TRANSPORT AND TELECOMMUNICATION INSTITUTE

DATA MINING

Report Computer Practice Nr: 2

Cluster Analysis

Riga 2024

Contents

Introduction

This report provides a detailed examination of data mining techniques, specifically focusing on cluster analysis, to study California's housing market. The primary goal is to apply hierarchical and K-means clustering methods to uncover patterns and segments within the housing data. The study begins with importing and examining the dataset, followed by more detailed analyses using different clustering techniques.

Data Import and Initial Analysis

Importing Libraries

The analysis begins by importing several essential Python libraries, each serving a specific role in the data analysis workflow:

Pandas (pandas) and NumPy (numpy) are used for data manipulation and numerical calculations.

Matplotlib (matplotlib.pyplot) and Seaborn (seaborn) provide extensive capabilities for data visualization.

Statsmodels (statsmodels.api and statsmodels.formula.api) offers classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests.

SciPy (scipy.cluster.hierarchy) is used for hierarchical clustering algorithms. Scikit-learn (sklearn.cluster, sklearn.metrics, sklearn.preprocessing) provides tools for data mining and data analysis, including clustering and scaling techniques.

Kneed (kneed.KneeLocator) is utilized to detect the knee point on curves, which is particularly useful for determining the optimal number of clusters in K-means clustering.

Collections (collections.Counter) helps with counting the occurrences of items, useful in tallying up labels or results.

Loading the Dataset

The dataset, sourced from a CSV file named 'ST83519_california_housing.csv', is loaded into a DataFrame called housing_df using Pandas' read_csv function. This step initializes the dataset for subsequent operations.

Preliminary Data Inspection

Following the loading of the data, two key functions are executed to gain an initial understanding of the dataset's structure:

housing_df.info(): This method is called to output a concise summary of the DataFrame, providing a quick overview of the total number of entries, the existence of null values in each column, and the data types of each column.

housing_df.shape: This attribute of the DataFrame returns a tuple representing the dimensionality of the DataFrame, showing the total number of rows and columns.

Data Overview

The dataset contains the following columns with their respective data types:

housing_median_age: int64 (median age of housing in the area)

total_rooms: int64 (total number of rooms in the area)

total_bedrooms: int64 (total number of bedrooms in the area)

population: int64 (total population in the area)

households: int64 (total number of households in the area) median_income: float64 (median income of households in the area)

median_house_value: int64 (median house value in the area)

The dataset has 300 entries and no missing values

Statistical Summary and Initial Visualization

To better understand the distribution of key numerical features within the Californian housing dataset, histograms are utilized. Histograms provide a visual summary of the data, allowing for the identification of patterns, anomalies, or deviations from expected behavior. This initial visualization helps in comprehending the spread and central tendencies of the dataset's variables, which is crucial for subsequent analyses.

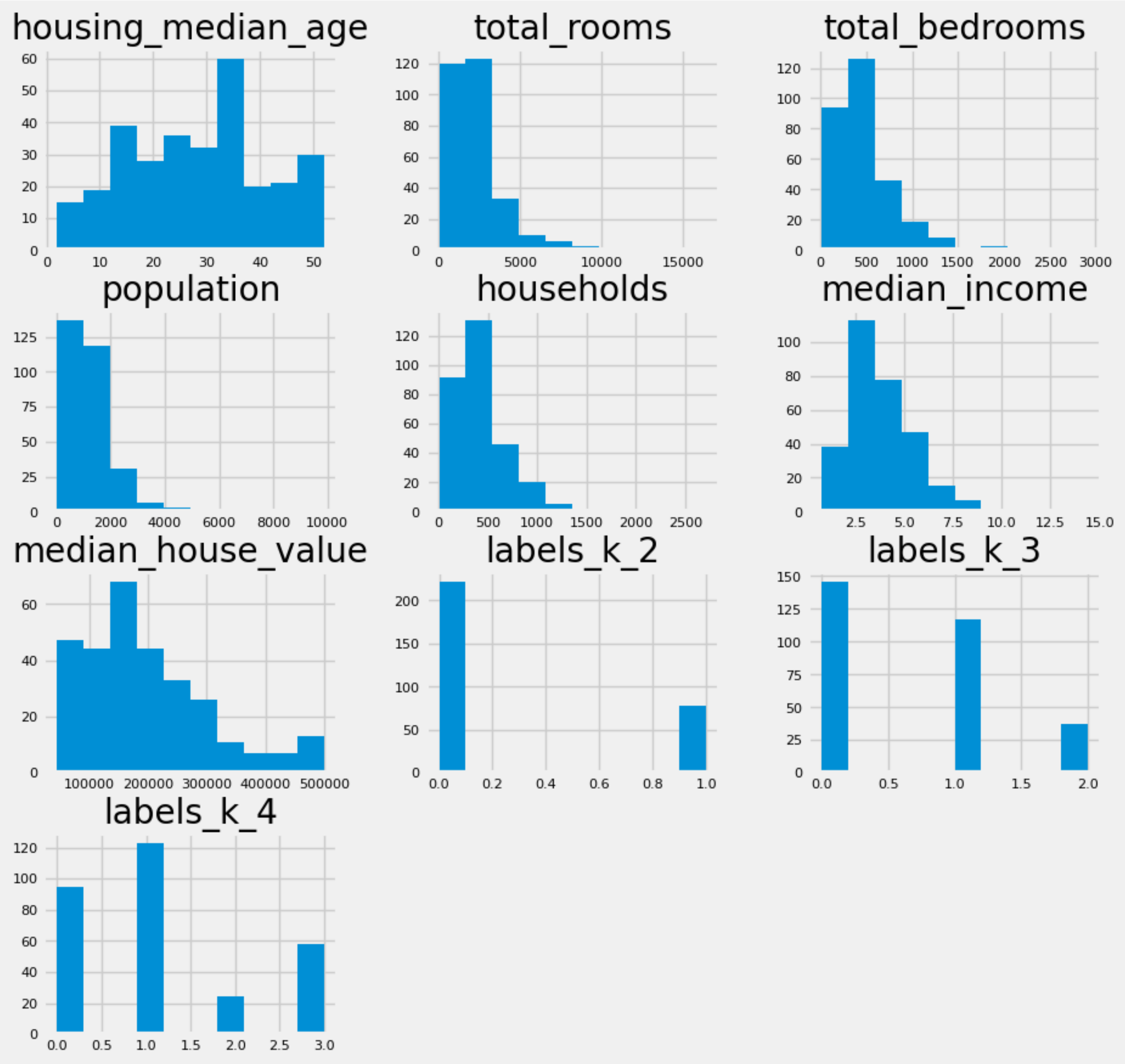


Figure 1. Visualization of Data Distribution.

The generated histograms represent the following features:

Housing Median Age: Most of the housing units fall within a certain age range, indicating typical periods of building booms.

- Housing Median Age: Most of the housing units fall within a certain age range, indicating typical periods of building booms.
- Total Rooms: The distribution is heavily right-skewed, suggesting that most houses have a smaller number of rooms, with few houses having a very large number of rooms.
- Total Bedrooms: Similar to total rooms, this feature is also right-skewed, reflecting a commonality in smaller household sizes.
- Population: This feature shows a right skew, indicating that most blocks have a relatively small population, with a few densely populated blocks.
- Households: Also right-skewed, showing that most blocks contain fewer households, aligning with the patterns seen in population distribution.
- Median Income: The distribution of median income is somewhat less skewed than rooms or population, but it still shows that higher incomes are less common.
- Median House Value: The values are somewhat normally distributed but with a tail on the higher end, suggesting that while most homes are moderately priced, there are some significantly more expensive homes.

These histograms provide a foundational understanding of the data's structure and distribution, which is instrumental in guiding further detailed analysis, such as clustering or predictive modeling.

Hierarchical Clustering Analysis Using Dendrogram

Implementation of Hierarchical Clustering

To explore potential groupings within the Californian housing dataset, hierarchical clustering was applied using various linkage methods.

Each method represents a different strategy for measuring distances between clusters:

- Single Linkage: Distance between nearest neighbors.
- Complete Linkage: Distance between farthest neighbors.
- Average Linkage: Average distance between all members.
- Centroid Linkage: Distance between the centroids of the clusters.
- Ward's Method: Minimization of variance within the clusters.

When evaluating the dendrograms produced by various hierarchical clustering methods using the original, unscaled data, it is important to consider the visibility of clusters, the appropriateness of each method for the specific characteristics of the data. Below, we analyze the results to determine the number of visible clusters for each.

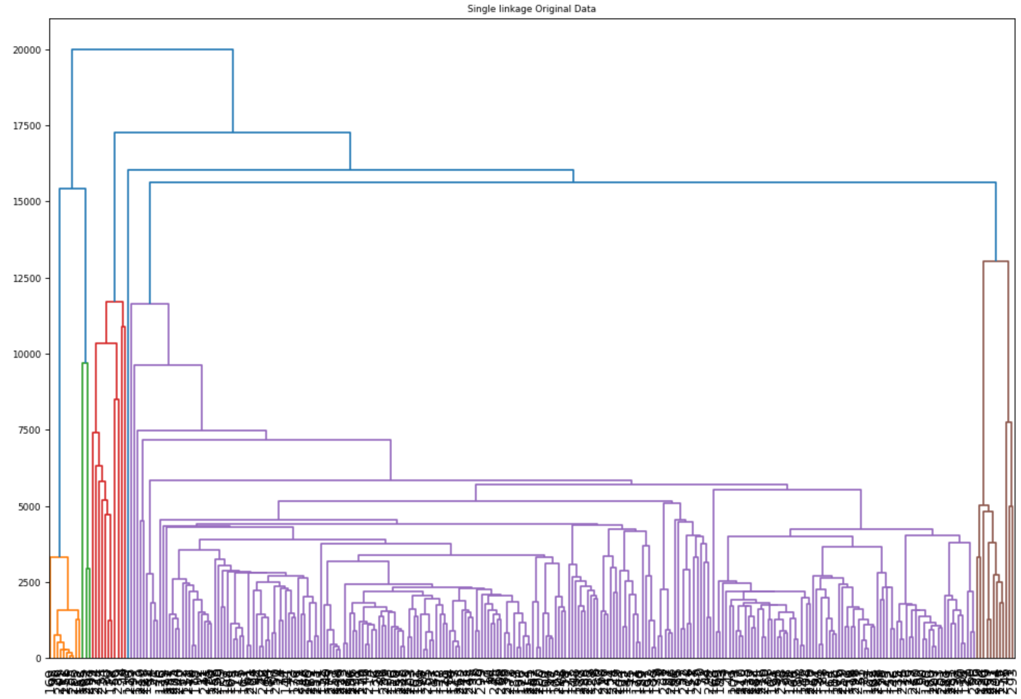


Figure 2. Dendrogram of Single Linkage Clustering on Original Data.

The dendrogram shows a tendency to form long chains, with one or two major clusters and several smaller or outlying groupings evident at high linkage distances. The sensitivity to outliers in single linkage can distort the true cluster structure, often resulting in an overestimation of the number of clusters due to the chaining effect.

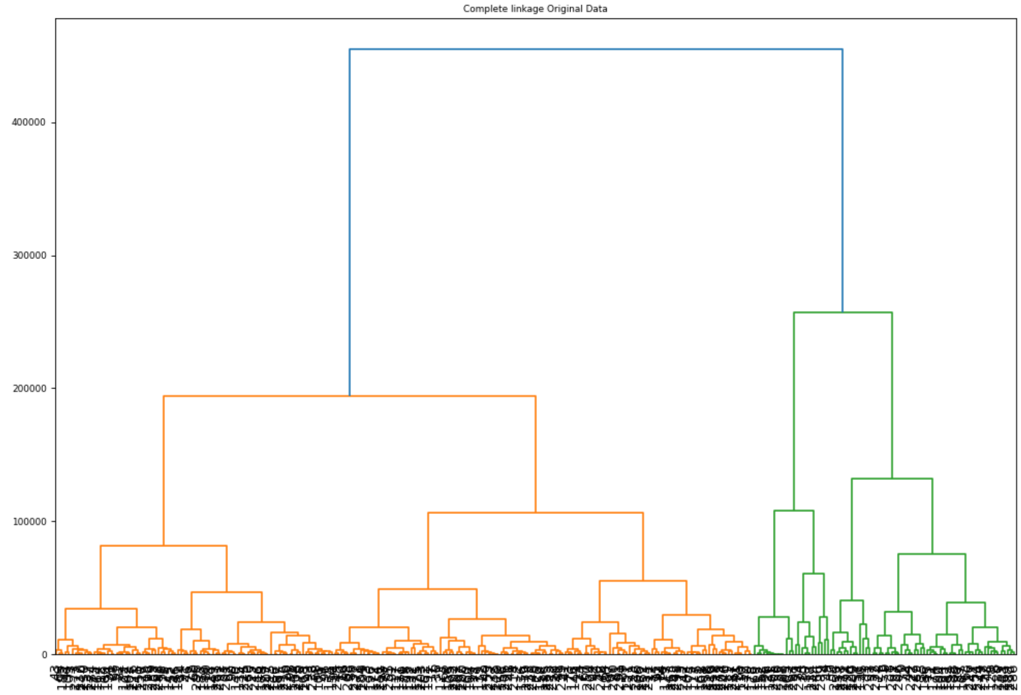


Figure 3. Dendrogram of Complete Linkage Clustering on Original Data.

This method shows a more balanced structure with roughly 2-3 major clusters, where data points are more uniformly distributed within each cluster. Complete linkage, by considering the maximum distance, tends to avoid chaining and thus forms more meaningful, compact clusters.

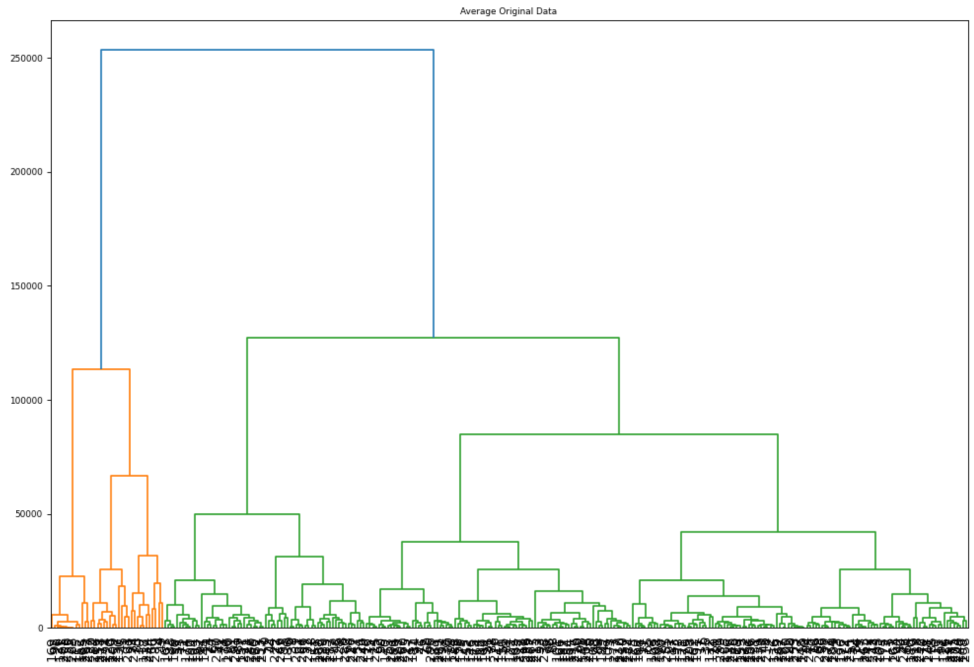


Figure 4. Dendrogram of Average Linkage Clustering on Original Data.

Average linkage suggests about 3-4 clusters, presenting a balanced view between the sensitivity of single linkage and the robustness of complete linkage. It offers a middle ground, which might be more representative of the actual data structure for many datasets, especially those without extreme outliers.

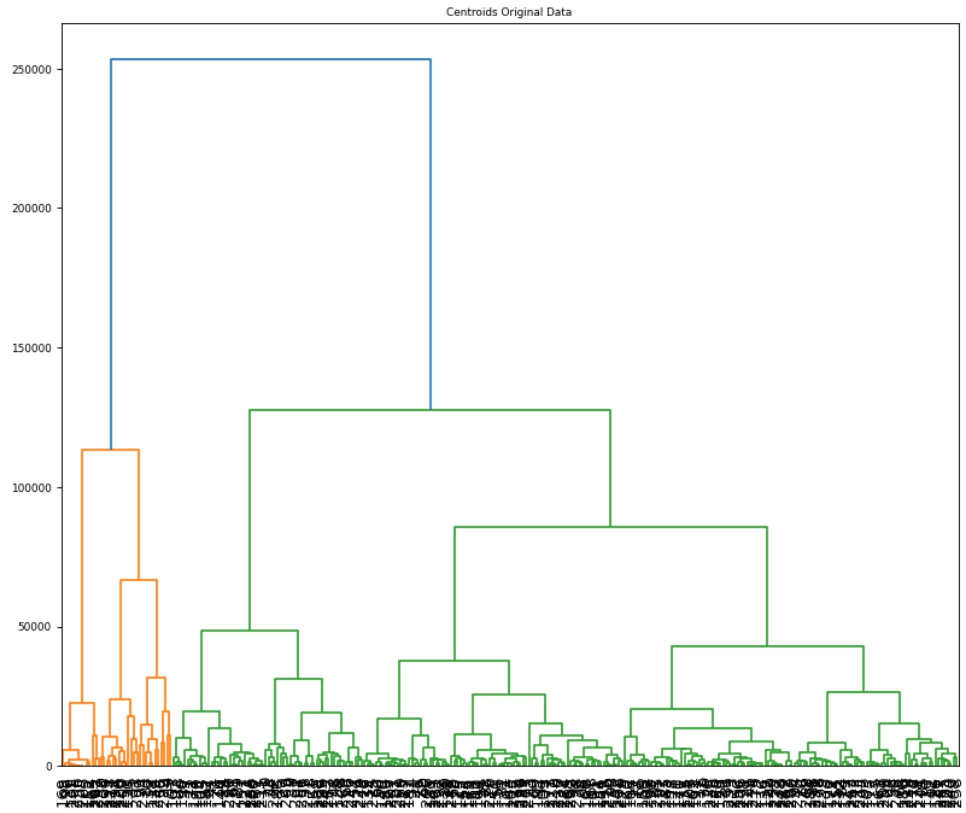


Figure 5. Dendrogram of Centroid Linkage Clustering on Original Data.

The centroid method illustrates about 3-4 clusters, similar to average linkage, with a tendency to merge clusters based on the proximity of their centroids. This method can be influenced by the presence of outliers which shift the centroid significantly, potentially leading to less accurate clustering.

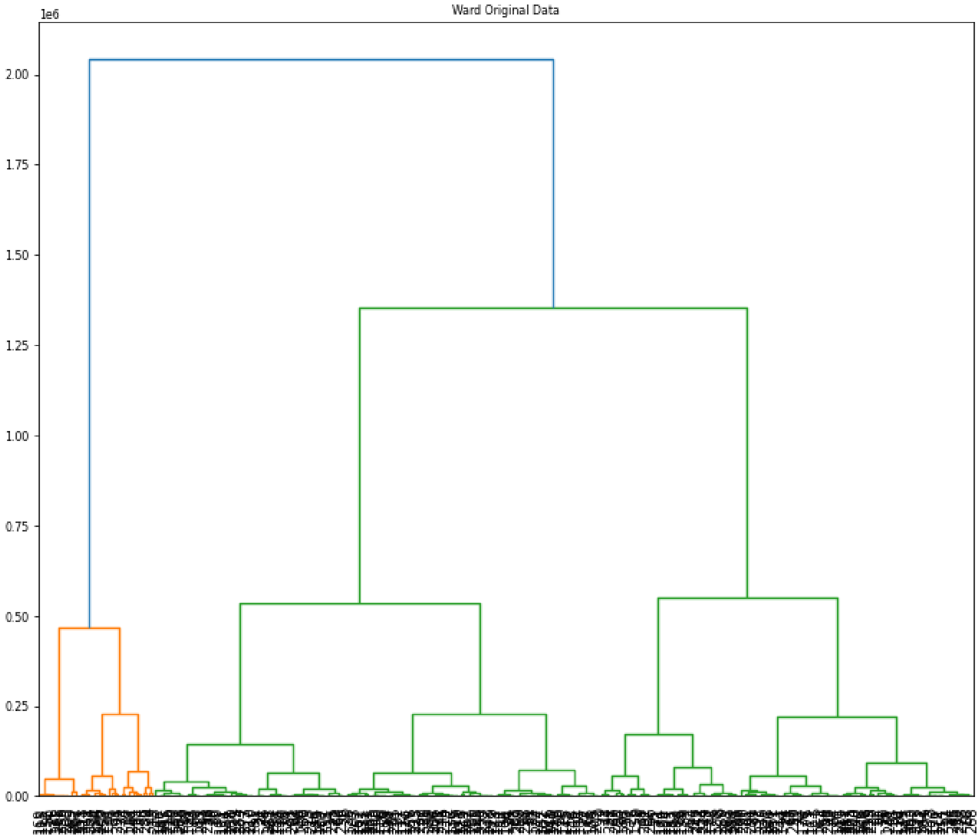


Figure 6. Dendrogram of Ward Linkage Clustering on Original Data.

Ward's method clearly shows about 2-3 major clusters, with a very structured and hierarchical integration of clusters. This method is particularly useful for creating clusters with a similar number of observations and minimizing within-cluster variance, often producing more actionable clusters.

Conclusion

Number of Clusters: Based on the analyzed dendrograms, choosing between 2 to 4 clusters seems reasonable across the different methods, depending on the method and specific data characteristics.

Use of Original Data: Clustering on the original, unscaled data can be influenced by variables with larger scales dominating the distance calculations.

Let's consider normalizing or standardizing data when clusters formed seem biased towards certain high-magnitude variables.

Standardization and Hierarchical Clustering Analysis Using Normalized Data

Standardization Process

Before applying clustering techniques, the data was standardized using the StandardScaler from scikit-learn, which normalizes each feature to have zero mean and unit variance. This standardization ensures that each variable contributes equally to the distance calculations used in clustering, preventing any one variable with a larger scale from dominating.

Clustering Techniques and Dendrograms

After standardizing the data, various hierarchical clustering methods were applied.

Each method provides a different strategy for clustering:

- Single Linkage: Clusters formed by the minimum distance between individual members.
- Complete Linkage: Clusters formed by the maximum distance between individual members.
- Average Linkage: Clusters formed by the average distance between all members.
- Centroid Linkage: Clusters based on the centroids of the clusters.
- Ward's Method: Clusters formed by minimizing the total within-cluster variance.

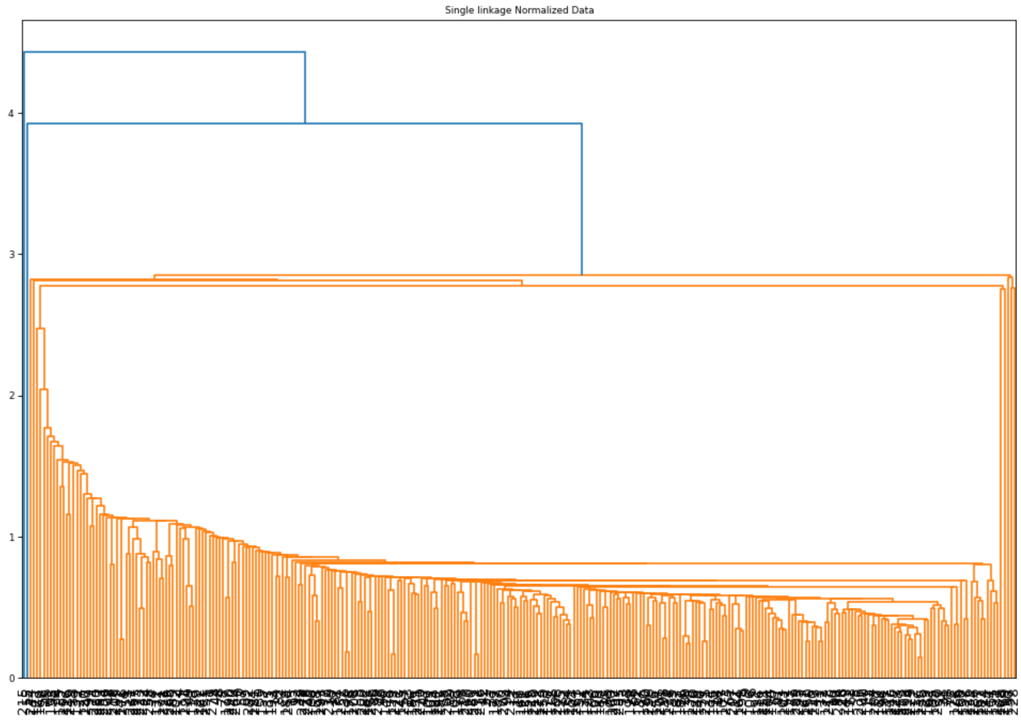


Figure 7. Dendrogram of Single Linkage Clustering on Normalized Data.

Single Linkage: Displays a clear chaining effect typical of this method, where many individual data points form a large cluster. This method shows two main clusters but with a high degree of chaining.

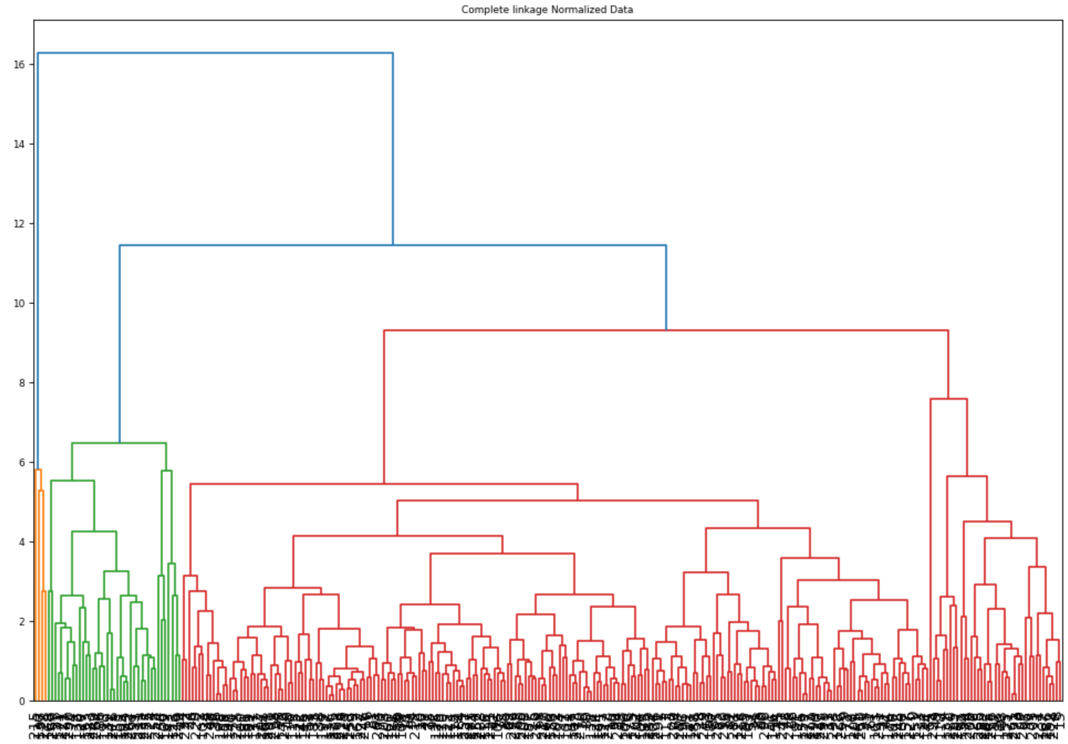


Figure 8. Dendrogram of Complete Linkage Clustering on Normalized Data.

Complete Linkage: Shows a more structured approach with about 3-4 distinct clusters visible. The clusters are well-separated, indicating robust grouping.

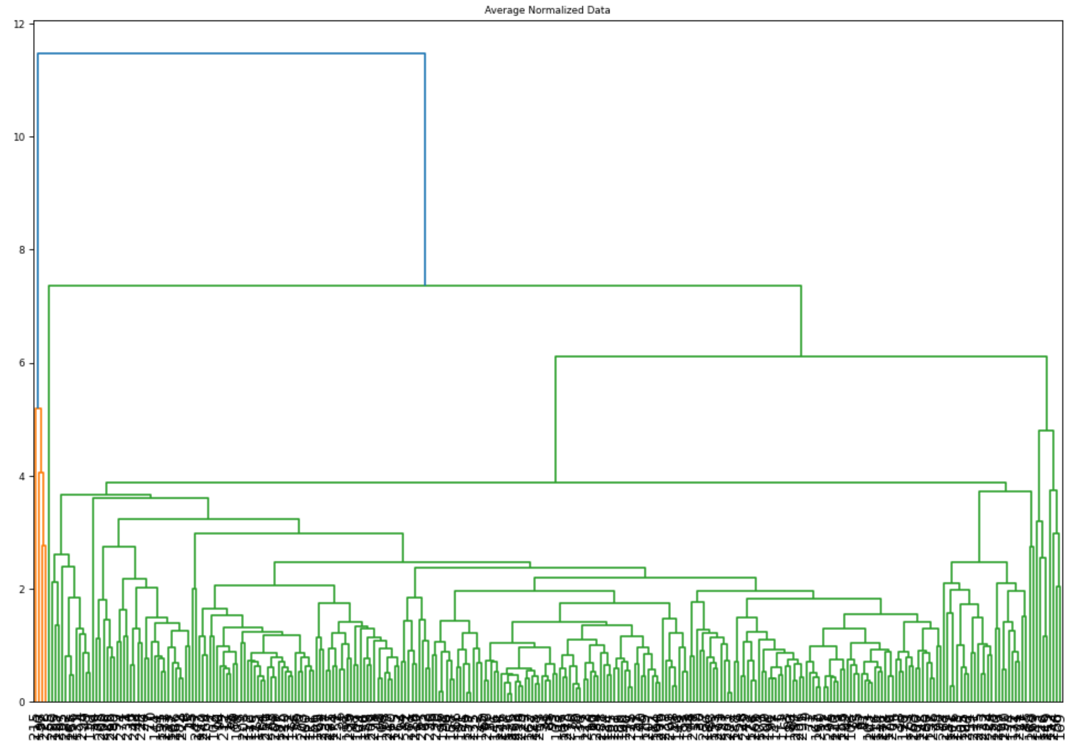


Figure 9. Dendrogram of Average Linkage Clustering on Normalized Data.

Average Linkage: Similar to complete linkage, it shows 3-4 main clusters with good separation and less susceptibility to outliers compared to single linkage.

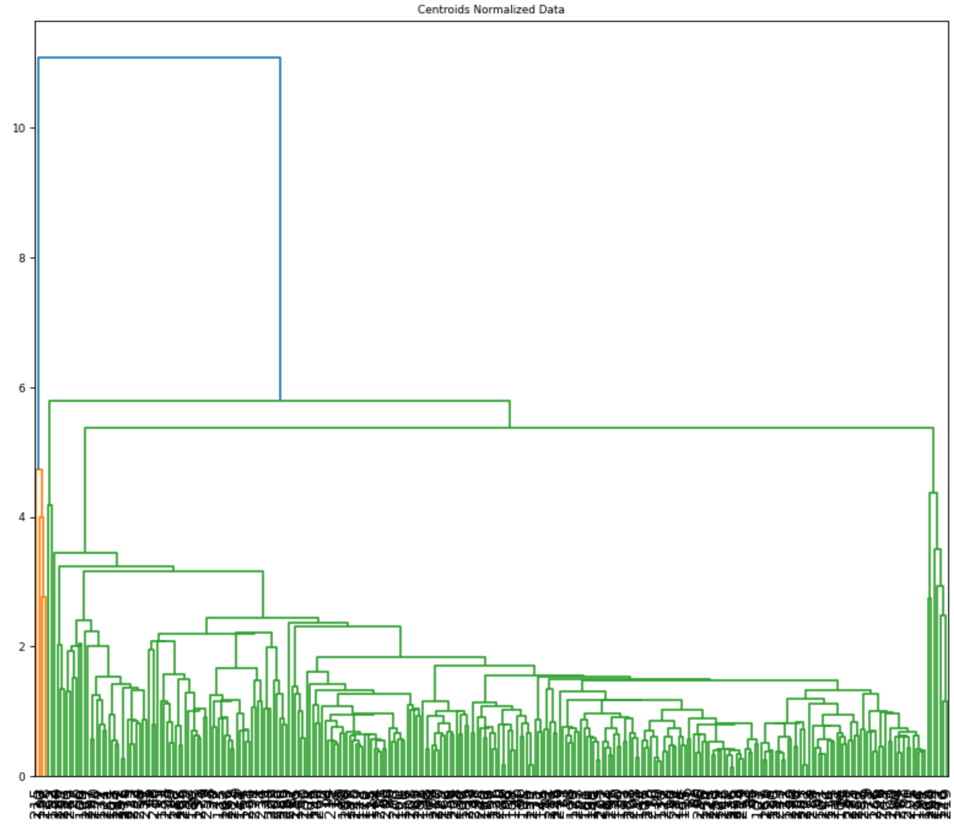


Figure 10. Dendrogram of Centroid Linkage Clustering on Normalized Data.

Centroid Linkage: Similar to average linkage, with 3-4 clusters visible. This method shows clusters that are somewhat more merged at lower heights, suggesting a moderate sensitivity to cluster centroids.

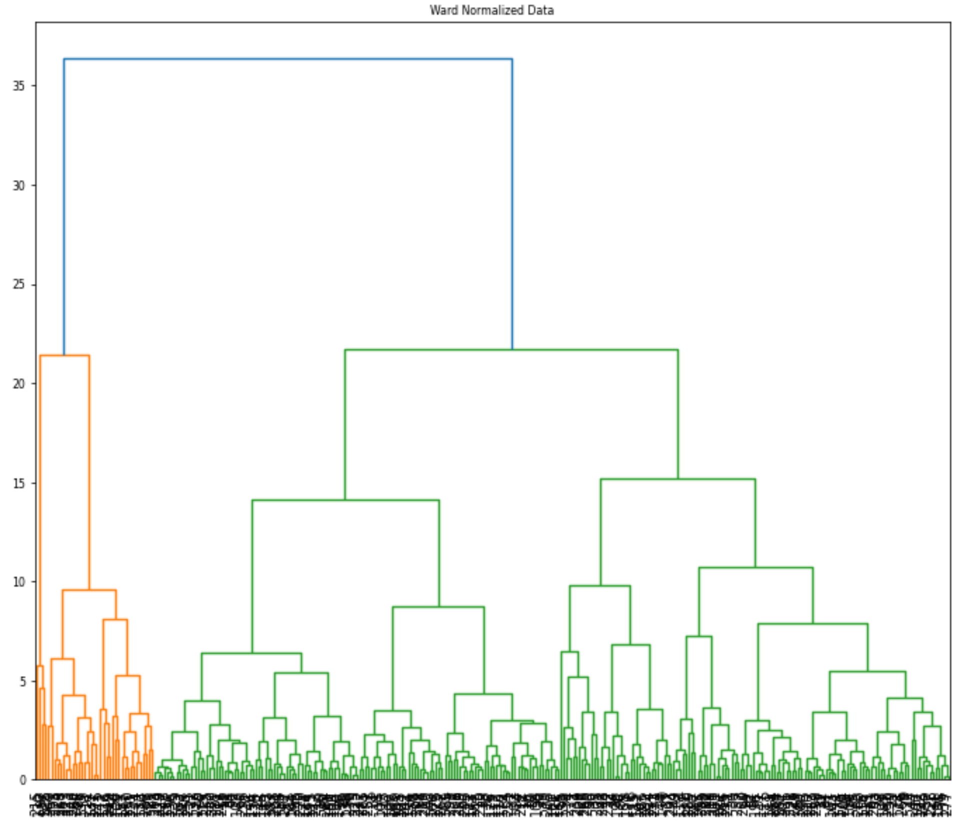


Figure 11. Dendrogram of Ward Linkage Clustering on Normalized Data.

Ward’s Method: Shows a very distinct structure with 3-4 main clusters and clear separation between them. This method effectively minimizes within-cluster variance, creating balanced clusters.

Analysis of K-Means Clustering

Data Description

The dataset was processed using K-Means clustering, both on the original scale and with normalized features, to identify distinct groups or segments within the market.

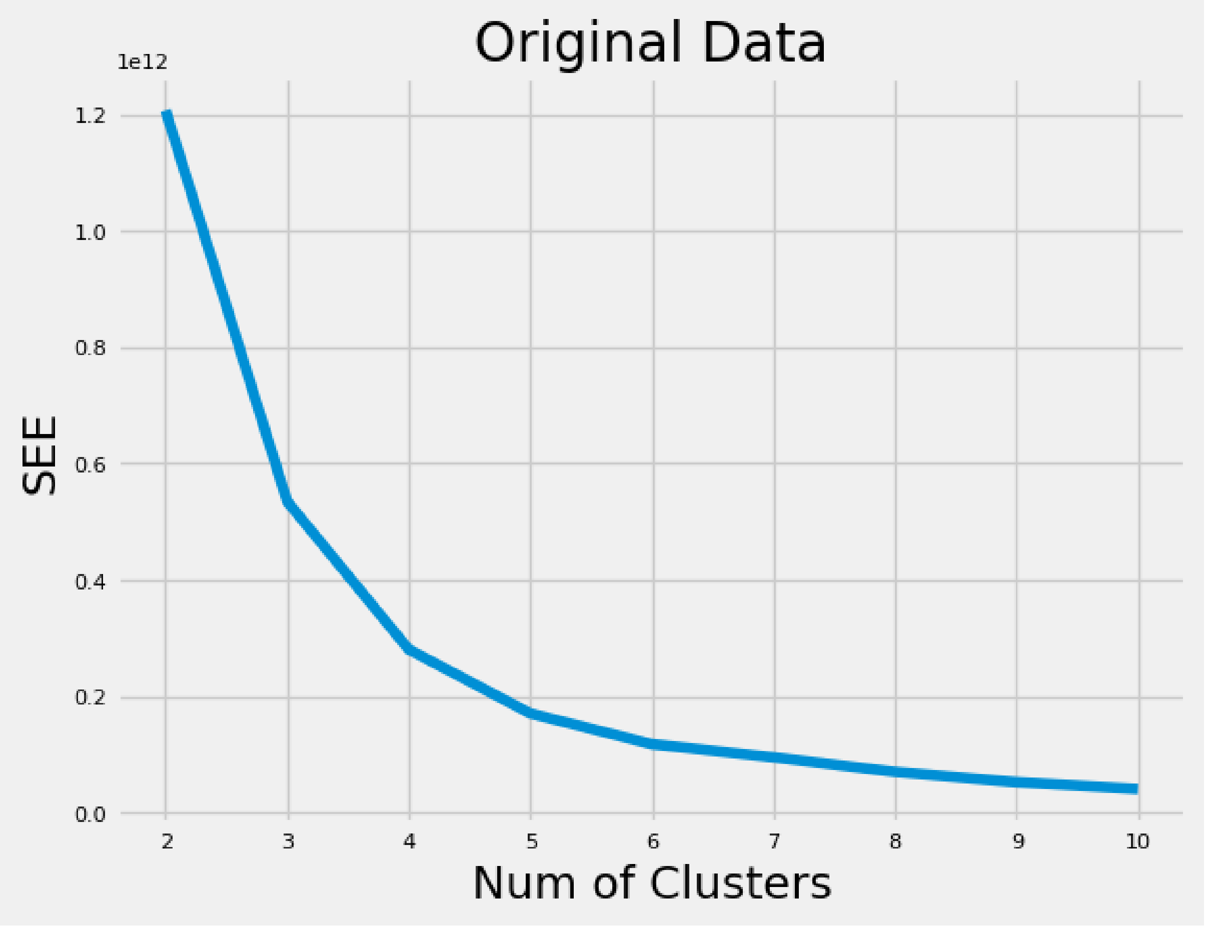
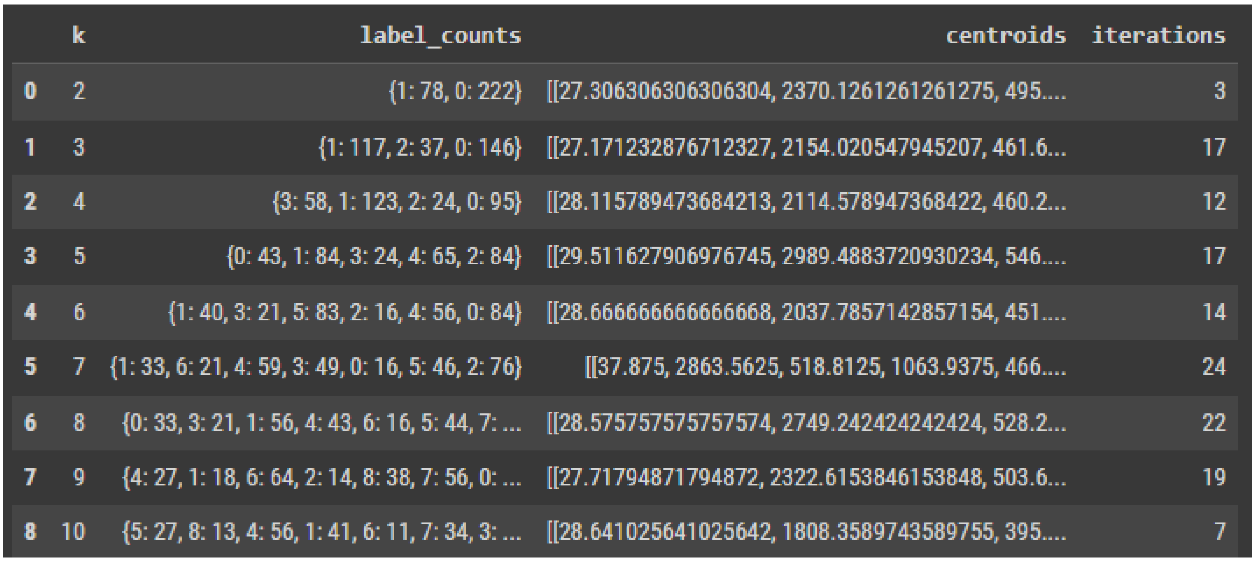


Figure 12. SEE Num of Clusters on Original Data

Original Data: The SSE graph shows a rapid decline as the number of clusters increases from 2 to around 5, suggesting significant improvement in cluster cohesion up to 5 clusters. Beyond this point, the decrease in SSE becomes less pronounced, indicating diminishing returns on increasing the number of clusters.

Table 1. K-Means clustering analysis on Original Data



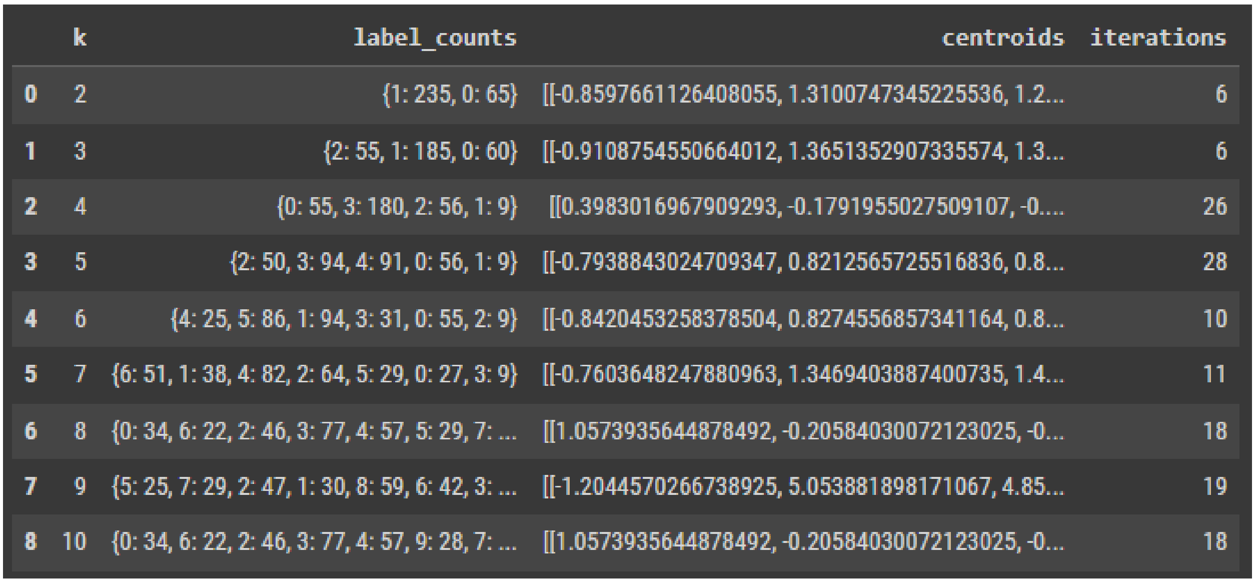
	k	label_counts	centroids	iterations
0	2	{1: 78, 0: 222}	[[27.306306306306304, 2370.1261261261275, 495....	3
1	3	{1: 117, 2: 37, 0: 146}	[[27.171232876712327, 2154.020547945207, 461.6...	17
2	4	{3: 58, 1: 123, 2: 24, 0: 95}	[[28.115789473684213, 2114.578947368422, 460.2...	12
3	5	{0: 43, 1: 84, 3: 24, 4: 65, 2: 84}	[[29.511627906976745, 2989.4883720930234, 546....	17
4	6	{1: 40, 3: 21, 5: 83, 2: 16, 4: 56, 0: 84}	[[28.666666666666668, 2037.7857142857154, 451....	14
5	7	{1: 33, 6: 21, 4: 59, 3: 49, 0: 16, 5: 46, 2: 76}	[[37.875, 2863.5625, 518.8125, 1063.9375, 466....	24
6	8	{0: 33, 3: 21, 1: 56, 4: 43, 6: 16, 5: 44, 7: ...	[[28.575757575757574, 2749.242424242424, 528.2...	22
7	9	{4: 27, 1: 18, 6: 64, 2: 14, 8: 38, 7: 56, 0: ...	[[27.71794871794872, 2322.6153846153848, 503.6...	19
8	10	{5: 27, 8: 13, 4: 56, 1: 41, 6: 11, 7: 34, 3: ...	[[28.641025641025642, 1808.3589743589755, 395....	7

The table provided details the results from a K-Means clustering analysis on a dataset, across a range of cluster sizes from 2 to 10.

Key Components: k (Number of Clusters): Indicates how many clusters the data was divided into for each run of the algorithm, ranging from 2 to 10. Label Counts: Shows the distribution of data points across the clusters for each k. For example, at k=5, the distribution is {0: 43, 1: 84, 2: 34, 3: 24, 4: 65}, suggesting that cluster 1 is the largest and cluster 3 is the smallest.

Centroids: These are arrays representing the mean (or centroid) values of the features for each cluster. The centroids help characterize the typical attributes of the data points grouped within each cluster. For instance, the centroid for k=2 is [27.306306306306304, 2370.1261261261275, 495...], which might include average values for features such as price, area, etc.

Table 2. K-Means clustering analysis on scaled version..



	k	label_counts	centroids	iterations
0	2	{1: 235, 0: 65}	[[-0.8597661126408055, 1.3100747345225536, 1.2...	6
1	3	{2: 55, 1: 185, 0: 60}	[[-0.9108754550664012, 1.3651352907335574, 1.3...	6
2	4	{0: 55, 3: 180, 2: 56, 1: 9}	[[0.3983016967909293, -0.1791955027509107, -0....	26
3	5	{2: 50, 3: 94, 4: 91, 0: 56, 1: 9}	[[-0.7938843024709347, 0.8212565725516836, 0.8...	28
4	6	{4: 25, 5: 86, 1: 94, 3: 31, 0: 55, 2: 9}	[[-0.8420453258378504, 0.8274556857341164, 0.8...	10
5	7	{6: 51, 1: 38, 4: 82, 2: 64, 5: 29, 0: 27, 3: 9}	[[-0.7603648247880963, 1.3469403887400735, 1.4...	11
6	8	{0: 34, 6: 22, 2: 46, 3: 77, 4: 57, 5: 29, 7: ...	[[1.0573935644878492, -0.20584030072123025, -0...	18
7	9	{5: 25, 7: 29, 2: 47, 1: 30, 8: 59, 6: 42, 3: ...	[[-1.2044570266738925, 5.053881898171067, 4.85...	19
8	10	{0: 34, 6: 22, 2: 46, 3: 77, 4: 57, 9: 28, 7: ...	[[1.0573935644878492, -0.20584030072123025, -0...	18

The table provided offers detailed results from a K-Means clustering analysis applied to a scaled version of dataset, with clusters ranging from 2 to 10.

Key Components: k (Number of Clusters): Specifies the number of clusters used in each analysis, ranging from 2 to 10. Label Counts: Displays the distribution of data points across the clusters for each k. For example, at k=3, the distribution is {2: 55, 1: 185, 0: 60}, which shows that most data points are in cluster 1, with fewer in clusters 0 and 2.

Centroids: These are the mean values of the scaled features for the data points in each cluster, giving insight into the characteristic profile of each cluster. For instance, the centroid for k=2 is [0.8597661126408055, 1.3100747344525536, 1.2...], indicating the average scaled values of features in this cluster.

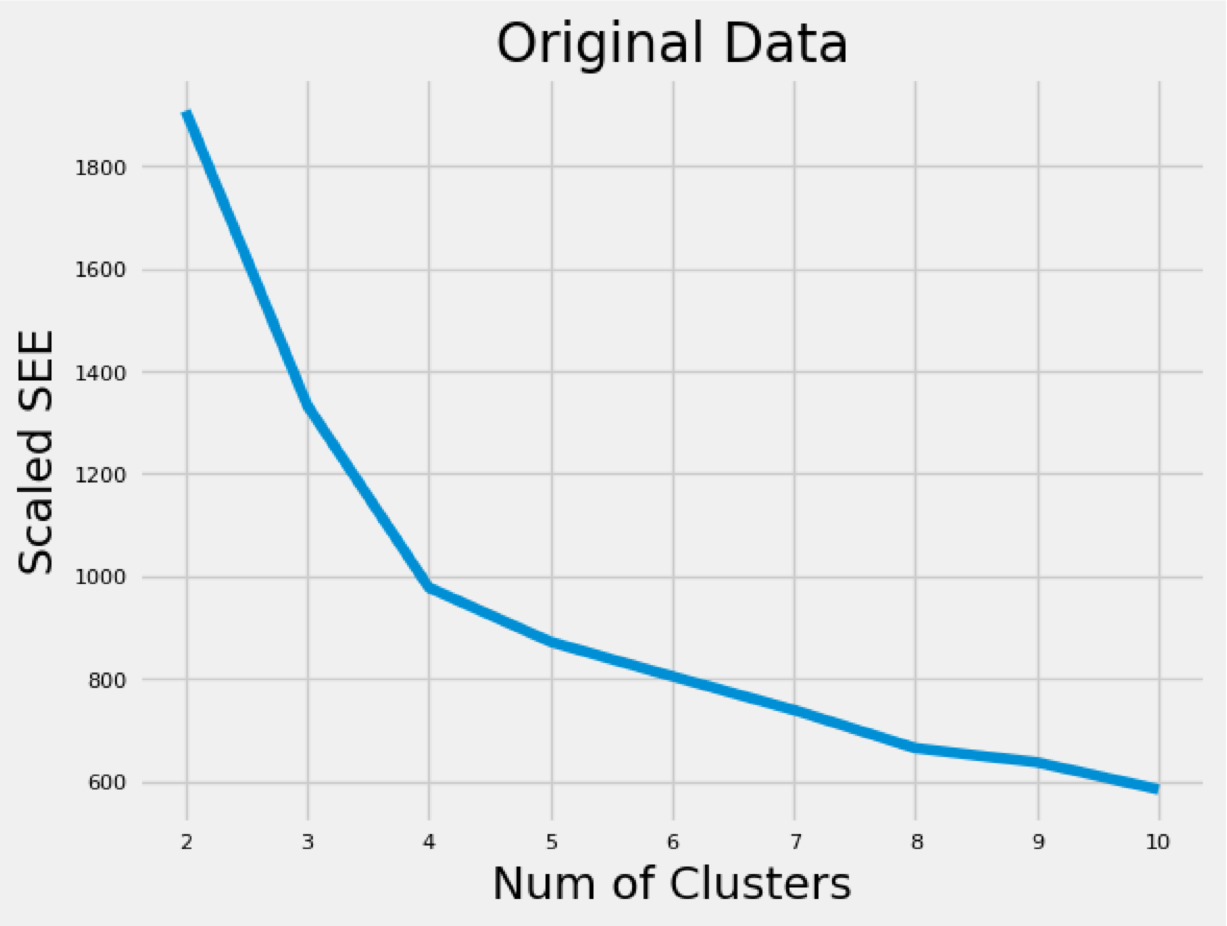


Figure 13. Scaled SEE Num of Clusters on Original Data.

Scaled Data: Similar to the original data, the SSE declines sharply up to about 4 clusters and then plateaus, suggesting that 4 clusters may be an optimal choice for the scaled data in terms of error minimization.

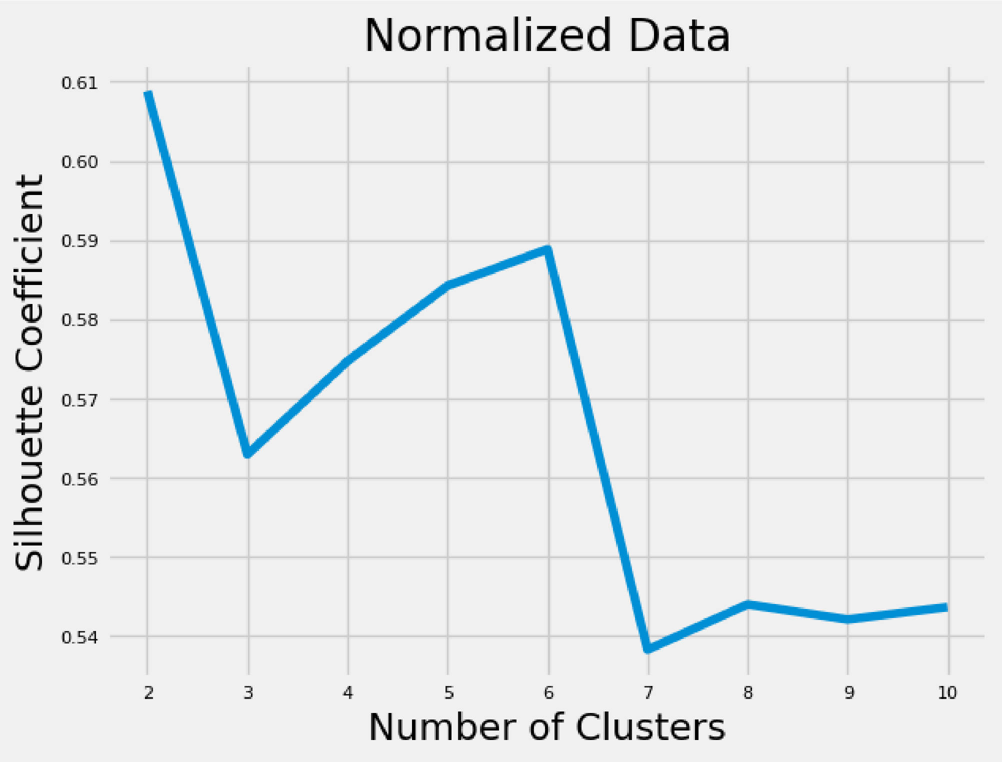


Figure 14. Silhouette Coefficient Number of Clusters on Normalized Data.

Normalized Data: The silhouette scores peak at 5 clusters, decreasing significantly as more clusters are added. This indicates that 5 clusters provide the best balance between within-cluster similarity and between-cluster differences in the normalized dataset.

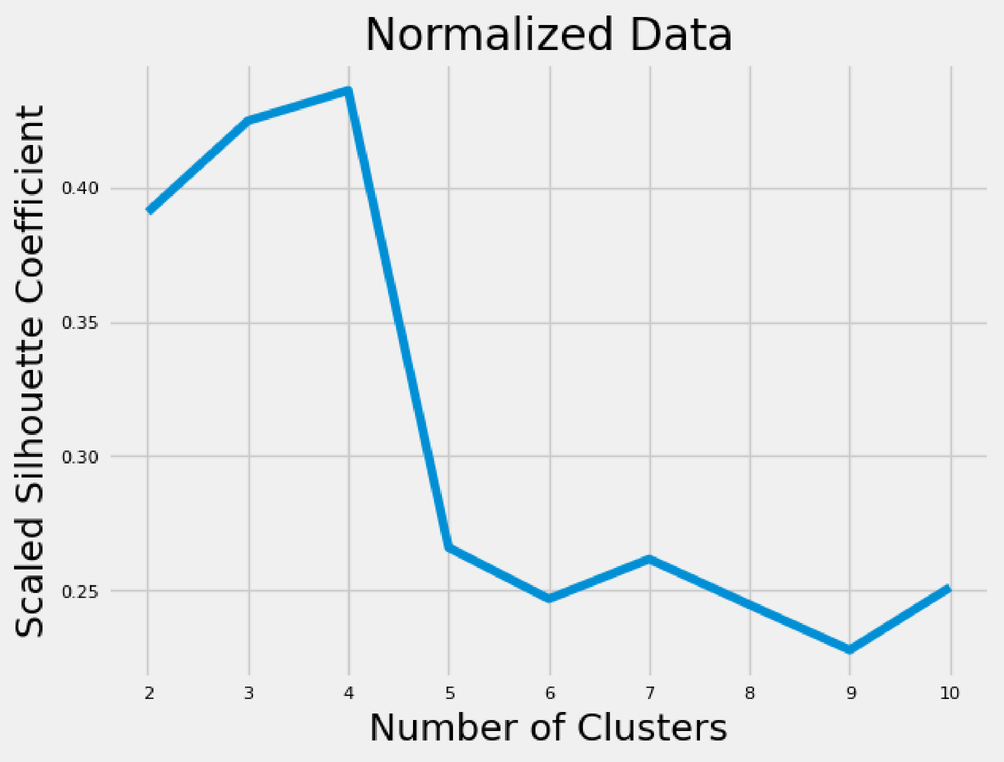


Figure 15. Scaled Silhouette Coefficient Number of Clusters on Normalized Data.

Scaled Data: The silhouette scores are highest at 4 clusters, suggesting that clustering the scaled data into 4 groups maximizes the distinction between clusters while maintaining homogeneity within them.

Conclusions

Optimal Number of Clusters: Based on the elbow method observed in the SSE plots and the peak values in the Silhouette Coefficients, 5 clusters are recommended for the original data and 4 for the scaled data.

ANOVA Analysis

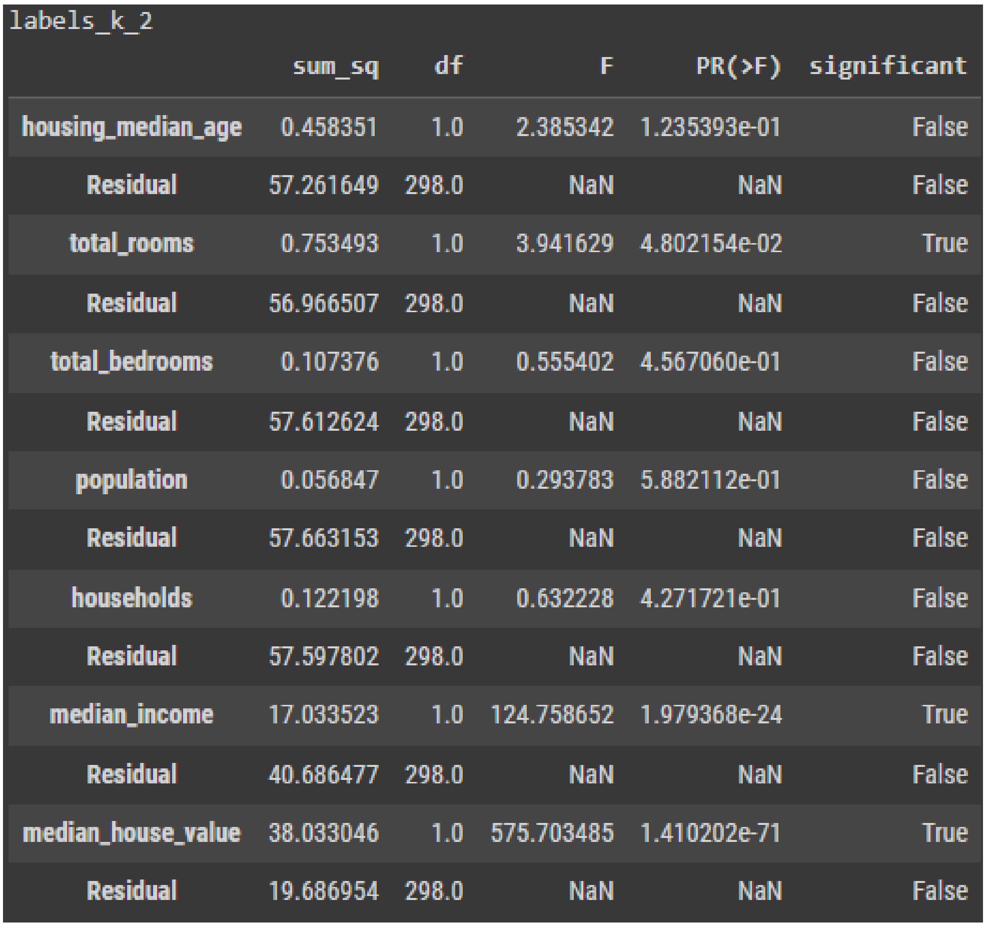
ANOVA analysis provides a detailed analysis of the clustering characteristics. The objective was to identify which variables significantly differentiate clusters formed from both unscaled and scaled data.

The dataset comprises various housing metrics such as median age, total rooms, bedrooms, population, households, median income, and house value. Clusters were generated using K-Means clustering, and ANOVA was used to determine if the means of these variables are significantly different across clusters.

Two versions of the data were analyzed:

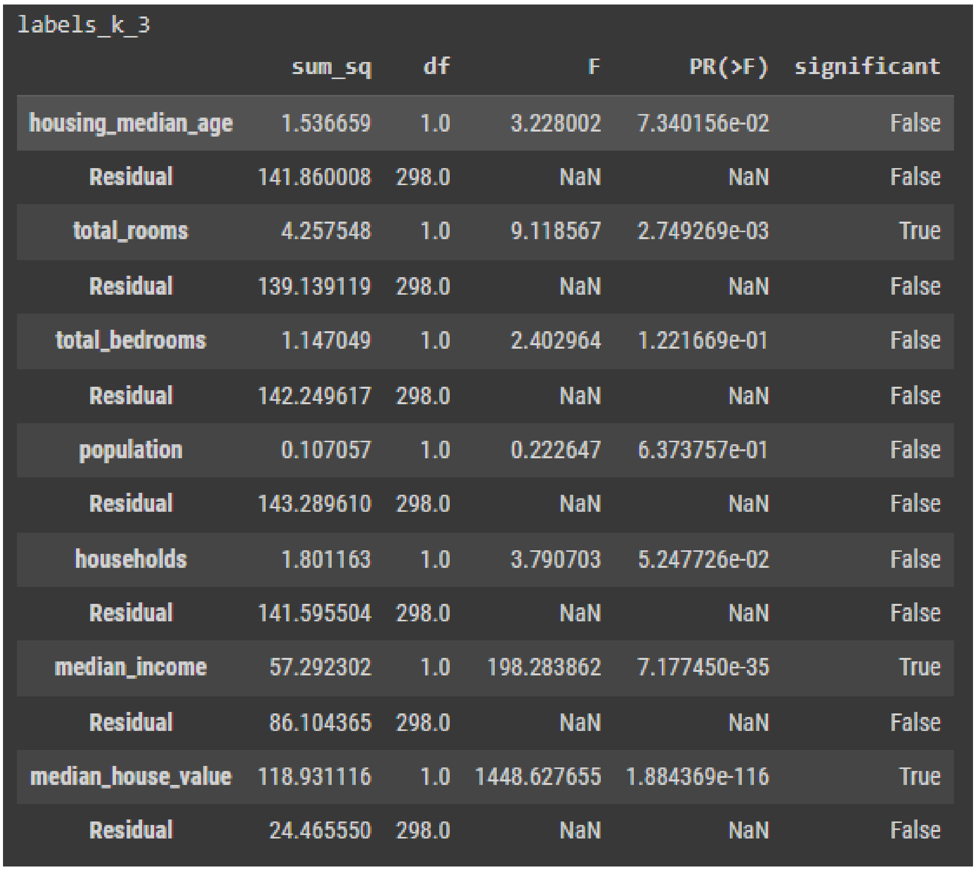
1. Unscaled Data (housing_df): Original data without any scaling.
2. Scaled Data (housing_scaled_df): Data where features were scaled to have zero mean and unit variance.

Table 3. ANOVA Table Cluster 2 Unscaled Data.



labels_k_2	sum_sq	df	F	PR(>F)	significant
housing_median_age	0.458351	1.0	2.385342	1.235393e-01	False
Residual	57.261649	298.0	NaN	NaN	False
total_rooms	0.753493	1.0	3.941629	4.802154e-02	True
Residual	56.966507	298.0	NaN	NaN	False
total_bedrooms	0.107376	1.0	0.555402	4.567060e-01	False
Residual	57.612624	298.0	NaN	NaN	False
population	0.056847	1.0	0.293783	5.882112e-01	False
Residual	57.663153	298.0	NaN	NaN	False
households	0.122198	1.0	0.632228	4.271721e-01	False
Residual	57.597802	298.0	NaN	NaN	False
median_income	17.033523	1.0	124.758652	1.979368e-24	True
Residual	40.686477	298.0	NaN	NaN	False
median_house_value	38.033046	1.0	575.703485	1.410202e-71	True
Residual	19.686954	298.0	NaN	NaN	False

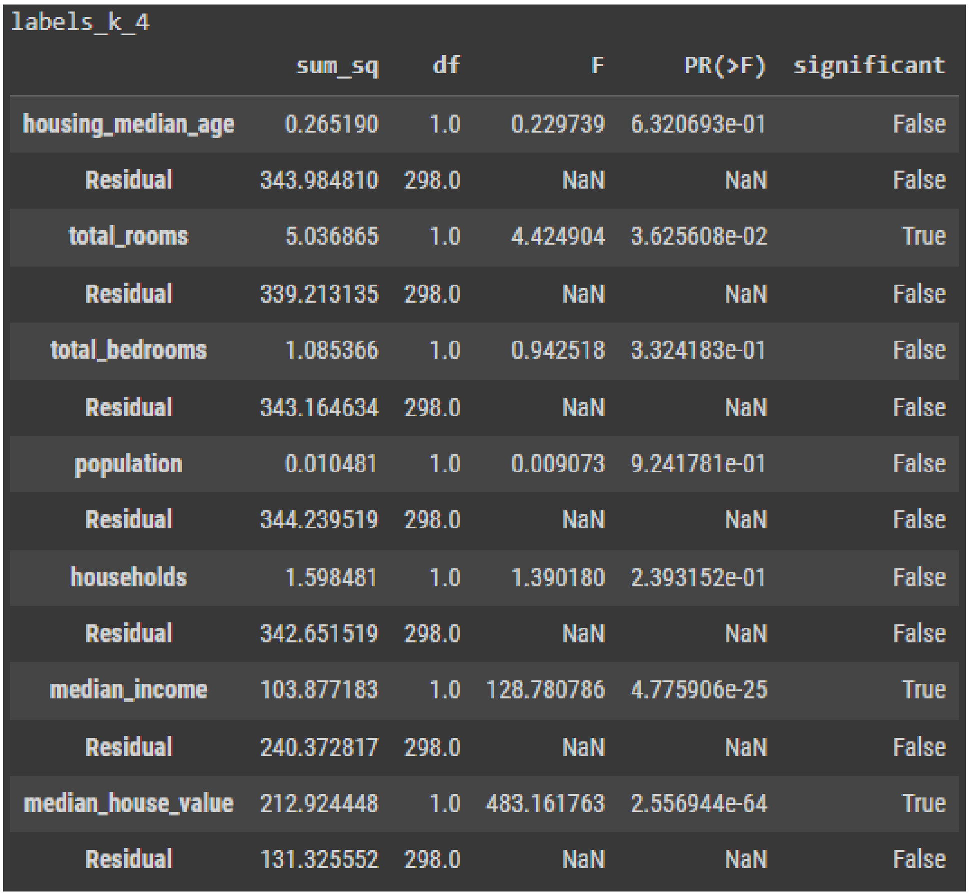
Table 4. ANOVA Table Cluster 3 Unscaled Data.



labels_k_3

	sum_sq	df	F	PR(>F)	significant
housing_median_age	1.536659	1.0	3.228002	7.340156e-02	False
Residual	141.860008	298.0	NaN	NaN	False
total_rooms	4.257548	1.0	9.118567	2.749269e-03	True
Residual	139.139119	298.0	NaN	NaN	False
total_bedrooms	1.147049	1.0	2.402964	1.221669e-01	False
Residual	142.249617	298.0	NaN	NaN	False
population	0.107057	1.0	0.222647	6.373757e-01	False
Residual	143.289610	298.0	NaN	NaN	False
households	1.801163	1.0	3.790703	5.247726e-02	False
Residual	141.595504	298.0	NaN	NaN	False
median_income	57.292302	1.0	198.283862	7.177450e-35	True
Residual	86.104365	298.0	NaN	NaN	False
median_house_value	118.931116	1.0	1448.627655	1.884369e-116	True
Residual	24.465550	298.0	NaN	NaN	False

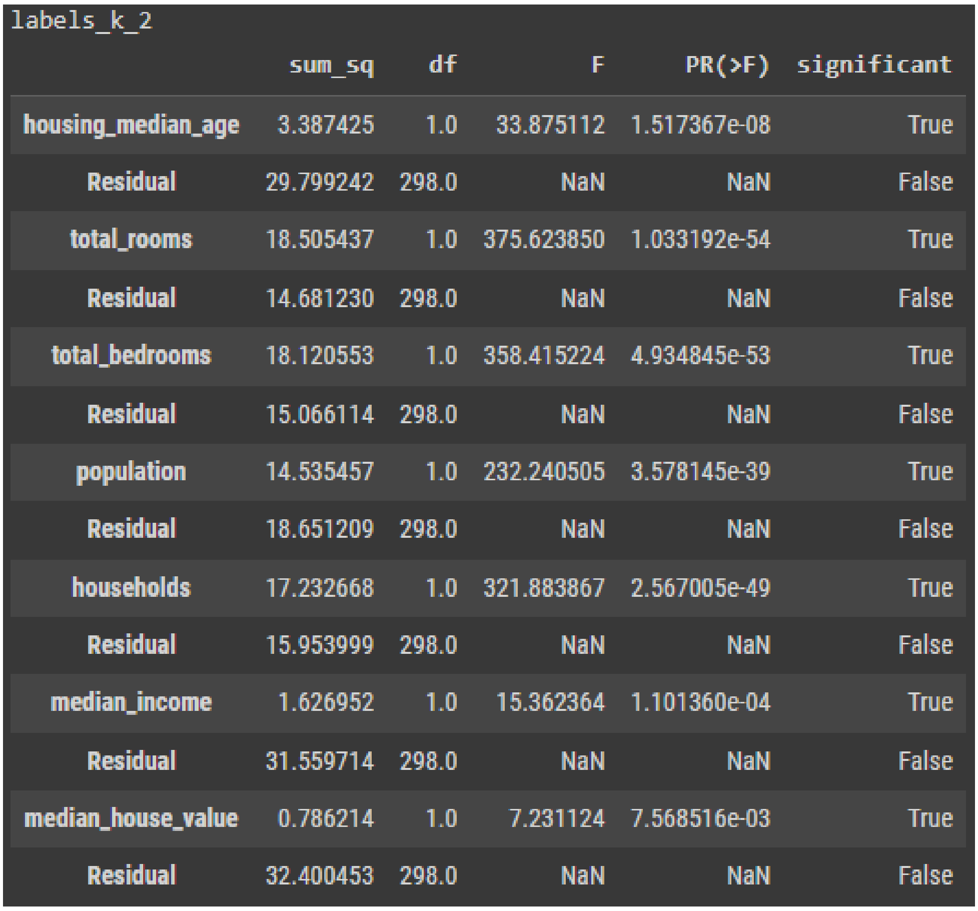
Table 5. ANOVA Table Cluster 4 Unscaled Data.



labels_k_4

	sum_sq	df	F	PR(>F)	significant
housing_median_age	0.265190	1.0	0.229739	6.320693e-01	False
Residual	343.984810	298.0	NaN	NaN	False
total_rooms	5.036865	1.0	4.424904	3.625608e-02	True
Residual	339.213135	298.0	NaN	NaN	False
total_bedrooms	1.085366	1.0	0.942518	3.324183e-01	False
Residual	343.164634	298.0	NaN	NaN	False
population	0.010481	1.0	0.009073	9.241781e-01	False
Residual	344.239519	298.0	NaN	NaN	False
households	1.598481	1.0	1.390180	2.393152e-01	False
Residual	342.651519	298.0	NaN	NaN	False
median_income	103.877183	1.0	128.780786	4.775906e-25	True
Residual	240.372817	298.0	NaN	NaN	False
median_house_value	212.924448	1.0	483.161763	2.556944e-64	True
Residual	131.325552	298.0	NaN	NaN	False

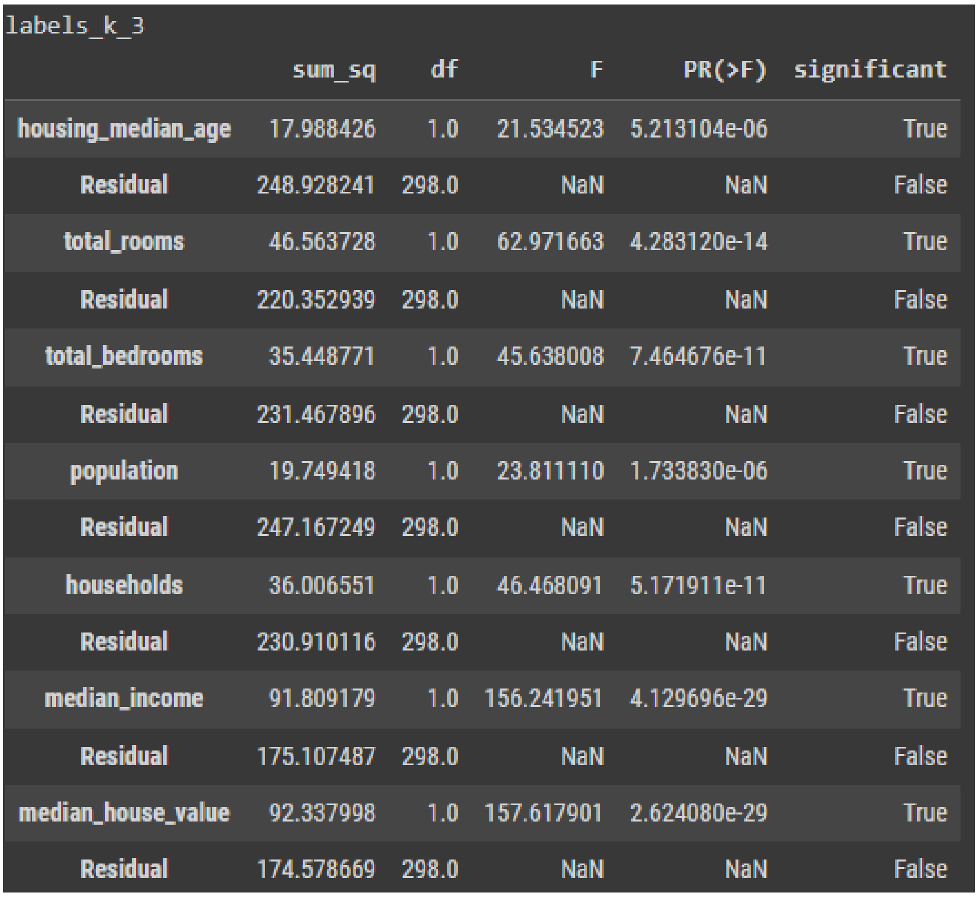
Table 6. ANOVA Table Cluster 2 Scaled Data.



labels_k_2

	sum_sq	df	F	PR(>F)	significant
housing_median_age	3.387425	1.0	33.875112	1.517367e-08	True
Residual	29.799242	298.0	NaN	NaN	False
total_rooms	18.505437	1.0	375.623850	1.033192e-54	True
Residual	14.681230	298.0	NaN	NaN	False
total_bedrooms	18.120553	1.0	358.415224	4.934845e-53	True
Residual	15.066114	298.0	NaN	NaN	False
population	14.535457	1.0	232.240505	3.578145e-39	True
Residual	18.651209	298.0	NaN	NaN	False
households	17.232668	1.0	321.883867	2.567005e-49	True
Residual	15.953999	298.0	NaN	NaN	False
median_income	1.626952	1.0	15.362364	1.101360e-04	True
Residual	31.559714	298.0	NaN	NaN	False
median_house_value	0.786214	1.0	7.231124	7.568516e-03	True
Residual	32.400453	298.0	NaN	NaN	False

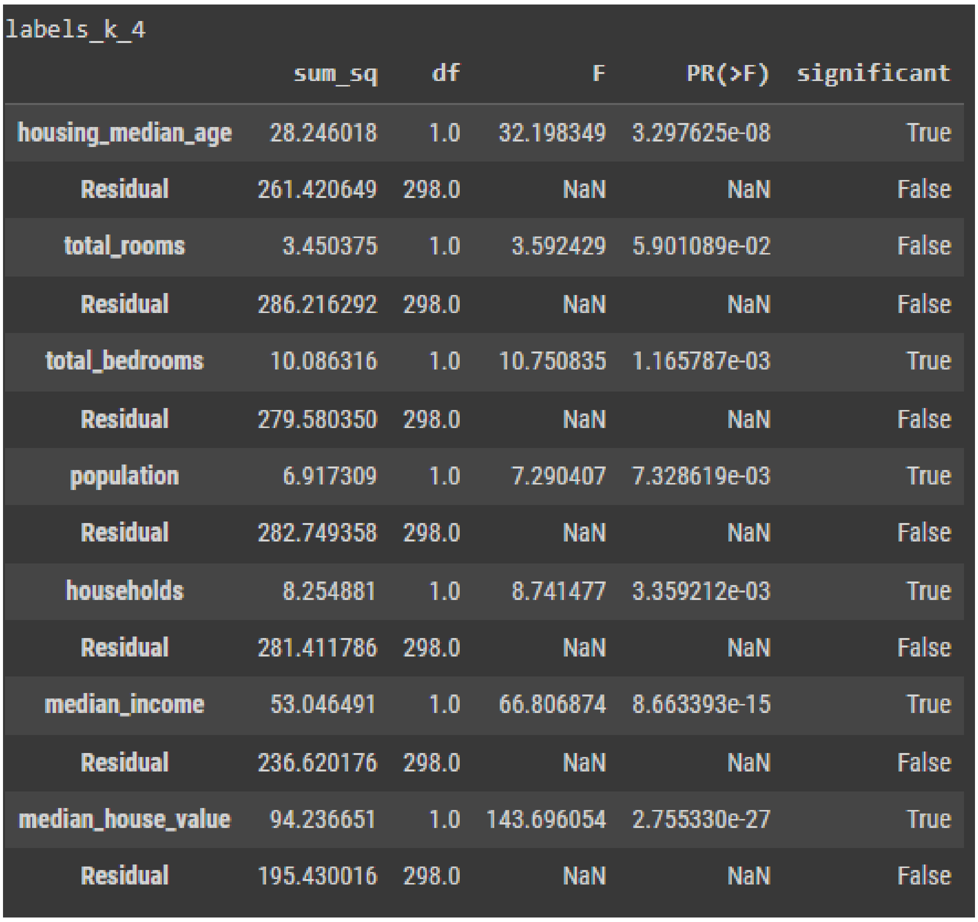
Table 7. ANOVA Table Cluster 3 Scaled Data.



labels_k_3

	sum_sq	df	F	PR(>F)	significant
housing_median_age	17.988426	1.0	21.534523	5.213104e-06	True
Residual	248.928241	298.0	NaN	NaN	False
total_rooms	46.563728	1.0	62.971663	4.283120e-14	True
Residual	220.352939	298.0	NaN	NaN	False
total_bedrooms	35.448771	1.0	45.638008	7.464676e-11	True
Residual	231.467896	298.0	NaN	NaN	False
population	19.749418	1.0	23.811110	1.733830e-06	True
Residual	247.167249	298.0	NaN	NaN	False
households	36.006551	1.0	46.468091	5.171911e-11	True
Residual	230.910116	298.0	NaN	NaN	False
median_income	91.809179	1.0	156.241951	4.129696e-29	True
Residual	175.107487	298.0	NaN	NaN	False
median_house_value	92.337998	1.0	157.617901	2.624080e-29	True
Residual	174.578669	298.0	NaN	NaN	False

Table 8. ANOVA Table Cluster 4 Scaled Data.



labels_k_4

	sum_sq	df	F	PR(>F)	significant
housing_median_age	28.246018	1.0	32.198349	3.297625e-08	True
Residual	261.420649	298.0	NaN	NaN	False
total_rooms	3.450375	1.0	3.592429	5.901089e-02	False
Residual	286.216292	298.0	NaN	NaN	False
total_bedrooms	10.086316	1.0	10.750835	1.165787e-03	True
Residual	279.580350	298.0	NaN	NaN	False
population	6.917309	1.0	7.290407	7.328619e-03	True
Residual	282.749358	298.0	NaN	NaN	False
households	8.254881	1.0	8.741477	3.359212e-03	True
Residual	281.411786	298.0	NaN	NaN	False
median_income	53.046491	1.0	66.806874	8.663393e-15	True
Residual	236.620176	298.0	NaN	NaN	False
median_house_value	94.236651	1.0	143.696054	2.755330e-27	True
Residual	195.430016	298.0	NaN	NaN	False

Conclusion

Analysis of Unscaled Data (housing_df) The analysis of the unscaled data from the California housing dataset reveals that certain variables consistently differentiate clusters across various configurations (k=2 to k=4). Specifically, total_rooms, median_income, and median_house_value have shown significant variability across clusters. These variables maintain p-values below 0.05, highlighting their strong influence in defining distinct groups within the housing market. This consistency suggests that these metrics are pivotal in understanding and clustering housing characteristics based on size, economic status, and property value.

Analysis of Scaled Data (housing_scaled_df) When examining the scaled version of the dataset, a wider array of variables emerges as significant in distinguishing between clusters. Notably, total_bedrooms, population, and households are among the variables that show significant differences among clusters post-scaling. This observation suggests that scaling enhances the sensitivity of the ANOVA test, allowing for a more nuanced detection of differences among clusters. Scaling helps to highlight the roles of variables that might otherwise be obscured due to their original scales in the unscaled dataset.

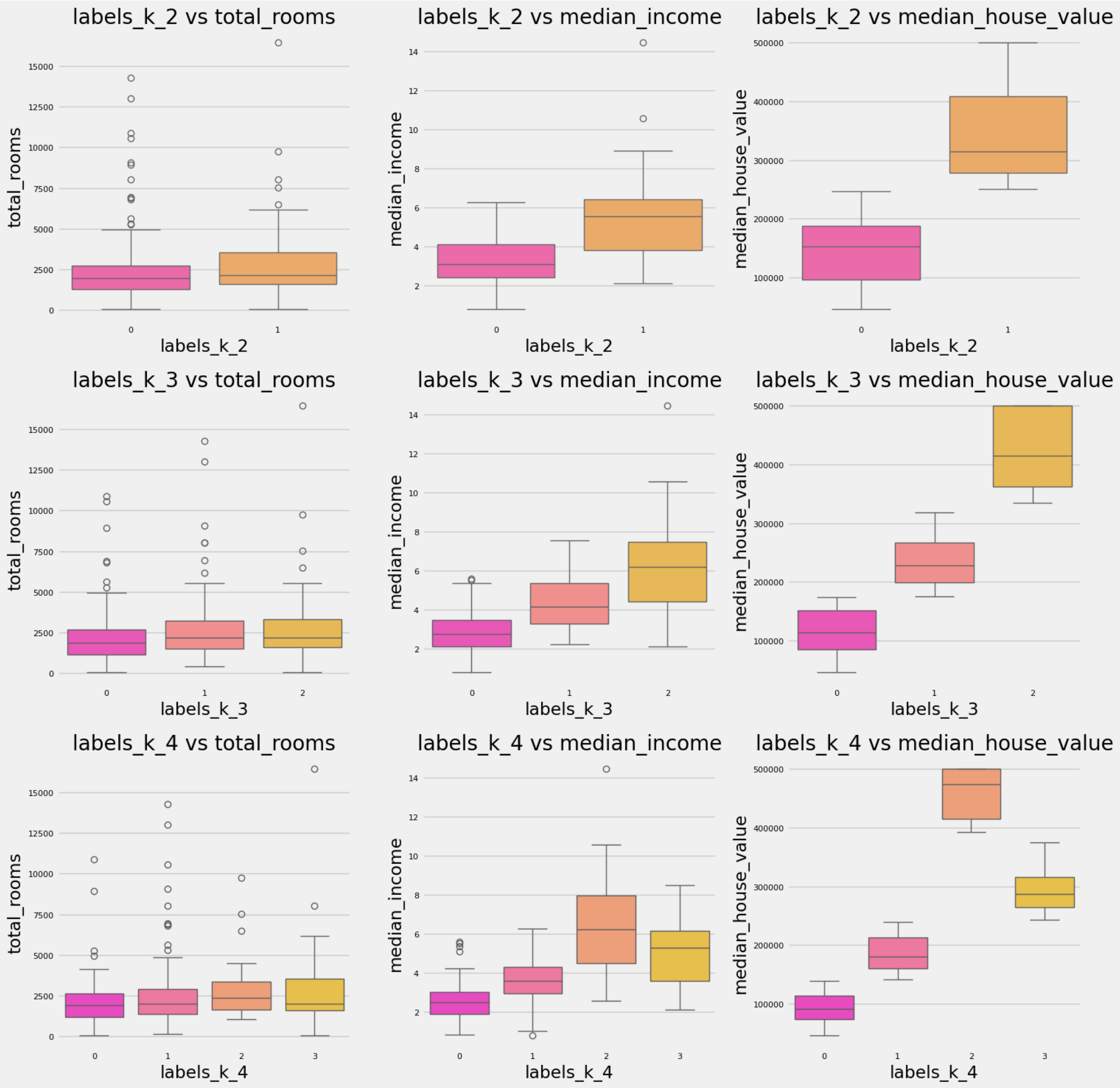
Impact of Scaling The process of scaling not only affects the sensitivity of statistical tests but also alters the scope of significant variables. It reveals how the inherent scale of variables, such as the number of rooms compared to median income, can either mask or amplify their influence on the clustering outcomes. By standardizing the scale across all variables, scaling ensures a more equitable evaluation of each variable's impact, revealing subtler distinctions across clusters that are not detectable in the unscaled data.

Visual Analysis of Different Clusters

The visualization of California housing data explores the distribution of three key variables: total_rooms, median_income, and median_house_value across clusters created with K-Means clustering for k=2, k=3, and k=4. These variables were identified as significant differentiators of the clusters through ANOVA tests.

The dataset on the Original Data used includes various attributes of housing data such as the number of rooms, median income of the area, and house values. The data is segmented into different clusters using the K-Means algorithm, and the clusters are visualized through box plots to compare the distribution of these key variables across different cluster labels.

Table 9. Visual Analysis of Different Clusters on the Original Data



Results

Total Rooms:

k=2: The box plots show a distinct difference in the distribution of total rooms between the two clusters. Cluster 1 generally has more rooms than cluster 0.

k=3 and k=4: As more clusters are introduced, the variability within clusters decreases, but differences between clusters become more nuanced. For instance, one of the clusters in k=4 displays a higher median but also a wider range of values, indicating a more heterogeneous grouping.

Median Income:

k=2: There's a clear disparity in median income between the two clusters, with cluster 1 displaying a significantly higher median income.

k=3 and k=4: The disparity in median income continues to be evident, with one cluster in k=3 and k=4 showing distinctly higher income levels. This suggests that income is a strong segregating factor in the clustering.

Median House Value:

k=2: Median house values are notably higher in cluster 1 compared to cluster 0.

k=3 and k=4: The trend in house values shows more diversification with increasing k-values. Particularly for k=4, one cluster clearly stands out with much higher house values, while the range of values in other clusters suggests different housing markets.

Conclusion

Cluster Characteristics: The analysis of box plots across cluster labels indicates that as the number of clusters increases, the clusters tend to specialize around key features. For instance, one cluster might represent high-income, high-value areas, while another focuses on more moderate economic areas.

Variable Impact: Total rooms, median income, and median house value are consistently shown to influence how the data segments into clusters, corroborating the results from the ANOVA tests.

Visual Analysis of Different Clusters on the Scaled Data

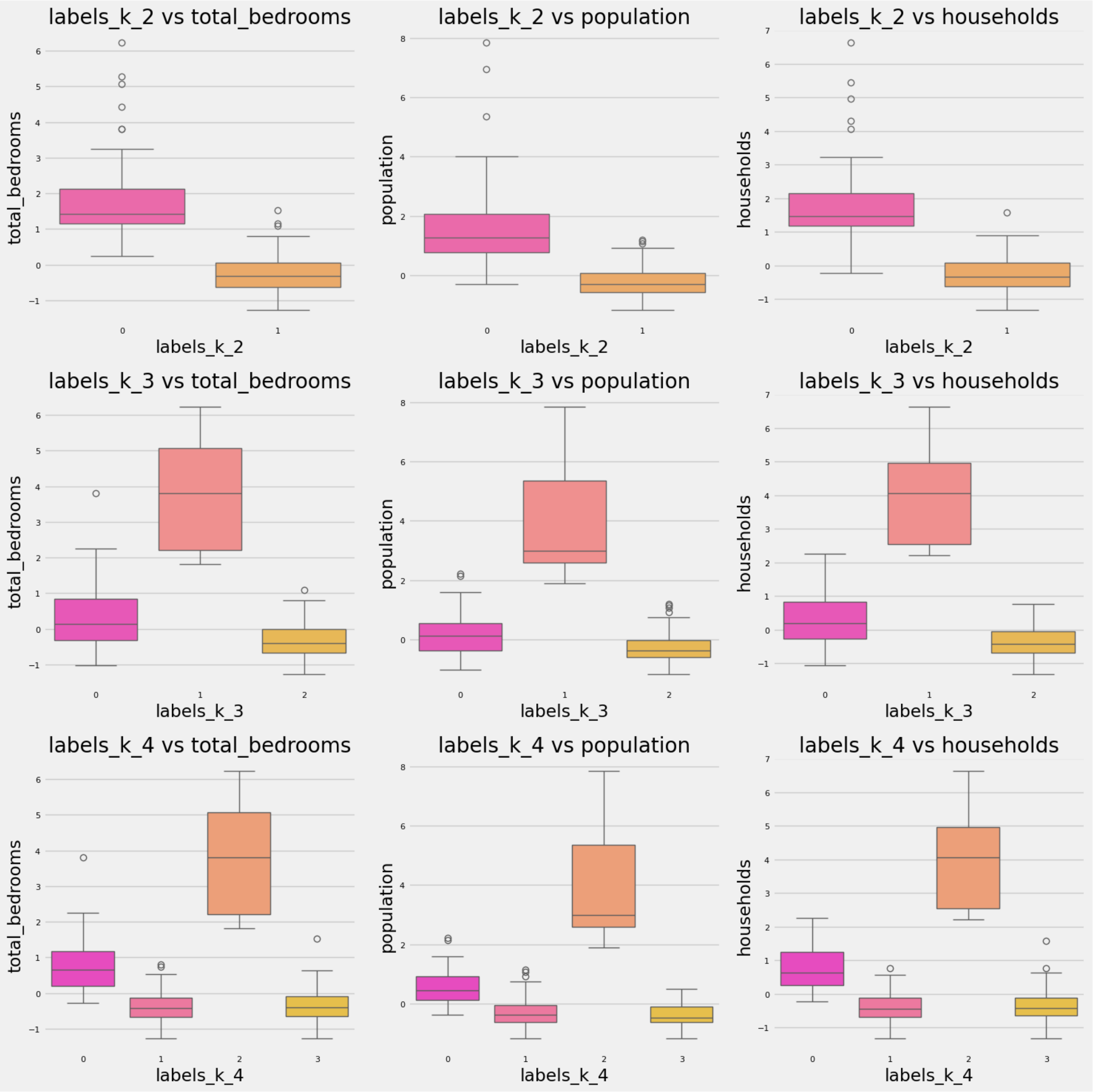
This analysis presents visual comparisons of key housing variables (total_bedrooms, population, households) from a normalized version of the California housing dataset, clustered using K-Means for k=2, k=3, and k=4. These variables were highlighted as significant in distinguishing cluster characteristics, particularly in the scaled data which enhances sensitivity to variances across clusters.

Data and Methodology

Data Used: The dataset consists of various standardized housing metrics, scaled to have zero mean and unit variance, ensuring that each feature contributes equally to the analysis.

Visualization Approach: Box plots illustrate the distribution of total_bedrooms, population, and households across clusters, providing insights into the spread and central tendency of these variables within each cluster.

Table 10. Visual Analysis of Different Clusters on the Standardized Data



Results

Total Bedrooms:

k=2: Clear differences are observed between the two clusters, with Cluster 0 typically showing fewer bedrooms compared to Cluster 1.

k=3 and k=4: The introduction of more clusters reveals more granular distinctions. Cluster 0 consistently shows fewer bedrooms, while other clusters vary more widely in their bedroom counts.

Population:

k=2: Cluster 0 has a noticeably lower population than Cluster 1, suggesting it may represent less densely populated areas.

k=3 and k=4: The population spread becomes more differentiated, with one cluster (e.g., Cluster 1 in k=4) showing significantly higher population figures, indicating high-density regions.

Households:

k=2: Similar to population, households in Cluster 0 are notably fewer than those in Cluster 1, aligning with the lower population and possibly smaller or less dense housing areas.

k=3 and k=4: Distributions show varied household sizes, with some clusters indicating larger household sizes possibly reflective of family-dense or communal living areas.

Conclusion

The normalization of data prior to clustering enhances the ability to detect subtle differences across clusters, which might be masked in unnormalized data due to scale disparities among variables.

Conclusion

The cluster analysis conducted on the California housing market data using both hierarchical and K-means clustering methods has provided key insights into the regional housing sector's characteristics and economic aspects. Our analysis identified distinct clusters that varied significantly in crucial variables such as total rooms, median income, and median house value. These variables were instrumental in distinguishing different groups within the housing market, each characterized by specific traits.

In our hierarchical clustering analysis on the original data, we identified approximately 2-3 significant clusters using Ward's method, which was particularly effective in minimizing within-cluster variance and producing actionable clusters. However, when we applied hierarchical clustering to the standardized data, the results indicated a more refined structure, typically showing 3-4 well-defined clusters, especially in methods like average linkage and Ward's method. This suggests that standardization of data helps in achieving a more nuanced differentiation between clusters.

Similarly, the K-means clustering analysis on the original data suggested that an optimal number of clusters ranged from 2 to 5, based on the elbow method observed in the SSE plots. When we shifted to the standardized data, the silhouette coefficients suggested that clustering the data into 4 groups maximized the distinction between clusters while maintaining homogeneity within them. This indicates that scaling the data changes the perception of cluster cohesion and separation, highlighting the impact of feature scale on the clustering outcomes.

References

1. Scikit-learn developers, no date. KMeans. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html [Accessed date: 27/04/2024].
2. Patel, K., no date. Detecting Knee/Elbow Points in a Graph. *Towards Data Science*. Available at: https://towardsdatascience.com/detecting-knee-elbow-points-in-a-graph-d13fc517a63c [Accessed date: 27/04/2024].].
3. Arvai, K., no date. Knee Elbow Point Detection. *Kaggle*. Available at: https://www.kaggle.com/code/kevinarvai/knee-elbow-point-detection [Accessed date: 27/04/2024].