Unsupervised Learning on California Housing Dataset - Project Overview

Project Objectives

This notebook implements **unsupervised learning techniques** (Practice 2, Group I) by Itziar Lopez Almagro and Adrian Carlos Skaczylo Sawicka to analyze the California Housing dataset using scikit-learn clustering algorithms. The main goals are:

1. **Geospatial Pattern Discovery**: Identify natural groupings of housing districts based on demographic and economic characteristics
2. **Clustering Analysis**: Apply K-Means and Hierarchical (Agglomerative) clustering to discover market segments
3. **Feature Engineering**: Create meaningful features from raw housing data (rooms/household, etc.)
4. **Model Evaluation**: Use Davies-Bouldin score and silhouette analysis to validate cluster quality
5. **Geographic Interpretation**: Map clusters to California regions and understand socioeconomic patterns
6. **Algorithm Comparison**: Contrast K-Means vs Hierarchical clustering performance and characteristics

Dataset Description

The **California Housing dataset** contains **20,433 housing blocks** from the 1990 California census with 10 features:

Feature	Description	Range
`longitude`, `latitude`	Geographic coordinates	CA statewide
`housing_median_age`	Median house age (years)	1-52
`total_rooms`	Total number of rooms	2-39,320

`total_bedrooms`	Total number of bedrooms	1-6,445
`population`	Block population	3-35,682
`households`	Block households	1-6,082
`median_income`	Median income (scaled $10k)	0.50-15.00
Target: `median_house_value`	Median house value	$14,999-$500,001

ocean_proximity: Categorical proximity to ocean (NEAR BAY, NEAR OCEAN, INLAND, etc.)

Methodology and Workflow

Phase 1: Data Preprocessing

- **Null removal**: Clean dataset (20433 → clean records)
- **Feature engineering**: Create density ratios (rooms/household, population/household)
- **Encoding**: Convert `ocean_proximity` to numeric labels
- **Exploratory analysis**: Correlation matrices, geographic scatter plots

Phase 2: K-Means Clustering (9 Clusters)

K-Means(n_clusters=9) applied to engineered features reveals **9 distinct market segments**:

Cluster	Size	Latitude	Median Income	House Value	Key Characteristics

0	1,583	37.78°	$3.44	$219K	Bay Area, moderate income
1	1,927	37.84°	$2.58	$98K	North Bay suburbs
2	1,636	37.65°	$4.79	$252K	SF Peninsula
3	**3,085**	33.87°	$4.34	$215K	**LA Basin core**
4	1,664	34.96°	**$6.77**	**$447K**	**Orange County luxury**
5	**4,068**	34.01°	$3.31	$185K	**LA Valley inland**
6	2,795	38.22°	$3.24	$123K	North Coast
7	2,512	33.66°	$2.97	$137K	Inland Empire
8	1,163	35.58°	$5.34	$369K	Central Coast

Phase 3: Hierarchical Clustering (6 Clusters)

AgglomerativeClustering(n_clusters=6, linkage='ward') produces **6 geographic regions**:

Cluster	Size	Longitude	Population Density	House Value
0	4,443	-119.20°	Moderate	**$360K** (Central Valley)

1	2,806	-117.97°	Low	$145K (Desert areas)
2	**5,764**	**-121.38°**	**High**	**$147K** (**Bay Area inland**)
3	1,736	-122.27°	High	$242K (SF proper)
4	3,252	-118.20°	Moderate	$169K (LA suburbs)
5	2,432	-117.72°	High	$165K (San Diego)

Key Findings and Insights

Market Segmentation Patterns

1. **Geographic Clustering**: Strong latitude/longitude separation matching California regions
2. **Income-Housing Correlation**: High-income clusters (4,8) → high house values
3. **Bay Area Premium**: Clusters 0,2,3 show highest values despite inland locations
4. **LA Basin Density**: Largest clusters (3,5) in high-population areas
5. **Coastal Premium**: Proximity to ocean correlates with higher values

Clustering Quality Metrics

text

```
K-Means (9 clusters):

✓ Davies-Bouldin Score: Low (good separation)

✓ Clear geographic interpretation

✓ Income/housing value stratification


Hierarchical (6 clusters):

✓ Geographic regions recovered
```

✓ Larger, more stable clusters

✓ Ward linkage captures spatial structure

Feature Engineering Impact

Engineered ratios proved crucial:

text

```
rooms_per_household = total_rooms / households

bedrooms_per_room = total_bedrooms / total_rooms

population_per_household = population / households
```

These normalized features revealed socioeconomic patterns invisible in raw counts.

Algorithm Comparison

Aspect	K-Means (9 clusters)	Hierarchical (6 clusters)
Speed	Very fast	Slower (O(n²))
Interpretability	Market segments	Geographic regions
Cluster sizes	Varied (1.2K-4K)	More balanced
Spatial structure	Good	**Excellent**
Optimal K	Requires elbow method	Dendrogram analysis

Business Applications

Real Estate Insights

text

```
1. **Targeted Marketing**: Luxury campaigns → Clusters 4,8

2. **Investment Opportunities**: Undervalued inland → Cluster 7

3. **Development Planning**: High density → Clusters 3,5

4. **Pricing Strategy**: Regional benchmarks established

5. **Risk Assessment**: Income-housing mismatches identified
```

Urban Planning Applications

- **Infrastructure allocation** by population density
- **Housing policy** targeting low-income clusters
- **Transit planning** matching regional patterns
- **Environmental impact** assessment by geography

Technical Implementation Highlights

Libraries Used

text

```
Core: pandas, numpy, scikit-learn, matplotlib

Clustering: KMeans, AgglomerativeClustering

Evaluation: davies_bouldin_score

Preprocessing: MinMaxScaler (if applied)
```

Key Parameters

text

```
KMeans: n_clusters=9, random_state=fixed

Agglomerative: n_clusters=6, linkage='ward'

All models: reproducible random_state
```

Learning Outcomes

Unsupervised Learning Skills

- **Cluster interpretation**: Translating math → business insights
- **Algorithm selection**: K-Means vs hierarchical use cases
- **Evaluation rigor**: Quantitative validation beyond visuals
- **Feature engineering**: Creating meaningful ratios
- **Geospatial analysis**: Latitude/longitude clustering patterns

Domain Knowledge

- **California housing dynamics**: Coastal vs inland premiums
- **Socioeconomic segmentation**: Income correlates with geography
- **Market structure**: 6-9 natural housing segments exist

Challenges and Solutions

High Dimensionality

Problem: 10 features + engineered ratios → curse of dimensionality
Solution: Geographic features (lat/lon) naturally dominate spatial clustering

Mixed Data Types

Problem: Numeric + categorical (ocean_proximity)
Solution: Numeric encoding + focus on continuous features

Scale Differences

Problem: Rooms (39K) vs Income (15) vastly different scales
Solution: Engineered ratios normalize population effects

Extensions and Future Work

text

```
1. **Optimal K Selection**: Elbow method, silhouette analysis

2. **Dimensionality Reduction**: PCA before clustering

3. **Density-based**: DBSCAN for irregular shapes

4. **Time Series**: Track segment evolution 1990→2026

5. **Spatial Autocorrelation**: Moran's I for validation

6. **Ensemble Clustering**: Combine K-Means + Hierarchical
```

Conclusion

This project successfully applies **unsupervised learning** to uncover **9 socioeconomic housing segments** and **6 geographic regions** in the California housing market. K-Means reveals granular market segments while hierarchical clustering captures broader regional patterns, both validated through statistical measures and geographic interpretability.

The analysis demonstrates practical business value through targeted marketing opportunities, investment insights, and urban planning applications. By systematically preprocessing, clustering, evaluating, and interpreting results, the notebook provides a complete unsupervised learning workflow applicable to real-world housing market analysis.