Melbourne Housing Market Segmentation: A Cluster-Based Analysis

Student: Yunzhen Wu

MSDS411 Unsupervised Learning Methods

Instructor: Dr. Jamie Riggs

Feb. 9, 2025

Abstract

This study applies clustering techniques to analyze the Melbourne housing market and segment properties based on price, size, and location attributes. After preprocessing the data from the Melbourne housing market in 2016, K-means and hierarchical clustering were implemented. The analysis revealed four distinct clusters of properties: affordable suburban homes, centrally located expensive properties, suburban homes with balanced attributes, and luxury or outlier properties. The clustering solutions were somewhat consistent with the predefined housing types, while also revealing better market segmentation. In addition, geographic mapping emphasized the impact of location on housing segmentation. The results show that housing characteristics, price, and location have a strong influence on clustering patterns, with economic level determining whether suburban clustering occurs and premium properties concentrated in high-demand areas. These insights provide valuable guidance for property valuation, market segmentation, and investment strategies.

1. Introduction

In today's competitive real estate market, understanding different types of properties can bring huge benefits. Identifying potential housing clusters can guide targeted marketing strategies, allocate sales areas more effectively, or improve the properties' valuation. At the same time, both owners and investors could gain a comprehensive understanding via better housing clusters. To meet these needs, this report performs a cluster analysis on the Melbourne Housing Market Dataset for January 2016. Through unsupervised learning methods, we are able to find structure in the data without relying on predefined property segmentation. The results will highlight meaningful groupings of houses that can enlighten more targeted management and investment strategies.

2. Literature Review

Bourassa, Hoesli, and Peng (2003) showed how real estate valuation models can be improved by defining submarkets to reduce pricing errors and acknowledge location differences. Goodman and Thibodeau (2007) extended this research, showing how neighborhood-level clustering captured through spatial econometric models can enhance the interpretability of submarkets and inform urban development decisions. The advent of advanced visualization techniques such as t-SNE (Van der Maaten & Hinton, 2008) has allowed researchers to identify clusters in high-dimensional real estate data. Building on this, this study integrates data normalization and clustering techniques to gain a deeper understanding of the Melbourne real estate market.

3. Methods

3.1 Exploratory Data Analysis (EDA)

(1) Dataset Overview

The dataset used for this analysis is from **Melbourne Housing Market in January 2016**. It contains real estate transaction records, including property characteristics, location details, and price information. The dataset originally consisted of **34,857 rows and 21 columns**, but only a subset of relevant attributes was selected for analysis.

The primary features considered include: **Rooms**, **Price**, **Distance** (Distance from the CBD), **Bedroom2** (Number of bedrooms), **Bathroom**, **Car** (Number of car parking spaces), **Landsize**, **BuildingArea**, and **YearBuilt**.

(2) EDA Insights

- **Missing Data:** Initially, 60.58% of entries in BuildingArea and 55.39% in YearBuilt were missing. These rows were removed for consistency before clustering.
- **Correlation Analysis:** A heatmap was generated to examine relationships between variables. Price showed strong correlations with BuildingArea (**0.51**), indicating that larger properties tend to be more expensive (Appendix A: Correlation Heatmap).
- **Outlier Detection:** Boxplots were used to detect extreme values. Price outliers included properties priced above **$8 million**, and some landsizes exceeded 3**0,000 sqm**, which could distort clustering (Appendix B: Boxplots of Key Features).

3.2 Data Preparation

- **Missing Data Handling:** All rows containing missing values in key variables were removed, reducing the dataset from **34,857** to **8,887** rows.

- **Outlier Treatment:** Columns are processed as outliers using common sense, such as **Rooms**, **Cars** should be greater than **0**, and **YearBuilt** should be greater than **1800**.
- **Log Transformations:** Applied to Price, Landsize, BuildingArea, and Distance to address skewness and mitigate the impact of outliers.
- **Standardization:** All numerical features were standardized using Z-score normalization to ensure equal weighting in clustering.

After data preparation, the dataset was reduced to **7,771 observations** and **9 columns**.

3.3 Clustering Techniques

To segment the housing market effectively, we implemented two clustering techniques: **K-Means Clustering** and **Hierarchical Agglomerative Clustering**.

(1) K-Means Clustering

- The optimal number of clusters was determined using the **Elbow Method**, which analyzed the sum of squared errors (SSE) for different values of K (Appendix C: Elbow Method Plot). According to the plot, **four clusters** were selected.

(2) Hierarchical Agglomerative Clustering

- A **dendrogram** was used to determine the number of clusters by examining the largest gap between hierarchical merges (Appendix D: Dendrogram for Hierarchical Clustering). The **Ward linkage** method was selected to minimize intra-cluster variance. According to the diagram, **four clusters** were selected.
- This method was ultimately chosen as the final clustering solution due to a higher **Silhouette Score** compared to K-Means.

3.4 Clustering Evaluation

To interpret the clustering results, various techniques were employed to evaluate:

- **Silhouette Score:** K-Means was **0.3319**, while Hierarchical Clustering was **0.3374**, indicating slightly better-defined clusters for the latter method.
- **t-SNE (t-distributed Stochastic Neighbor Embedding):** the plot revealed distinct groupings of properties, confirming that the four clusters were meaningful and not arbitrary (Appendix E: t-SNE Visualization of Clusters).
- **Geographical Mapping:** Properties were plotted based on latitude and longitude to observe spatial distributions of housing clusters. It showed that clusters were **strongly influenced by location**, with distinct patterns emerging across different suburbs (Appendix F: Geospatial Distribution of Clusters).

4. Results

4.1 Cluster Characteristics

The clusters revealed distinct groupings of properties based on their characteristics:

- **Cluster 0 (3431 counts):** Houses located in **outer suburban areas** with **moderate prices**, **moderate building areas**, **moderate land sizes**, and **longer distances from the CBD** (median 13.8 km, largest among clusters). Properties in this cluster exhibit balanced attributes in terms of size and price.
- **Cluster 1 (1715 counts): Affordable properties** located in **suburban areas** with **smaller land sizes** (median 259 sqm, smallest among clusters), **smaller building areas** (median 87 sqm), and **mid-range proximity to the CBD**. This cluster primarily consists of compact, budget-friendly homes.

- **Cluster 2 (1416 counts): High-priced properties** found in **central locations** (median distance 5.6 km, smallest among clusters) with **moderate building areas** and **smaller land sizes**. These properties are typically situated in prime urban areas where space is limited but property values are high.
- **Cluster 3 (1209 counts):** Outliers or unique properties, including **luxury homes** (median price $1.65 million, highest among clusters), **largest building areas** (median 228 sqm), and **largest land sizes** (median 650 sqm). These properties tend to be in well-established or high-end neighborhoods, often offering larger living spaces with premium amenities.

4.2 Comparison with Housing Type and Location Analysis

To evaluate the clustering solution, we compared the generated clusters with existing housing types in the dataset. The clustering solution aligned well with the predefined categories of houses (e.g., **houses, units, townhouses**), but also revealed further segmentation within these types based on size and price.

Additionally, when analyzing cluster distribution by suburb and geographical location, we observed:

- Cluster 2 was more prevalent in **central and high-demand** areas.
- Cluster 0 and 1 were primarily located in the **suburbs**, reflecting affordability constraints.
- Cluster 3 consisted of anomalies, including either **luxury homes** or **highly underpriced properties**.

The mapping visualization further confirmed the geographical distinctiveness of the clusters (Appendix F: Geospatial Cluster Map).

5. Conclusion

This study uses clustering techniques to segment the Melbourne housing market, revealing key patterns in property attributes, pricing, and location. The results show that housing characteristics, price, and location play a key role in market segmentation. Affordable housing tends to cluster in suburban areas, while high-priced housing is concentrated in prime locations. In addition, luxury estates and some outlier properties form unique clusters that may attract the interest of investors and policymakers.

For the industry, real estate companies can use these clusters to develop targeted marketing strategies and optimize the tasks of real estate agents. Investors can use cluster segmentation to identify undervalued properties or high-growth areas, thereby providing effective investment strategies. Urban planners can analyze the spatial distribution of clusters to inform housing policies, zoning decisions, and infrastructure development.

Future research can further refine housing classification, such as incorporating variables such as community facilities and transportation accessibility. At the same time, trying other clustering methods (such as DBSCAN or Gaussian mixture models) may also provide deeper insights into property segmentation. This study demonstrates the effectiveness of unsupervised learning in housing market analysis and highlights its contribution to advancing data-driven decision-making in the real estate industry.

References

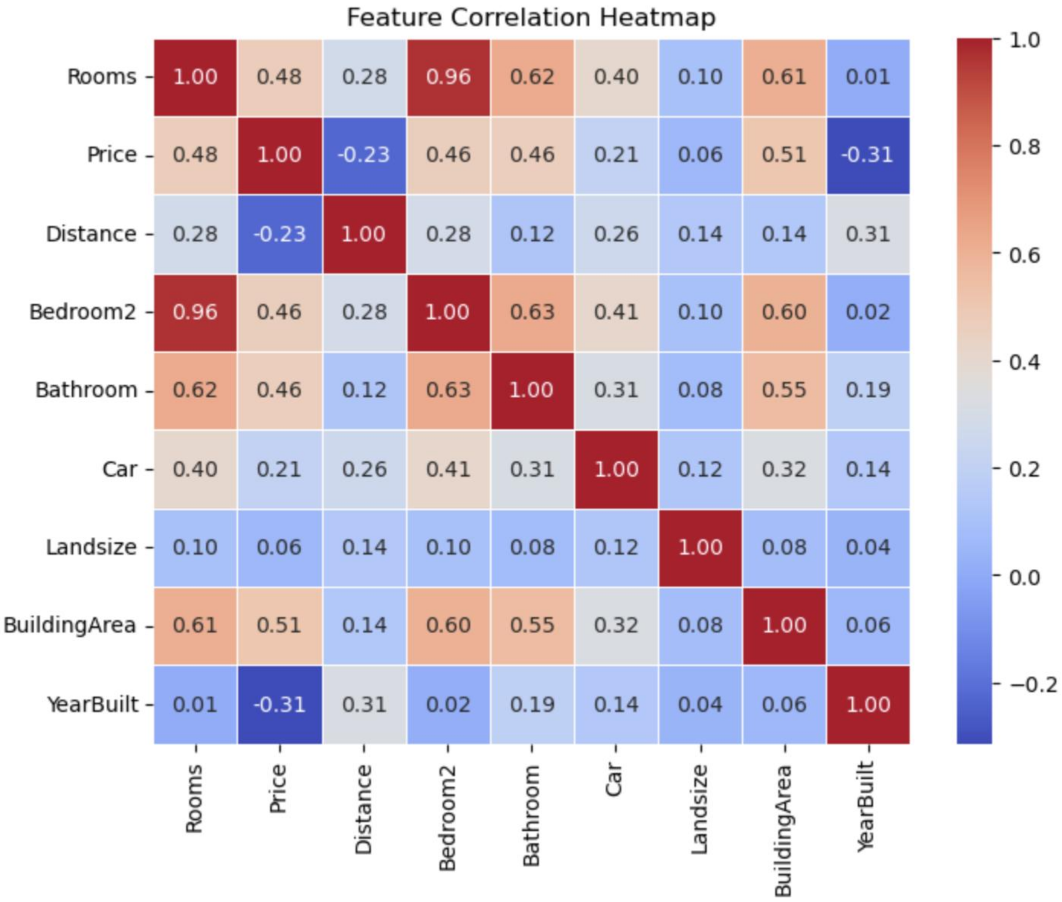
Bourassa, Steven C, Martin, Hoesli, and Vincent S, Peng. "Do housing submarkets really matter?". *Journal of Housing Economics 12*, no.1 (2003): 12–28.

Goodman, Allen C, and Thomas G, Thibodeau. "The spatial proximity of metropolitan area housing submarkets". *Real Estate Economics 35*, no.2 (2007): 209–232.

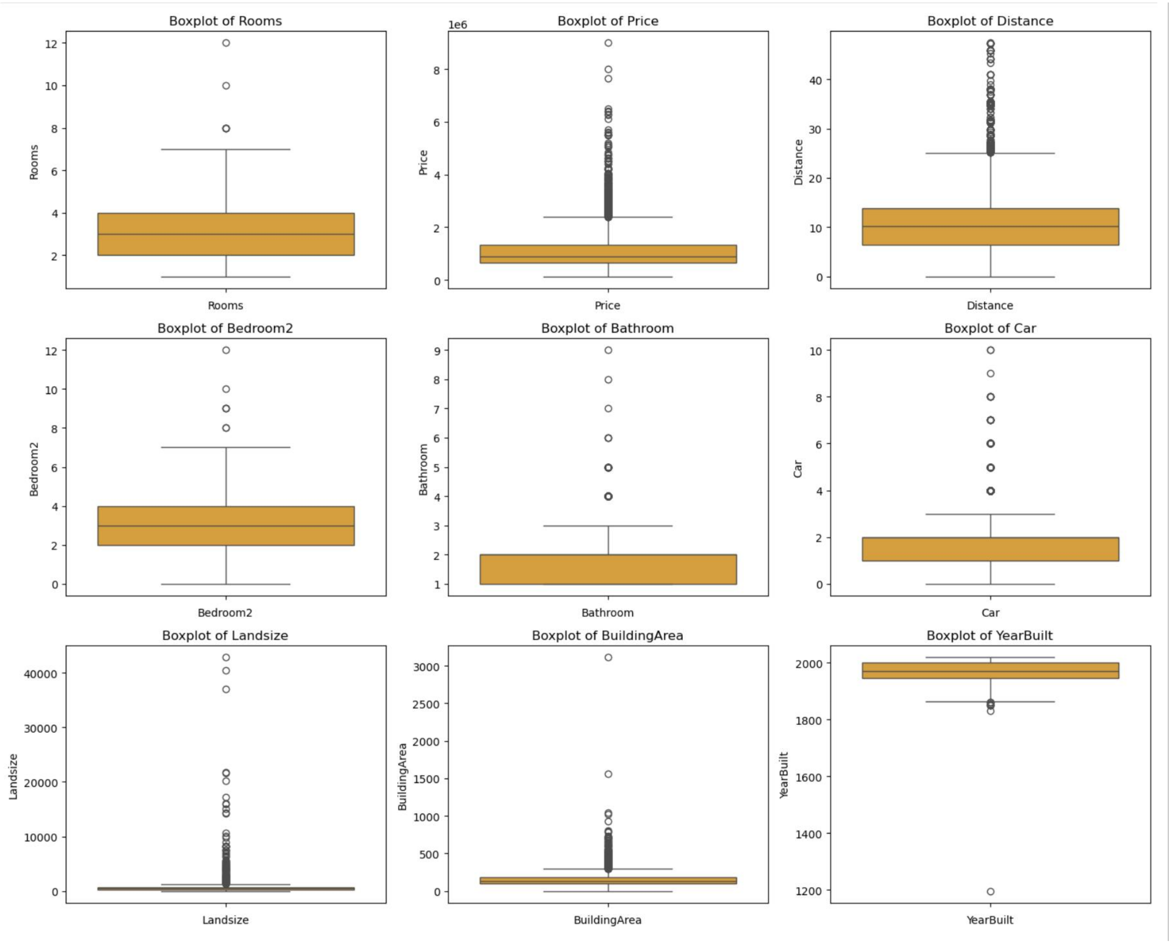
Maaten, Laurens, and Geoffrey, Hinton. "Visualizing data using t-SNE". *Journal of machine learning research 9*, no.11 (2008).

Appendix

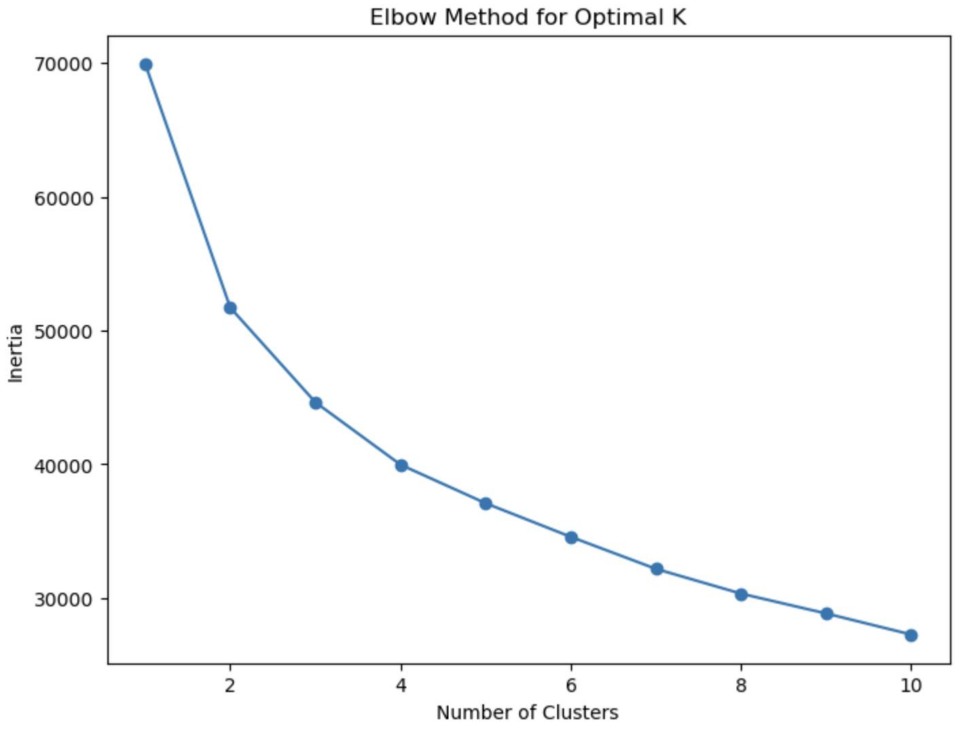
Appendix A: Correlation Heatmap – Displays the relationships between key variables such as price, building area, and land size, highlighting strong correlations that influence property clustering.



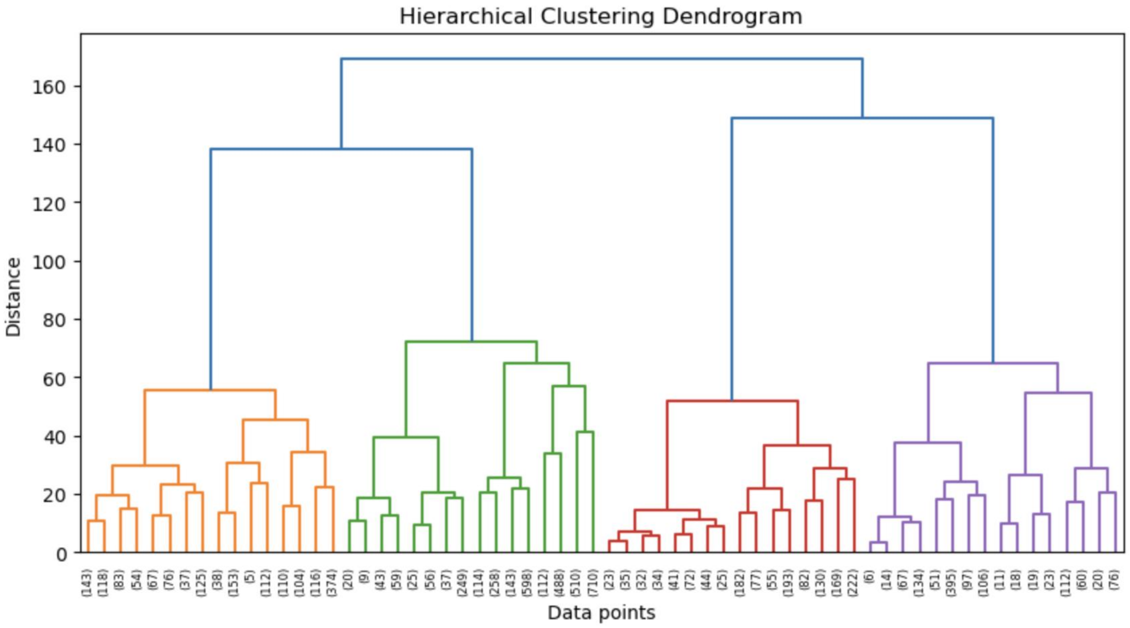
Appendix B: Boxplots of Key Features – Visualizes outliers in price, building area, and land size to assess potential distortions in clustering.



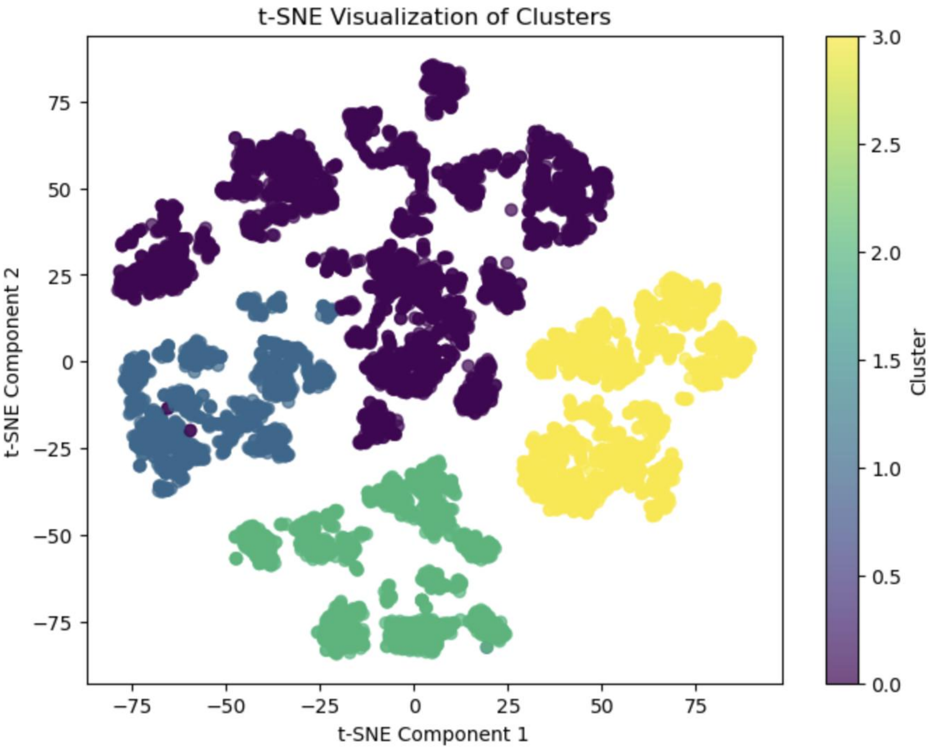
Appendix C: Elbow Method Plot – Demonstrates the selection process for the optimal number of clusters in K-Means clustering.



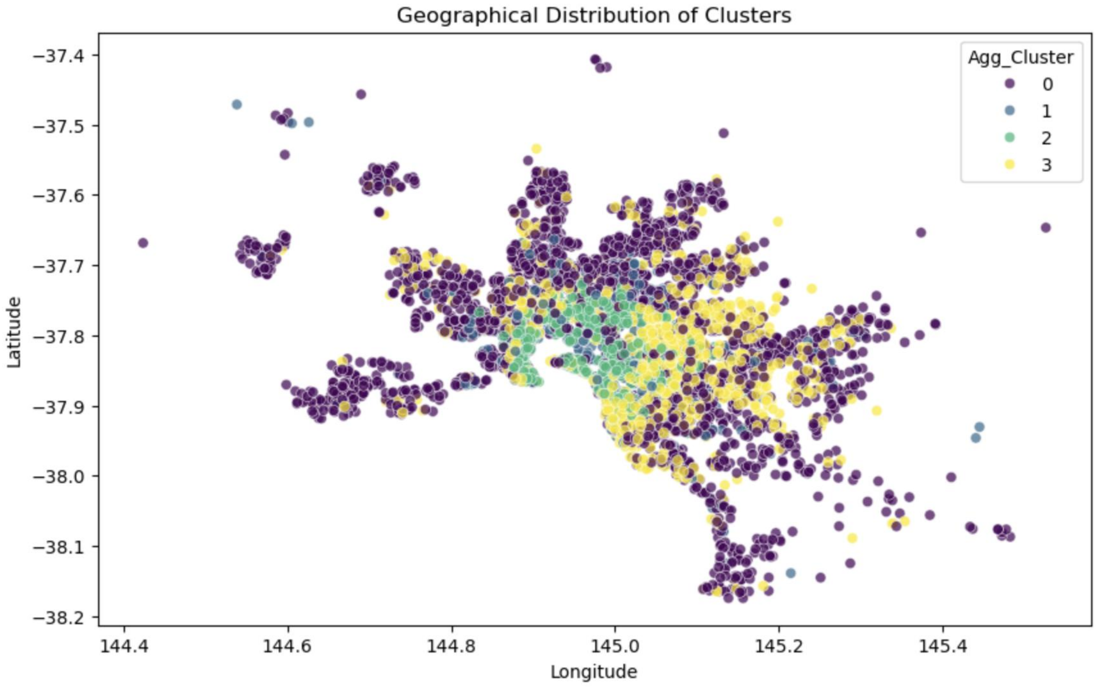
Appendix D: Dendrogram for Hierarchical Clustering – Depicts hierarchical relationships between housing properties and guides the determination of cluster numbers.



Appendix E: t-SNE Visualization of Clusters – Provides a two-dimensional representation of high-dimensional property attributes, confirming distinct group formations.



Appendix F: Geospatial Distribution of Clusters – Maps housing clusters based on latitude and longitude, demonstrating the influence of location on property segmentation.


Geographical Distribution of Clusters
Agg_Cluster
● 0
● 1
● 2
● 3
Latitude
Longitude