Melbourne Housing Market Segmentation Analysis

Vincent Pun

Abstract

Real estate firms have found a growing need to remain agile in understanding market trends as warnings of a property bubble in the Melbourne housing market continue to increase. Understanding how homes are priced in relation to properties with similar characteristics is crucial when managing large investment portfolios, and this can be accomplished through the segmentation of the housing market based on key features such as building size, property age, price, and proximity from the Central Business District (CBD).

In this paper, clustering algorithms, machine learning techniques used to group similar data points, are applied on housing data that is sourced from Domain.com.au. This is used to segment 7,735 unique properties into homogeneous subgroups.

Keywords Melbourne housing market, segmentation, clustering, factor analysis, principal component analysis, K-means algorithm, hierarchical clustering, t-distributed stochastic neighbor embedding, interactive map

Introduction

Real estate investors would like to know when to buy and sell properties depending on how they are valued against comparable properties. Since many types of homes can be found even within a single suburb, multiple factors need to be considered to properly divide the housing market. Moreover, profiling properties in Melbourne is an iterative

task that investors need to perform to both follow market trends and effectively assign territories to realtors.

Literature review

Clustering techniques have previously been used to identify relationships among United States housing markets; K-means is applied to the 1977-1992 returns to housing data to identify price fluctuations across cities and regions (Abraham, Goetzmann, and Wachter 1994).

Methods

The original Melbourne housing dataset contains 34,857 records and 21 variables that contain property information such as price, location, age, type, and configuration. A subset of 8,887 records is extracted from this source to work with complete data, and the following nine continuous variables are used in this study: Rooms, Price, Distance, Bedroom2, Bathroom, Car, Landsize, BuildingArea, and YearBuilt.

Outlier removal is performed to improve the performance of the statistical techniques and machine learning algorithms that are used in this paper. Skewness and kurtosis is found for all nine variables, and we see that many of these attributes are highly asymmetric.

Figure 1: BuildingArea - Distribution, Quantile-Quantile Plot, Barplot

Histograms, quantile-quantile plots, and boxplots are generated for each variable, and this is used to both find and remove records that contain extreme outliers such as property sizes that are equal to zero and homes that were built before 1899. However, we retain a generous amount of realistic outliers in this study, for it appears that there

may be a large homogenous subset of properties with large building areas and high prices that may represent high-end luxury homes. After removing outliers, a common scale from 0 to 1 is used to normalize the data to prevent certain variables such as Price from having too much impact on the model results; we are left with 7,735 observations.

The optimum amount of factors or components to include in clustering algorithms is found by analyzing the scree plot of eigenvalues. They appear to level off at four components for both principal components and factor analysis.

Figure 2: Scree plot with parallel analysis

Also, performing principal component analysis on the normalized data shows that the cumulative proportion of variance explained by these components is equal to 91.97%, which confirms that this is an appropriate cutoff to reduce data complexity.

Figure 3: Heatmap - Factor loadings for housing attributes

Figure 4: Path Diagram - Latent variables analysis

Factor analysis loadings are then visualized in both a heatmap and path diagram. The first latent variable demonstrates a strong correlation between the following variables: Rooms, Bedroom2, Bathroom, BuildingArea, and Price; the trait that these features may be describing is property value, and we assess this trait by comparing price per building size in meters for each of the clusters later in the paper. The second latent variable appears to solely focus on property age, which consists of YearBuilt.

The two clustering algorithms that are explored in this study are K-means and hierarchical clustering.

Figure 5: K-means Clusters

Four clusters are created with both models, and the goodness of fit is compared using both the coefficient of determination and the average silhouette width. In result, the K-means solution offers better performance on both tests. The large amount of data also makes hierarchical clustering dendrogram outputs more difficult to interpret, so we recommend the use of K-means for understanding housing types moving forward. Cluster results can also be visualized using a t-SNE model, which affirms that there are homogenous subsets in the solution.

Table 1: Goodness of Fit

Figure 6: t-SNE

Results

We learned that the four homogenous property subsets formed by the K-means clustering algorithm are most differentiated by their overall value (building price per square meter) and year built. When organizing properties by suburb, we find that each cluster has high concentrations in different general areas, but the presence of outliers may be a reason why this variable was not influential for the first four loadings of the factor analysis. Information about property types are available in the original data, and this is also used to find dissimilarities between each cluster. Descriptions based on median summary statistics of each group are found below are used instead of mean due to the presence of outliers in the dataset.

Table 2, Table 3, Table, 4, Figure 7, and Figure 8

- **Group 1** - Group 1 has the lowest median price, but it ranks in the middle for price per square meter. It has the highest density of properties in CBD, which may be a reason why its buildings are almost the smallest out of all groups; this group also has the largest proportion of unit/duplex properties and the smallest proportion of houses/cottages/villas/semis/terraces.

- **Group 2** - Group 2 has the highest price, but it ranks in the middle for price per square meter. It appears to be the furthest from CBD, which is heavily influenced by dense clusters of properties in satellite cities such as Melton, Sunbury, Wallan, and Seaford. This group contains the largest homes (also largest proportion of townhouses), and they are also the newest out of the four.
- **Group 3** - While Group 3's homes are not the most expensive, they are the most expensive per square meter. and it has the smallest proportion of townhouses relative to other clusters. While this group has the oldest properties, they are likely located in prime locations and have experienced high appreciation in value. This group is also the least dispersed in geographical location; as seen on the interactive map, almost all of these properties are located right outside of CBD, which makes it an ideal location to live for professionals who want to live outside of CBD while maintaining a short commute.
- **Group 4** - Group 4 is ranked in the middle for price, but it is the least expensive when it comes to price per square meter. While this group is the only other group that has a concentration of properties in the epicenter of CBD, most of these observations are located furthest away from CBD (aside from the outliers in Group 2). Group 4 has the largest proportion of houses/cottages/villas/semis/terraces, which aligns with the observation that it also has the largest building areas.

Conclusions

Strategic planning is crucial for real estate investors to make fast decisions when appraising properties amidst the housing bubble. Differentiating properties based on overall value, while effective, is more interpretable when compared on an actual map. Through robust clustering techniques and interpretable visualizations, we are able to formulate credible insights when evaluating properties in addition to optimizing assignments of territories to realtors.

Appendix

Table 1 - Goodness of Fit:

Method	R-Squared	Average Silhouette Width
K-means	0.5363	0.37
Hierarchical Clustering	0.5069	0.21

Table 2 - Cluster Profile Summary Statistics (Median):

Cluster	Rooms	Price	Distance	Bedroom2	Bathroom	Car	Landsize	Buildingarea	YearBuilt	Pricesqm
1	3	772,000	41	3	1	1	557	114	1960	7,045.45
2	4	1,263,500	194.5	4	2	2	558	181	1994	7,207.10
3	3	1,060,000	185	3	1	1	367	113	1925	9,235,07
4	4	950,000	51	4	2	2	597	182	1990	5,391.30

Table 3: Properties Per Cluster By Suburb:

Please refer to "output_clustersuburbcount.xlsx" located in the Deliverables folder.

cluster	suburb	freq
1	Reservoir	137
1	Richmond	89
1	Glenroy	76
1	Bentleigh East	60
1	Fawkner	45

	cluster	suburb	freq
79	2	Balwyn North	76
80	2	Kew	50
81	2	Glen Iris	47
82	2	Pascoe Vale	47
83	2	Maribyrnong	45

	cluster	suburb	freq
289	3	Brunswick	101
290	3	Coburg	84
291	3	Preston	82
292	3	Northcote	72
293	3	Yarraville	70

	cluster	suburb	freq
92	4	Craigieburn	75
93	4	Bentleigh East	73
94	4	Brighton East	53
95	4	Keilor East	53
96	4	Doncaster	47

Table 4: Property Types per Cluster

df_clu	df_h	df_t	df_u	cluster_total	prop_h	prop_t	prop_u
1	1599	131	319	2049	0.78	0.06	0.16
2	1132	212	52	1396	0.81	0.15	0.04
3	1675	64	262	2001	0.84	0.03	0.13
4	2022	221	46	2289	0.88	0.1	0.02

Figure 1 - BuildingArea - Distribution, Quantile-Quantile Plot, Barplot:

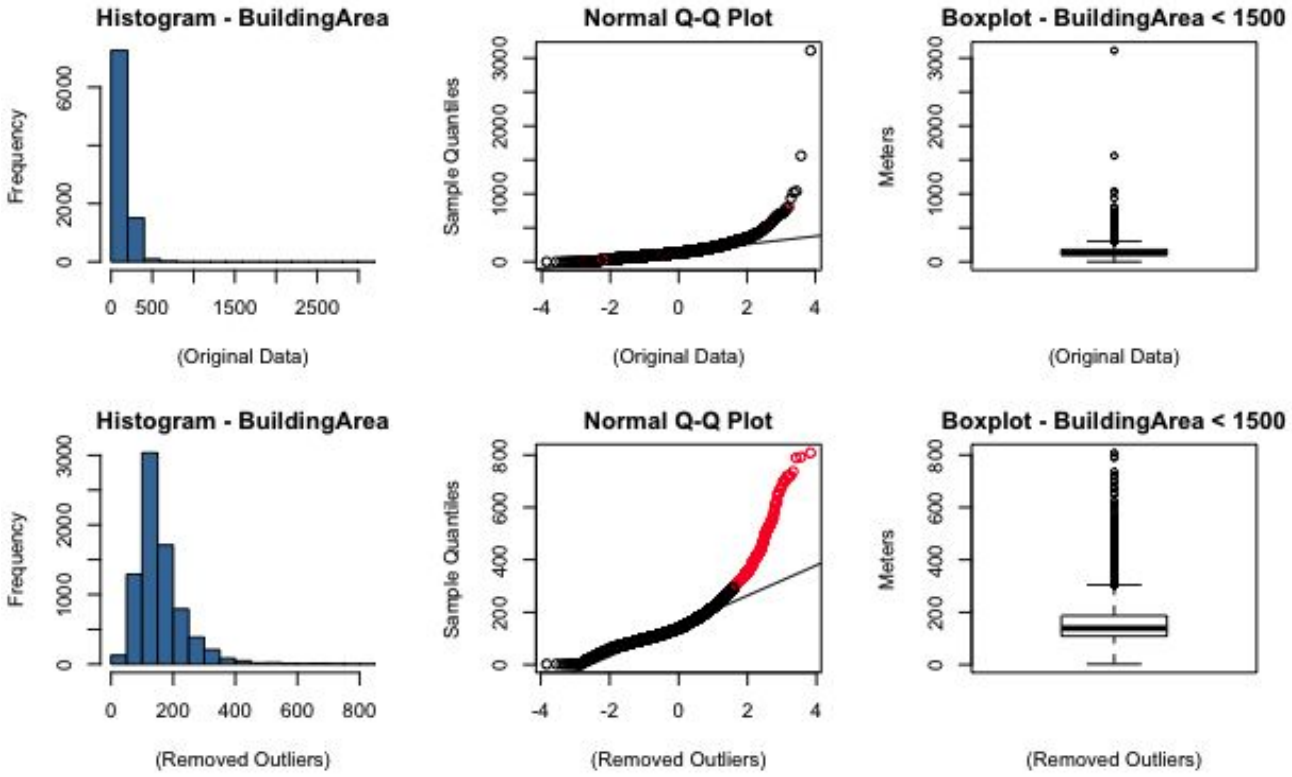


Figure 2: Scree plot with parallel analysis:

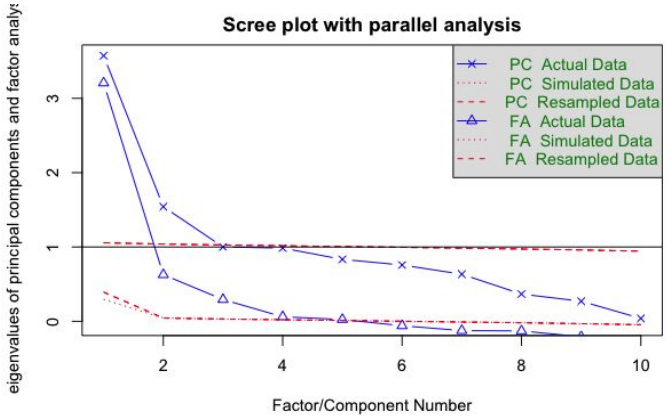


Figure 3: Heatmap - Factor loadings for housing attributes

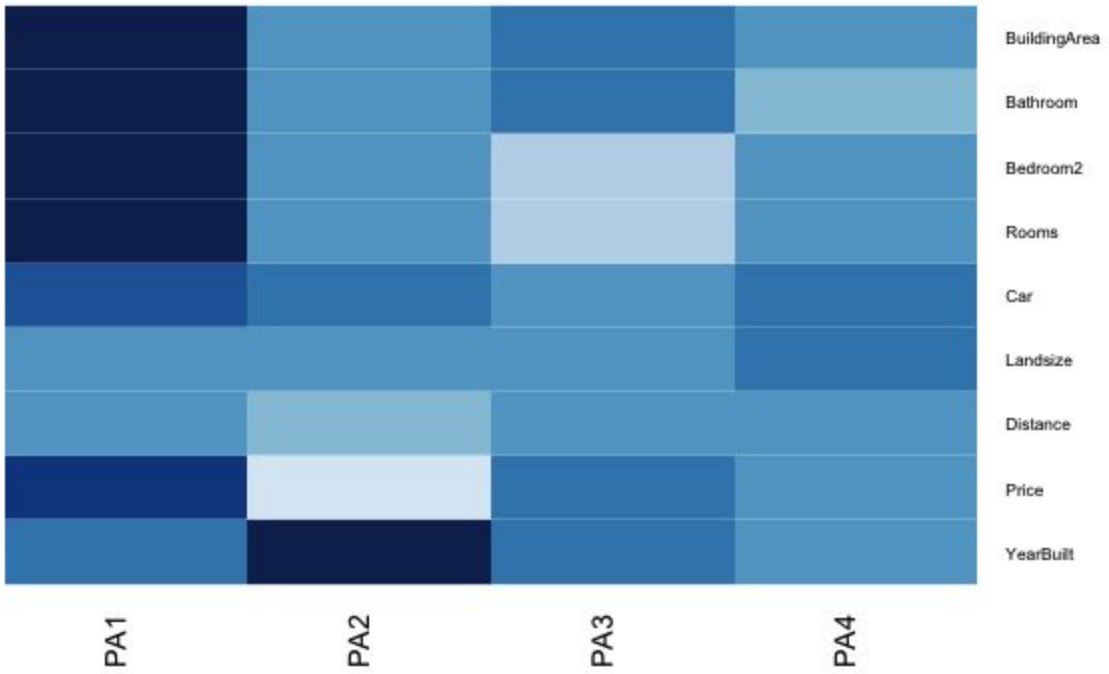


Figure 4: Path Diagram - Latent variables analysis

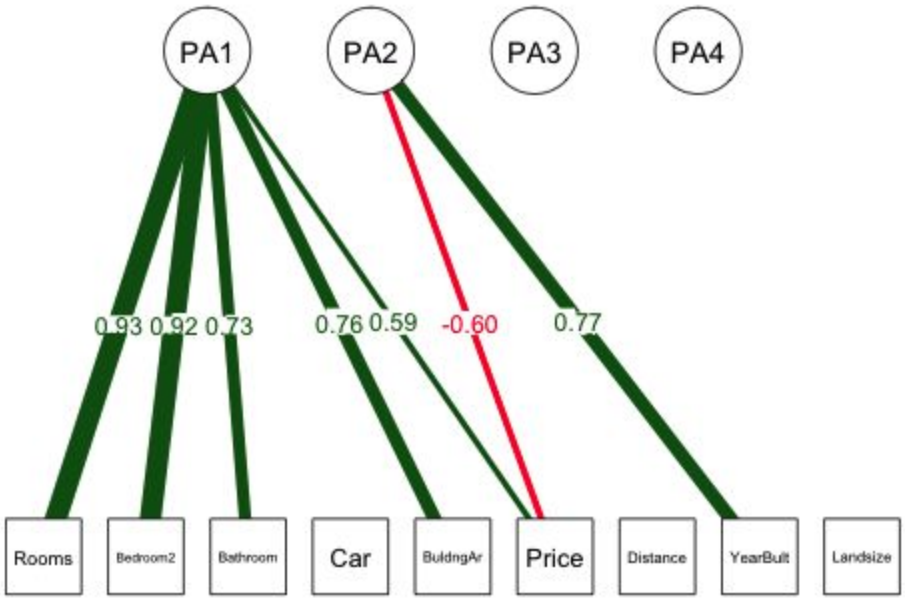


Figure 5: K-means Clusters

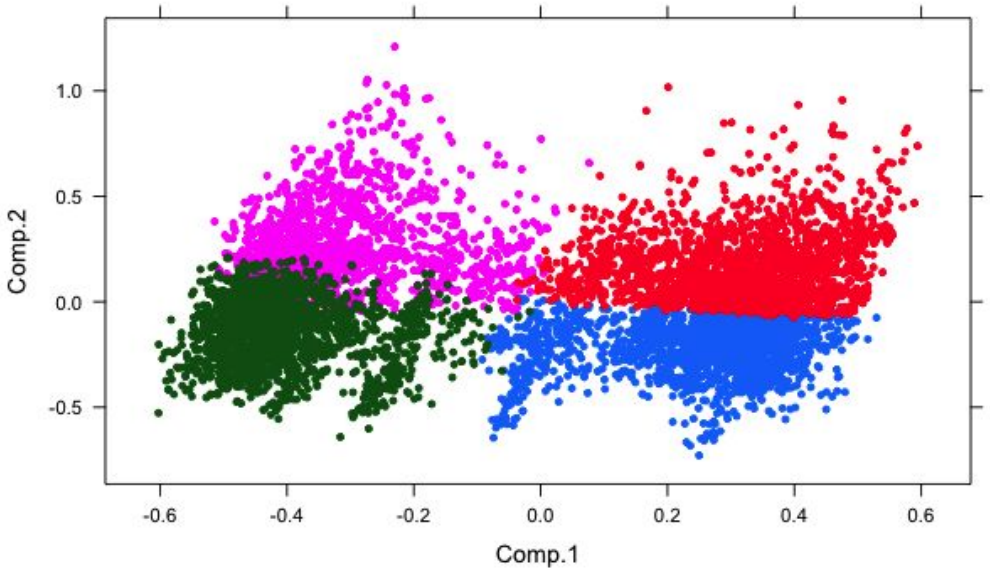


Figure 6: t-SNE:

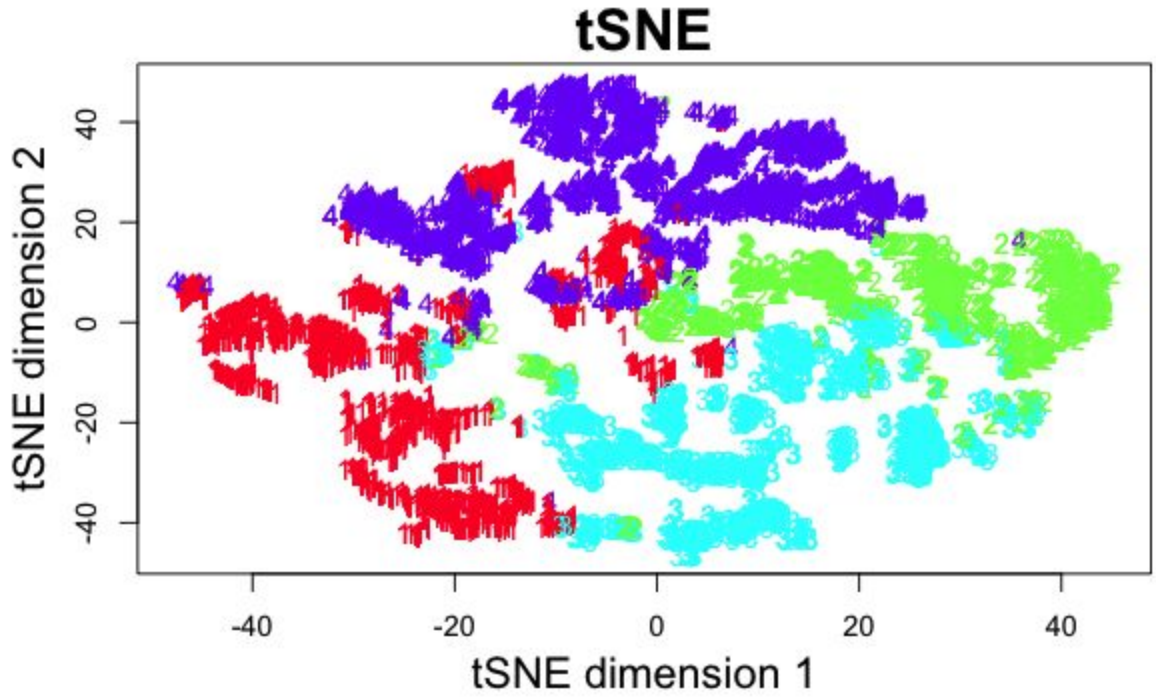


Figure 7: Barplot of Property Type and Cluster:

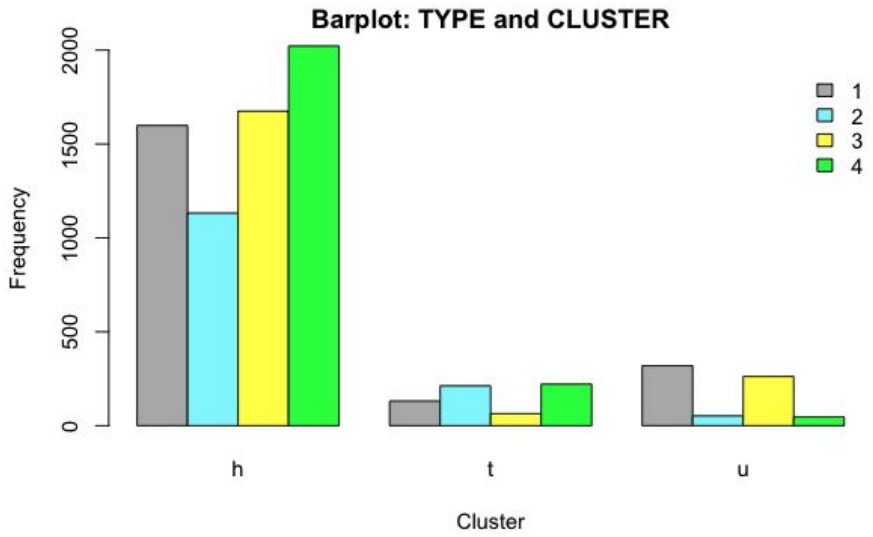


Figure 8: Interactive Map (Leaflet.js) of Property Clusters:

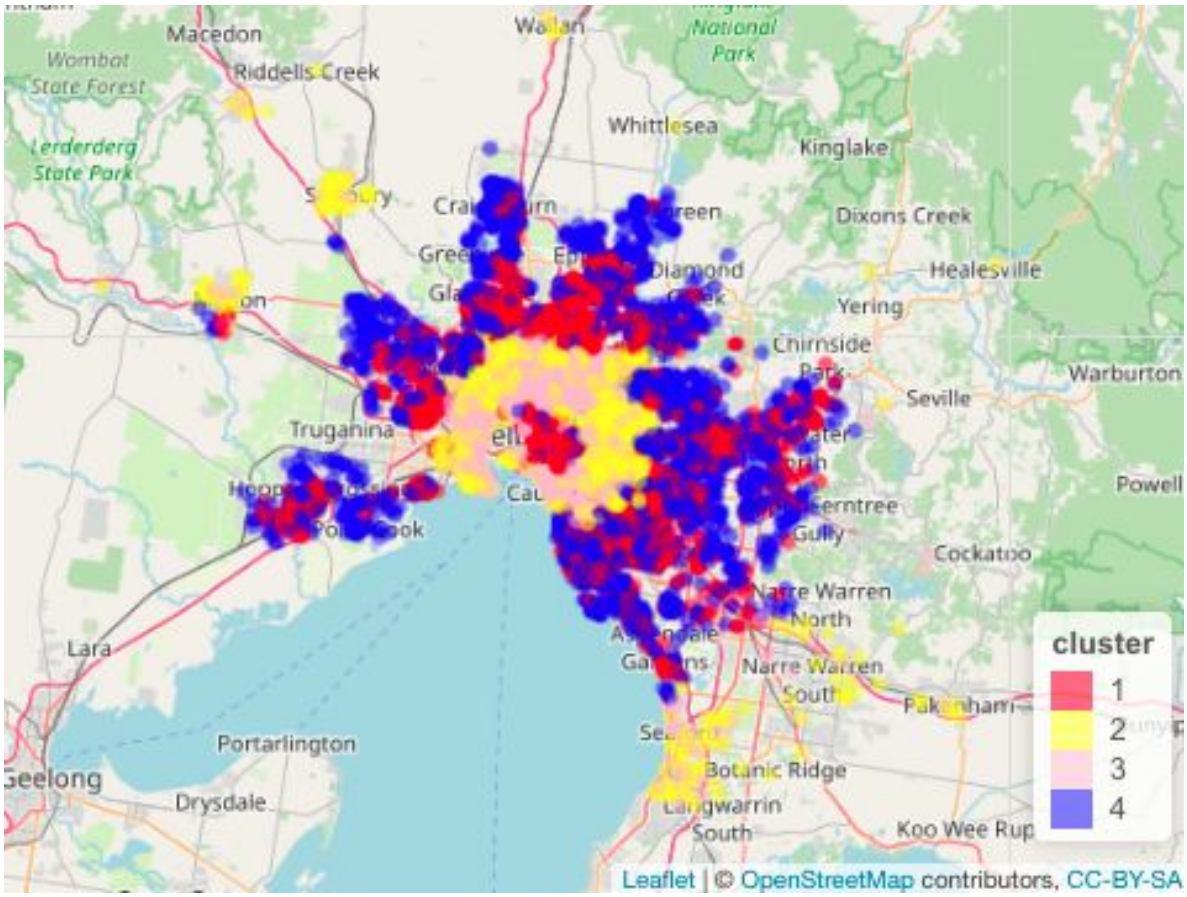
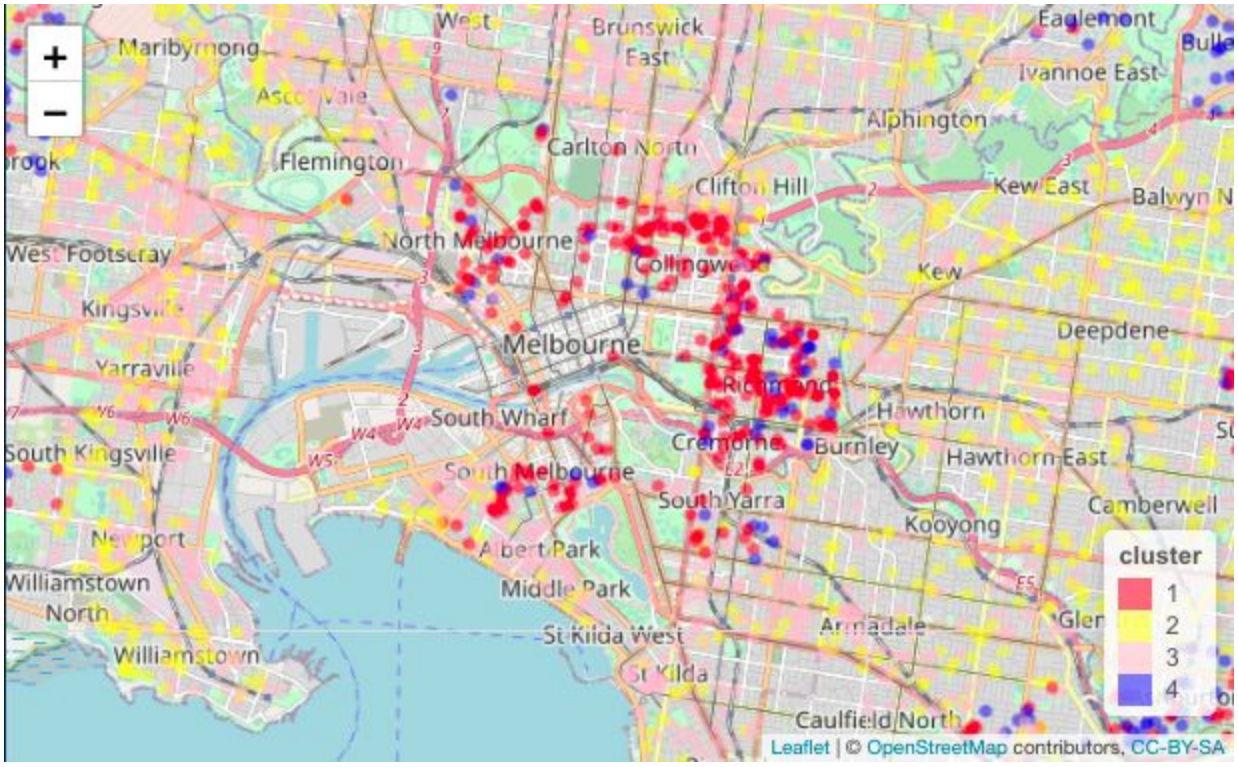


Figure 8 Continued (Zoomed In):



References

Abraham, Jesse M, William N Goetzmann, and Susan M Wachter. “Homogeneous Groupings of Metropolitan Housing Markets.” Journal of Housing Economics 3, no. 3 (1994): 186–206. https://doi.org/10.1006/jhec.1994.1008.

Pino, Tony. “Melbourne Housing Market.” Kaggle, October 14, 2018.

Izenman, Alan Julian. 2013. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. New York, NY: Springer. [ISBN-13: 978-0-387-78189-1] Chapter 12: Cluster Analysis, pages 407-420. Available from the Springer collection: https://link-springer-com.turing.library.northwestern.edu/book/10.1007%2F978-0-387-78189-1Links to an external site.

Chapman, Chris. & Feit, E. R For Marketing Research and Analytics. Cham: Springer, 2015.

About the Author

Vincent Pun is a Master’s candidate in the data science program at Northwestern University, experienced in financial services process automation, tax reporting, and strategy consulting.