# Unsupervised Learning Using K-Means Clustering on California Housing Data

Omid Saberi
*Dept. of Industrial Automation*
*University West*
Trollhättan, Sweden
https://orcid.org/0009-0005-3596-9182

***Abstract*—The goal of this assignment was to apply K-Means clustering to the California Housing dataset to identify meaningful segments based on geographic and economic features, specifically longitude, latitude, and median income. By conducting essential preprocessing steps and experimenting with various cluster configurations, we aimed to optimize the number of clusters using the Elbow Method and Silhouette score. The analysis indicated that a four-cluster configuration offered the most informative segmentation, capturing distinct regional patterns across California. Key insights included the differentiation of higher-priced coastal clusters from more affordable inland regions, revealing nuanced housing markets within the state. This clustering approach provides actionable insights for real estate professionals, enabling targeted development and investment strategies. Additionally, policy-makers could leverage these findings to address regional housing demands more effectively, guiding infrastructure and resource allocation. Overall, K-Means clustering proved valuable in segmenting the housing data, uncovering geographic and socioeconomic patterns that could inform future decisions in real estate and urban planning.**

***Index Terms*—Unsupervised Learning, K-Means clustering, California Housing Dataset, Silhouette Score, Geographic Segmentation, Machine Learning, Elbow Analysis, Inertia**

## I. Introduction

K-Means clustering is a fundamental algorithm in unsupervised learning, designed to partition data into distinct groups or clusters based on their inherent similarities. Unlike supervised learning, where labeled data is used to train models, unsupervised learning aims to discover hidden patterns within unlabeled data [2]. The K-Means algorithm accomplishes this by minimizing the variance within each cluster while maximizing the variance between clusters, ultimately providing a segmentation of the dataset [3].

In this assignment, K-Means clustering is applied to the California Housing dataset, focusing on geographic (longitude and latitude) and economic (median income) features. The goal is to identify meaningful clusters in the data that correspond to different segments of the housing market. By optimizing the number of clusters using the Silhouette score, this analysis aims to reveal patterns in housing characteristics, providing valuable insights into how various regions and income levels impact the housing landscape in California.

## II. K-Means Clustering and Evaluation

In this section, we describe the preprocessing steps, the application of the K-Means clustering algorithm, and the methods used to optimize the number of clusters for the California Housing dataset.

### A. Preprocessing of Data

The dataset used in this experiment consists of several features related to California housing, including the geographical coordinates (longitude and latitude) and the median income. Before applying the K-Means algorithm, we performed necessary preprocessing steps to ensure the data was suitable for clustering. These steps included handling missing values, normalizing the data to ensure equal weight for all features, and selecting the relevant features—longitude, latitude, and median income—since these were deemed important for identifying geographical and economic segments within the dataset.

### B. K-Means Clustering Algorithm

K-Means clustering is a widely used unsupervised learning algorithm that partitions data into a predefined number of clusters by minimizing the within-cluster variance. The algorithm works by assigning each data point to the nearest centroid and then recalculating the centroids based on the mean of the assigned points. This process is repeated until convergence, i.e., when the centroids no longer move significantly.

In this experiment, we applied the K-Means algorithm to the California Housing dataset using the longitude, latitude, and median income features. We experimented with various values of $k$, ranging from 2 to 19 clusters, to observe how the clustering performance changed with different configurations.

### C. Optimizing the Number of Clusters

To determine the optimal number of clusters, we employed several metrics, including the Silhouette score, which measures how similar each data point is to its own cluster compared to other clusters. The Silhouette score ranges from -1 to +1, with higher values indicating better-defined clusters. We calculated the Silhouette score for various values of $k$ (ranging from 2 to 10 clusters) and evaluated clustering quality for each configuration. Although the Silhouette score suggested that smaller values of $k$ (such as $k = 2$) might offer the best

cluster cohesion, the clustering results for $k = 4$ demonstrated the best overall segmentation, particularly in differentiating the data based on geographical distribution and median income.

Additionally, we employed the Elbow Method, which involves plotting the inertia (within-cluster sum of squares) against the number of clusters. The "elbow" point, where the curve begins to level off, often indicates the optimal number of clusters. Based on this analysis, we observed that while the inertia continued to decrease with increasing $k$, the rate of decrease slowed significantly after $k = 4$, suggesting that 4 clusters provided the best trade-off between model complexity and clustering performance.

### D. Final Configuration

Thus, after evaluating these multiple metrics, $k = 4$ was concluded to provide the most insightful differentiation of the data, as confirmed by the Elbow analysis and further validated by the clustering results.

## III. Results and Analysis

The clustering results for the California Housing dataset were examined for two configurations of the K-Means algorithm: $k = 2$ and $k = 4$. Each configuration was assessed using the Silhouette Score and Elbow Analysis. To illustrate the effectiveness of the clustering, line graphs and scatter plots were used to visualize the results.

### A. Clustering Results and Visualizations

As shown in Fig. 1, silhouette analysis suggests that $k = 2$ may be a suitable choice, as it produces the highest average silhouette score. This implies that, on average, data points within these two clusters are well-separated from others while maintaining internal similarity.

In contrast, Fig. 2 presents the Elbow Analysis, which indicates an optimal cluster count near $k = 4$. The graph reveals a sharp initial drop in inertia, followed by a gradual decline, suggesting that adding more clusters beyond this point yields diminishing returns. Although the "elbow" is not sharply defined, this point represents a reasonable balance between model simplicity and capturing underlying data structures.

For $k = 2$, the resulting clusters, illustrated in Fig. 3, reveal two distinct groupings. One cluster is concentrated in the central and northern regions, possibly representing urban areas, while the other spans southern regions, potentially indicating less urbanized or rural areas. This clustering may reflect urban-rural distinctions or factors like population density and industrial presence. The silhouette score for $k = 2$ is 0.532, indicating relatively well-defined clusters with clear separation, which aligns with the visual representation.

On the other hand, the $k = 4$ configuration, shown in Fig. 4, provides a more detailed segmentation. The clusters are spread across the state, with Cluster 0 (Blue) along the southern coast, especially in Los Angeles and San Diego; Cluster 1 (Orange) centered around San Francisco; Cluster 2 (Green) distributed over Northern California; and Cluster 3 (Red) overlapping with Cluster 0 in the southern regions. Although the silhouette score for $k = 4$ is lower (0.400), it captures a more complex

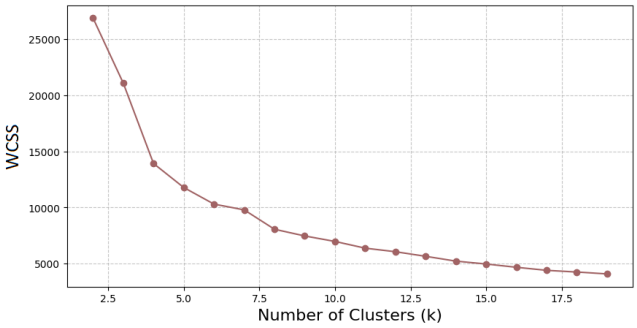


Fig. 2. Within-Cluster Sum of Squares (Inertia) vs Number of Clusters $k$

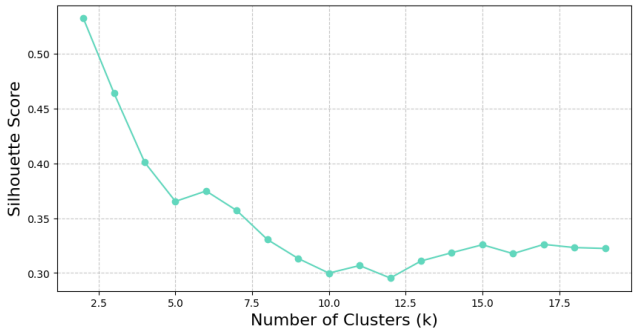


Fig. 1. Silhouette Score vs. Number of Clusters (k)

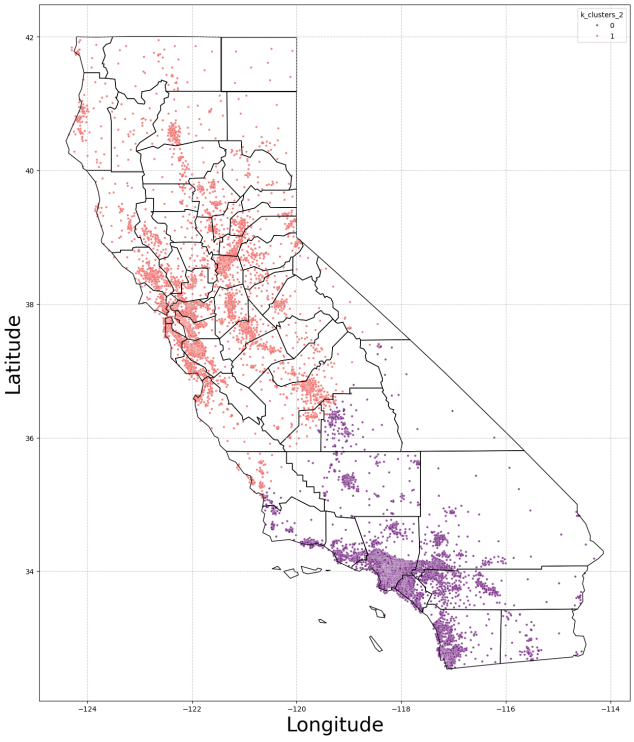


Fig. 3. K-Means Clustering Results with $k = 2$

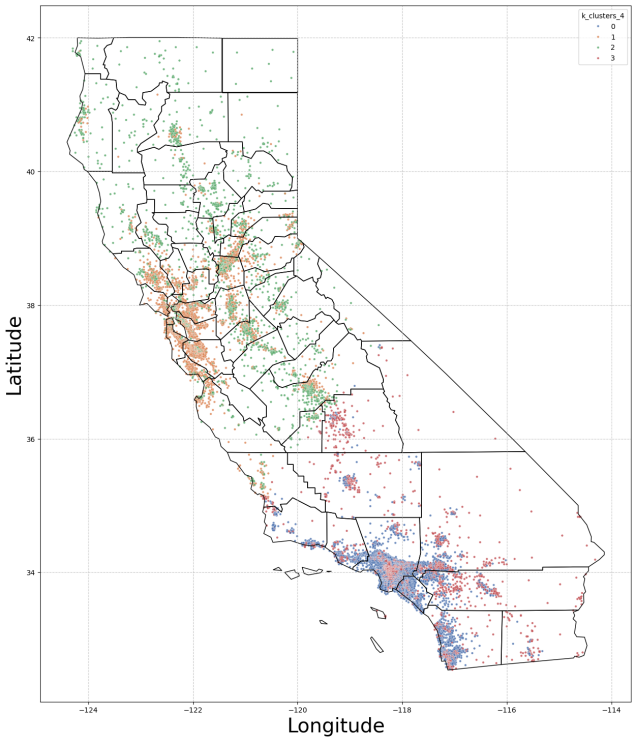


Fig. 4. K-Means Clustering Results with $k = 4$

segmentation, revealing distinctions that add interpretive value. The slight reduction in silhouette score reflects the increased complexity in segmentation, which provides a nuanced, albeit less compact, view of the data.

### B. Comparison of Clustering Configurations

The choice of $k = 4$ over $k = 2$ was driven by the need for more actionable insights. While $k = 2$ offers higher cohesion, it fails to provide meaningful distinctions between key regions, such as the coastal and inland areas, which could have different housing demands. On the other hand, $k = 4$ provides a clearer delineation of distinct geographical regions and housing markets, as evidenced by the spatial distribution of the clusters in the scatter plot.

For instance, Cluster 0 (Blue) is concentrated along the coast, notably in the San Francisco Bay Area and Southern California. This cluster, characterized by higher housing prices, could be a prime target for luxury housing developments or investments. Cluster 1 (Orange) in the Central Valley represents more affordable housing markets, ideal for first-time homebuyers. Cluster 2 (Orange), focused on Southern California, might attract rental property investments due to the region's high housing demand. Finally, Cluster 3 (Red), scattered across the state, includes diverse sub-regions that require more targeted approaches to address varying housing needs.

### C. Insights from Clustering and Silhouette Score Analysis

While the Silhouette score for $k = 4$ is slightly lower than for $k = 2$, the greater granularity provided by four clusters enables more targeted insights for housing market segmentation. The increased complexity is justified by the ability to identify specific regions with unique housing needs, such as the distinction between coastal luxury markets and inland affordable markets. This differentiation aligns with the objective of the assignment: to segment housing data for further insights and potential applications in real estate and development strategies.

## IV. Conclusion

In this analysis, we explored how K-Means clustering could segment the California Housing dataset to reveal meaningful patterns across geographical and economic features. By examining different configurations of clusters and applying metrics such as the Silhouette score and the Elbow Method, we identified that a configuration of k = 4 clusters provided the most insightful segmentation, balancing both complexity and interpretability.

K-Means clustering proved instrumental in identifying distinct regions within California that differ in median income and geographical distribution. This segmentation highlighted variations such as higher-priced coastal clusters and more affordable inland regions, offering a clear view of the diverse housing markets in the state.

The potential applications of this segmentation are vast. Real estate professionals, for instance, can use these insights to tailor development and investment strategies for specific regions, prioritizing luxury housing in high-cost areas and affordable housing in regions with lower median incomes. Additionally, policy-makers could leverage these clusters to address regional housing demands more effectively, guiding zoning and infrastructure planning according to the unique needs of each cluster. This clustering approach offers a foundation for further exploration into housing trends and socioeconomic patterns, enabling more informed decision-making across sectors involved in California's housing market.

Future work could include comparing these results with other clustering algorithms such as DBSCAN, or further exploring the impact of additional features like housing price and population density on the clustering outcome.

## References

[1] O. Saberi, "IAI600Lab6.ipynb," GitHub repository, GitHub, Nov. 12, 2024. [Online]. Available: https://github.com/omisa69/IAI600Lab6

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[3] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.