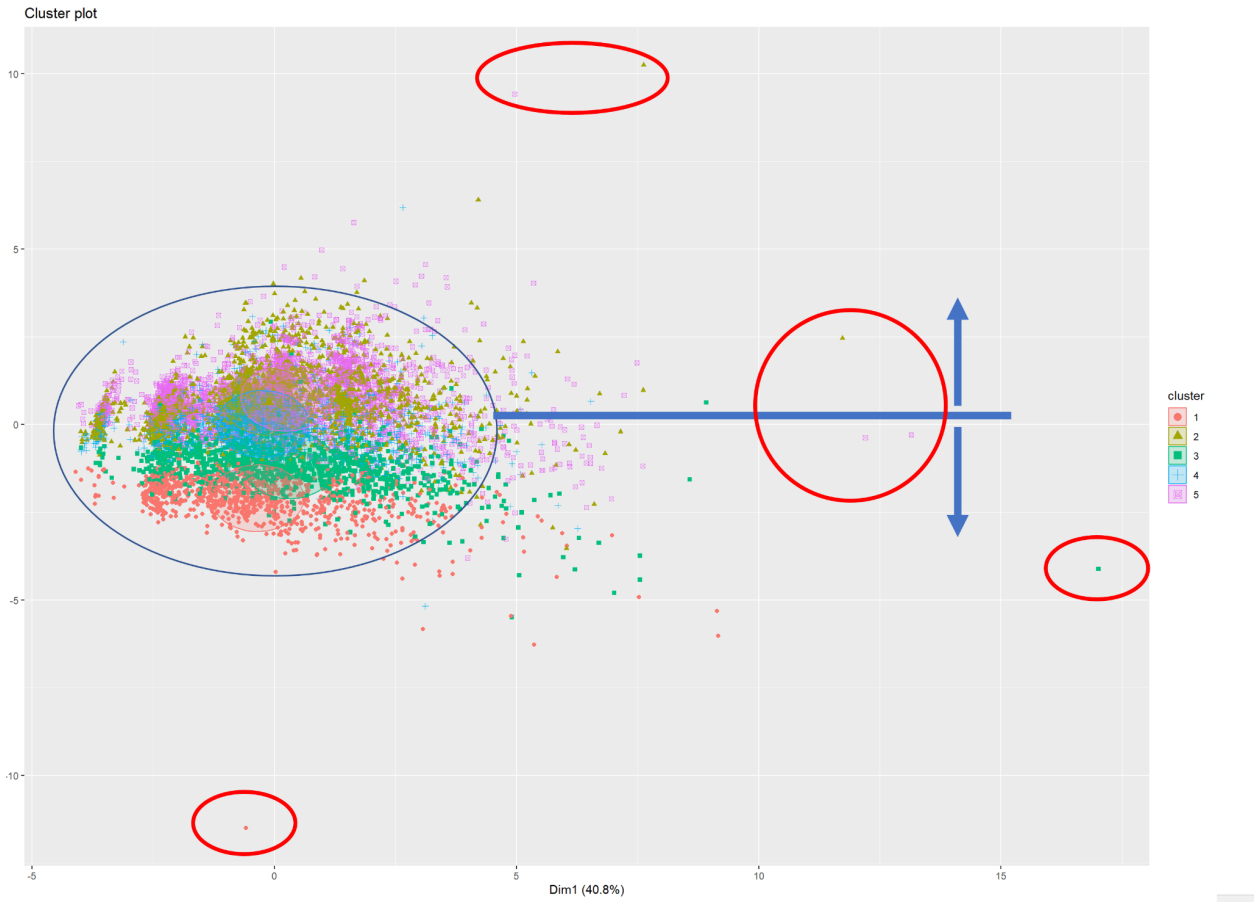


Figure 8 Specific resulting element for further investigation

Figure 8 provides insight into both elements for assessment with respect to current real estate assessment criteria and also a number of outliers identified in the red ellipses for consideration as anomalies compared to those clusters within the blue ellipses. An interesting observation is those in Cluster 5, new development may bring opportunities in both new construction in the geographic location and also opportunities to realize those under the horizontal blue line as consideration for future investment.

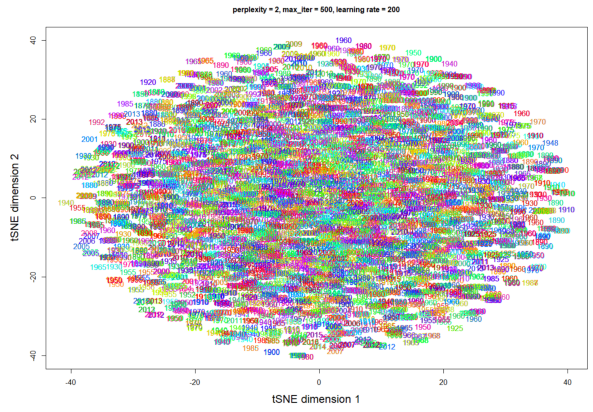
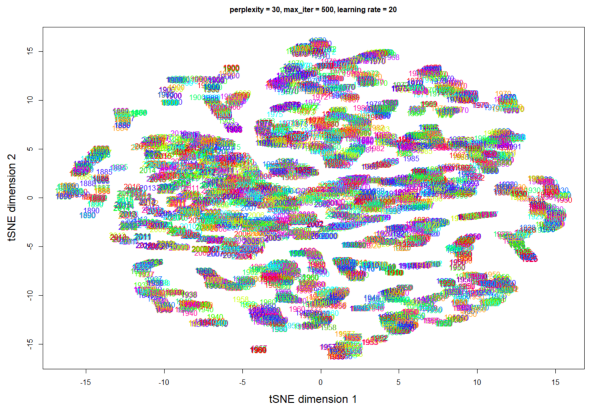
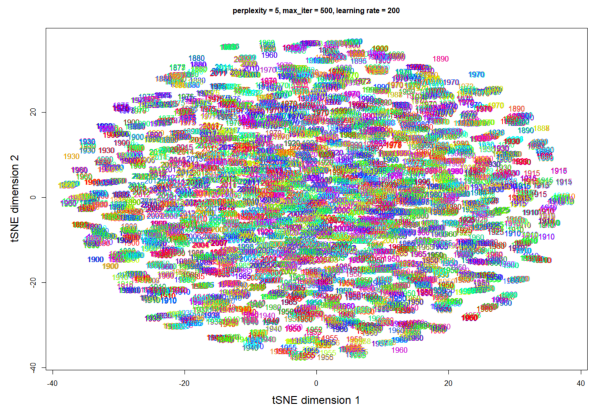
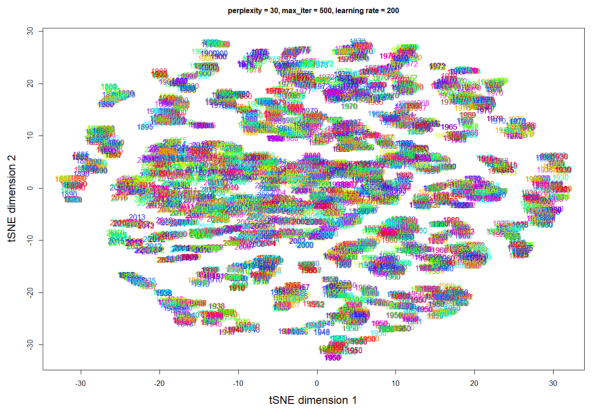
Conclusion

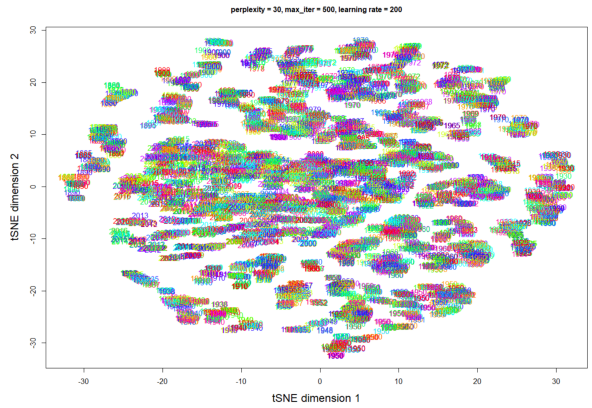
The analysis of the dataset provides a method to cluster specific property types and identifies those clusters for both the marketing of properties as well as determining the

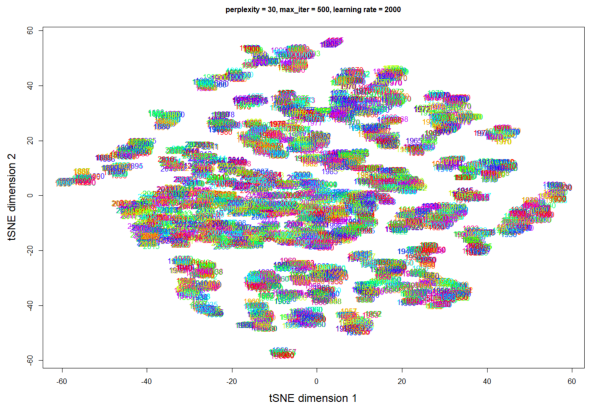
appropriate valuation. Outliers in the visualization will point to one of two elements. A property is undervalued allowing for specific consideration of reappraisal value or a property is overvalued and determining the appropriate valuations based on the k-means clustering method employed. The limitation of the method is to provide generalizations of specific characteristics in determining the associate groupings of housing types. Specific outliers will need to be reviewed in order to understand the opportunity or risk associated with its characterization in this data analysis methodology.

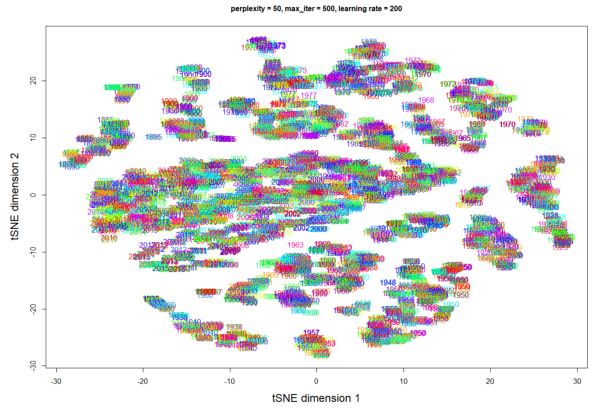
Appendix A

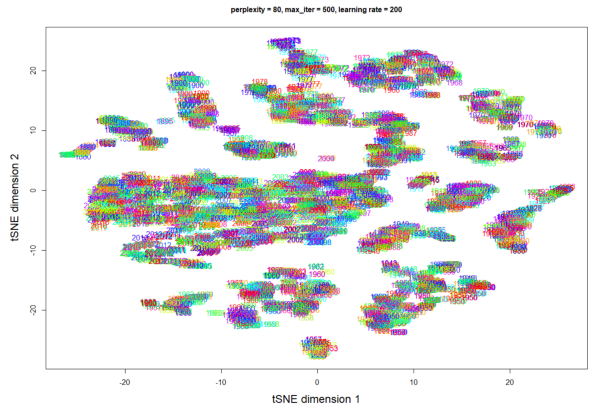
Hyperparameter tuning for tSNE - over several iterations the perplexity and learning rate provided correlation between patterns in the high-dimensional data and cluster like structures.

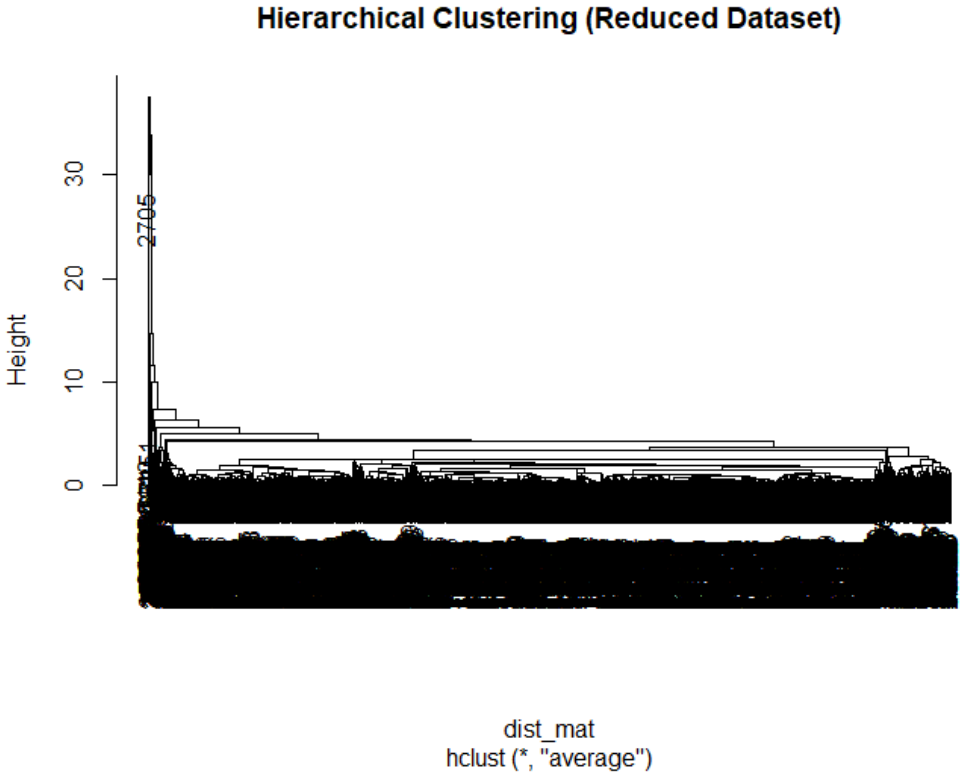
Perplexity (2,5,30,50,80)	Learning Rate (20,200,2000)
	
	

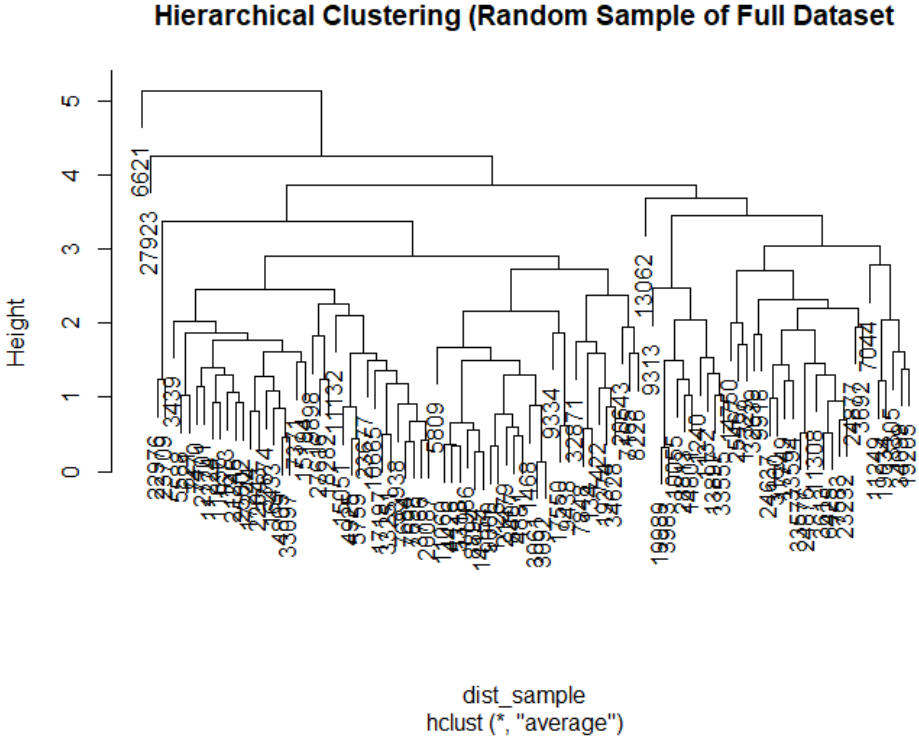

perplexity = 30, max_iter = 500, learning rate = 200
tSNE dimension 1
tSNE dimension 2


perplexity = 30, max_iter = 500, learning rate = 2000
tSNE dimension 1
tSNE dimension 2


perplexity = 50, max_iter = 500, learning rate = 200
tSNE dimension 1
tSNE dimension 2


perplexity = 80, max_iter = 500, learning rate = 200
tSNE dimension 1
tSNE dimension 2


Hierarchical Clustering (Reduced Dataset)
Height
0
10
20
30
2705
dist_mat
hclust (*, "average")


Hierarchical Clustering (Random Sample of Full Dataset
Height
0
1
2
3
4
5
6621
27923
13062
9313
7044
dist_sample
hclust (*, "average")

References

data4help. (2020, November 2). *Clustering Real Estate Data*. Medium. Retrieved February 8, 2022, from https://becominghuman.ai/clustering-real-estate-data-594894e24484

Hepsen, A., & Vatansever, M. (2012). Using hierarchical clustering algorithms for Turkish residential market. *International Journal of Economics and Finance, 4*(1), 138-150.

Nallathambi, J. (2018, June 20). *R series - K means clustering (silhouette)*. Medium. Retrieved February 11, 2022, from https://medium.com/codesmart/r-series-k-means-clustering-silhouette-794774b46586