Clustering Analysis in Real Estate Guide

Introduction and Business Understanding:

Redfin Corporation is a real estate company that organizes its listing in a structured data format. One form of analysis that can benefit the company would be clustering, which enables an analyst to discern significant variations among data points, avoiding the distraction of individual anomalies lacking discernible trends. Such analysis is a key component in pattern recognition within datasets. In this context, cluster analysis will be applied to Redfin housing listings to categorize them into distinct groups sharing common traits.

Data Understanding and Preparation:

The data set contains 1,286 listings that are of interest. The listings of focus are in the Dallas area, so other areas were labeled as OTHER and excluded. In preparation of the data, many variables were excluded for only having one value or for being heavily skewed and not of interest. Also, the "Zip or Postal Code" column was incorrectly labeled as continuous so it was changed to nominal. The variables of interest are property type (excluding anything besides "Single Family Residential," "Condo/Co-op," and "Townhouse"), location, beds, baths, square feet, price, and $/ square feet. The variables of interest were analyzed for missing values and recoded appropriately to account for them. Five clusters were created. The data set was then screened for outliers and then any outliers found were excluded along with the missing values.

Analysis:

Clustering:

The real estate data set was analyzed through k-means clustering and hierarchical clustering. K-means clustering was applied to group the data according to five variables: price, beds, bath, square feet, and $/square feet. A local filter was applied as to focus on the Dallas area, results can be seen in Table 1. At the top end, the first cluster captures the essence of luxury with its staggering average prices and expansive living spaces, including numerous bedrooms and bathrooms. The second cluster still signifies a high-value market but with less extravagance than the first. Clusters three and four show a notable dip in both price and size, indicating mid-tier properties. Finally, the fifth cluster represents the most economical options, with the smallest average square footage and the lowest cost per square foot, suggesting entry-level or budget-friendly housing options.

Hierarchical clustering with the use of the Ward methods was used to obtain five clusters, as seen in Table 2. Cluster 1 has the highest count of properties, suggesting these are common in the dataset, with moderate prices, sizes, and the lowest price per square foot, which could indicate a standard housing segment. Cluster 2 offers slightly smaller and less expensive homes. In contrast, Cluster 3 shows a leap in both price and size, moving towards a more upscale market. Cluster 4 has a significant increase in average price and square footage, hinting at luxury properties. Finally, Cluster 5, with the fewest properties, showcases exceptionally high prices and large spaces, representing the most exclusive and expansive properties in the dataset. A constellation plot was then created using the hierarchical plot, as seen in Figure 1. It visually represents the relationships between the different clusters. The dense clusters suggest groupings of properties with similar characteristics, while the lines connecting them may represent hierarchical relationships or similarities between the clusters. For example, the close proximity of some clusters might indicate slight differences between property types or features, whereas clusters that are farther apart might signify more distinct differences in property characteristics. This plot aids in visualizing the multi-dimensional data in a two-dimensional space, providing insights into how the clusters are related to each other.

The geographic distribution of property listings from the K-Means clustering is visualized in Figure 2, revealing the spatial relationships and market segmentation within the Dallas area. The five different colors represent the clusters identified in the K-Means clustering table, each corresponding to groupings of properties with similar characteristics. The map shows a diverse spread of these clusters throughout the city, indicating how property features correlate with geographical location. For instance, more expensive properties might be concentrated in specific neighborhoods, while more affordable options are distributed across wider areas. This spatial analysis is crucial for understanding real estate dynamics and can guide targeted marketing and investment strategies.

Hierarchical clustering was then utilized again for all variables previously analyzed in Table 2 along with property type and location, as seen in Table 3. This clustering mirrors the patterns seen in Table 3, aligning larger, more expensive properties with a lower count, and more modest, affordable homes with a higher count.

Lastly, a crosstab of cluster versus property type was created and organized in Table 4. The crosstab aligns clusters with property types, revealing distinct patterns in housing categorization. Single Family Residential homes are predominant across all clusters, with particularly high concentrations in Clusters 2 and 5. Cluster 3 is notable for its significant proportion of Condo/Co-op properties, which could suggest a clustering strategy that captures urban living spaces. Townhouses are the least represented across clusters but have a slightly higher presence in Cluster 1. The diversity in property types within each cluster points to varying lifestyle preferences and market demands within the real estate landscape.

Conclusion:

In conclusion, Redfin's application of k-means and hierarchical clustering to Dallas area listings has segmented the market effectively, with each method offering unique insights. K-means clustering provided clear market segments based on property attributes, though it may oversimplify by imposing spherical clusters. Hierarchical clustering offered nuanced groupings that reflect a natural order, advantageous for identifying subtle market hierarchies but potentially complex in interpretation. The constellation plot enriched this analysis by visualizing data relationships, yet may present challenges in deciphering complex structures. Together, these techniques provide a comprehensive market overview, highlighting segmentation useful for targeted strategies, despite some methodological limitations.

Table 1: K Means Clustering

Cluster	PRICE	BEDS	BATHS	SQUARE FEET	$/SQUARE FEET
1	19,200,000	7	11	17,597	1,097
2	4,585,540	5	6	8,449	532
3	365,991	2	2	1,464	246
4	3,293,000	3	3	3,113	1,049
5	713,380	4	3	2,769	248

Table 2: Hierarchical Clustering

Cluster	Count	PRICE	BEDS	BATHS	SQUARE FEET	$/SQUARE FEET
1	579	453,170	3	2	1,927	233
2	279	343,300	2	2	1,295	256
3	171	1,690,475	4	4	4,596	336
4	30	3,830,667	3	3	3,589	1,053
5	17	12,676,053	6	8	13,106	960

Figure 1: Constellation Plot

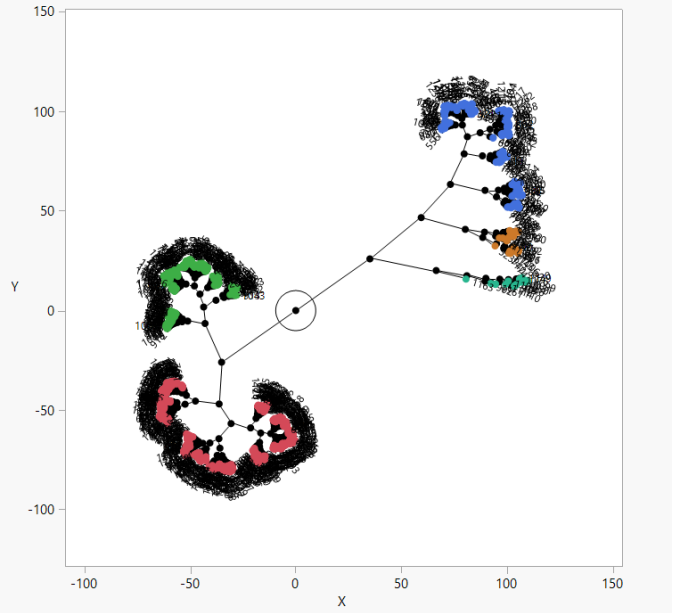

Figure 2: Five Geographic Clusters using K-Means Clustering

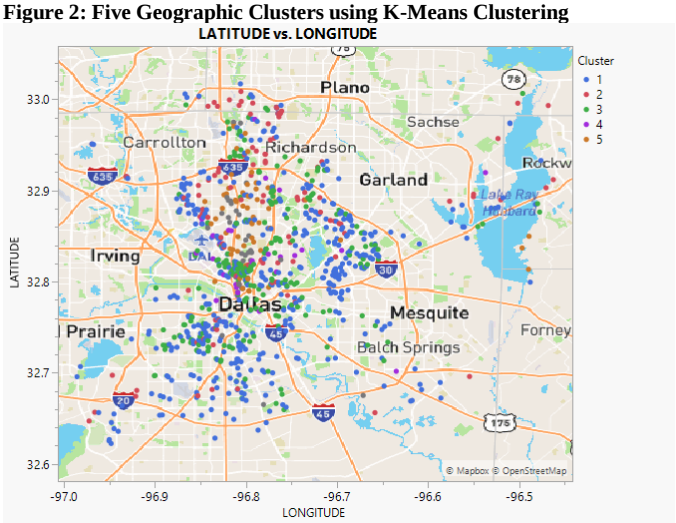

Table 3: Hierarchical Clustering with Location and Property Type

Cluster	Count	PROPERTY TYPE	PRICE	BEDS	BATHS	LOCATION	SQUARE FEET	$/SQUARE FEET
1	600	6	453,096	3	2	84,977	1,971	227
2	121	6	943,470	4	4	113,370	3,541	265
3	288	4	373,829	2	2	91,258	1,245	284
4	60	6	2,965,843	5	6	92,551	6,248	482
5	9	2	4,110,889	3	3	103,338	3,788	1,095

Table 4: Crosstab of Cluster versus Property Type

	PROPERTY TYPE		
	Condo/Co-op	Single Family Residential	Townhouse
Cluster	Row %	Row %	Row %
1	6.10%	86.02%	7.87%
2	0.74%	97.06%	2.21%
3	61.87%	31.10%	7.02%
4	35.00%	62.50%	2.50%
5	18.75%	76.56%	4.69%