Assignment #2: Segmentation of Melbourne Housing Data

MSDS 411 Unsupervised Learning Methods

Alison Au, James Maxwell, Jhansi Munagala

10/31/2021

Abstract

In today's super competitive property market, real estate companies must employ any and all available technology to gain an edge that enables them to succeed. Unsupervised learning algorithms can be used in this domain by segmenting the market to gain insight into transaction data, provide better service to our clients, and enable efficiencies across the business.

Introduction

We are a real estate company located in Melbourne, Australia that provides sales and rental services to our residential customers in the surrounding area. Data drives our business model as we continue to enhance our business to understand our client needs and match them with appropriate housing. Our goal is to remain the leading revenue-producing real estate firm. We have initiated this project utilizing unsupervised learning methods to segment housing stock in this city based on available sales data. These results will be leveraged for targeted marketing of potential home buyers, showcasing homes that match their style of preference. This will also help guide the assignment of our realtors to properties, aligning their experience with particular types of homes. This project demonstrates the use of clustering algorithms to divide the housing market into distinct groups to enhance our knowledge base and help optimize our business.

Keywords

Unsupervised learning, Exploratory data analysis, Normalization, Hierarchical agglomerative clustering, K-means clustering, Data visualization, Dendrograms, t-SNE

Literature Review

The literature reviewed for this study was exclusively regarding clustering methodology and implementation, and the references are listed later in this report. Clustering essentially constructs groups based on distance where observations within a group are as similar as possible, and groups must be as different from each other as possible. A multitude of clustering techniques exist with k-means and hierarchical clustering being some of the most common and easily interpretable. Both approaches will be explored in this analysis.

Methods

The raw data used for this study is a .csv file of transactions regarding Melbourne housing sales from 2016 to 2018 and contains 34,857 observations and 21 variables. After inspection of this initial data set, we decided to retain 11 of the most relevant variables and only observations with complete data given the risk of skewing with imputing many missing data points. Following this decision, our working data set retains 8,887 observations and 11 variables (Type, Suburb, Rooms, Price, Distance, Bedroom2, Bathroom, Car, Landsize, BuildingArea, Age). A data dictionary is included as Figure 1 in the appendix. All figures referred to hereafter are in the appendix. The Age variable is a transformation of the original YearBuilt variable where Age = (2021 - YearBuilt). Only 2 variables, Type and Suburb, are non-numeric data.

Exploratory data analysis (EDA) was performed on 9 variables, excluding Type and Suburb, to understand the nature of the data using box plots, histograms, QQ-plots, and a scatter plot matrix. It was noted that the distributions were not normal distributions, containing both skew and outliers. These outliers have been retained given the business' interest in exploring the entire

Melbourne housing market. Clustering was also tested with retaining versus removing outliers, and models with the former explained higher variability (sum of squares) of the data. EDA output is included in Figure 2.

Additionally for EDA, T-distributed Stochastic Neighbor Embedding (t-SNE), a machine learning algorithm for data reduction and visualization was applied on our working data set. The goal here was to reduce the dimensionality of the data to visualize the data in 2 dimensions and see if any distinct clustering became evident. An example of the output is in Figure 3. The attempts applying t-SNE were unsuccessful and this can be revisited in the future.

To solve this business problem, we have chosen clustering analysis. Clustering analysis will discover the patterns in the data by grouping the similar data groups together. For this we have applied two of the most popular techniques in hierarchical agglomerative and k-means.

We chose to normalize the features in the working data set with a min-max scaler prior to applying our cluster analysis because, as noted in the EDA, the distributions were not normal. The min-max scaler also produced better results than the standard scaler in terms of more distinct clusters and higher total variance explained.

In hierarchical clustering, we tried an agglomerative method in which each object is considered as a single element cluster. It is based on Euclidean distance measures. In each step of the algorithm the 2 clusters which are most similar are combined. Once the similarity matrix is calculated, we tried different methods of linkage functions such as Ward's, average, and complete linkage in order to group pairs of objects into clusters. These clusters can be

represented graphically using dendrograms. Only the Ward method produced clear divisions with 4 clusters being a suitable cut. This dendrogram and cluster plot are shown in Figure 4.

We found k-means clustering to produce more logical and applicable segments while being more intuitive to work with. The optimal number of clusters for this data set was determined by plotting the total within-cluster sum of squares (WCSS) by “k” clusters. As shown in Figure 5 “k” bends and stops significantly decreasing WCSS around 3 to 5. Using a combination of this plot and business sense, we landed on 5 clusters, ensuring enough segments for a targeted strategy while having distinct groupings.

PCA prior to k-means clustering was tested, however, this led to lower variability explained by the model. A comparison of explained variance by different k-means models tested is in Figure 6. To summarize, the best approach used normalized data with outliers retained and no PCA.

Results

After evaluating various methods, five final clusters have been defined using k-means clustering on the normalized data with outliers retained. A graph of the clusters and a table of their property metric means are in Figures 7 and 8. While the clusters are not very distinct graphically, indicating some overlap, the mean values show distinguishable characteristics. A plot of the segments on a map shown in Figure 9 illustrates that the segments are closely related to their location, which logically makes sense since properties in the same area tend to be similar. This also aligns strategically given that location is critical in home purchasing.

The five clusters are detailed as follows. Cluster 1 homes are farther from the Central Business District (CBD) at 15 km with an approximately average price for Melbourne ($861,543) and size

(558 m^2 land size, 3 bedrooms). This is the largest cluster and contains typical suburban homes. This is suitable for people wanting space and some accessibility to downtown at a more affordable cost.

Cluster 2 is the most centrally located (7 km from CBD) but also smallest properties, mainly consisting of apartment units. On average, they have just 1-2 bedrooms and 1 bathroom but are affordable at $690,681. These fit owners who value the hustle and bustle of downtown or convenient work commute over living space. Typically, these tend to be young adults and couples who do not have kids.

Homes in segment 3 are central (10 km from CBD) but also large with 4 bedrooms, 2-3 bathrooms, and 691 m^2 land size. This combination of highly desirable traits make them the most expensive properties at $2,070,698 on average. They cluster around the central eastern part of Melbourne such as Balwyn North, Brighton East, and Glen Iris neighborhoods. Owners are expected to be very high income families who can afford space in a prime location.

The fourth cluster is also central (7 km from CBD) but a step down from cluster 3, as they are slightly smaller, less expensive, and the oldest properties. They are still pricey at an average of $1.2 M with 3 bedrooms and 1-2 bathrooms. This is appropriate for high income owners wanting to reside near downtown but not needing the space or not able to afford the homes in cluster 3.

The fifth and final segment are the farthest homes in the outskirts of town but also the largest (916 m^2 land size), newest, and most affordable ($675,755). Target owners are those seeking affordability, space, and/or newer homes even if that means being farther from the city center.

Conclusions

Utilizing clustering models, the Melbourne housing market has been segmented into five distinct groups: 1) typical suburban homes outside downtown, 2) small, central apartments, 3) priciest homes that are large and central, 4) central homes a tier below group 3, and 5) outskirt properties which are the largest, newest, and most affordable. Potential home buyers should be associated with the segment most suitable for them, enabling marketing and home viewing to be better tailored to their preference. Real estate agents can also be assigned properties based on the housing segment where they hold the most experience. These clusters can be leveraged for additional analyses such as determining which ones drive the highest profitability for the company. Strategic effort can then be focused on selling these home types more in the future.

While these clusters are valuable for the business already, we want to continue exploring other unsupervised techniques to further differentiate housing stock. Other clustering algorithms may lead to more distinct clusters with less overlap. Additionally, retaining just the properties with complete data for this analysis resulted in dropping 75% of the records. We can look to fill in the missing data from more complete sources or collect more sales data moving forward. Overall, segmentation is a naturally subjective and iterative process, so we can look to refine the clusters to meet ongoing business needs.

References

Kabacoff, Robert. 2015. *R In Action*: *Data Analysis and Graphics in R (2nd edition)*. Shelter Island, NY: Manning.

"K Means Algorithms in R ". *learnbymarketing,* https://www.learnbymarketing.com/tutorials/k-means-clustering-in-r-example/

"Hierarchical Clustering in R". 2016. *r-bloggers,* https://www.r-bloggers.com/2016/01/hierarchical-clustering-in-r-2/

"How to Perform Hierarchical Clustering using R". 2017. *r-bloggers,* https://www.r-bloggers.com/2017/12/how-to-perform-hierarchical-clustering-using-r/

"The complete guide to clustering analysis: k-means and hierarchical clustering by hand and in R". 2020. *statsandr,* https://statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r/

"Clustering Data in R". 2021. *rpubs,* https://rpubs.com/pjmurphy/599072

"Cluster Analysis". *statmethods,* https://www.statmethods.net/advstats/cluster.html

"How to interpret the clusplot in R".2017. *stackexchange,* https://stats.stackexchange.com/questions/274754/how-to-interpret-the-clusplot-in-r

"Comprehensive Guide on t-SNE algorithm with implementation in R & Python". 2017. *analyticsvidhya,* https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/

"R Programming Live - Lecture 6 | Cluster Analysis - Concepts and Application". 2020. *youtube,* https://www.youtube.com/watch?v=otjWCaMcVaA

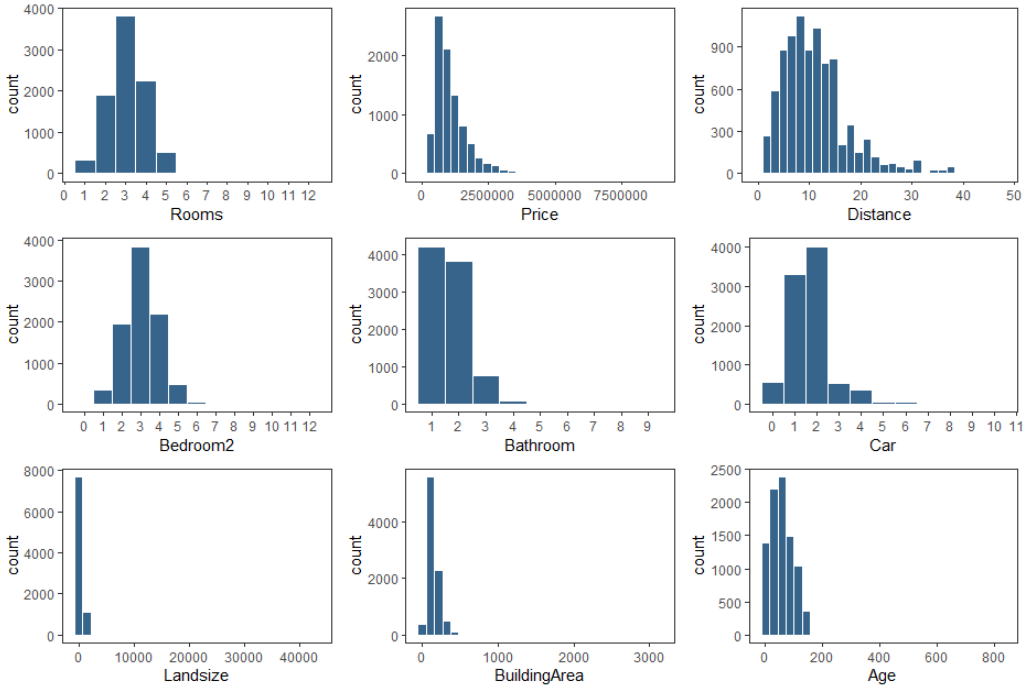
"Data normalization in machine learning". 2020. *towardsdatascience,* https://towardsdatascience.com/data-normalization-in-machine-learning-395fdec69d02

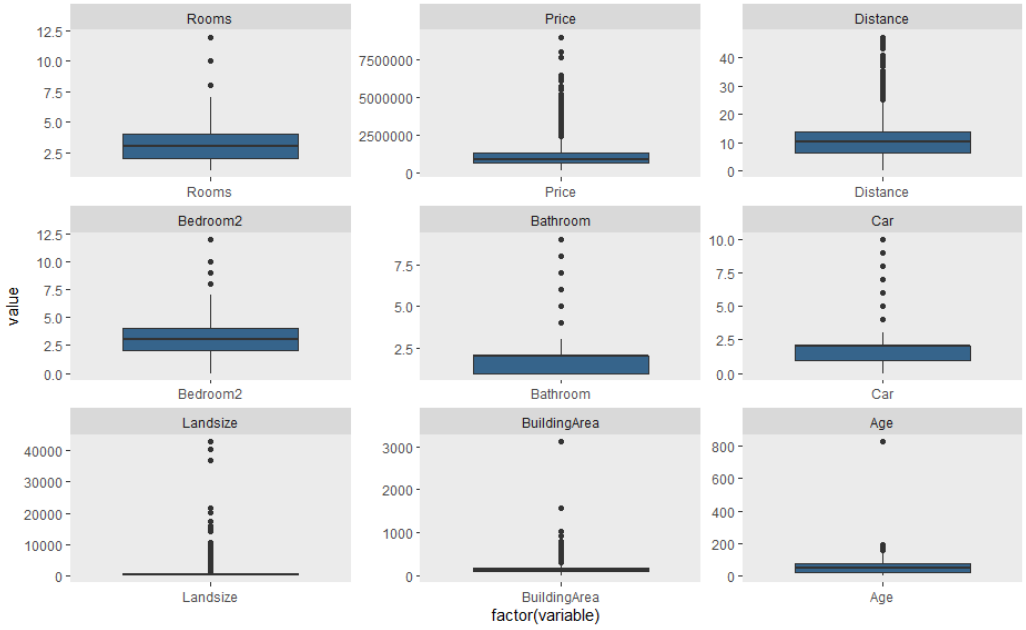
Appendix

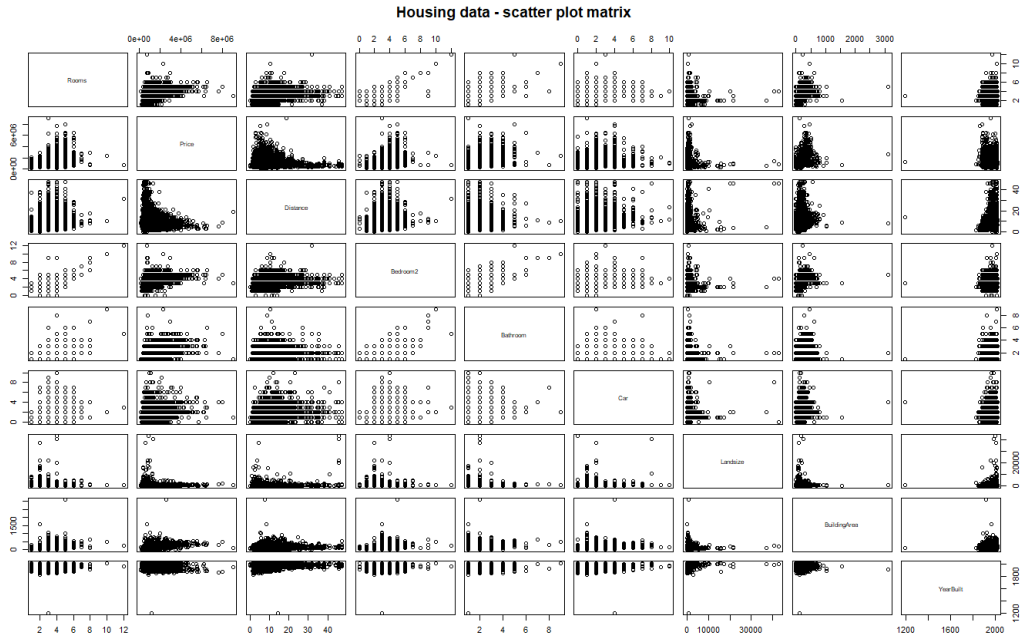
Figure 1. Variables in the data set.

Variable	**Meaning**
Rooms	Number of rooms
Price	Price in Australian dollars
Distance	Distance from CBD in Kilometres
Bedroom2	Scraped # of Bedrooms
Bathroom	Number of Bathrooms
Car	Number of carspots
Landsize	Land Size in Metres
Building Area	Building Size in Metres
YearBuilt	Year the house was built
Type	House type: h - house,cottage,villa, semi,terrace u - unit, duplex t - townhouse
Suburb	Suburb Name

Figure 2. EDA output.






Housing data - scatter plot matrix
Rooms
Price
Distance
Bedroom2
Bathroom
Car
Landsize
BuildingArea
YearBuilt

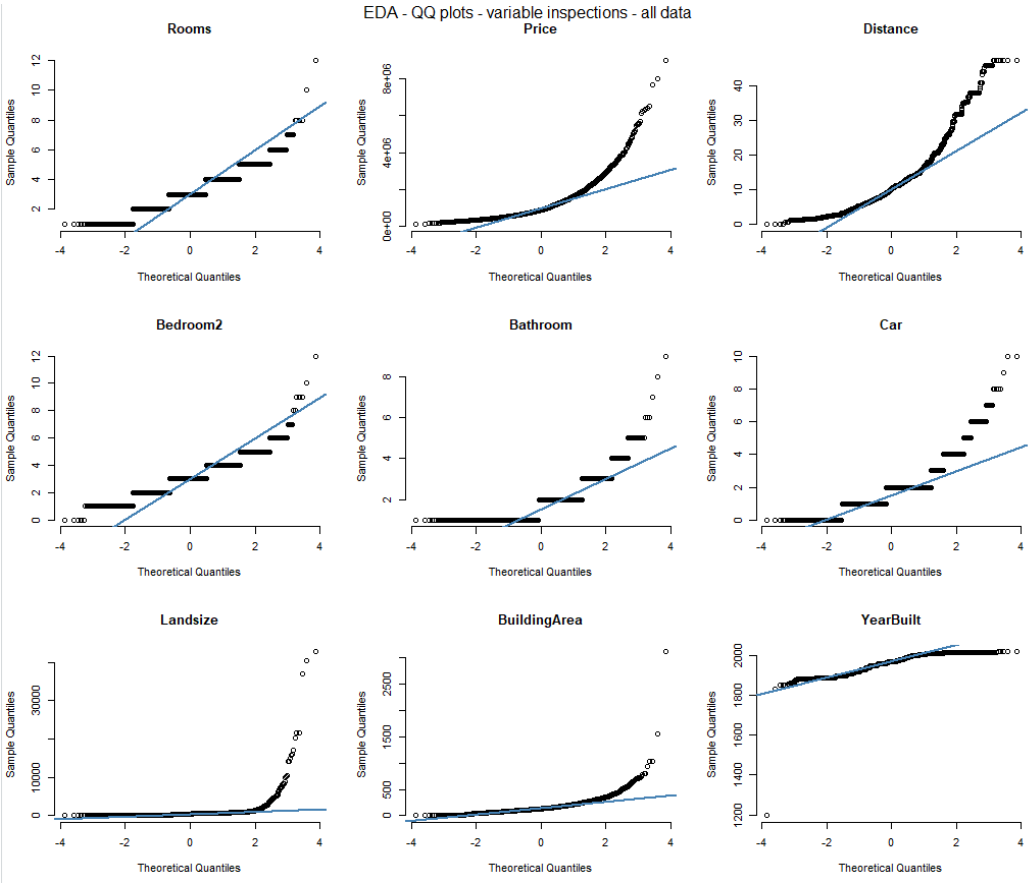

EDA - QQ plots - variable inspections - all data
Rooms
Price
Distance
Bedroom2
Bathroom
Car
Landsize
BuildingArea
YearBuilt
Sample Quantiles
Theoretical Quantiles

Figure 3. Example t-SNE output (unsuccessful).

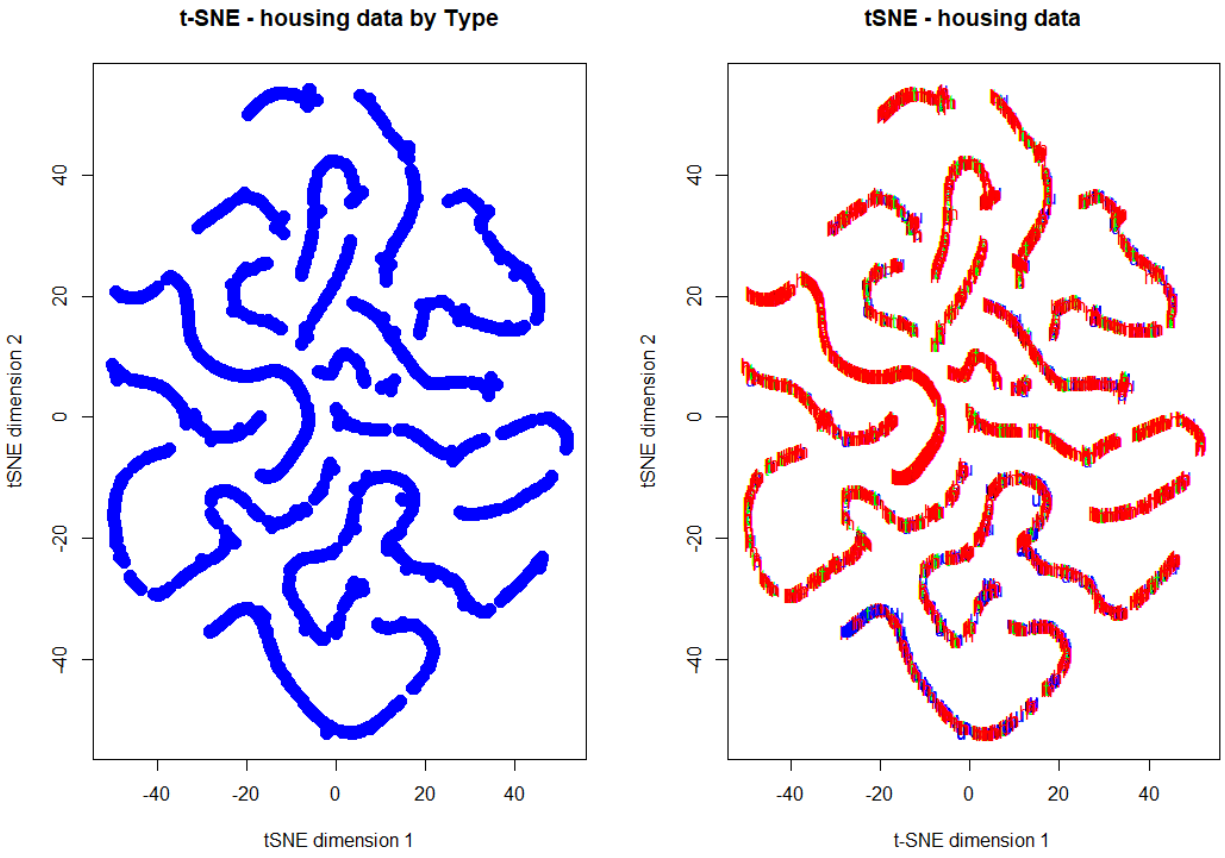
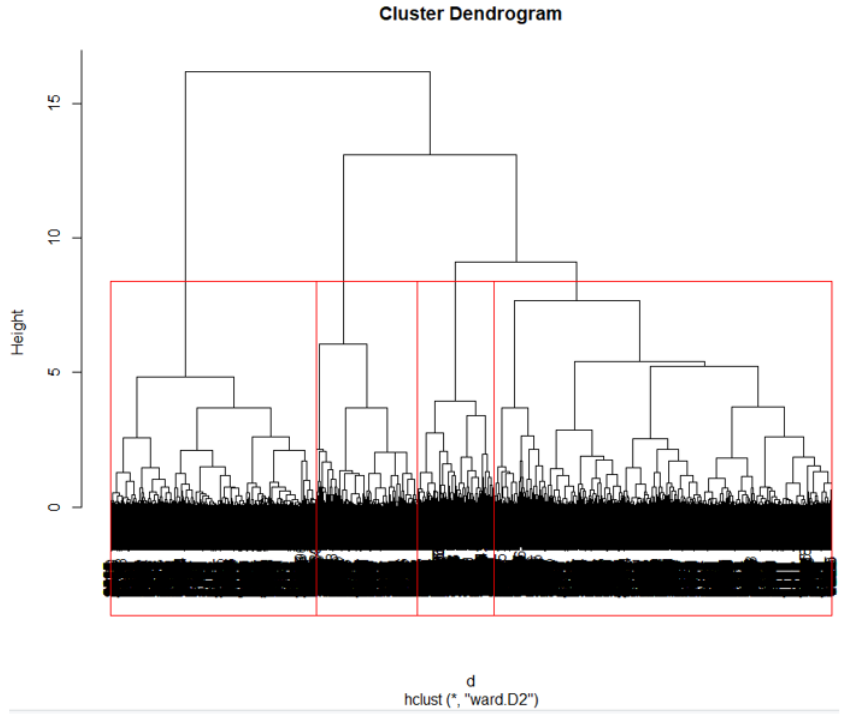


Figure 4. Output dendrogram and cluster plot from Wards method and K=4.



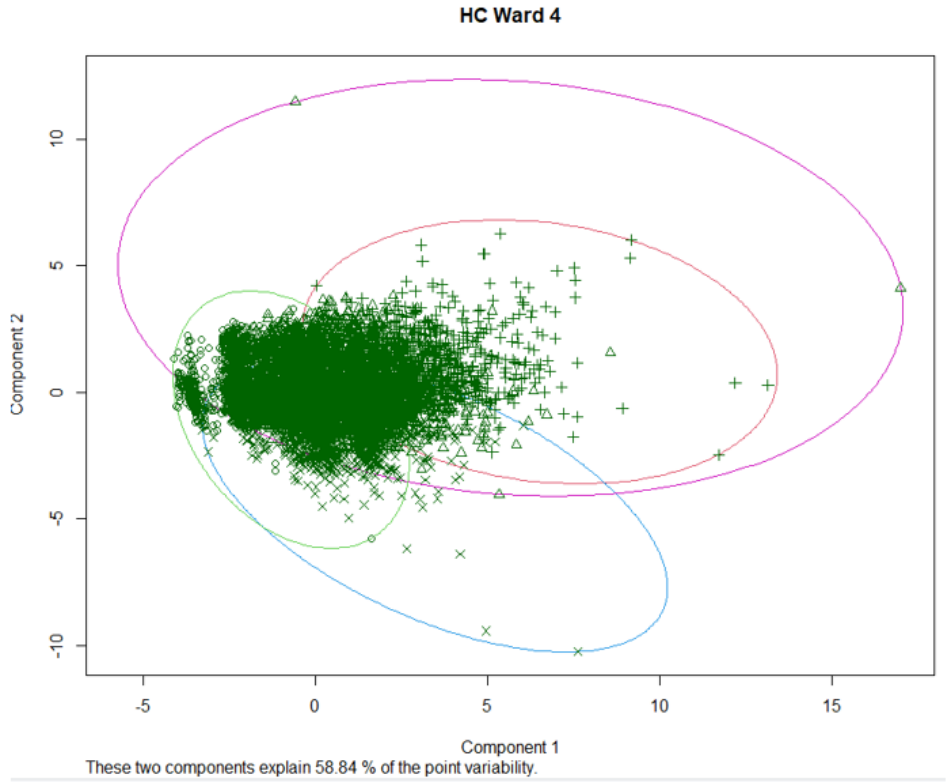


Figure 5. Determining the number of clusters in K means.

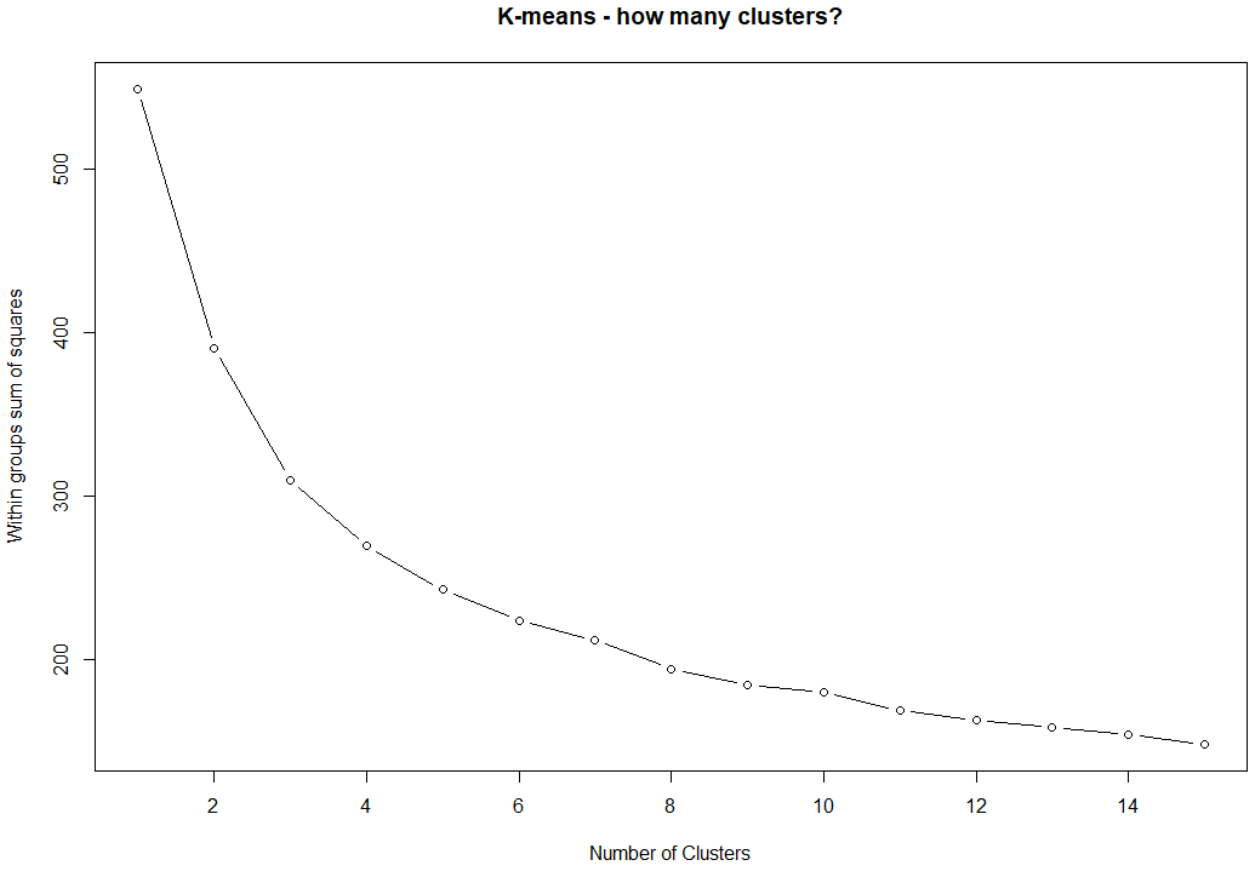


Figure 6. K-means model comparison.

K-means model comparisons

Model	# of Clusters	Data Scaling	Outliers	PCA Prior	Between SS / Total SS
1	3	Standardized	Retained	N	0.35
2	3	Normalized	Retained	N	0.44
3	3	Normalized	Removed	N	0.4
4	4	Standardized	Retained	N	0.42
5	4	Normalized	Retained	N	0.51
6	4	Normalized	Removed	N	0.48
7	4	Normalized	Retained	Y	0.38
8	5	Standardized	Retained	N	0.48
9	5	Normalized	Retained	N	0.56
10	5	Normalized	Removed	N	0.53
11	5	Normalized	Retained	Y	0.45

Figure 7. K-means cluster plot for k=5.

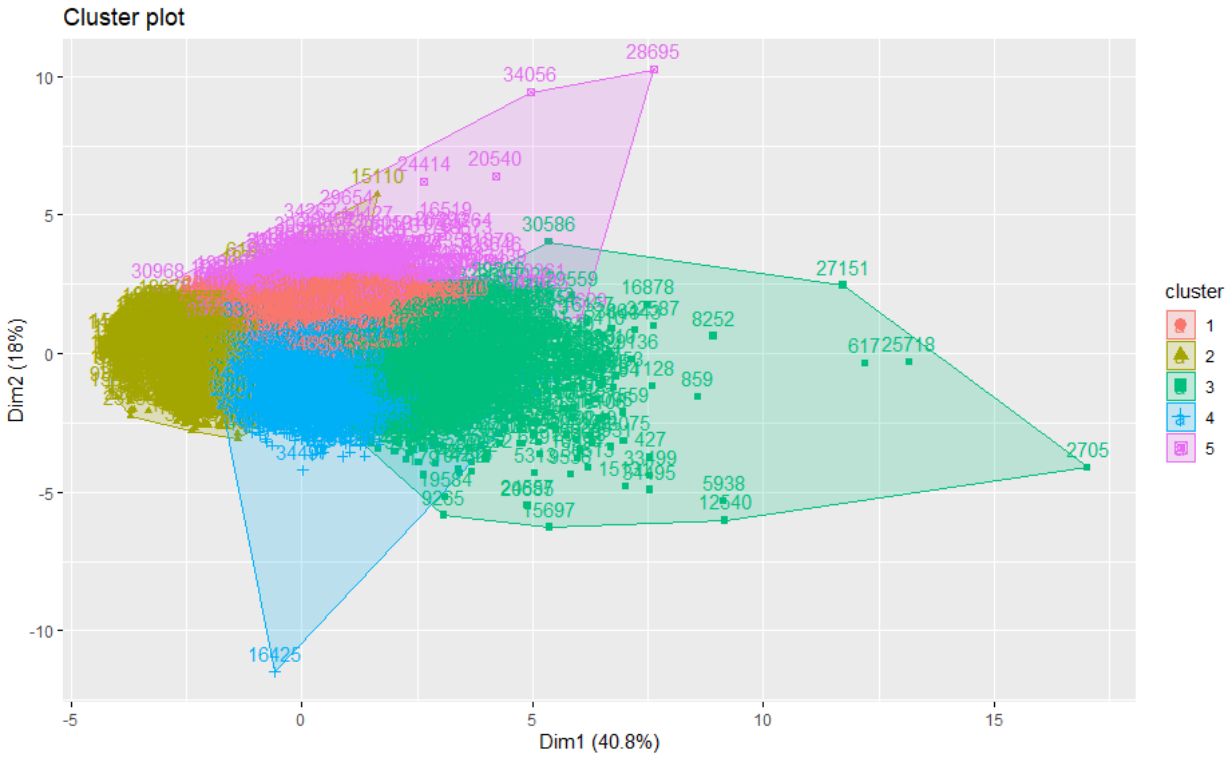


Figure 8. Mean values by Cluster.

	Cluster					Total
Type	**1**	**2**	**3**	**4**	**5**	
h	2,180	704	1,236	1,979	526	**6,625**
t	220	197	57	240	8	**722**
u	123	1,270	1	133	13	**1,540**
Total	**2,523**	**2,171**	**1,294**	**2,352**	**547**	**8,887**

h - house,cottage,villa, semi,terrace
t - townhouse
u - unit, duplex

Figure 9. Mapped Clusters.

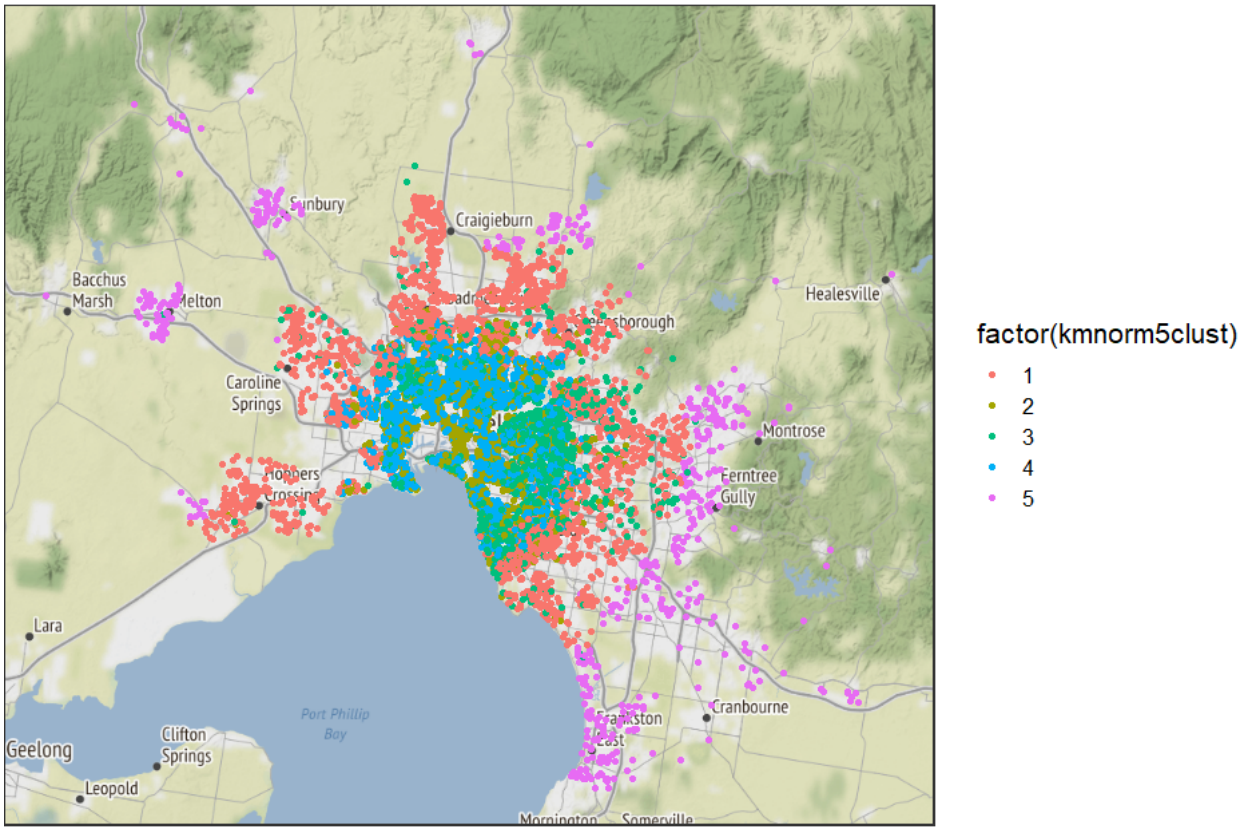


Figure 10. Count of Properties by Cluster and Type.

	Cluster					Total
Type	**1**	**2**	**3**	**4**	**5**	
h	2,180	704	1,236	1,979	526	**6,625**
t	220	197	57	240	8	**722**
u	123	1,270	1	133	13	**1,540**
Total	**2,523**	**2,171**	**1,294**	**2,352**	**547**	**8,887**

h - house,cottage,villa, semi,terrace
u - unit, duplex
t - townhouse