Identifying Distinct Property Buyers using K-means and Hierarchical Clustering Analysis

Patricia Sarmiento

Abstract

Real estate development companies depend on reliable market analyses to build properties that appeal to specific market segments. Property characteristics are critical to understanding the preferences of different target groups. Conducting a comparative study using K-Means Clustering and Hierarchical Clustering Analysis (HCA) methods on the online property sales data from Melbourne, Australia, helped identify four distinct market segments and their corresponding property preferences: high-price near the city, high-price far from the city, low price near the city, and low-price far from the city.

Keywords: K-Means clustering, Hierarchical clustering analysis, Dendrogram

1 Introduction

Online real-estate directories such as Australia's *Domain.com.au* provide extensive information about property market trends through a continual stream of property sales data. The correct analysis can transform these market trends into actionable insights that can help key industry players such as property development firms and bank loan departments understand the property preferences of different buyers. As such, the goal of this study was to identify the major property buyer groups in the Melbourne region of Australia and their property preferences.

2 Literature Review

Given the significance of the real estate market, there is an abundance of literature describing models that predict home value trends. Liu et al. (2016) used a Hierarchical Spatial Functional Model (HFSM) to identify real-estate price fluctuations with an added layer for understanding land areas of high desirability in general and within subgroups in the City of Edmonton, Canada. The resulting mapped clusters highlighted that geo-partitions were meaningful in mapping and learning local structures that produced subregions of different prices based on parks, rivers, and other physical characteristics usually missed in common property descriptions.

The idea that latent features affect property prices beyond what is obvious was applied to this study by recognizing that geography could have a more significant impact on the Melbourne property data than assumed.

3 Methods

The Melbourne property data was created by scraping the online real estate directory *Domain.com.au.* The data described the features of properties sold.

3.1 Data Processing

Out of the original 21 property features, the following nine features combined the ideal minimum property descriptions needed for the clustering analysis: Price (Australian dollars), Rooms, Bedoom2, Bathroom, BuildingArea (meters), Landsize (meters), Car, YearBuilt, and Distance (in kilometers from the city center). Further cleaning of the data excluded NAs.

Running boxplots for each feature showed distributions that contained many outliers that could degrade the clustering performance and provide a misleading analysis given the clustering algorithms' dependence on the mean (ur Rehman and Belhaouari 2021). Therefore, most outliers were manually removed, except for those in close range to the maximum value of the boxplot.

A correlation analysis revealed that the features for the number of rooms (Room) and bedrooms (Bedroom2) were 96% correlated. On closer inspection, these two features mostly had the same values. Since the number of rooms is a better property descriptor, the feature Bedroom2 was removed to decrease redundancy.

Scaling the data was an essential next step due to the significant difference in value ranges such as Price in millions and number of rooms of up to 7, which could compromise the K-means algorithm results. Furthermore, measuring the distance of the scaled data was necessary for computing clusters using the Hierarchical Clustering Analysis method. The method used was the Euclidean distance.

3.2 Exploratory Data Analysis

Visualizing scatterplots revealed some preliminary associations between property features. Firstly, the scatter plot between BuildingArea and Price, with clusters of colored dots representing YearBuilt, revealed that newer properties tended to be bigger and cheaper (see Fig. 1). Secondly, the scatter plot between Distance and Price, with clusters of colored dots representing YearBuilt, revealed that properties far from the city center tended to be newer and cheaper (see Fig. 2). Finally, the scatter plot between YearBuilt and Price, with clusters of

colored dots representing the Rooms, revealed that newer properties had more rooms at a lower cost (see Fig. 3).

Comparing small groups of features revealed some potential property buyer clusters based on characteristics such as distance from the city center and price sensitivity. However, a clustering algorithm would have to be used to confirm clear clusters based on all the features.

3.3 Clustering

Clustering algorithms work by organizing data into homogenous subgroups. The two main types of clustering are agglomerative, such as the hierarchical clustering algorithm, and partitioning, such as the K-means clustering algorithm.

3.3.1 K-means

When using the K-means clustering algorithm, a set number of k clusters are chosen, after which all data points are assigned to their closest cluster centroid. Then, the centroids are recalculated as the average of all data points in a cluster, and all the data points are re-assigned to their closest recalculated cluster centroid (Kabacoff 2015, Ch 16). This process repeats until no more data points are re-assigned, or the algorithm reaches its maximum number of iterations.

This study used the elbow method, which computes the total within-cluster sum of squares, to determine the best number of ks Plotting this method showed that the optimal k clusters were 3 to 5 (see Fig. 4).

Running the K-means algorithm based on 3, 4, and 5 clusters and plotting the results helped discard the k=5 clusters since two of its five clusters intersected heavily (see Fig. 5).

The final number of clusters used was determined by grouping the k-means cluster results for k=3 and k=4 by the mean of each property feature. Examining the resulting tables revealed that k=4 displayed more specific property preferences within each cluster (see images A6, A7). A boxplot of all features for clusters k=3 and k=4 reflected the abovementioned findings and exposed intergroup dynamics that helped further identify each group's property preferences. (see Fig. 8, 9).

A hierarchical clustering analysis helped to further validate the k=4 clusters chosen above.

3.3.2 Hierarchical clustering analysis

When using the hierarchical clustering algorithm, all observations begin as individual clusters and are combined two clusters at a time until all cluster distances match the clustering method of choice. This study used the complete method, which computes the longest distance between a point in one cluster and another (Kabacoff 2015, Ch 16).

Hierarchical clusters can be plotted using dendrograms, diagrams representing a tree that illustrate the cluster arrangements created during the analysis. This study cut the dendrogram into k=3 and k=4 groups (see Fig. 9, 10). However, the high number of values in the data makes interpreting the dendrogram difficult beyond inferring if the number of clusters chosen seems reasonable.

A different approach was to group the hierarchical cluster results for k=3 and k=4 by the mean of each property feature (see Fig. 12, 13). The resulting tables resembled the results found using the K-means algorithm.

4 Results

Since comparable results were found between the K-means and the Hierarchical Clustering Analysis methods, this section focuses on the K-means results. The tables from Fig. 6 and Fig. 7 show the mean value of each property feature per cluster for k=3 and k=4, respectively.

Fig 6. shows a clear distinction between 3 clusters of low-priced, mid-priced, and high-priced properties. However, a few inconsistencies stand out; cluster 1, the low-priced properties, has the highest mean distance value, even though Fig. 2 shows low-priced properties at low distances from the city center. Furthermore, the low-priced properties in cluster 1 show higher values than the mid-priced properties of cluster 3 for all features except Price. This could be attributed to the Distance factor since properties far from the city are generally larger and cheaper. Therefore, we seem to be missing low-priced properties close to the city.

The table in Fig. 7 shows more complexity in describing the groups of property buyers. It shows a clear distinction between low-priced properties far from the city (cluster 1), low-priced properties near the city (cluster 4), high-priced properties far from the city (cluster 3), and high-priced properties near the city. Furthermore, the boxplots in Fig. 9 help visualize each group's

preferences. In particular, the wiggle room available for each group given the presence of a few outliers.

5 Conclusions

The K-means and Hierarchical clustering methods helped uncover the preferences of four distinct property buyers based on price and distance from the city center. The four market segments identified, and their property preferences are the following:

Market 1: Low-priced properties far from the city (red)

- Lowest budget
- Large land and property. Priority on land size.
- Large number of rooms.
- Two cars.
- New construction.

Market 2: Low-priced properties near the city (purple)

- Small land and property.
- Small number of rooms.
- One or two cars.
- New to mid-century construction.

Market 3: High-priced properties far from the city (blue)

- Highest budget
- Large land and property. Priority on property size.
- Large number of rooms.

- Two cars.
- New or old construction.

Market 4: High-priced properties near the city (green)

- Nearest to the city.
- Small land and property.
- Small number of rooms.
- Prefers old construction but is very flexible.

This information can potentially spearhead the construction of more specific property development opportunities and target the correct customers based on a clear understanding of the different buyer preferences regarding budget, distance from a city center, and property attributes. Future work could incorporate changes in sales value over time within each market group to further understand the buying trends.

Appendix

Fig. 1: Scatterplot of BuildingArea vs Price by YearBuilt

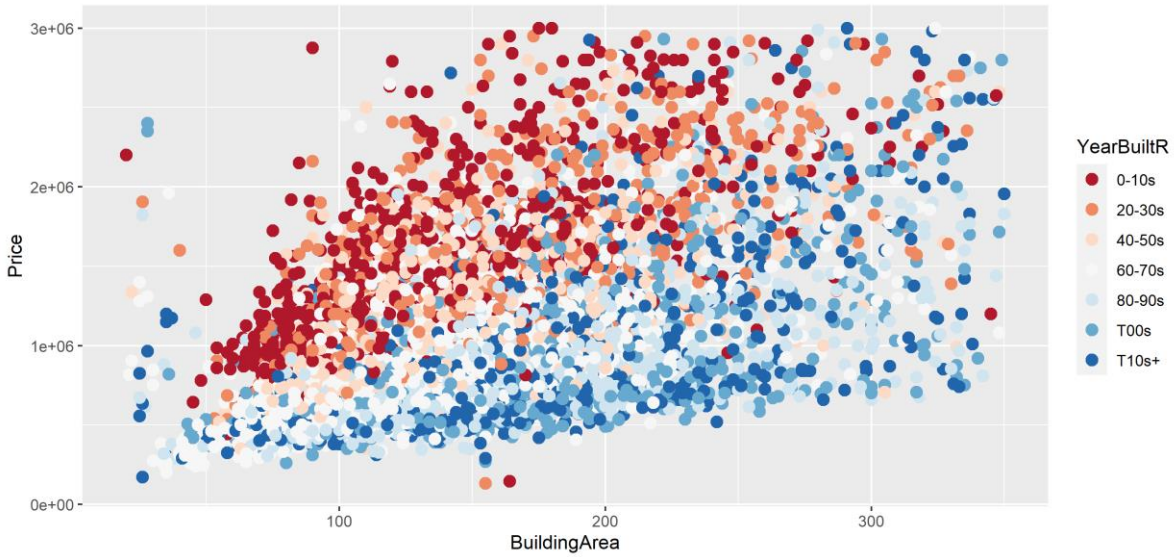

Fig. 2: Scatterplot of Distance vs Price by YearBuilt

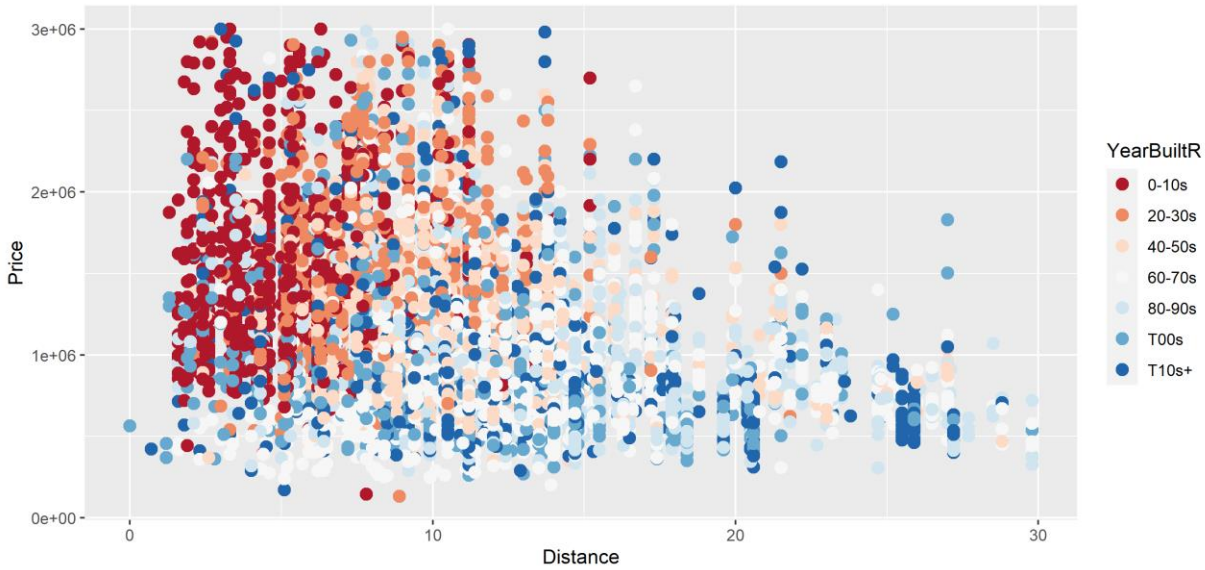

Fig. 3: Scatterplot of YearBuilt vs Price by Rooms

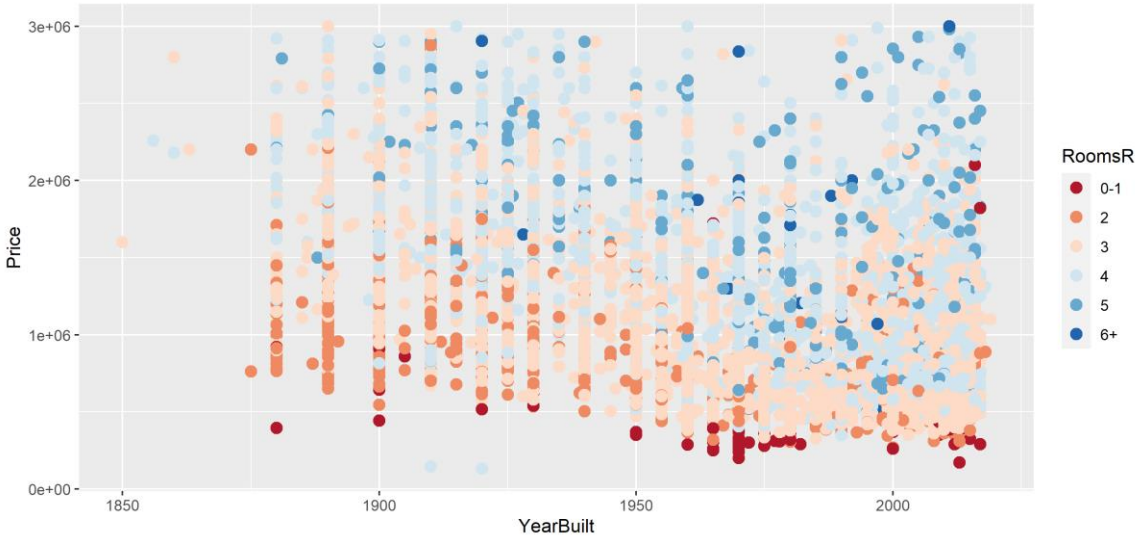

Fig. 4: Plot of elbow method

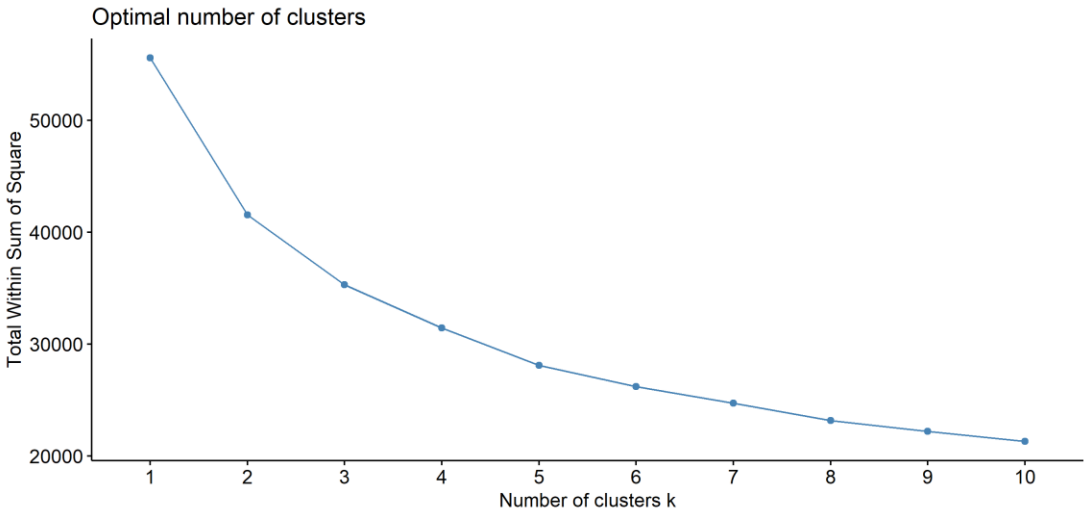

Fig. 5: K-means, Scatterplot of the clusters created for k=3, k=4 and k=5

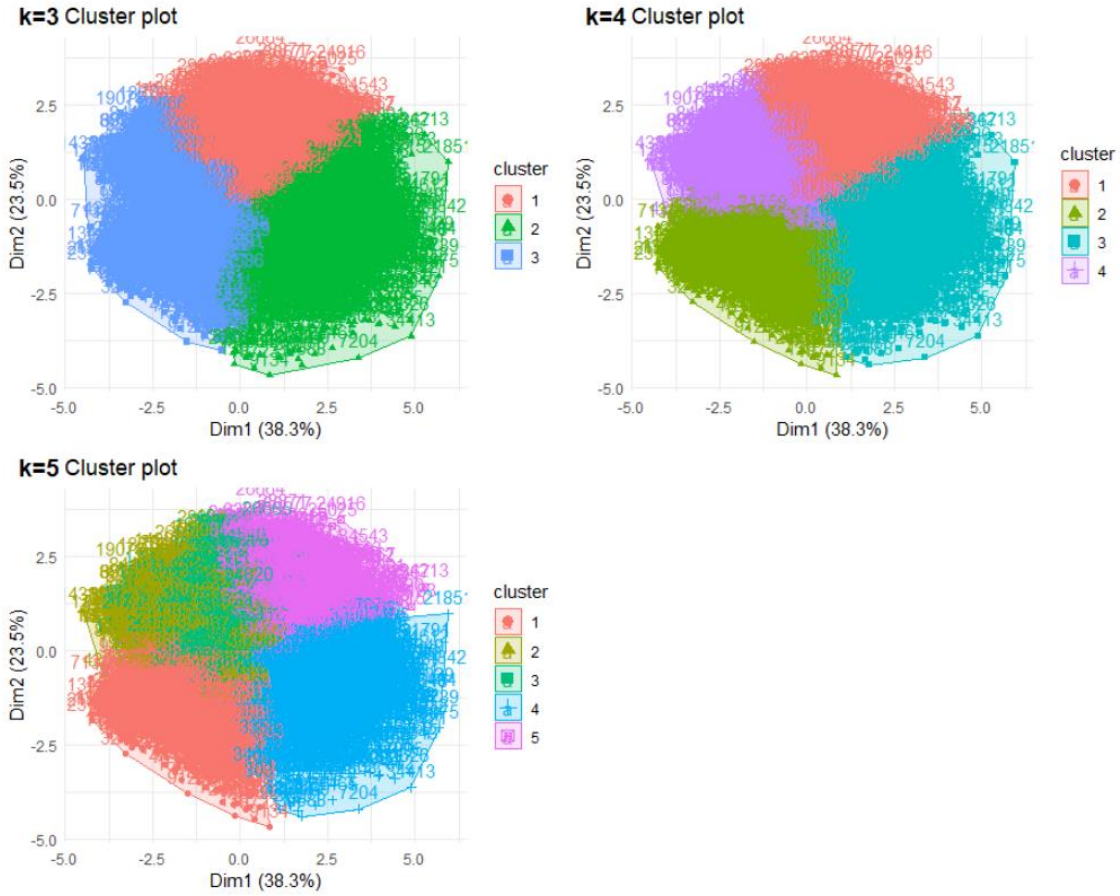

Fig. 6: K-means k=3, Mean value of each property feature per cluster

km3_cl	Price	Rooms	Bathroom	BuildingArea	Landsize	Car	YearBuilt	Distance
<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	819999.	3.30	1.67	148.	544.	1.98	1982.	15.5
2	1723988.	4.05	2.31	220.	600.	2.08	1958.	9.98
3	1003469.	2.57	1.22	110.	337.	1.13	1946.	7.53

Fig. 7: K-means k=4, Mean value of each property feature per cluster

km4_cl	Price	Rooms	Bathroom	BuildingArea	Landsize	Car	YearBuilt	Distance
<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	852784.	3.52	1.88	163.	573.	2.12	1985.	16.7
2	1341783.	2.84	1.33	124.	299.	0.962	1911.	5.57
3	1770711.	4.08	2.35	225.	618.	2.14	1959.	9.85
4	790832.	2.63	1.23	110.	418.	1.45	1975.	10.5

Fig. 8: K-means k=3, Boxplot of each k-means cluster by property feature

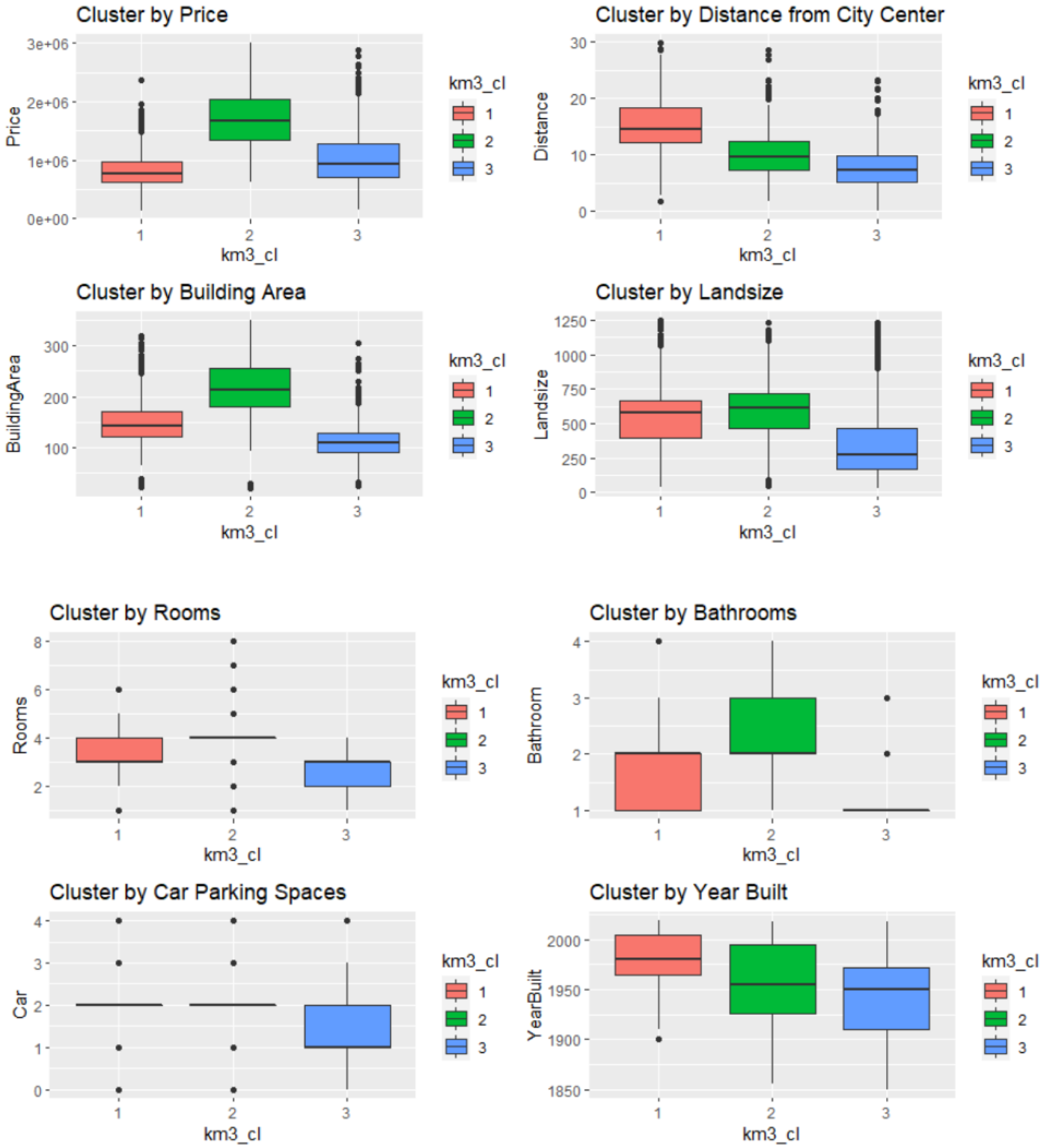

Fig. 9: K-means k=4, Boxplot of each k-means cluster by property feature

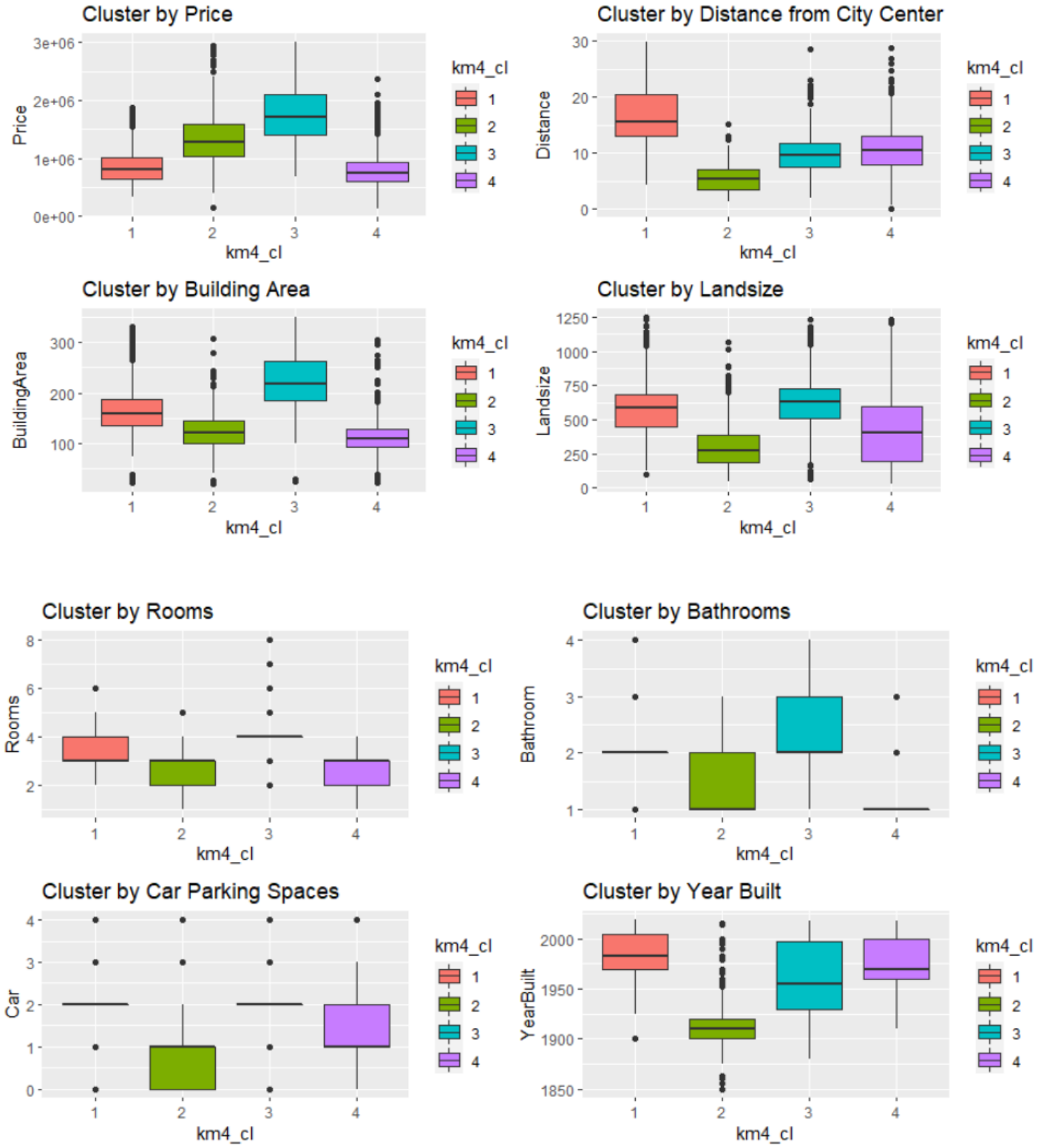

Fig. 10: HCA dendrogram, k=3

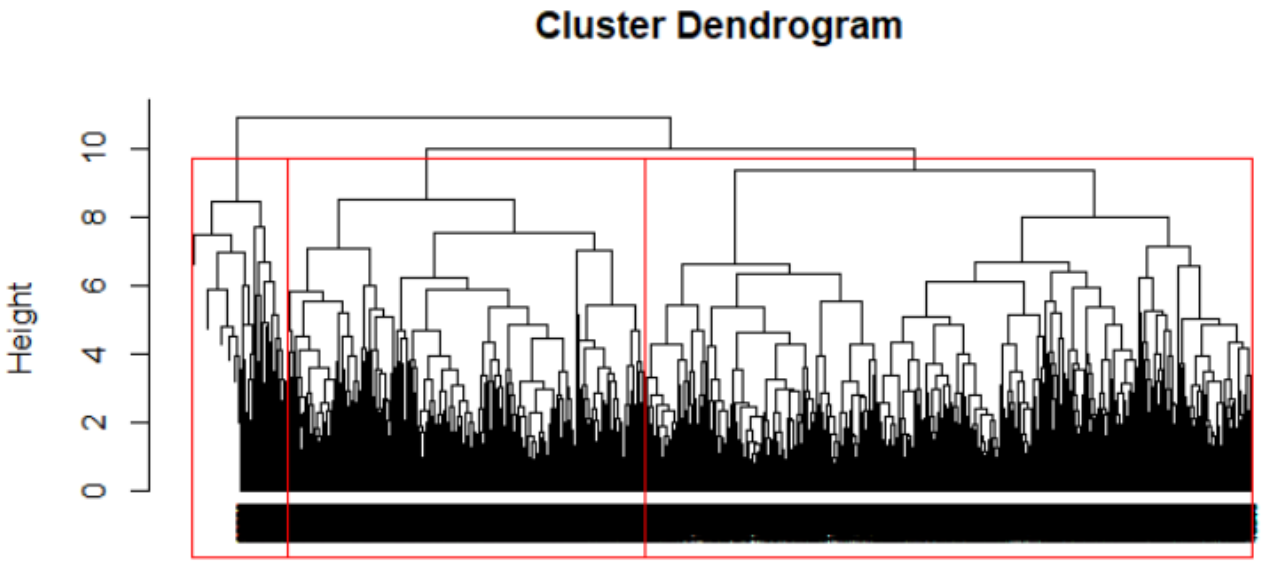

Fig. 11: HCA dendrogram, k=4

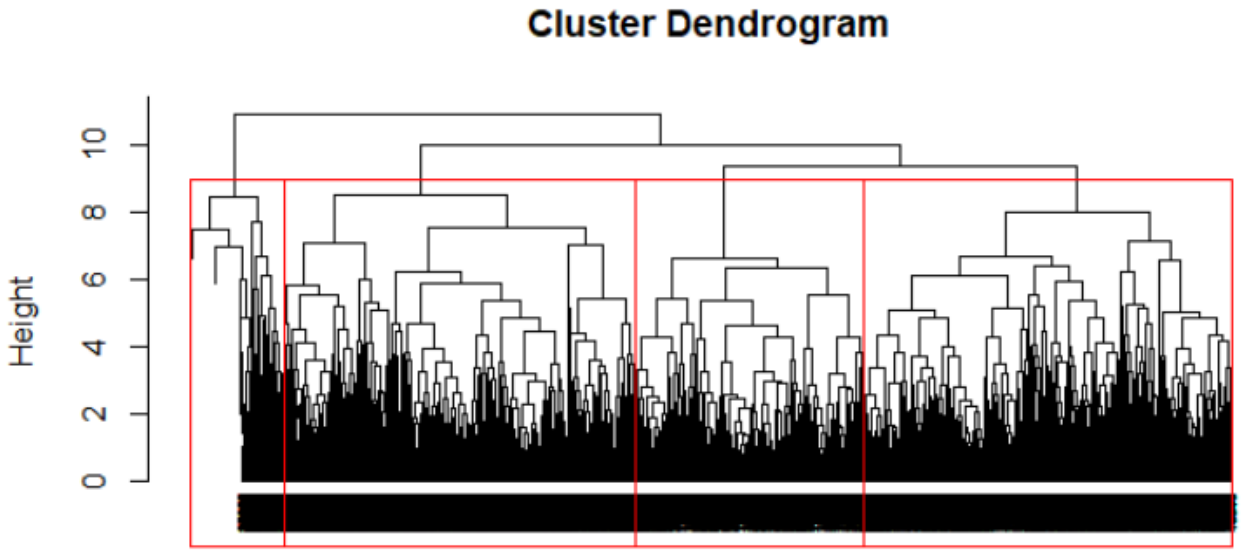

Fig. 12: HCA k=3, Mean value of each property feature per cluster

hc3_cl	n	price	rooms	bathr	buildA	landS	car	yearB	dist
<fct>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	3973	1092130.	2.87	1.48	130.	377.	1.29	1956.	8.62
2	2349	906351.	3.45	1.66	157.	602.	2.18	1973.	15.7
3	626	1870162.	4.31	2.65	252.	652.	2.26	1973.	10.8

Fig. 13: HCA k=4, Mean value of each property feature per cluster

hc4_cl	n	price	rooms	bathr	buildA	landS	car	yearB	dist
<fct>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1520	800893.	2.27	1.02	96.1	356.	0.970	1956.	8.76
2	2453	1272594.	3.25	1.76	151.	391.	1.49	1955.	8.53
3	2349	906351.	3.45	1.66	157.	602.	2.18	1973.	15.7
4	626	1870162.	4.31	2.65	252.	652.	2.26	1973.	10.8

References

Liu, Bang, Borislav Mavrin, Di Niu, and Linglong Kong. “House Price Modeling over Heterogeneous Regions with Hierarchical Spatial Functional Analysis.” *2016 IEEE 16th International Conference on Data Mining (ICDM)*, March 1, 2016. https://doi.org/10.1109/icdm.2016.0134.

Kabacoff, Robert I. “R In Action, Second Edition. 2015.” O'Reilly Online Learning. Manning Publications, May 2015. https://learning.oreilly.com/library/view/r-in-action/9781617291388/kindle_split_028.html.

ur Rehman, Atiq, and Samir Brahim Belhaouari. 2021. “Unsupervised Outlier Detection in Multidimensional Data.” *Journal of Big Data* 8, no. 1. https://doi.org/10.1186/s40537-021-00469-z.