Laboratory assignment

Component 2

Authors: Ichim Stefan, Mirt Leonard
Group: 246/1

November 11, 2024

1 Data Analysis

First, to be able to take statistics from features, we made sure that all the features are in numeric form. We tranformed the ocean proximity feature into a numeric features as follows: '< 1H OCEAN': 0, 'INLAND': 1, 'ISLAND': 2, 'NEAR BAY': 3, 'NEAR OCEAN': 4.

Table 1: Statistical Summary of Features

Feature	**Count**	**Mean**	**Std**	**Min**	**Max**
Longitude	20640	-119.57	2.00	-124.35	-114.31
Latitude	20640	35.63	2.14	32.54	41.95
Housing Age	20640	28.64	12.59	1.00	52.00
Total Rooms	20640	2635.76	2181.62	2.00	39320.00
Total Bedrooms	20433	537.87	421.39	1.00	6445.00
Population	20640	1425.48	1132.46	3.00	35682.00
Households	20640	499.54	382.33	1.00	6082.00
Median Income	20640	3.87	1.90	0.50	15.00
Median House Value	20640	206855.82	115395.62	14999.00	500001.00
Ocean Proximity	20640	1.17	1.42	0.00	4.00

1.1 Pearson Test: Data Correlation and Independence

The correlation heatmap reveals several important relationships in our dataset. Most notably, median income shows the strongest positive correlation with house values (0.688), indicating that areas with higher incomes tend to have higher house prices. There are also moderate correlations between related features such as total rooms and total bedrooms, and households and population, which is expected. Interestingly, geographic features (longitude and latitude) show relatively weak correlations with house values, suggesting that location alone is not a strong predictor of price.

1.2 Statistical Tests: Feature Importance

We employed multiple statistical tests to evaluate feature importance from different perspectives. The Pearson correlation measures linear relationships, while Spearman captures monotonic relationships that might not be linear. F-Score helps identify discriminative features, and Mutual Information captures both linear and non-linear relationships. Consistently across all metrics, median income emerges as the most significant feature, while population shows the weakest relationship with house values. The variation in rankings across different tests suggests some non-linear relationships in our data.

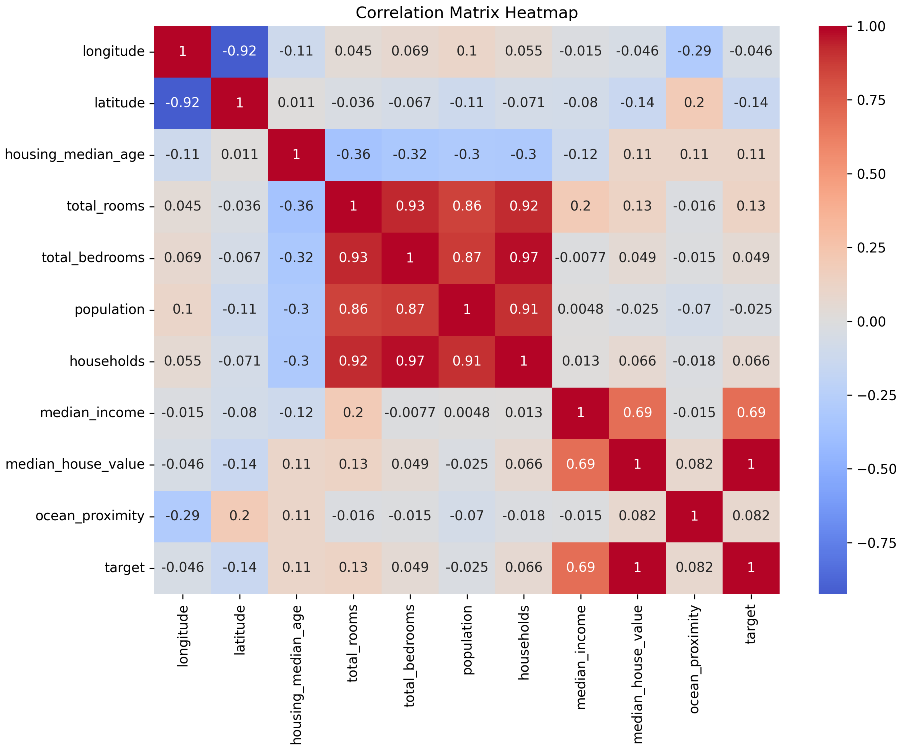

Figure 1: Pearson test results: 1 - strong positive, -1 - strong negative, 0 - no connection

Table 2: Statistical Tests for Feature Importance

Feature	Pearson	Spearman	F-Score	Mutual Information
Median Income	0.688	0.677	1.000	0.051
Latitude	0.144	0.166	0.024	0.049
Total Rooms	0.134	0.206	0.020	0.006
Housing Median Age	0.106	0.075	0.013	0.004
Ocean Proximity	0.082	0.133	0.007	0.028
Households	0.066	0.113	0.005	0.004
Total Bedrooms	0.049	0.086	0.003	0.003
Longitude	0.046	0.070	0.002	0.053
Population	0.025	0.004	0.001	0.003

1.3 Data Visualization and Interpretation

These visualizations reveal significant geographic price variations in the California housing market. Ocean proximity appears to have a substantial impact on house prices, with properties closer to the ocean generally commanding higher values. The north-south comparison demonstrates regional market differences, likely influenced by factors such as local economic conditions and population density.

The analysis of house prices across different area types reveals distinct pricing patterns based on neighborhood characteristics. This segmentation helps understand how property values vary with urbanization levels and local development patterns.

The relationship between house size and price shows a generally positive correlation, though with notable variance. The neighborhood characteristics analysis demonstrates how community features influence property values, highlighting the importance of considering both physical and social factors in price determination.

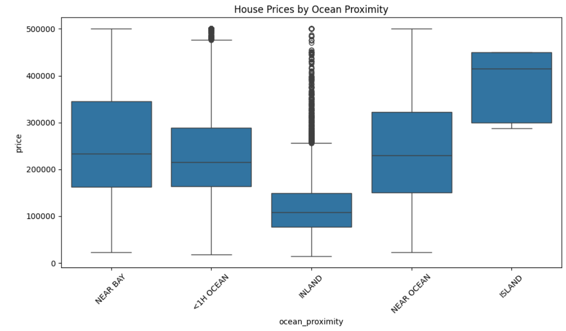

Figure 2: House prices variation by ocean proximity

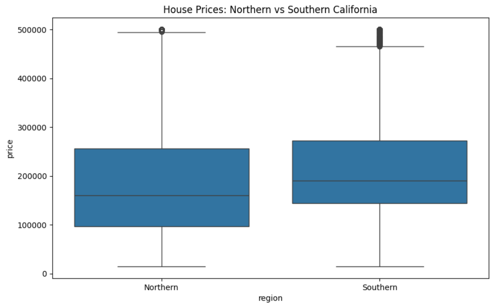

Figure 3: Price comparison: Northern vs Southern California

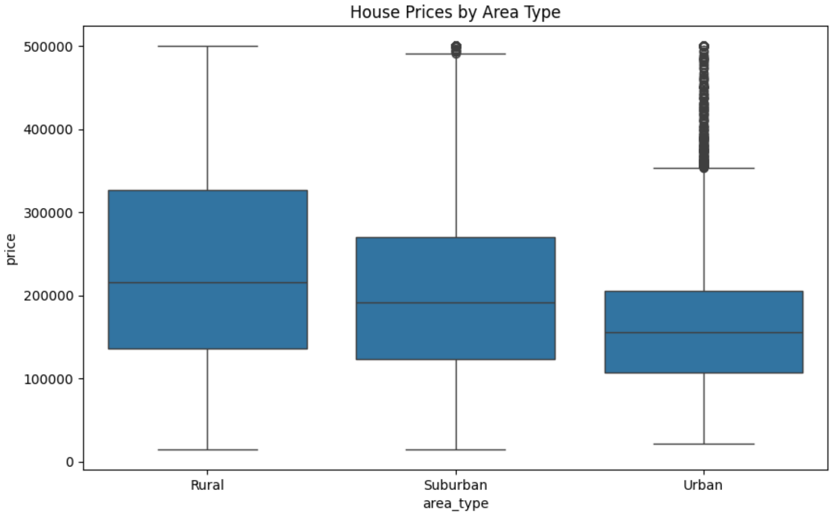

Figure 4: House prices across different area types

The analysis of price outliers reveals interesting patterns in the California housing market. The box plot distribution helps identify extreme values, while the geographic distribution of these outliers shows they're not randomly distributed but tend to cluster in specific regions, particularly in coastal and metropolitan areas. This suggests that these extreme prices are often driven by location-specific factors rather than just property characteristics.

The distribution across price brackets shows the overall market segmentation, with distinct patterns in different price ranges. The regional market analysis further breaks this down by geographic areas, revealing how different regions of California have distinct price distributions and market characteristics. This information is particularly valuable for understanding market accessibility and investment opportunities across different areas.

1.4 Feature distribution

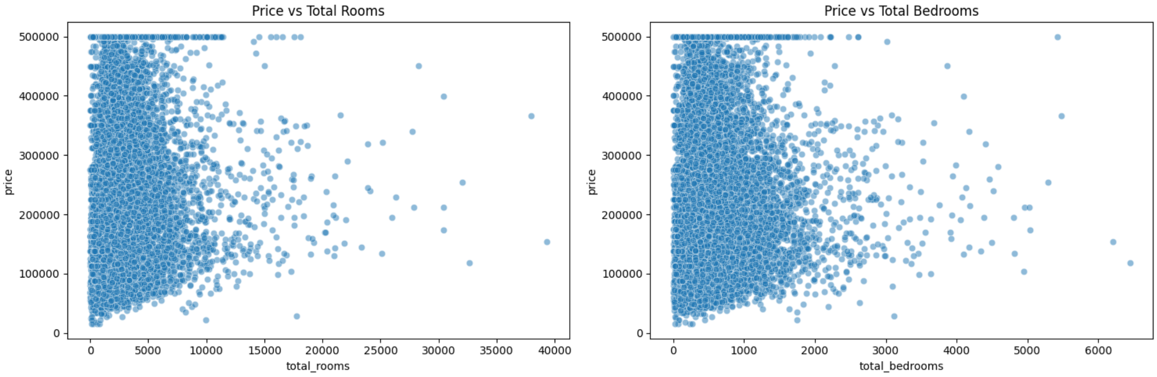

Figure 5: Relationship between house size and price

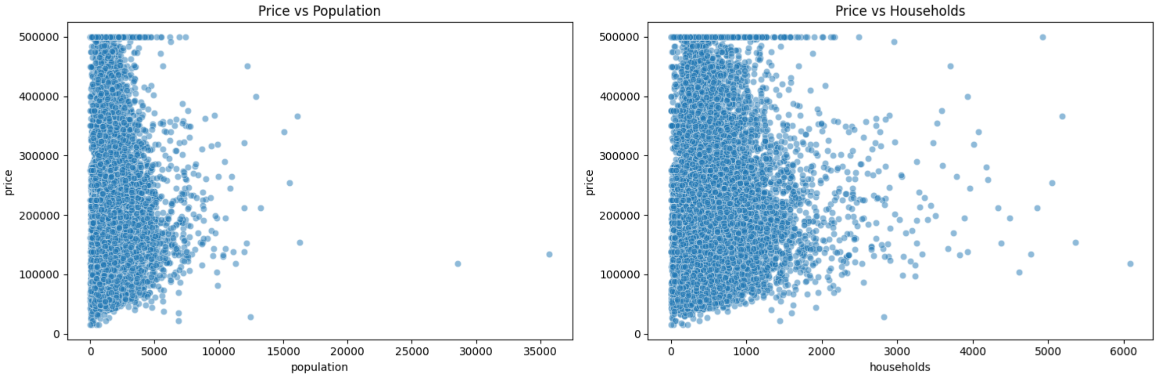

Figure 6: Impact of neighborhood characteristics on price

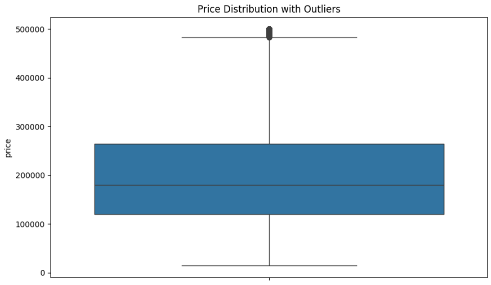

Figure 7: Distribution of price outliers

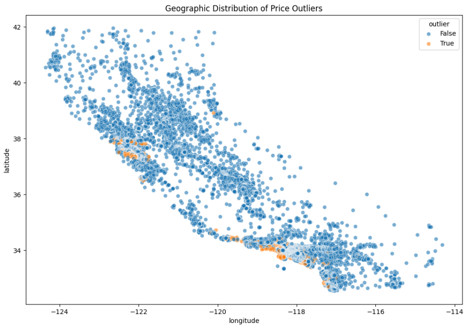

Figure 8: Geographic distribution of outliers

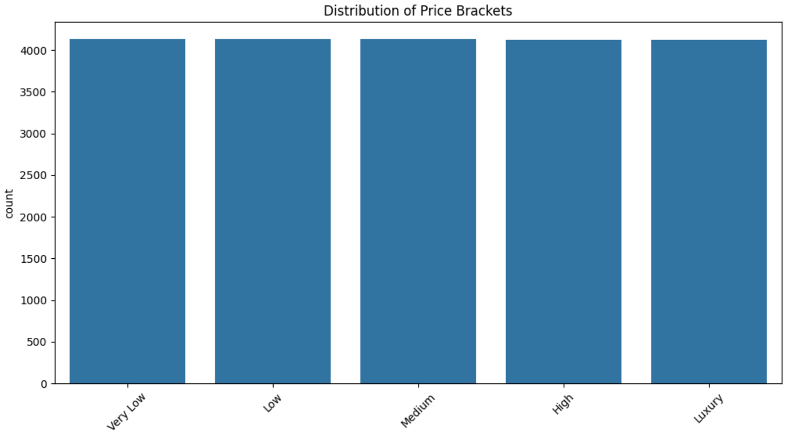

Figure 9: Distribution of properties across price brackets

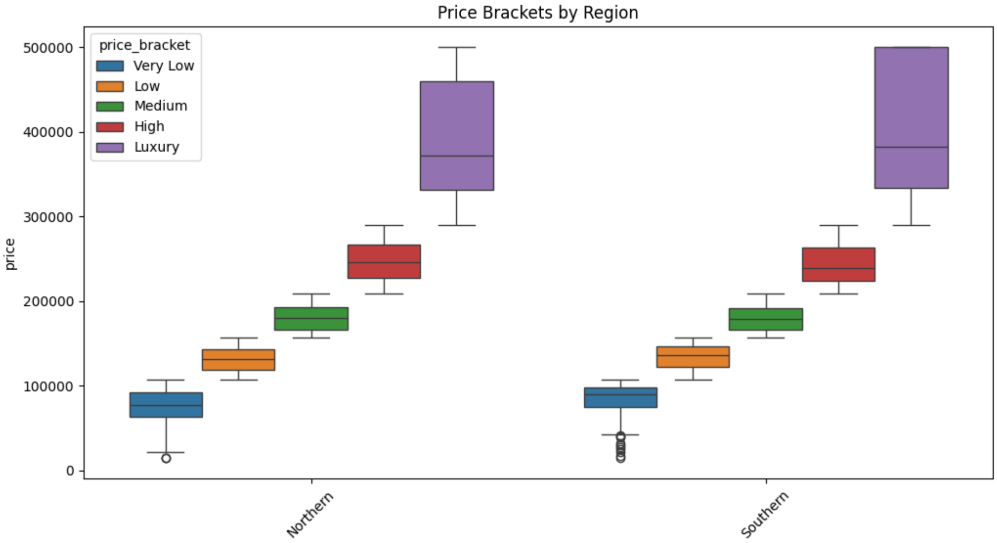

Figure 10: Regional market characteristics by price bracket

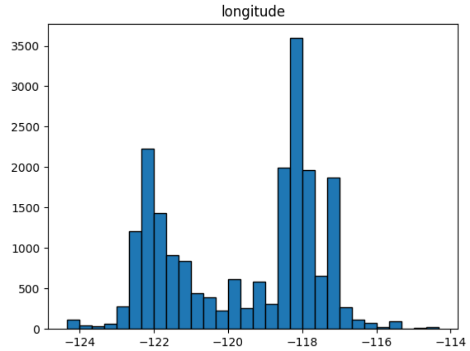

Figure 11: Longitude distribution

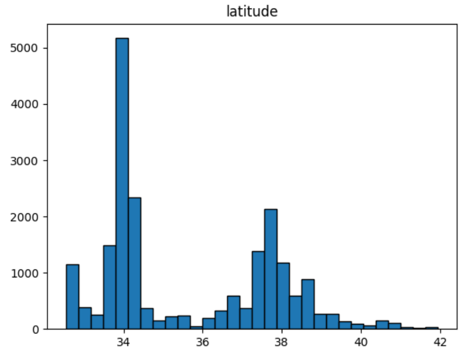

Figure 12: Latitude distribution

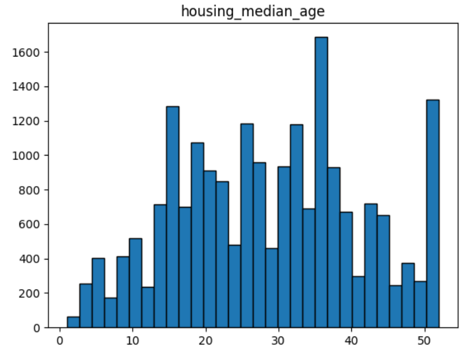

Figure 13: Housing median age distribution

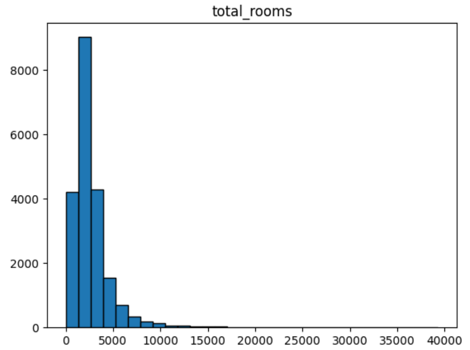

Figure 14: Total rooms distribution

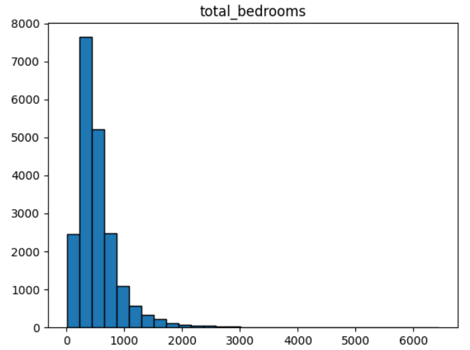

Figure 15: Total bedrooms distribution

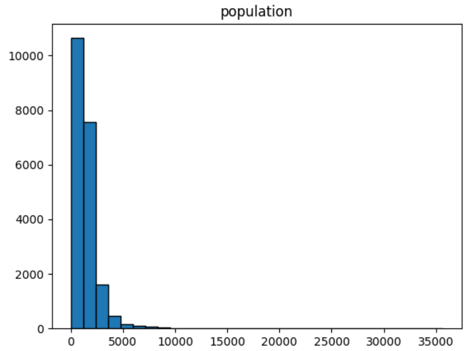

Figure 16: Population distribution

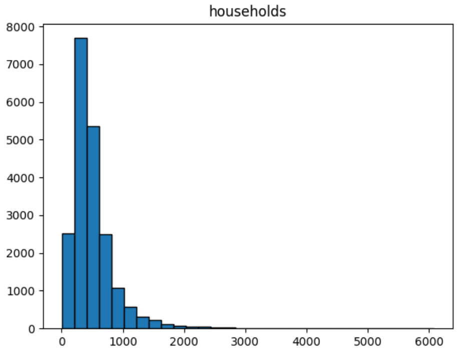

Figure 17: Households distribution

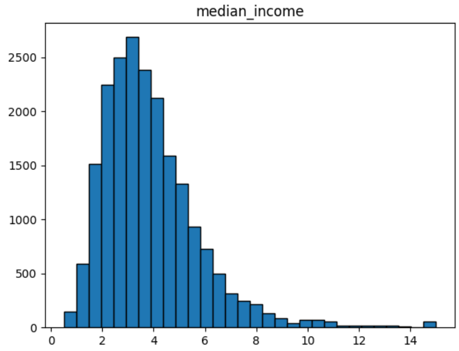

Figure 18: Median income distribution

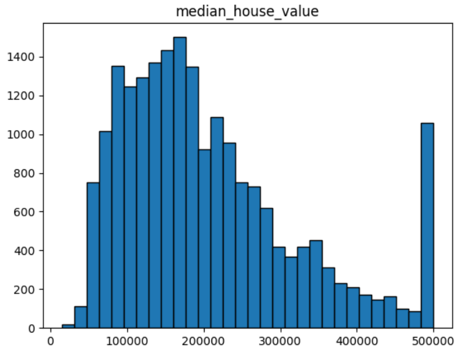

Figure 19: Median house value distribution

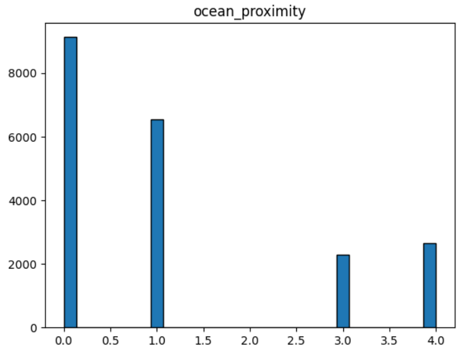

Figure 20: Ocean proximity distribution