Housing Data Clustering Analysis Report

Objective of the Project

This project's main goal is to use clustering techniques on a housing dataset in order to find patterns and put related observations in one group. The objective is to learn more about the housing market by investigating the ways in which various features influence the data's segmentation. The purpose of this analysis is to inform decision-making processes and assist stakeholders in the fields of real estate, urban planning, and economic policy by illuminating the variables influencing housing clusters.

Data Set Description

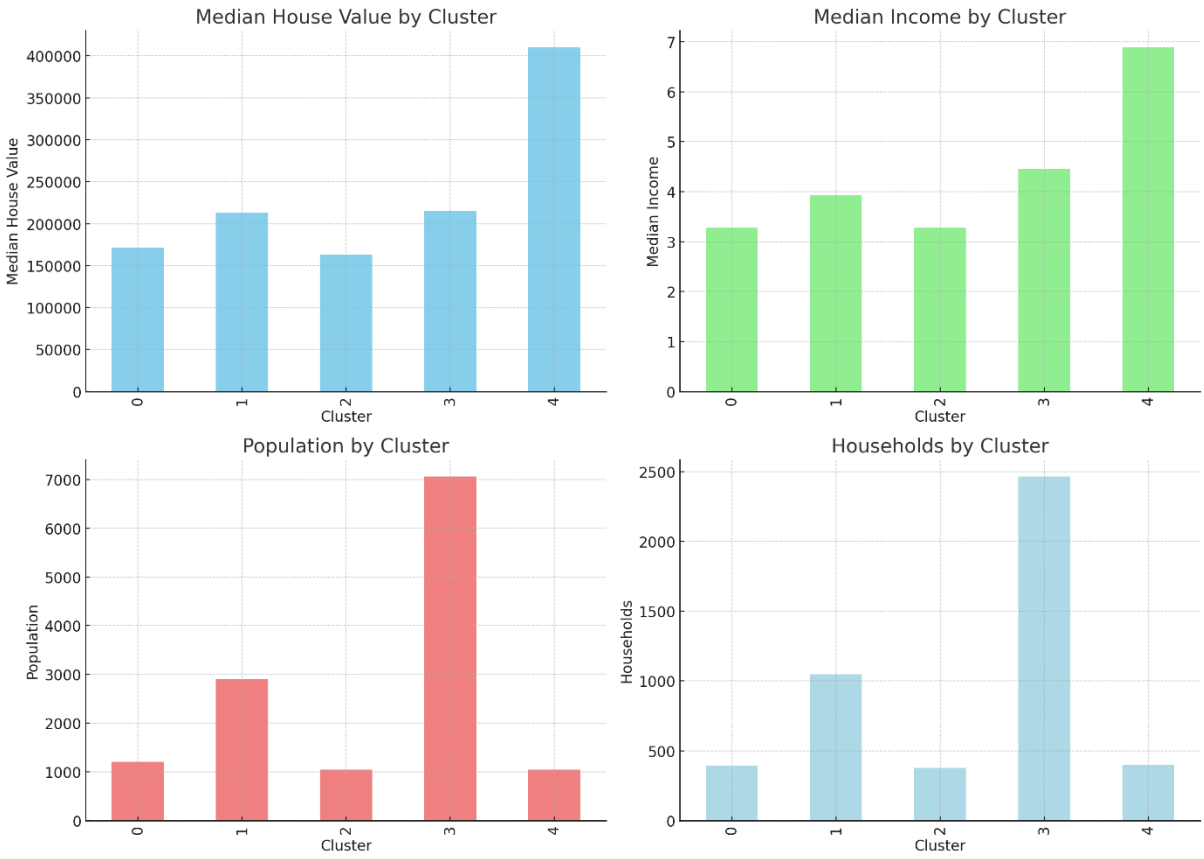
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

- Longitude: Geographic coordinate, longitude of the location.
- Latitude: Geographic coordinate, latitude of the location.
- Housing Median Age: Median age of the houses in the area.
- Total Rooms: Total number of rooms in the area.
- Total Bedrooms: Total number of bedrooms in the area.
- Population: Population in the area.
- Households: Number of households in the area.
- Median Income: Median income of the households.
- Median House Value: Median house value in the area.
- Ocean Proximity: Proximity to the ocean.

Training Summary for Clustering Algorithms

To divide the housing data into various clusters, we used the K-Means clustering algorithm. In order to handle missing values and normalize the features, the data underwent preprocessing. Five unique clusters were found by the clustering analysis, and they had the following traits:

- Cluster 0: median home value of $171,000, mid-range longitude and latitude.
- Cluster 1: A median home value of $213,000, with slightly lower longitude and higher latitude.
- Cluster 2 has a median home value of $163,000 and higher longitude and latitude.
- Cluster 3: Densely populated, lower latitude, median house value of $215,000.
- Cluster 4: Wealthy neighborhoods with a $410,000 median home value.,



Model Recommendation

The following features were determined to be the most significant in determining the clusters based on the feature importance analysis performed using a Random Forest classifier:

Latitude

Median House Value

Longitude

Median Income

Households

Summary/Key Findings

Geographical Features: The most important features are latitude and longitude, which show how important location is in the formation of clusters.

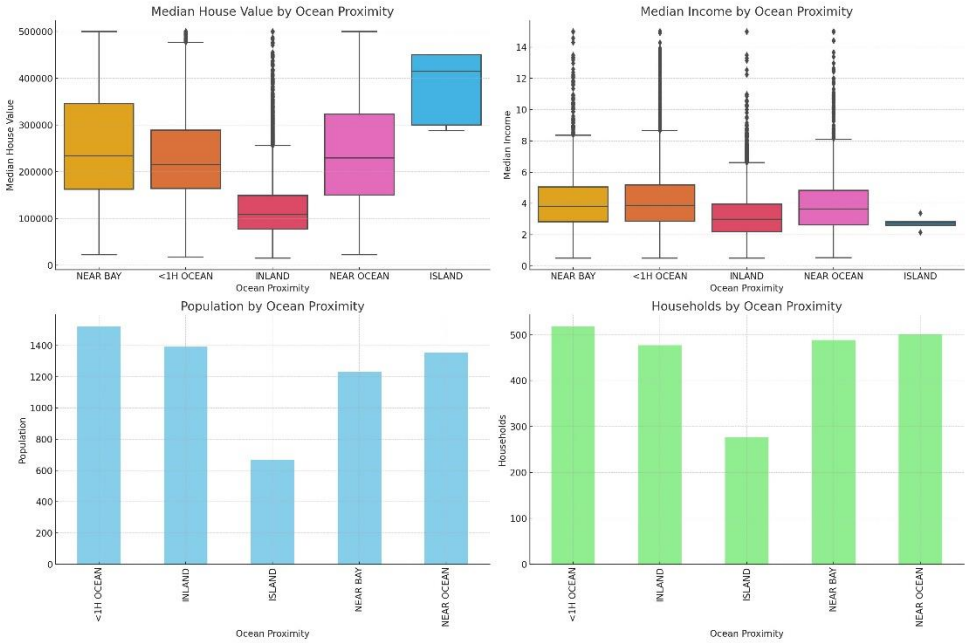
Economic Features: The significance of both median income and median house value points to the importance of economic factors in identifying clusters.

Housing Characteristics: A house's size and density are reflected to a moderate extent by the number of households, total bedrooms, and total rooms.

Clustering by Ocean Proximity

Properties nearer the bay or the ocean typically have higher median incomes and house values, indicating more affluent and desirable neighborhoods.

Despite having lower median incomes, inland properties are more reasonably priced, indicating distinct market niches.



Next Steps

Further Analysis: Investigate the factors influencing the high median house values and incomes in affluent clusters.

Segmentation Strategy: Utilize the identified clusters to develop targeted marketing or policy strategies for housing and urban development.

Dimension Reduction: Apply PCA or t-SNE to visualize the data in a lower-dimensional space for additional insights.

Model Refinement: Consider combining clustering with supervised learning techniques to improve predictive accuracy and tailor models to specific clusters.