Kay Quiballo | MSDS 411 | 04.17.2023

MSDS 411 Assignment 2

Abstract. An executive summary of the research.

In the housing market, sellers and buyers are always trying to get their money's worth in such an important investment. Sellers want to make the biggest profit possible, but don't want their properties to get passed up on for similar properties that are priced lower. There are many aspects that impact how a property is valued including the type of property (house, townhouse, or apartment unit), location, and size of the property.

In this paper, we analyzed the Melbourne Housing Dataset from January 2016 which consisted of 9 variables and 8,399 records. Our goal was to perform a cluster analysis to determine which properties exhibited similar traits and the distinctions between each cluster.

Introduction. Why are you conducting this research?

As a real estate firm, it is important to identify the different types of houses to assign properties to different realtors. A real estate firm may properly appraise a property by identifying similarities and differences between other comparable, neighboring properties. With such a large dataset, we made sure that properties were allocated by general region. It is also important to conduct this research because the housing market is constantly changing and must be repeatedly analyzed for current trends and relevant pricing.

Literature review. Who else has conducted research like this?

Similar research has been conducted by Gabrielli et al in the textbook, Appraisal: From Theory to Practice. In the chapter, *Gaps and Overlaps of Urban Housing Sub-market: Hard Clustering and Fuzzy Clustering Approaches*, they speak on the topic of the spatially divided housing submarkets and thus its segmentation. Using cluster analysis, they investigate the multidimensionality of the housing market and determine which attributes are significant predictors of sale price.

Methods. How are you conducting the research?

1. Data Cleaning

The full Melbourne Housing Dataset from January 2016 contains 34,857 records and 21 variables ranging from Address to Price to Sale Date. The data was then subsetted to only include records without missing data resulting in 8,887 records and the following 9 variables:

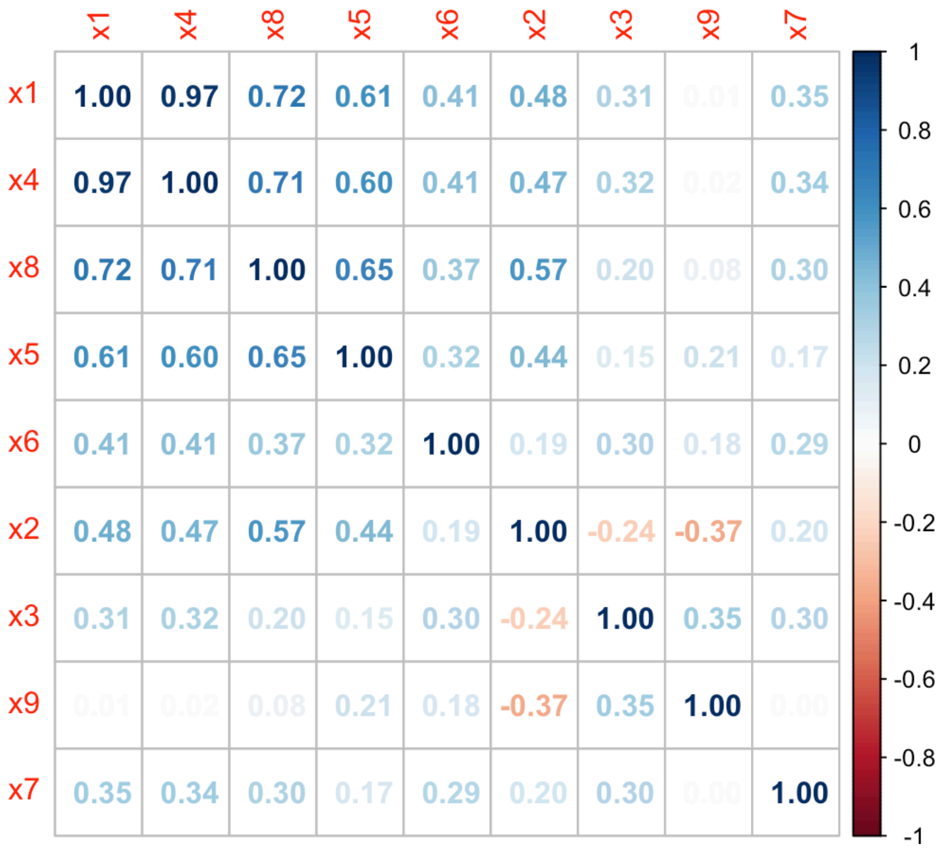
Rooms	**(x1 – number of rooms),**
Price	**(x2 – price in Australian dollars),**
Distance	**(x3 - distance from CBD in km),**
Bedroom2	**(x4 - scraped number of bedrooms from a different source),**
Bathroom	**(x5 – number of bathrooms),**
Car	**(x6 – number of car spots),**
Landsize	**(x7 – land size in meters),**
Building Area	**(x8 – building size in meters), and**
YearBuilt	**(x9 – year the house was built).**

In order to prepare the data for cluster analysis, all 9 continuous variables were scaled into z-scores in order to easily compare across data points as well as identify outliers. Any data points that had standardized z-scores that were less than -3.29 or greater than 3.29 were considered a very rare occurrence (1 in 1000) and were excluded as outliers from the dataset leaving 8,399 datapoints in the cleaned subset of data.

2a. Exploratory Data Analysis – Correlation Plot

In the initial EDA we first looked at a correlation plot between variables (this correlation plot has been rearranged according to the Primary Component Analysis in the following section)

Correlation Plot



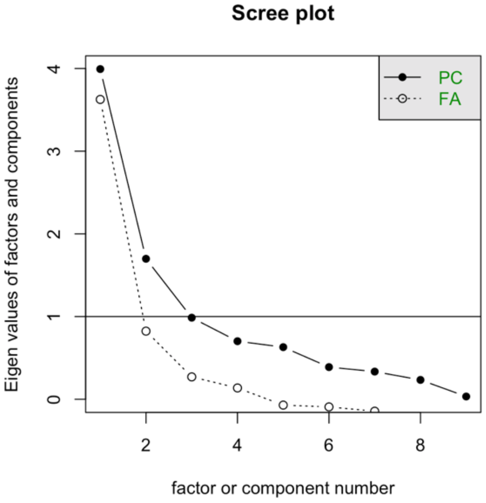
	x1	x4	x8	x5	x6	x2	x3	x9	x7
x1	1.00	0.97	0.72	0.61	0.41	0.48	0.31	0.01	0.35
x4	0.97	1.00	0.71	0.60	0.41	0.47	0.32	0.02	0.34
x8	0.72	0.71	1.00	0.65	0.37	0.57	0.20	0.08	0.30
x5	0.61	0.60	0.65	1.00	0.32	0.44	0.15	0.21	0.17
x6	0.41	0.41	0.37	0.32	1.00	0.19	0.30	0.18	0.29
x2	0.48	0.47	0.57	0.44	0.19	1.00	-0.24	-0.37	0.20
x3	0.31	0.32	0.20	0.15	0.30	-0.24	1.00	0.35	0.30
x9	0.01	0.02	0.08	0.21	0.18	-0.37	0.35	1.00	0.00
x7	0.35	0.34	0.30	0.17	0.29	0.20	0.30	0.00	1.00

From the correlation plot, we see the strongest correlations in the top left corned which are between x1, x4, x8, and x5. This means that there is a positive association between number of rooms (x1), scraped number of rooms (x4), building area (x8), and number of bathrooms (x5). It makes sense that larger properties may have a larger number of bedrooms and bathrooms.

There also seems to be an inverse relationship between SalePrice with Distance and Yearbuilt. This indicates that properties that are closer to the Central Business District tend to sell for higher, and older units tends to sell for higher.

2b. Exploratory Data Analysis – Primary Component Analysis and Factor Analysis

The next part of the Exploratory Data Analysis was to conduct a primary component analysis.



Loadings and Goodness of Fit from PCA	
2 Primary Components	3 Primary Components
PC1 PC2	PC1 PC2 PC3
x1 **0.92** -0.04	x1 **0.92** -0.04 0.00
x2 0.60 **-0.67**	x2 0.60 **-0.67** 0.00
x3 0.35 **0.73**	x3 0.35 **0.73** 0.24
x4 **0.91** -0.02	x4 **0.91** -0.02 0.00
x5 **0.75** -0.02	x5 **0.75** -0.02 -0.42
x6 **0.56** 0.29	x6 **0.56** 0.29 0.14
x7 0.47 0.16	x7 0.47 0.16 **0.70**
x8 **0.86** -0.10	x8 **0.86** -0.10 -0.16
x9 0.08 **0.77**	x9 0.08 **0.77** -0.46
Cumulative Var: 0.44 0.63	Cumulative Var: 0.44 0.63 0.74
(RMSR) is 0.09	(RMSR) is 0.08
Fit based upon off diagonal values = 0.95	Fit based upon off diagonal values = 0.96

From the scree plot, we can tell that the Primary Component Analysis (PCA) calls for 2-3 components and the Factor Analysis calls for 1 component indicated by the eigenvalues greater than 1. A Factor Analysis with only one component will not provide with any insight on how the variables relate to each other, so we will continue with only the PCA.

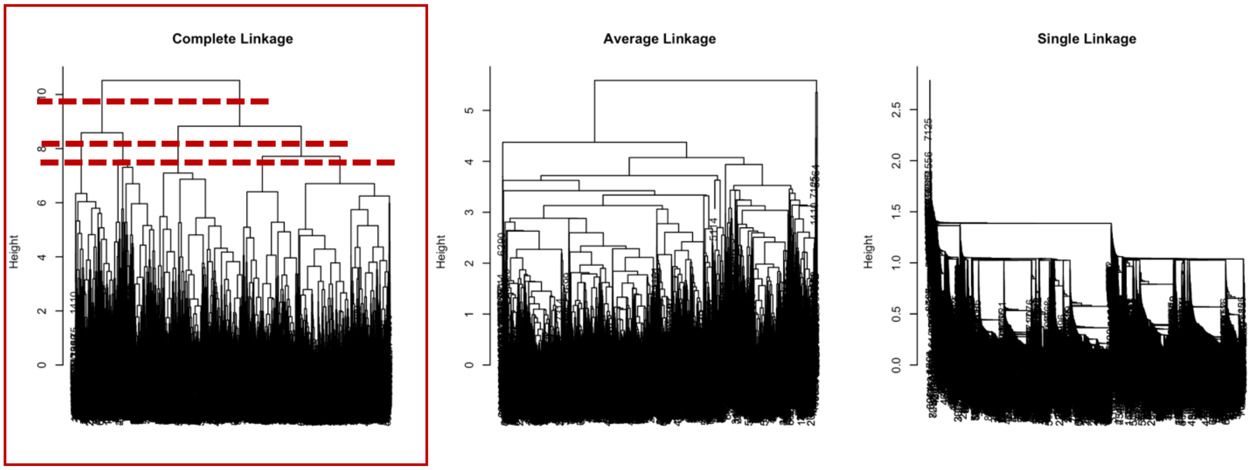
With the third eigenvalue very close to 1, we decided to compare both the PCA with 2 and 3 components. From the loading values, the components have very similar groupings of variables which are deemed significant by magnitudes greater than 0.50. The only difference is that x7 (land size) which is insignificant in the 2 PCA can stand alone in a 3 PCA. The PCA's also perform similarly in terms of RMSR (0.09 and 0.08 respectively – both of which indicate a fair fit being less than 0.10) and Fit (0.95 and 0.96 – both of which indicate a high goodness of fit being close to 1.00). The 2 PCA has a 0.63 cumulative variance which means that the 2 components explain 63% of the variance in the dataset, and the addition of another component bumps the cumulative variance from 63% to 74%.

We organized the correlation plot in the previous section according to the clusters identified here. Component 1 contains x1, x4, x8, x5, x6 (respectively – number of rooms, scraped number of bedrooms, building size, number of bathrooms, and car spots), and Component 2 contains x2, x3, x9 (respectively - price, distance from CBD, and year built) in order of heaviest to lightest loadings values (which indicate level of explained variance).

Component 1 seems to indicate size of property. This makes sense, as larger properties have more rooms, bathrooms, and car spots. Component 2 indicates location and age. This makes sense as more expensive properties are closer from the Central Business District which is a more desired location of living due to accessibility. More expensive properties were older which may be due to houses being older than apartments and having a higher price point.

From the PCA, we will continue to use 2 components as we may only use 2 dimensions in cluster visualizations.

3a. Cluster Analysis – Hierarchical Clustering

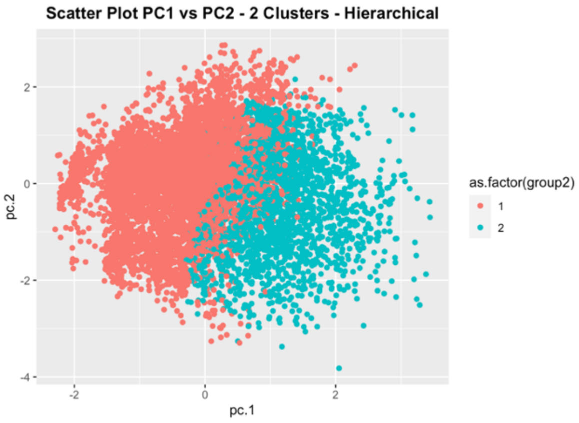


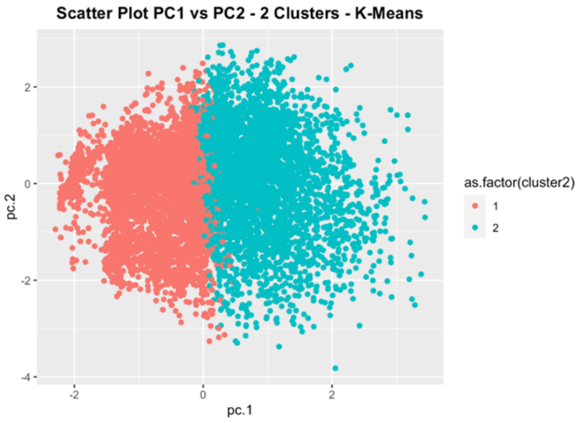
From the Hierarchical Clustering, we ran 3 different methods of linkage. We see the most even distribution between potential clusters from the Complete Linkage tree and use this to determine a potential amount of cluster. The first dotted line indicates a potential break between 2 clusters (A/B). The second dotted line indicates a potential break between 4 clusters (2 subsets of 2 clusters from the first break - A1/A2/B1/B2). The third dotted line indicates a potential break between 5 clusters (another subcluster of B2 - A1/A2/B1/B2i/B2ii).

The Hierarchical Clustering indicates homogenous sized clusters with potentially 2, 4, and 5 clusters. We consider each of these options in the K-Means cluster analysis to determine which one is best for the dataset.

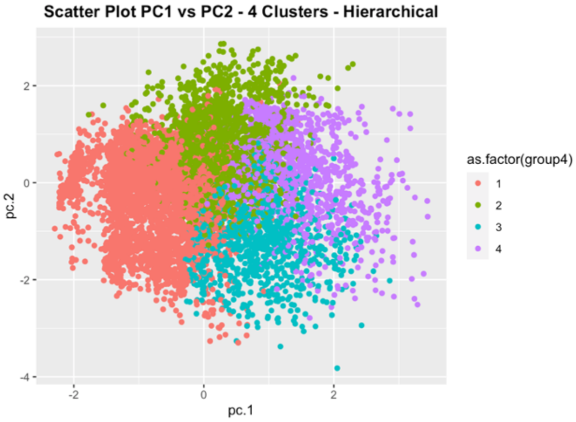
3b. Cluster Analysis – K-Means Clustering

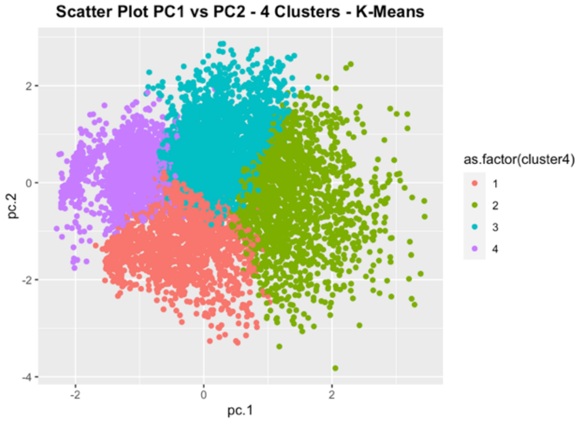
2 Clusters – Left and Right



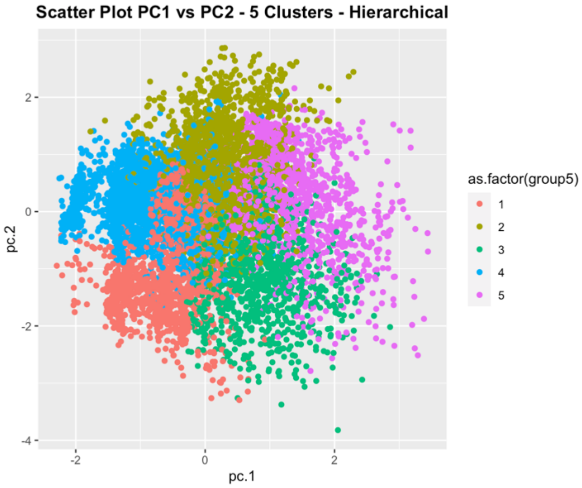


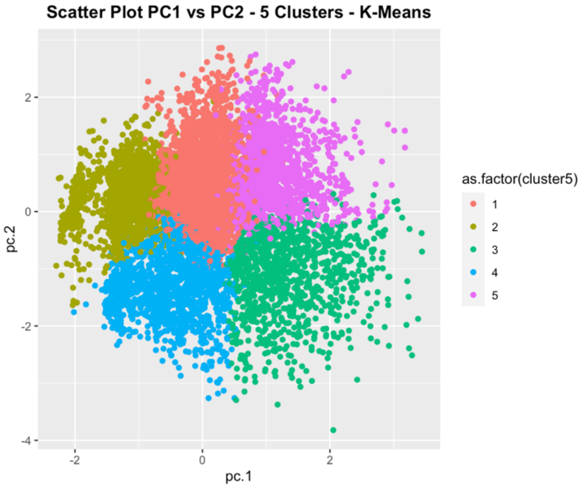
4 Clusters – Left, Right, Upper Middle (taken from Left), Lower Middle (taken from Right)





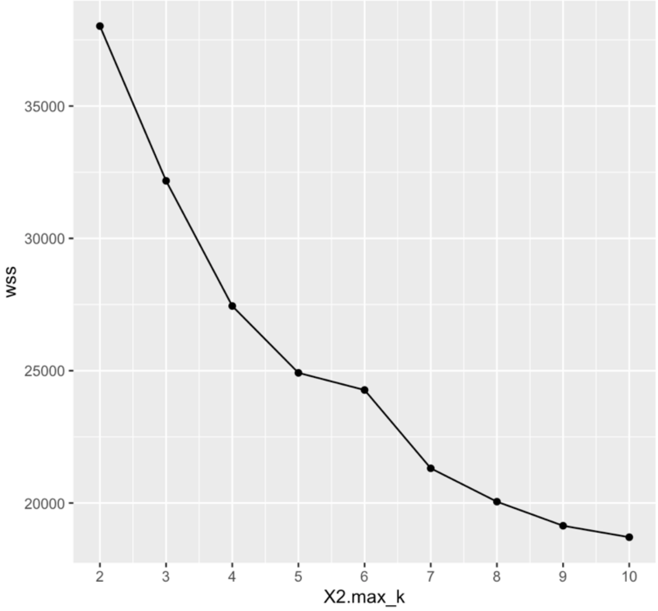
5 Clusters – Upper Left, Lower Left, Right, Upper Middle (taken from Left), Lower Middle (taken from Right)





The 6 plots above consider the 2 clustering methods in combination of the 3 different cluster numbers (2, 4, 5). Although the Hierarchical Clustering helped us determing how many potential clusters to use, the K-Means clustering has much more distinct groups. K-means also provides us with useful statistics to look at.

Total Within Cluster Sum of Squares vs Number of K



From the plot on the left, we see the most diminishing returns of wss at the point k=5. From here we can confirm that 5 clusters is the optimal amount of clusters to use.

# Clusters	K-Means	CSS
2 Clusters	Kmeans output – 2 clusters x1 x2 x3 x4 x5 x6 x7 x8 x9 1 -0.6759790 -0.3620388 -0.3236237 -0.6735266 -0.6610255 -0.4691575 -0.15512133 -0.5411928 -0.1766063 2 0.6852617 0.2851648 0.2467841 0.6821184 0.6568648 0.3772342 0.03119069 0.4513835 0.1919793	Within cluster sum of squares by cluster: [1] 17254.96 20767.13 (between_SS / total_SS = 30.7 %)
4 Clusters	Kmeans output – 4 clusters x1 x2 x3 x4 x5 x6 x7 x8 x9 1 -1.2923019 -0.7162167 -0.500098815 -1.2850371 -0.72592260 -0.5828420 -0.23403412 -0.7931918 0.37739598 2 0.1282040 -0.3895495 0.520433544 0.1397497 -0.03226101 0.1784587 -0.01948711 -0.1053942 0.46096289 3 1.1668066 0.8328901 -0.002036095 1.1506370 1.05889021 0.4988790 0.09364994 0.9164338 0.03362446 4 -0.2252969 0.3510594 -0.702640402 -0.2388525 -0.45516650 -0.5705442 -0.14261301 -0.2611006 -1.30651830	Within cluster sum of squares by cluster: [1] 3934.433 9925.607 8235.155 5349.221 (between_SS / total_SS = 49.9 %)
5 Clusters	Kmeans output – 5 clusters x1 x2 x3 x4 x5 x6 x7 x8 x9 1 1.07031840 -0.1215249 0.6795259 1.07617171 0.8374022 0.4320565 0.06456189 0.6421044 0.6830254 2 1.05497506 1.4068768 -0.3866424 1.03312577 0.9629147 0.3880190 0.08917351 0.8611138 -0.6305112 3 -0.05291859 -0.3372832 0.3354507 -0.04347467 -0.1945884 0.1361466 -0.03350785 -0.2132566 0.3437024 4 -0.34358179 0.2865572 -0.7561327 -0.36205622 -0.5272089 -0.6909021 -0.17593744 -0.3188808 -1.4003684 5 -1.32067330 -0.7330484 -0.5115393 -1.31308079 -0.7279883 -0.5960803 -0.23725413 -0.8037035 0.4010784	Within cluster sum of squares by cluster: [1] 4444.884 4887.050 8201.917 3870.646 3513.961 (between_SS / total_SS = 54.6 %)

We will only draw information from the 5 cluster K-means output. The following section converts the means from z-scores to unstandardized values for better interpretation and annotates them based on relative value and variables from the primary component analysis.

Kay Quiballo | MSDS 411 | 04.17.2023

Results. What did you learn from the research?

Unstandardized K-Means for 5 Clusters

Number Clusters	X1 Rooms	X2 Price	X3 Distance	X4 Bedroom2	X5 Bathroom	X6 Car	X7 Landsize	X8 Building Area	X9 YearBuilt
1 (n=1372)	4.05	982,793	15.38	4.04	2.17	1.98	474	185	1990
2 (n=1046)	4.04	1,843,306	8.12	4.00	2.25	1.94	482	199	1943
3 (n=2818)	3.01	861,318	13.04	3.00	1.48	1.73	439	128	1978
4 (n=1421)	2.74	1,212,550	5.60	2.70	1.26	1.04	388	121	1915
5 (n=1742)	1.83	638,496	7.27	1.81	1.13	1.12	367	89	1980

High | Medium | Low
PCA 1 | PCA 2

Clusters 1 and 2 have an above average amount of rooms, bathrooms, car spaces, building area (which are all part of PC1), and land size. However, they differ in terms of PC2 variables. Cluster 1 is more average priced, further from CBD, and are newer units. Cluster 2 is more expensive, older houses, that are closer to CBD.

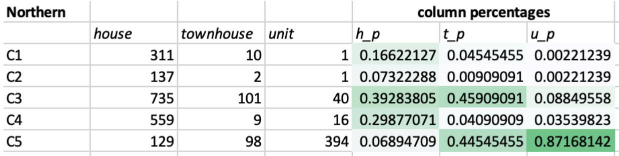
Clusters 3 and 4 have an average amount of rooms, bathrooms, and building area (which are all part of PC1). However they differ in terms of PC2 variables. Cluster 3 is more average priced, further from CBD, and are newer units (just like cluster 1). Cluster 4 is more expensive, older houses that are closest to CBD (just like cluster 2).

Cluster 5 has a below average about of rooms, bathrooms, car spaces, and building area (which are all part of PC1). Cluster 5 is also the cheapest and newest with an average distance from CBD.

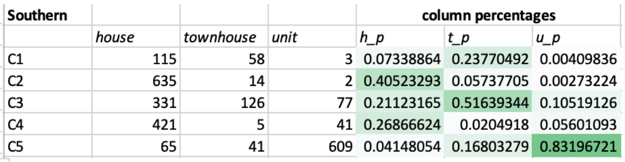
Cluster 3 has the highest number of records (n=2818) whereas Cluster 2 has the lowest number of records (n=1046).

Clusters by Region

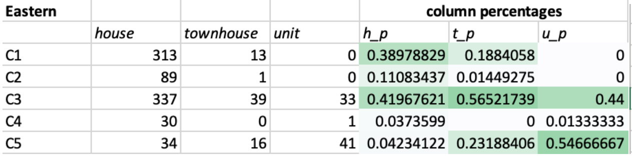
Here we consider the categorical variable Region with the results of our cluster analysis. We will only look at North/South/East/West Metropolitan regions because they have much larger sample sizes than any other regions.



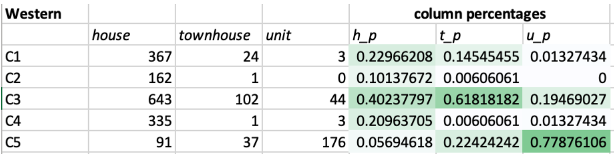
Northern				column percentages		
	house	*townhouse*	*unit*	*h_p*	*t_p*	*u_p*
C1	311	10	1	0.16622127	0.04545455	0.00221239
C2	137	2	1	0.07322288	0.00909091	0.00221239
C3	735	101	40	0.39283805	0.45909091	0.08849558
C4	559	9	16	0.29877071	0.04090909	0.03539823
C5	129	98	394	0.06894709	0.44545455	0.87168142



Southern				column percentages		
	house	*townhouse*	*unit*	*h_p*	*t_p*	*u_p*
C1	115	58	3	0.07338864	0.23770492	0.00409836
C2	635	14	2	0.40523293	0.05737705	0.00273224
C3	331	126	77	0.21123165	0.51639344	0.10519126
C4	421	5	41	0.26866624	0.0204918	0.05601093
C5	65	41	609	0.04148054	0.16803279	0.83196721



Eastern				column percentages		
	house	*townhouse*	*unit*	*h_p*	*t_p*	*u_p*
C1	313	13	0	0.38978829	0.1884058	0
C2	89	1	0	0.11083437	0.01449275	0
C3	337	39	33	0.41967621	0.56521739	0.44
C4	30	0	1	0.0373599	0	0.01333333
C5	34	16	41	0.04234122	0.23188406	0.54666667



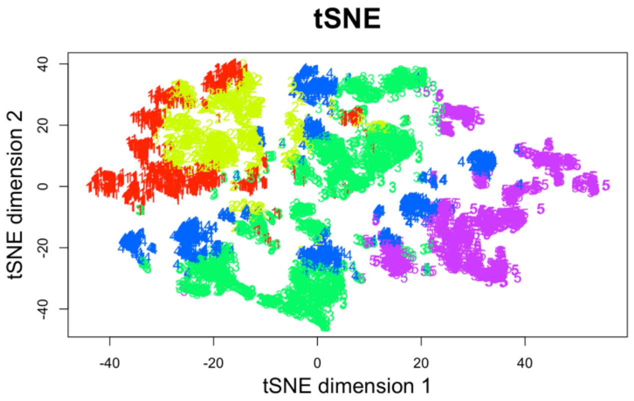
Western				column percentages		
	house	*townhouse*	*unit*	*h_p*	*t_p*	*u_p*
C1	367	24	3	0.22966208	0.14545455	0.01327434
C2	162	1	0	0.10137672	0.00606061	0
C3	643	102	44	0.40237797	0.61818182	0.19469027
C4	335	1	3	0.20963705	0.00606061	0.01327434
C5	91	37	176	0.05694618	0.22424242	0.77876106

From the the K-means analysis and the Clusters by region table, we can determine the following:

Townhouses typically fall in Cluster 3 and sometimes Cluster 5 and share the following in common: cheap to average pricing and newer buildings, average to below average space. The same can be said of Units primarily falling in Cluster 5 and sometimes Cluster 3

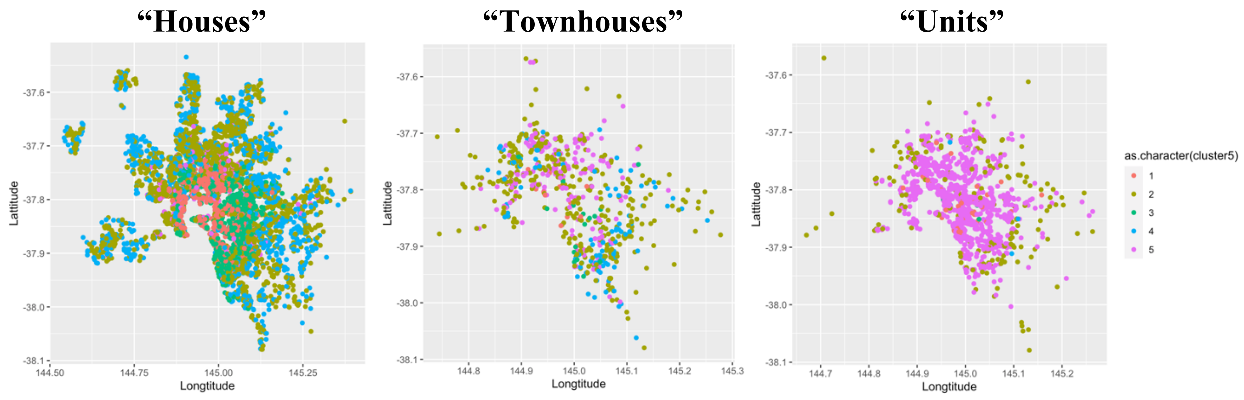
Houses tend to vary in cluster based on location. Northern and Western houses are typically Clusters 1, 3 and 4, Southern are Clusters 2, 3, and 4, and Eastern is Clusters 1 and 3. This just goes to show that there may be some interaction between type of home and the primary components of this analysis. We already know this is true based of the distance from CBD variable that affects Sale Price.

tSNE



The tSNE visualization also confirms the exististence of homogenous subsets of data. Clusters 1, 2, and 5 are alll distinctly separated. There is a bit of mix up between clusters 3 and 4, but these clusters are distinctly seperated from Clusters 1, 2, and 5. We may reflect on the Unstandardized K-Means for 5 Clusters Table that shows similar means for all PC1 variables across clusters 3 and 4.

Latitude and Longitude Maps of Clusters



Note that the center of each map corresponds to the Central Business District which can be considered the hub of the city. The Houses Map shows a pretty clear depiction of how distance from CBD differs between clusters. Clusters 2 and 4 are furthest away whereas Clusters 3 and 5 are closer, and Cluster 1 is the closest.

Townhouses show a spread of Clusters 2, 4 and 5 in various regions. Units are predominantly Cluster 5 which are closest to the CBD.

Kay Quiballo | MSDS 411 | 04.17.2023

Conclusions. So, what does it all mean?

In conclusion there are a few things we need to take into consideration as a real estate firm that has conducted a cluster analysis.

Primary Component 1: The highest influencing primary component for cluster distinction is the size of a building (rooms, bathrooms, car spaces, and building area). This is typically definitive of Cluster 1 and 2. Units with more space tend to sell higher. Smaller units like Cluster 5 tend to sell for cheaper. *When pricing properties, it is important to compare to units similar in size.*

Primary Component 2: The other primary component for cluster distinction is Distance and Year built. Older buildings tend to sell more like Clusters 2 and 4, and new building sell for cheaper like Clusters 1, 3 and 5. *When pricing properties, it is important to compare units similar in age and/or similar in distance from CBD.*

The Map of Clusters by Building type show us that different types of buildings are sold based on their proximity to the CBD. We already know from Principal Component 2 to compare units based on their distance from CBD, but *we must also compare to similar types of units* (houses, townhouses, or units). Using the maps, we can also help determine what each type of unit we should use in comparison. Here are 3 examples:

1. I am selling a House that is far from the CBD. How should I price it?
Well, we know Houses far from CBD fall in Clusters 2 and 4. A 4 bed 2 bath typically sells for 1.8 million whereas a 3 bed 1 bath typically sells for 1.2 million. These prices are typical of older homes built between 1915-1943. The base quote can be adjusted as we evaluate more information about the house.

2. I am selling Townhouse. How should I price it?
A Townhouse is typically Cluster 2 and 5 and is typically located a moderate distance away from the CBD. Based on the size of the Townhouse, a 4 bed 2 bath typically sells for 1.8 million whereas a 2 bed 1 bath typically sells for 640k. These prices are based off average cluster values and can be adjusted based on further information about the property.

3. I am selling a Unit close to CBD. How should I price it?
Units are typically a Cluster 5. We know that a 2 bed 1 bad typically sells for 640k. We can adjust the base price based on size of unit and distance from CBD. The base quote can be adjusted from further information on the unit.

This is just an example of the starting point for some pricing quotes for a real estate firm. These prices are broad because they are starting points and need to be adjusted when comparing different attributes across similar clusters.

We may also note that properties should be allocated to realtors who have experience in a certain area (based on the maps) and have experience in working with the respective Type of property. There are many variables at play when it comes to the housing market, but the above points are the few takeaways we advise based off the cluster analysis.

Kay Quiballo | MSDS 411 | 04.17.2023

Source:

Gabrielli, L., Giuffrida, S., Trovato, M.R. (2017). Gaps and Overlaps of Urban Housing Sub-market: Hard Clustering and Fuzzy Clustering Approaches. In: Stanghellini, S., Morano, P., Bottero, M., Oppio, A. (eds) Appraisal: From Theory to Practice. Green Energy and Technology. Springer, Cham. https://doi.org/10.1007/978-3-319-49676-4_15