NEEDLE IN HAYSTACK: CAN UNSUPERVISED CLUSTERING UNLOCK HIDDEN HOUSING SECRETS?

Northwestern University, Unsupervised Learning Methods

February 10th, 2022

Abstract:

The largest expense for the average family is housing or shelter. On average, a mortgage makes up approximately 18 to 30 percent of the family's income. Being such a large family expense or a significant investment opportunity, it is reasonable to analyze housing market fundamentals with the goal of identifying properties that are undervalued when evaluated against comparable properties. In most jurisdictions, significant amounts of housing data are available and can be used to help identify pricing outliers. This paper looks specifically at housing data from Melbourne, Australia and uses unsupervised learning techniques to create clusters that are compared to the housing types and their locations. Different forms of scaling, K-means clustering, hierarchical agglomerative clustering, PCA, and T-SNE were tested and evaluated for accuracy. Standard scaling in combination with k-means clustering seemed to provide the most accurate unsupervised learning results.

Keywords:
k-means, hierarchical agglomerative clustering, unsupervised learning, housing market

I. Introduction

Housing market is an interesting area to study due to the constant booms and crashes that can occur. What makes a house worth the price to pay? In many real estate firms, there are countless attributes to determine the price of a house. Sometimes these attributes can counteract with the actual value of a house and its worth. As a real estate firm, in order to market houses to the right population, there is a lot to discover in terms of the types of houses that are available. Once determining more information about a house's attributes and characteristics to its value, it is extremely helpful to organize the housing territories and market them to appropriate buyers.

II. Literature Review

Cluster-based analyses using k-means, hierarchical clustering, and Principal Component Analysis for housing market understanding have been used extensively in the literature. Calka (2019) estimated the value of residential property using k-means clustering. Tomal (2020) used k-means clustering for county housing markets in Poland. Going further, expert-based biases combined with statistical clustering algorithms have also been used. Berna and Watkins (2017), for example, were inspired by previous authors in using the different statistical methods such as k-means and hierarchical clusterings. They, however, incorporated an additional step in bringing in real-estate agents to improve on the methodology and have shown that expert guidance leads to higher accuracy in prediction.

III. Methods

In order to grasp a high level understanding of the housing dataset, EDA was conducted with a subset of variables (Rooms, Price, Distance, Bedroom2, Bathroom, Car, Landsize, BuildingArea, YearBuilt, Type, Suburb, Latitude, Longitude, Propertycount). Both boxplots and scatterplots were used to establish basic analysis on the dataset. The boxplots (example in Figure 1) for each of the variables displayed various outliers which can later be resolved through scaling. Figure 1 also displays the scatterplot of houses grouped by geographical location, the results displayed prominent groupings of each house type.

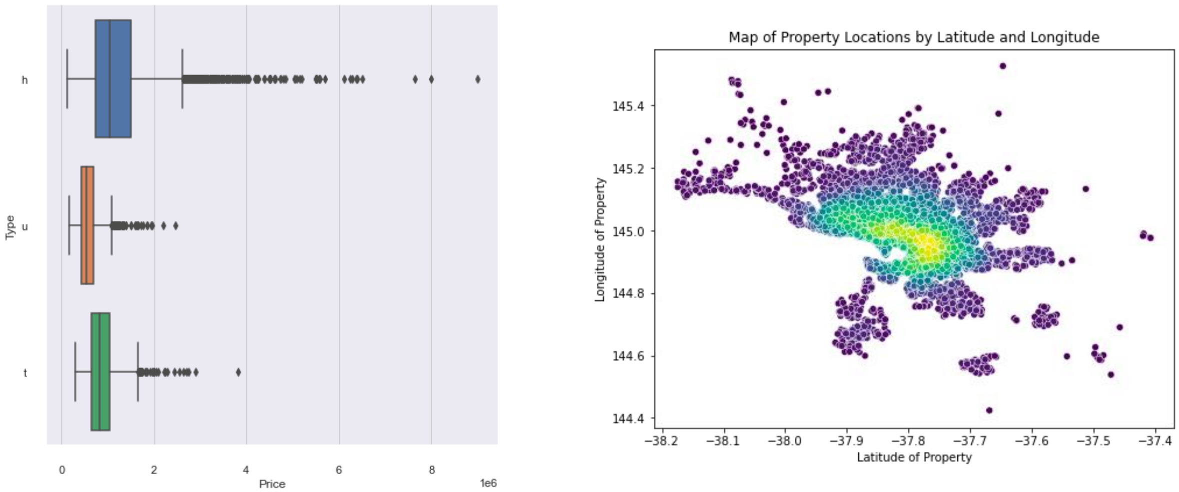


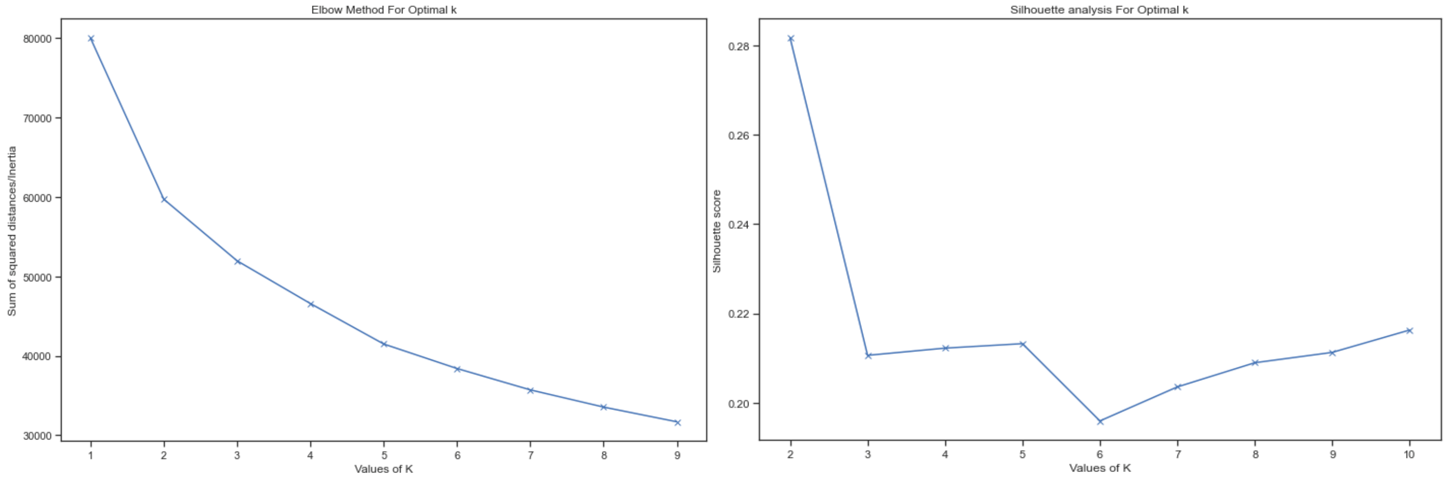
Figure 1: EDA plots of Type

To replicate these potential clusters, certain variables (Type, Suburb, Latitude, Longitude, Propertycount) were dropped in order to investigate deeper relationships. The scatterplot demonstrated a baseline for a potential outcome to compare to once further analysis is done. From the new subset, we tested two types of scalar methods and two types of clustering methods (along with three types of linkages) to determine which scaling and clustering method combination yielded the best results. These methods and usage justifications are discussed below.

Clustering methods

The first clustering method we chose was K-Means because it works well on large datasets, is computationally efficient, and seems to work well in a variety of real-world situations. The clusters were predetermined through plotting interia and silhouette scores such as the example in Figure 2, to obtain the best number of clusters for the dataset.

Figure 2: Inertia and Silhouette Plots



The second clustering method we chose was Agglomerative Hierarchical Clustering because we did not need to pre-specify the number of clusters (which can be difficult to determine a priori). Additionally, we tested three different linkage methods to use with the AHC algorithm: single linkage, complete linkage, and ward linkage. We tested these methods separately because they determine the distance between pairwise points in different ways.

Scaling Methods

For our first scaling method, we chose StandardScaler because it forces the variance of each feature to be the same (equal to one), which is what the K-means clustering method assumes of the data to be true. However, we also noticed some outliers in the feature data (see the boxplots for a visual illustration). We decided to also test out the RobustScaler method because it helps reduce skewing the data towards outlier values by scaling to the interquartile range.

IV. Results

Clustering results from the four models are analyzed using dimensionality reduction plots, percentage of housing types in each cluster with respect to the original data breakdowns, and geographical locations of each cluster,

Grouping results are displayed in Figures 3 and 4 using t-SNE and Principal Component Analysis dimensionality reduction plots.

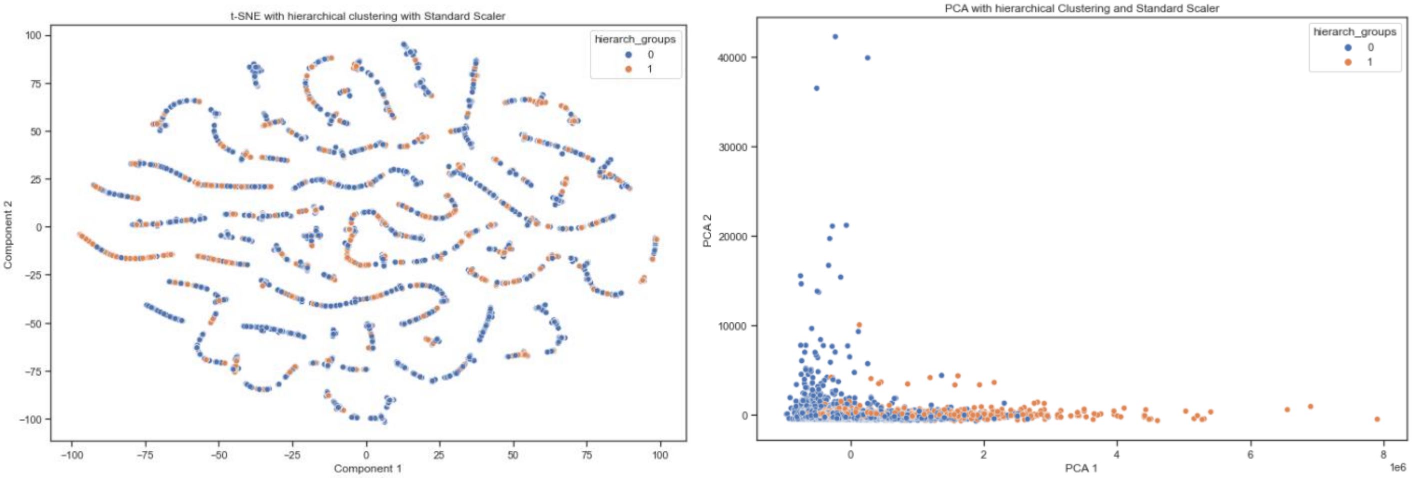


Figure 3: T-SNE & PCA Visualization on Hierarchical Clustering

Visualizations show either data imbalance in the agglomerative hierarchical clustering models and many overlaps using k-means Clustering. While hierarchical clustering in Figure 3 shows better differentiations, the results are too simplistic to form any conclusions. k-means clustering measurements shown in Figure 4 provide greater potential for more informative groupings.

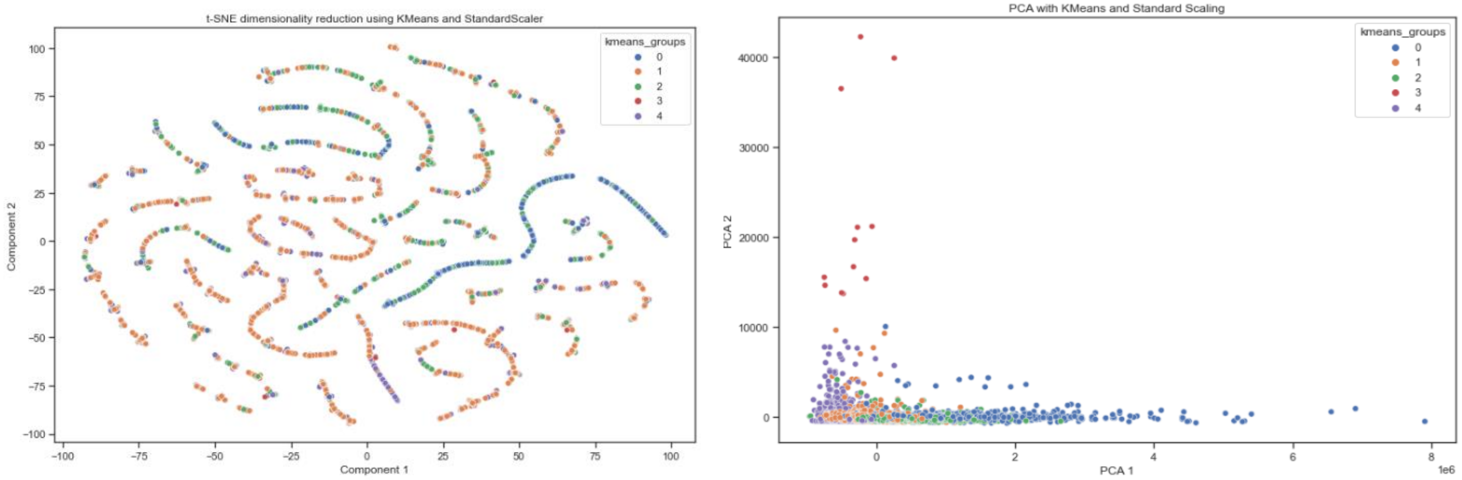
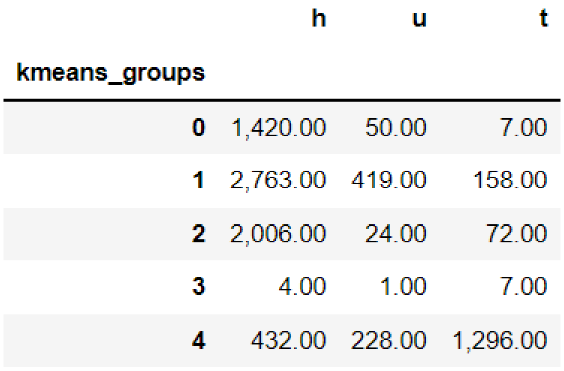
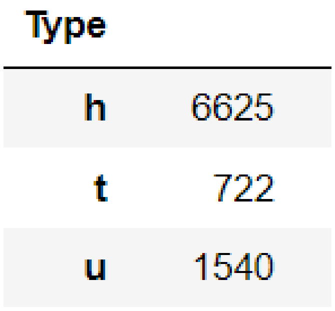


Figure 4: T-SNE & PCA Plots for K-means Clustering

Comparing the number of items in each cluster compared to the number of types in the data shows that the clusterings using both k-means and agglomerative hierarchical does not show the same proportion of types in each cluster. That said, k-means with standard scaling gets closest to real data proportions, albeit with most groups having a majority of type h. As displayed below in Tables 1 and 2, group 0 has more type h, group 4 has more type t, and group 2 and 3 have more type h. The representative type populations in each cluster as compared to real data proportions is shown below.



kmeans_groups	h	u	t
0	1,420.00	50.00	7.00
1	2,763.00	419.00	158.00
2	2,006.00	24.00	72.00
3	4.00	1.00	7.00
4	432.00	228.00	1,296.00

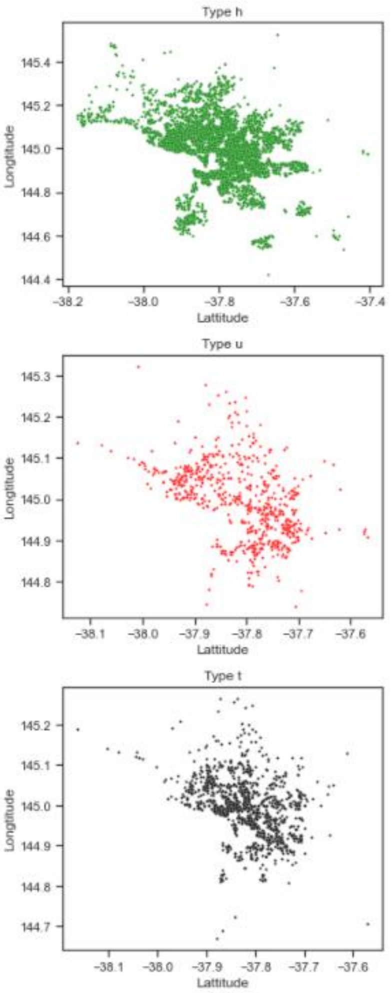


Type	
h	6625
t	722
u	1540

Tables 1 & 2: Clusters pertaining to House Types & Total Counts

It is interesting to note that groups 0, 1, and 2, all assign a dominant majority of items to type h. This is not surprising considering the actual data breakdown with a significant data imbalance in favor of type h.

Considering the better fit of standard scaling with k-means clustering to represent the data, we can now view the geographical location of each cluster with respect to the geographical location of the housing types in Figure 5.



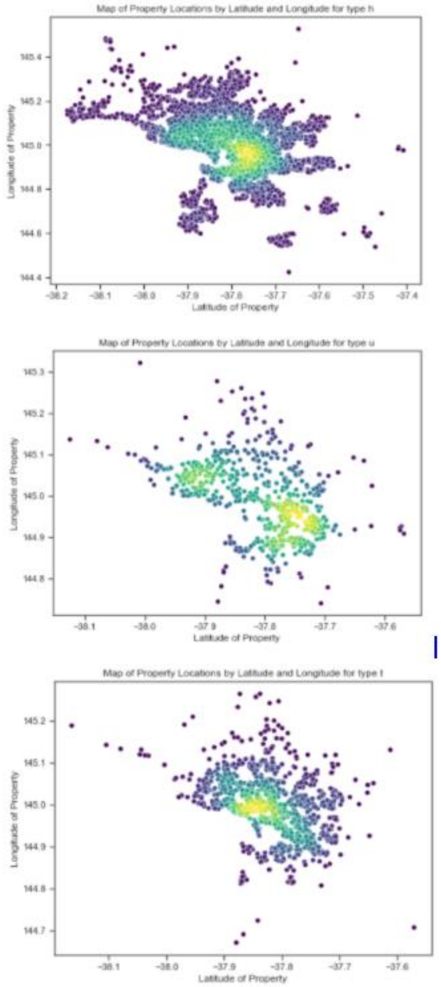


Figure 5a: Scatter Plots *Figure 5b: Gaussian Plots*

Plots in Figure 5a shows geographic plots for each type, Figure 5b shows density gaussian plots for each type, and Figure 5c shows a geographical plot of k-means with standard scaling clusters.

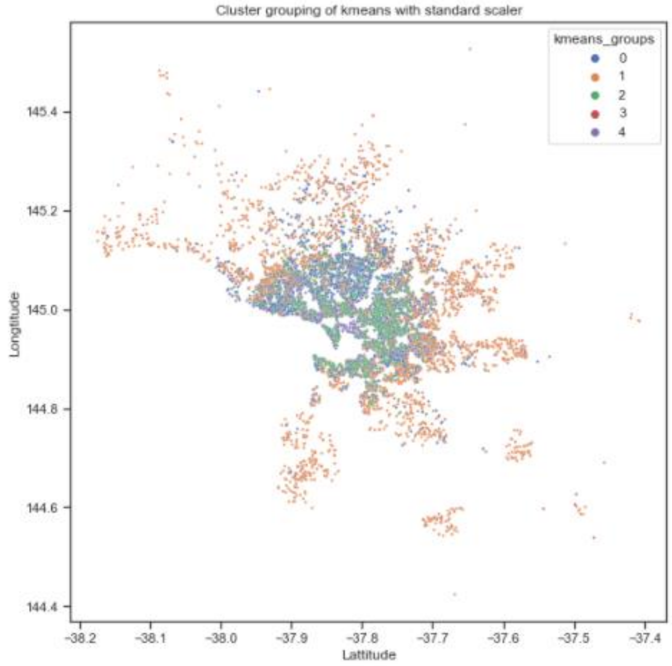


Figure 5c: Scatterplot of K-means groups

As can be seen, cluster group 0 appears to have data closest to the center of the plot, downtown area, equivalent to type t. Cluster group 4 surrounds the center, equivalent to type u, and cluster group 2 and 3 are equivalent to type h as they are located throughout the map.

Finally, looking at the densest population area for each cluster with respect to the cluster areas for each group (taking the 50 largest population items), we see that group 0 resembles type u in terms of neighborhood locations, and group 2 and 4 are similar to type h neighborhoods.

Combining the interpretation from all three visualization methodologies, we see that the clustering in all modeling algorithms is not exact. While clusters resemble some properties of type datas, there does not appear to be any consensus nor any distinct ways to adequately represent the dataset in lower dimensions. That being said, we can look as to whether resolving the data imbalance and the number of omitted variables with NA can help in clustering results. We ran a preliminary model imputing NAs for columns where less than thirty percent of the values were missing (see modeling results in python code). While the imputation increased our dataset from 8887 items to 11700, there is still much work to be done as clustering, while better, still lacks the distinct clustering separations.

V. Conclusions

This paper aims to create more meaningful clusters of housing data to provide real estate agents with more efficient ways to notice early housing deal outliers. Adequate clustering leads to more informed decision making and a higher return on investment. The methodology results show that, out of four models testing standard and robust scaling with respect to k-means and hierarchical agglomerative clustering, the k-means standard scalar groups performed best. Groups in said clusters provide interesting data separations between clusters as well as property count approximations resembling best the type and property breakdown in actual data.

VI. Reference

Calka, Beata (2019). “Estimating Residential Property Values on the Basis of Clustering and Geostatistics.” *Geosciences,* vol 9, 143, 1 - 12.

Keskin, Berna and Craig Watkins (2017). “Defining spatial housing submarkets: Exploring the case for expert delineated boundaries.” *Urban Studies Journal Limited* vol 54(6), 1446-1462.

Tomal, Mateusz (2020). “Housing market heterogeneity and cluster formation: evidence from Poland.” *International Journal of Housing Markets and Analysis,* vol 14., no. 5, 1166-1185.