Assignment #2

David Caban

Bronsin Jabraeili

Scott Jue

Dr. Joe Wilck

Northwestern University

MSDS 411 – Unsupervised Learning Methods

April 30, 2023

Abstract

This paper highlights the importance of using clustering analysis in the real estate industry to gain a competitive advantage. The Melbourne dataset, consisting of 21 columns and 34,857 rows of data, is used as an example of how clustering analysis can provide valuable insights into key activities such as marketing, valuation, and identification of outliers. The study reviews several research papers and articles demonstrating the effectiveness of clustering analysis in identifying housing submarkets, improving the accuracy of real estate appraisal, and identifying distinct submarkets based on various property characteristics and location factors. The paper also presents the methodology used to implement clustering techniques, including outlier search and scaling. Both hierarchical and k-means clustering techniques were used and compared to identify the optimal number of clusters to use. After comparing the results, it was determined that the k-means clustering model provides the most useful results, with either 4 or 5 clusters providing the best fit for the dataset. The study concludes that clustering analysis is a well-documented and effective tool for addressing real estate problems, and can help inform decision-making and improve outcomes in the real estate industry.

Keywords: *Hierarchical Clustering, K-means Clustering, Outlier Detection, Unsupervised Learning*

Introduction

The real estate industry is highly competitive, unpredictable, and vast in scope, resulting in a plethora of data readily available to synthesize. However, many real estate firms fail to capitalize on the opportunity to gain a competitive advantage by utilizing data science techniques

to obtain valuable insights regarding key activities such as: marketing their services to potential buyers, valuation and comparison of properties, and identification of outliers. A powerful yet intuitive unsupervised learning technique to answer the questions a real estate firm would be interested in answering as described above is clustering analysis. More specifically, the firm should implement both hierarchical clustering and k-means clustering because the best approach is to utilize multiple techniques and compare the results. The Melbourne dataset is an excellent proxy for real estate firms to explore the usefulness of clustering algorithms for important real estate considerations. The dataset consists of 21 columns and 34,857 rows of data, providing a holistic representation of important metrics for consideration including: Price, Rooms, Year Built, and Suburb to name a few. The vast amount of records provides no value in its raw form, clustering analysis will allow the firm to see the various subgroups that exist within the data. This better understanding of the different segments will allow for key strategic decisions to be made specifically for identified, targeted customers that optimize value for the firm.

Literature Review

Several research papers and articles have been published on the topic of clustering analysis for real estate problems, demonstrating that this method is well-documented and widely recognized in the field. For example, a study by John Steller and John C. H. Chiang (2006) applied clustering analysis to identify housing submarkets in Brisbane, Australia, based on various demographic and economic variables. The study found that clustering analysis was effective in identifying meaningful submarkets that could be used to inform real estate decision-making. Similarly, T. H. Jackson and P. M. Ortman (2001) discussed how cluster analysis can be

used for real estate appraisal in an article published in the Journal of Real Estate Appraisal and Economics. They described how cluster analysis can be used to identify similar properties for appraisal purposes, improving the accuracy and reliability of the appraisal process. In another example, a study by Chongyu Wang, Yu Chen, and Wei Huang (2017) implemented clustering analysis to the housing market in Toronto, Canada, to identify distinct submarkets based on various property characteristics and location factors. The authors found that clustering analysis was able to identify meaningful submarkets that were consistent with the local housing market. Finally, a study by Yanshan Chen, Geoffrey K. Turnbull, and Velma Zahirovic-Herbert (2015) used clustering analysis to identify distinct housing markets in Beijing, China, based on various housing and neighborhood characteristics. The authors found that clustering analysis was effective in identifying distinct markets that could be used to inform real estate decision-making in the region.

Overall, these studies demonstrate that clustering analysis is a well-documented and effective tool for addressing real estate problems. By using this technique, analysts and practitioners can gain valuable insights into housing submarkets, property appraisal, and the housing market as a whole, helping to inform decision-making and improve outcomes in the real estate industry.

Methodology

Prior to the application of clustering techniques, an outlier search was performed on the data. This search was conducted by inspecting each variable via a boxplot and calculating skewness and kurtosis. Based on this inspection, outliers were dropped to account for the fact

that k-means clustering is sensitive to their presence. Please refer to Exhibit A for more information.

With outliers removed, the data was scaled to ensure that all variables could contribute equally to the clustering processes for which two methods, hierarchical and k-means, were used. Scaling was achieved by subtracting the sample mean from each actual value, and dividing by the standard deviation of the particular variable.

Hierarchical clustering was performed by first computing the euclidean distance of all pairs of rows in the data to create a distance matrix. Euclidean distance was used as the variables were largely continuous in nature, the importance of the variables was believed to be equal, outliers had been removed, and it is a straight-line measure commonly used and easy to explain to users of various disciplines.

Three methods were used to create dendrograms; complete, average, and single linkage. For each method, the dendrogram was reviewed and the practicality, interpretability, and cluster shapes reviewed.

K-means clustering was performed by using the R package ‘factoextra’ and the within-cluster sum of squares method to generate a plot in which the ‘elbow-point’ could be used to identify the optimal number of clusters to use. Additionally, the gap statistic was calculated and a second plot generated to visualize the optimal number of clusters for the data. These plots can be found in exhibit B.

K-means clustering was performed, based on the results of the above analysis, using 4, 5, 6, and 7 clusters. Varying number of initial centroids were workshopped but it was determined that the results didn’t materially change across ranges. A rendered cluster plot looked largely

unaffected when using 5 or 1,000 initial centroids. For each cluster created, summary statistics were calculated and compared. T-SNE plots were created to further visualize the data, but it was determined that they did not add additional value when compared to the cluster plots.

Results

After comparing the results for each of the models, the k-means clustering model provides the most useful results for generating recommendations on housing cluster subsets. It was determined that using either 4 or 5 clusters provided the best fit for the data set. The final solution implements 4 clusters in order to avoid overfitting and to facilitate generalization of the model. Furthermore, having fewer clusters simplifies the interpretation of the results and makes it easier to draw meaningful insights from each cluster. A t-SNE plot is provided in Exhibit C; however, this visualization did not provide any meaningful investigatory value for the analysis. On the other hand, the visualization provided in Exhibit D shows how the data points have been grouped based on their similarity. This allows the firm to explore and understand the structure of the data in a more intuitive way. An interesting pattern seen in this visualization is that cluster 2 is more spread out and less dense than the other clusters. The original data set contains certain characteristics that differentiate each of the four clusters which can be seen Exhibit E and also summarized below.

Cluster 1: This cluster has the highest proportion of house/cottage/villa/semi/ terrace (31%) and low proportion of townhouses (5%) and a moderate proportion of units/duplex (16%). The mean price of properties in this cluster is relatively low compared to other clusters, and the mean number of rooms, bedrooms, and bathrooms is also low. The mean distance from the city center is relatively high, indicating that properties in this cluster are located further away from

the city center. The mean land size and building area are also relatively small. This cluster contains newer properties as the mean year built is around 1978.

Cluster 2: This cluster consists of primarily house/cottage/villa/semi/ terrace type properties. The mean price of properties in this cluster is the highest among all clusters, and the mean number of rooms, bedrooms, and bathrooms is also high. The mean distance from the city center is relatively low, indicating that properties in this cluster are located closer to the city center. The mean land size and building area are also relatively large. These properties are the third newest of the four clusters with the mean year built around 1950.

Cluster 3: This cluster has a relatively high proportion of house/cottage/villa/semi/ terrace (27%) and a low proportion of townhouses (3%) and units/duplex (1%). The mean price of properties in this cluster is moderate, and the mean number of rooms, bedrooms, and bathrooms is also moderate. The mean distance from the city center is relatively high, indicating that properties in this cluster are located further away from the city center. The mean land size and building area are also relatively small. These properties are the second newest with the mean year built around 1955.

Cluster 4: This cluster has a moderate proportion of house/cottage/villa/semi/ terrace (14%) and an insignificant amount of townhouses and units/duplex. The mean price of the properties in this cluster is relatively high, and the mean number of rooms, bedrooms, and bathrooms is also moderate. The mean distance from the city center is relatively low, indicating that properties in this cluster are located closer to the city center. The mean land size and building area are also relatively large. The properties in this cluster are the oldest as the mean year built is around 1945.

In summary, the k-means clustering analysis has identified four distinct clusters of properties based on their characteristics. These clusters can be interpreted as follows: Cluster 1 includes smaller, less expensive properties located further from the city center; Cluster 2 includes larger, more expensive properties located closer to the city center; Cluster 3 includes moderately sized and priced properties located further from the city center; and Cluster 4 includes larger, more expensive properties located at a moderate distance from the city center.

Conclusion

This analysis highlights the importance of utilizing data science techniques, specifically clustering analysis, in the real estate industry to gain valuable insights into housing submarkets, property appraisal, and the housing market as a whole. Using a k-means clustering approach, four different meaningful subgroups were identified within the Melbourn housing data set. The results of these clusters can help the firm identify key market segments based on various attributes such as location, property type, size, price range, and amenities. This information can be used to develop targeted marketing strategies and sales campaigns. By utilizing cluster analysis, the firm can better evaluate comparable properties and improve the precision and insight of their decisions regarding pricing, marketing, and other crucial factors of a real estate transaction. Moreover, when combined with other data science methods, clustering analysis can help the real estate firm gain a competitive edge, become more strategic, and achieve greater success in the market.

References

Chen, Yanshan, Geoffrey K. Turnbull, and Velma Zahirovic-Herbert. "Cluster Analysis of Housing Markets in Beijing, China." *Journal of Real Estate Finance and Economics 51, no. 1* (2015): 35-58.

Jackson, T. H., and P. M. Ortman. "Cluster Analysis for Real Estate Appraisal." *Journal of Real Estate Appraisal and Economics 23, no. 4* (2001): 217-234.

Steller, John, and John C. H. Chiang. "Identifying Housing Submarkets Using Cluster Analysis: A Case Study of the Brisbane Housing Market." *Journal of Real Estate Research 28, no. 4* (2006): 367-388.

Wang, Chongyu, Yu Chen, and Wei Huang. "Cluster Analysis of the Housing Market in Toronto, Canada." *Journal of Real Estate Literature 25, no. 2* (2017): 277-295.

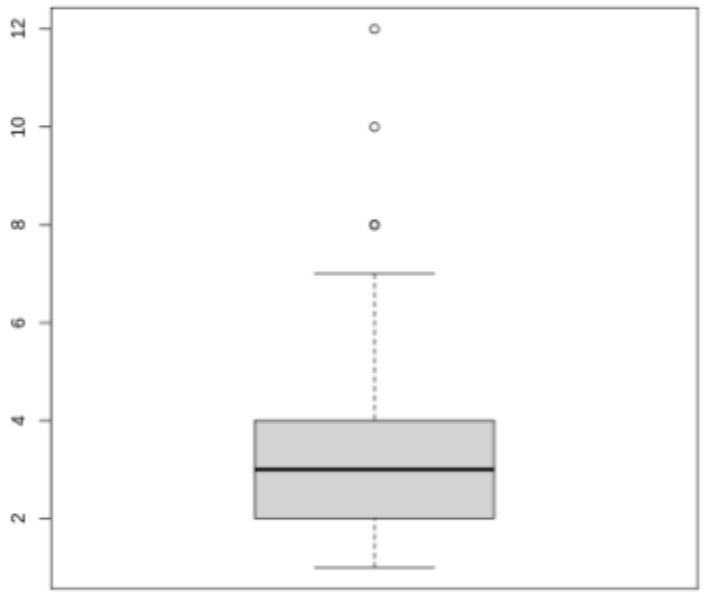
Appendix

Exhibit A - Outlier Detection

Rooms

```
# check outliers in Rooms
skewness(df3$Rooms)
kurtosis(df3$Rooms)
boxplot(df3$Rooms)
```

```
0.327254395510821
4.37945139343599
```




```
[ ] # identify the  outliers
    outliers <- boxplot.stats(df3$Rooms)$out
    head(outliers)
```

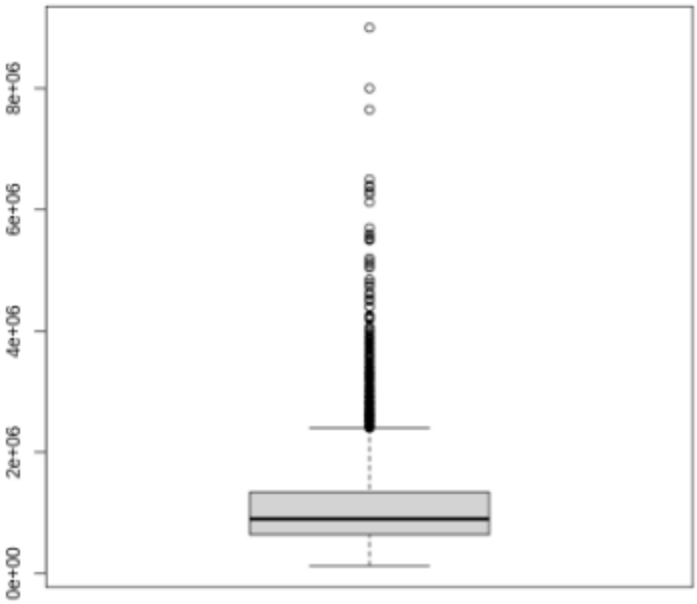
```
8 · 8 · 8 · 8 · 10 · 12
```

```
[ ] df3 <- subset(df3, df3$Rooms < 10)
```

Price

```
# check outliers in Price
skewness(df3$Price)
kurtosis(df3$Price)
boxplot(df3$Price)
```

```
2.41194694793075
14.044901647098
```




```
# identify the  outliers
outliers <- boxplot.stats(df3$Price)$out
head(sort(outliers, decreasing = TRUE))
```

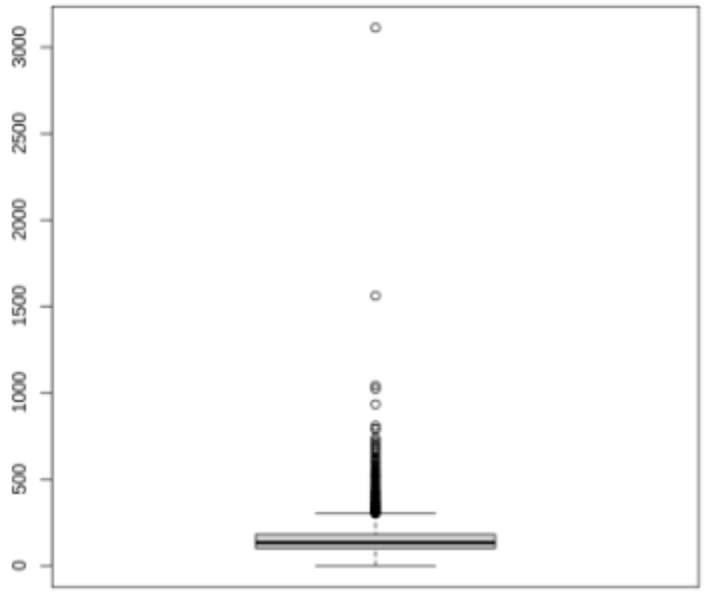
```
9000000 · 8000000 · 7650000 · 6500000 · 6400000 · 6370000
```

```
df3 <- subset(df3, df3$Price < 7650000)
```

Building Area

```
# check outliers in Building area
skewness(df3$BuildingArea)
kurtosis(df3$BuildingArea)
boxplot(df3$BuildingArea)
```

```
6.51940598555105
163.588226853036
```




```
# identify the  outliers
outliers <- boxplot.stats(df3$BuildingArea)$out
head(sort(outliers, decreasing = TRUE))
```

```
3112 · 1561 · 1041 · 1022 · 934 · 808
```

```
min(df3$BuildingArea)
```

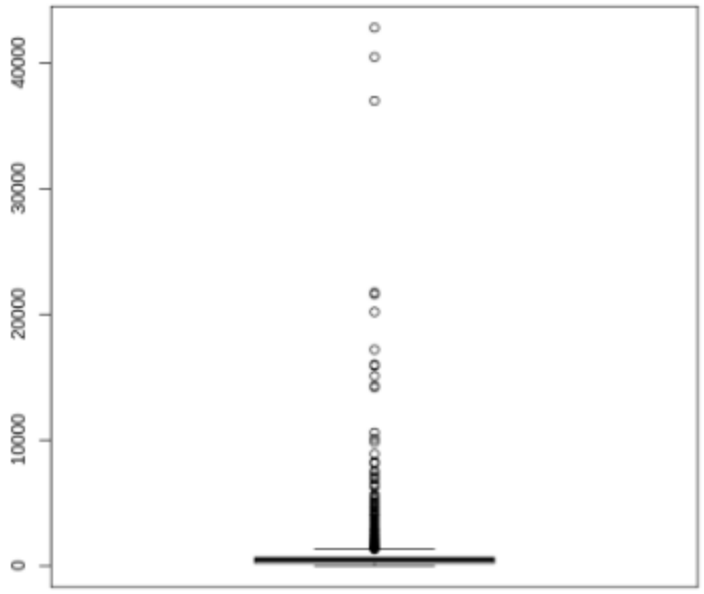
```
0
```

```
df3 <- subset(df3, df3$BuildingArea < 1500)
df3 <- subset(df3, df3$BuildingArea > 0)
```

Landsize

```
# check outliers in landsize
skewness(df3$Landsize)
kurtosis(df3$Landsize)
boxplot(df3$Landsize)
```

```
23.140626070367
749.202982634875
```




```
# identify the  outliers
outliers <- boxplot.stats(df3$Landsize)$out
head(sort(outliers, decreasing = TRUE))
```

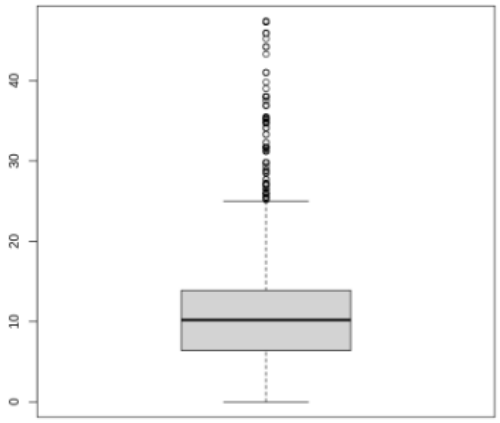
```
42800 · 40469 · 37000 · 21715 · 21600 · 20200
```

```
df3 <- subset(df3, df3$Landsize < 30000)
```

Distance

```
skewness(df3$Distance)
kurtosis(df3$Distance)
boxplot(df3$Distance)
```

```
1.53960650422239
6.58491902694624
```



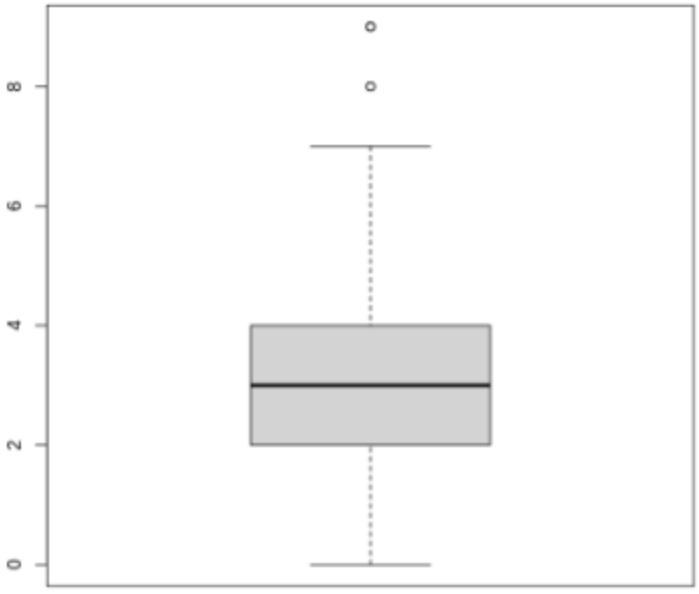

```
# identify the  outliers
outliers <- boxplot.stats(df3$Distance)$out
head(sort(outliers, decreasing = TRUE), n=10)

# didn't remove since there were no obvious outliers as there are quite a few data points for on the higher end
```

```
47.4 · 47.3 · 47.3 · 47.3 · 47.3 · 47.3 · 47.3 · 47.3 · 45.9 · 45.9
```

Bedroom2

```
skewness(df3$Bedroom2)
kurtosis(df3$Bedroom2)
boxplot(df3$Bedroom2)
```

```
0.243466576001683
3.67303956441622
```




```
[ ] # identify the  outliers
    outliers <- boxplot.stats(df3$Bedroom2)$out
    head(sort(outliers, decreasing = TRUE))
```

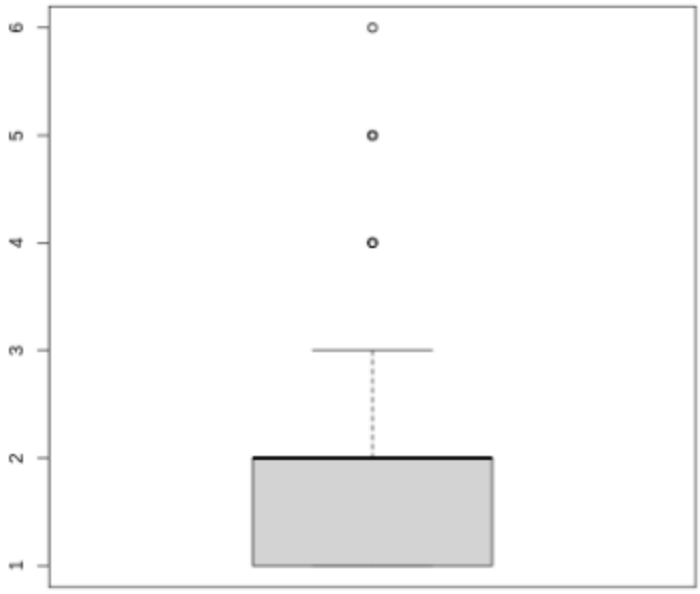
```
9 · 9 · 9 · 8 · 8
```

```
[ ] df3 <- subset(df3, df3$Bedroom2 < 8)
```

Bathroom

```
skewness(df3$Bathroom)
kurtosis(df3$Bathroom)
boxplot(df3$Bathroom)
```

```
1.04363490403561
4.51371753052369
```




```
# identify the  outliers
outliers <- boxplot.stats(df3$Bathroom)$out
head(sort(outliers, decreasing = TRUE), n=10)
```

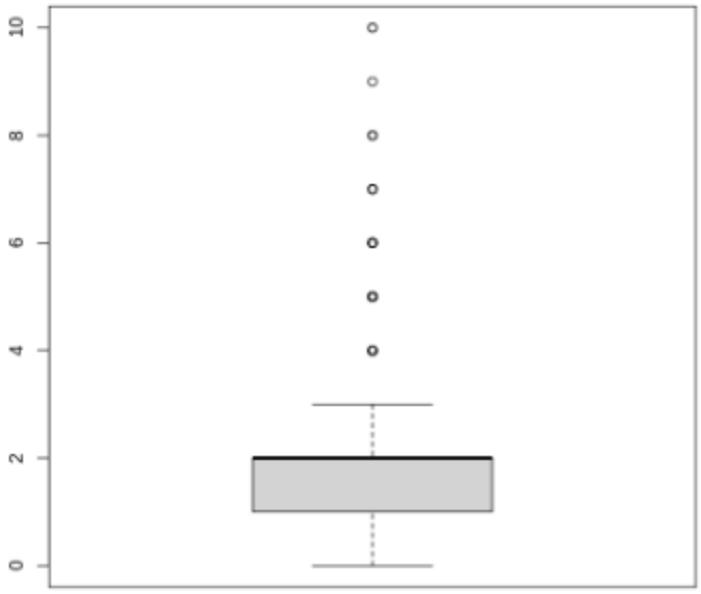
```
6 · 6 · 5 · 5 · 5 · 5 · 5 · 5 · 5 · 5
```

```
df3 <- subset(df3, df3$Bathroom < 6)
```

Car

```
skewness(df3$Car)
kurtosis(df3$Car)
boxplot(df3$Car)
```

```
1.3946450624063
8.17697248988732
```




```
# identify the  outliers
outliers <- boxplot.stats(df3$Car)$out
head(sort(outliers, decreasing = TRUE), n=20)
```

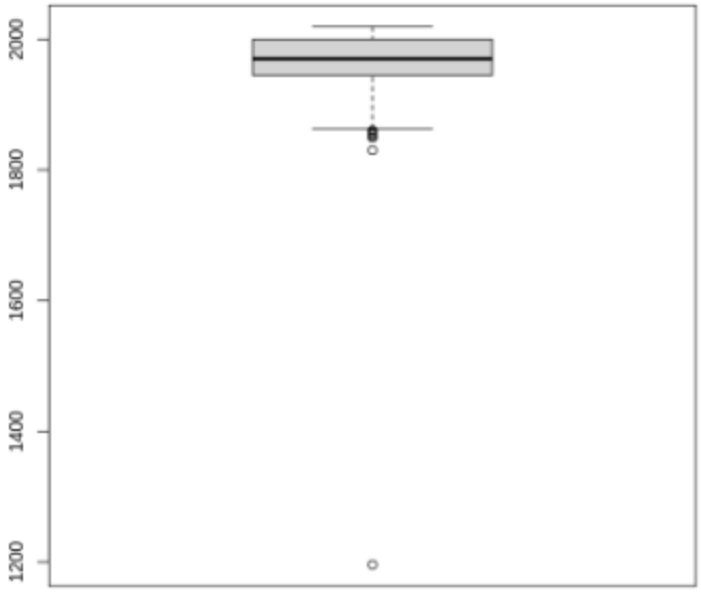
```
10 · 10 · 9 · 8 · 8 · 8 · 8 · 7 · 7 · 7 · 7 · 7 · 7 · 7 · 6 · 6 · 6 · 6 · 6 · 6
```

```
df3 <- subset(df3, df3$Car < 6)
```

Year Built

```
skewness(df3$YearBuilt)
kurtosis(df3$YearBuilt)
boxplot(df3$YearBuilt)
```

```
-1.49752769378888
23.3034075240361
```



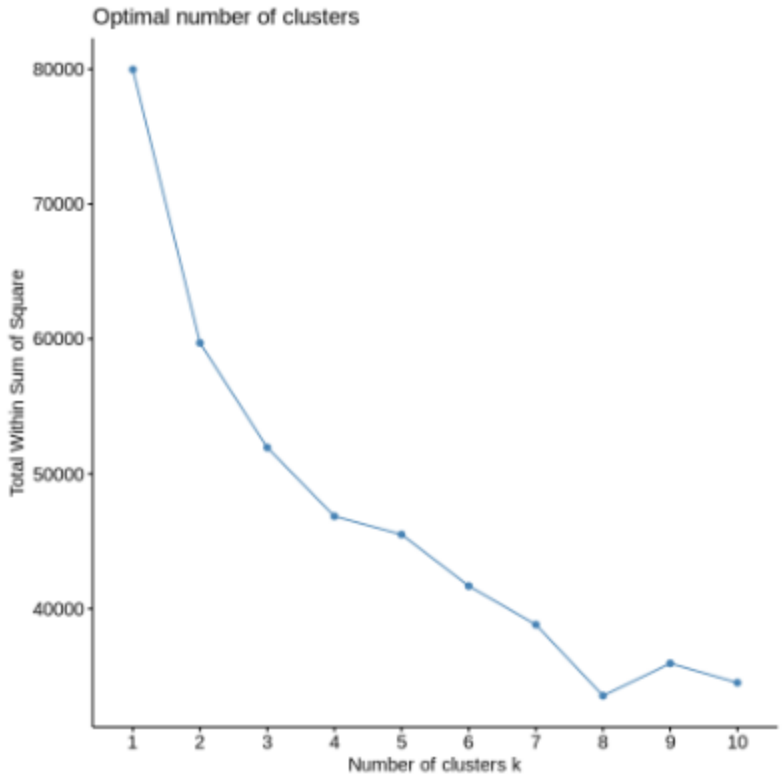

```
[ ]  # identify the  outliers
     outliers <- boxplot.stats(df3$YearBuilt)$out
     head(sort(outliers, decreasing = FALSE))
```

```
1196 · 1830 · 1850 · 1850 · 1850 · 1854
```

```
[ ]  df3 <- subset(df3, df3$YearBuilt > 1800)
```

Exhibit B - Optimal Number of Clusters Plots

```
#Get optimal number of clusters

fviz_nbclust(xsc, kmeans, method = "wss")
```



```
#calculate gap statistic based on number of clusters
gap_stat <- clusGap(df1,
                    FUN = kmeans,
                    nstart = 25,
                    K.max = 8,
                    B = 50)

#plot number of clusters vs. gap statistic
fviz_gap_stat(gap_stat)
```

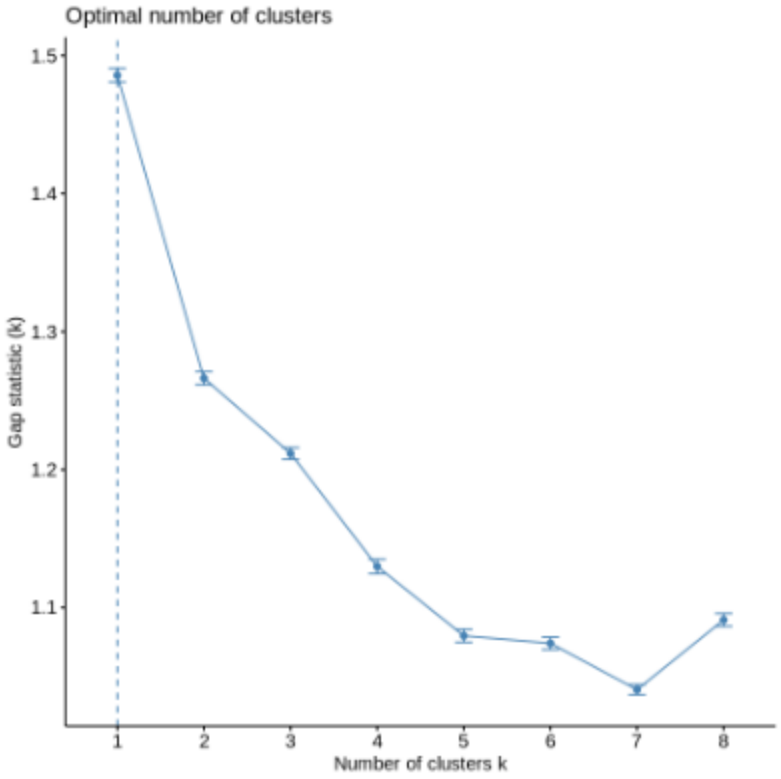


Exhibit C - t-SNE

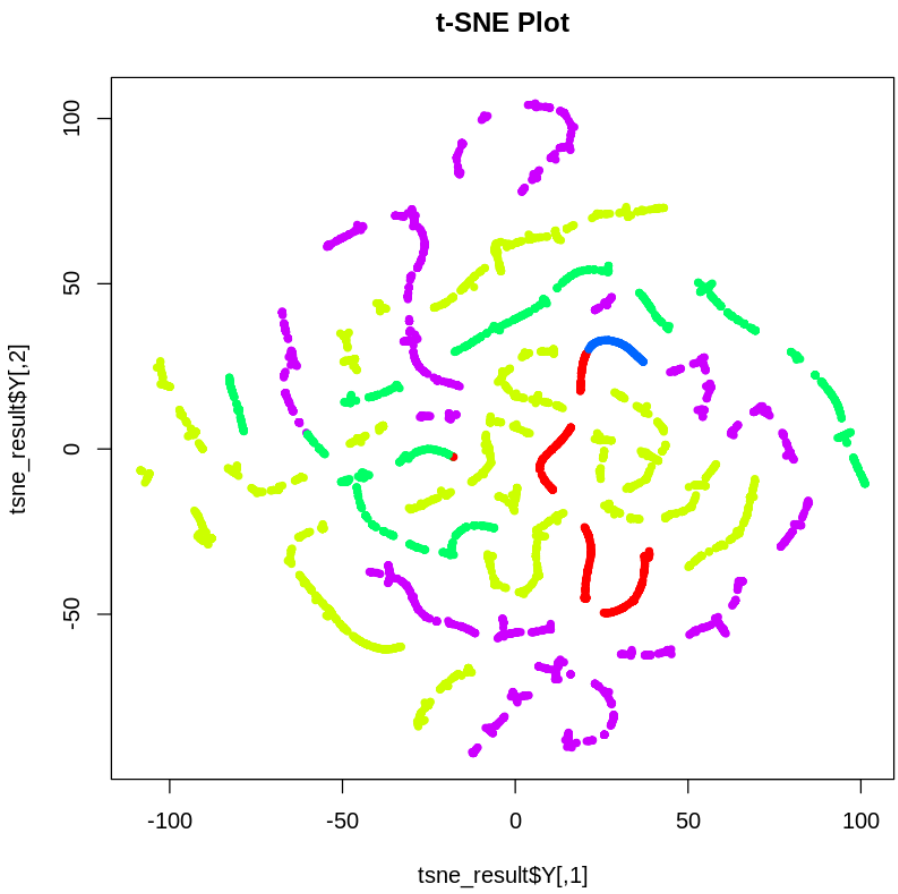


Exhibit D - K-means clustering

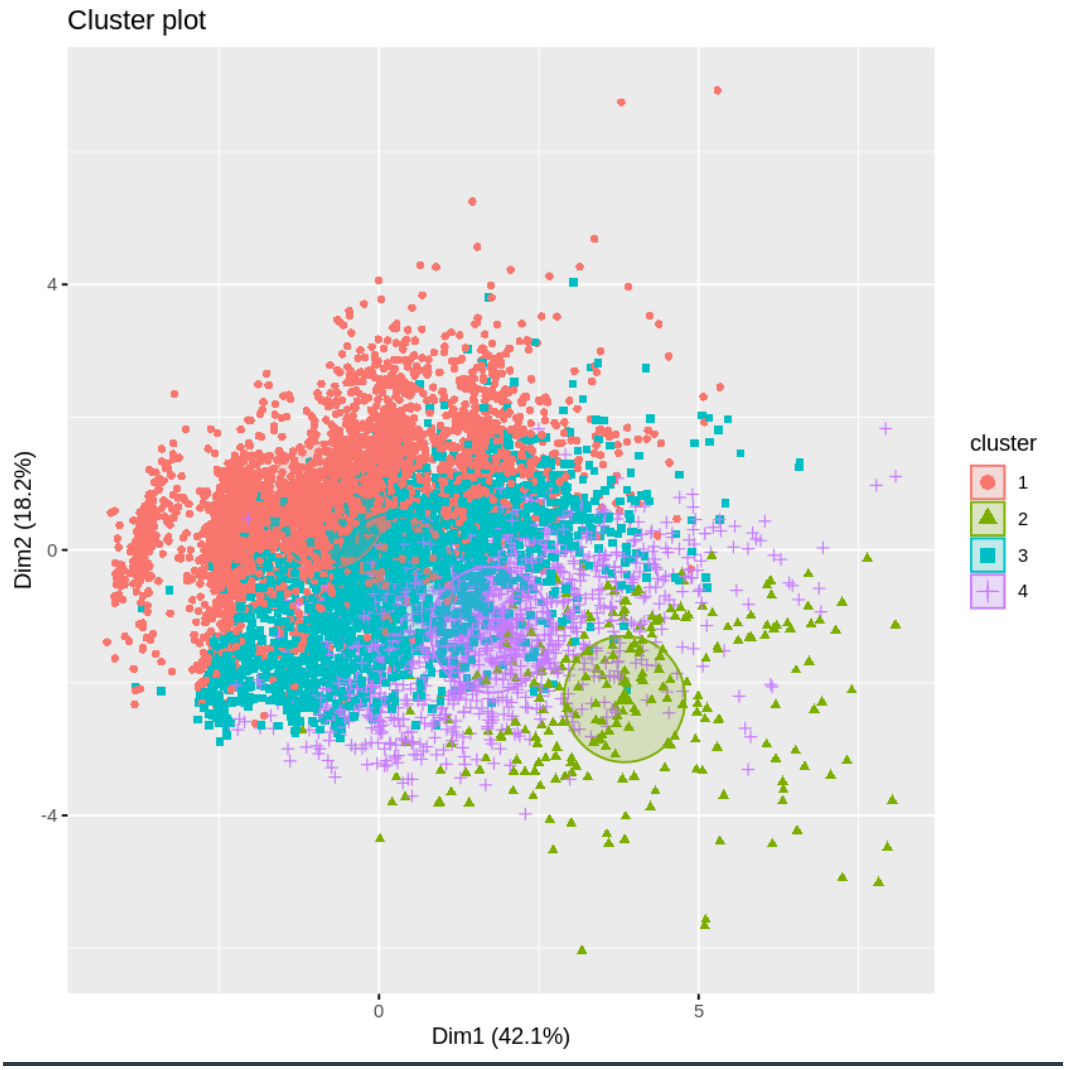


Exhibit E - Cluster Mean Values

cluster	mean_rooms	mean_price	mean_distance	mean_bedroom	mean_bathroom	mean_car	mean_landsize	mean_buildingarea	mean_yearbuilt	mean_price_area
1	2.734623	646072.5	12.806846	2.721365	1.415997	1.545316	475.4514	118.3563	1978.216	8225.774
2	4.271255	3488863.9	7.600405	4.198381	2.757085	2.307692	745.7571	288.7682	1950.421	28831.968
3	3.266617	1209534.5	9.805601	3.238237	1.691561	1.667662	498.8439	159.0305	1955.246	10021.535
4	3.777147	1985146.1	8.598136	3.742301	2.129660	1.901134	579.0292	213.4155	1945.254	11832.681