California Housing Prices Dataset
Analysis

Maria Vrana

Contents

1 Dataset Overview

The dataset, California Housing Prices, used in the book "Hands-On Machine Learning with Scikit-Learn and TensorFlow" by Aurélien Géron, is based on the 1990 California census. It provides a beginner-friendly introduction to machine learning with manageable data size and basic preprocessing needs.

It includes information about houses in California districts, with summary statistics like:

- Geographical data: longitude, latitude (numerical variables)
- Housing statistics: housing_median_age, total_rooms, total_bedrooms, population, households, median_income, median_house_value (numerical variables)
- Categorical variable: ocean_proximity

The dataset contains 20,640 entries, with 207 missing values in the total_bedrooms feature.

2 Main Analysis

2.1 Encoding Categorical Data and Handling Missing Values

The one-hot encoding method was chosen to encode the ocean_proximity feature because the dataset is relatively compact, and the feature contains only five distinct categories. This approach not only ensures that the

categorical data is effectively represented (ensuring that no ordinal relationships are implied between categories) without significantly increasing the dimensionality of the dataset but also makes it easier to navigate and analyze the dataset by breaking the categories into separate, clearly defined columns.

The one-hot encoding process introduced five new features to the dataset: <1H OCEAN, INLAND, ISLAND, NEAR BAY, and NEAR OCEAN. These features represent the categories of the original ocean_proximity column as separate binary columns, making the data easier to analyze and work with.

The rows containing NaN values were dropped because the dataset has a large number of entries, and removing these rows does not significantly affect the overall statistics. The changes in the mean and median of each feature are minimal, less than 0.04% and 0.05%, respectively. The ISLAND category, which has too few entries (only 5 entries), was not affected by the removal of NaN rows. However, had it been impacted, a different strategy, such as imputation, would have been considered to preserve the integrity of this underrepresented category. The total number of entries has been reduced to 20,433.

2.2 Data Visualizations

The geographical data were visualized using scatterplots, where the points were color-coded to represent the values of other features, making it easier to identify spatial patterns and trends. The results are illustrated in Figure 1.

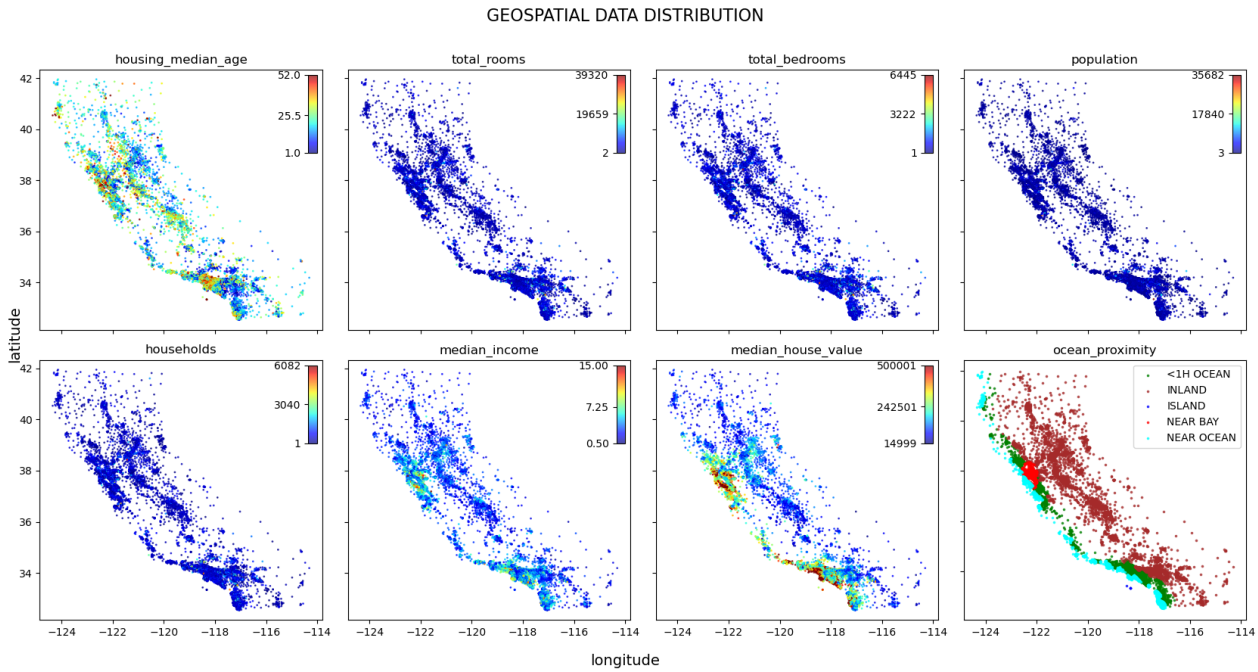

Figure 1: Geospatial distribution of data.

Examining our findings (see Figure 1), we observe that housing_median_age forms clusters of similar values, reflecting distinct periods of high development activity in different regions. Total_rooms, total_bedrooms, population, and households primarily take lower values, with no distinct clusters of large values. Median_income also leans toward the lower side, but there are noticeable groups of higher values near the bay and islands. Similarly, median_house_value is generally higher near the ocean, with the highest values concentrated near the bay and islands.

The data distribution for each feature was visualized using histograms, as presented in Figure 2.

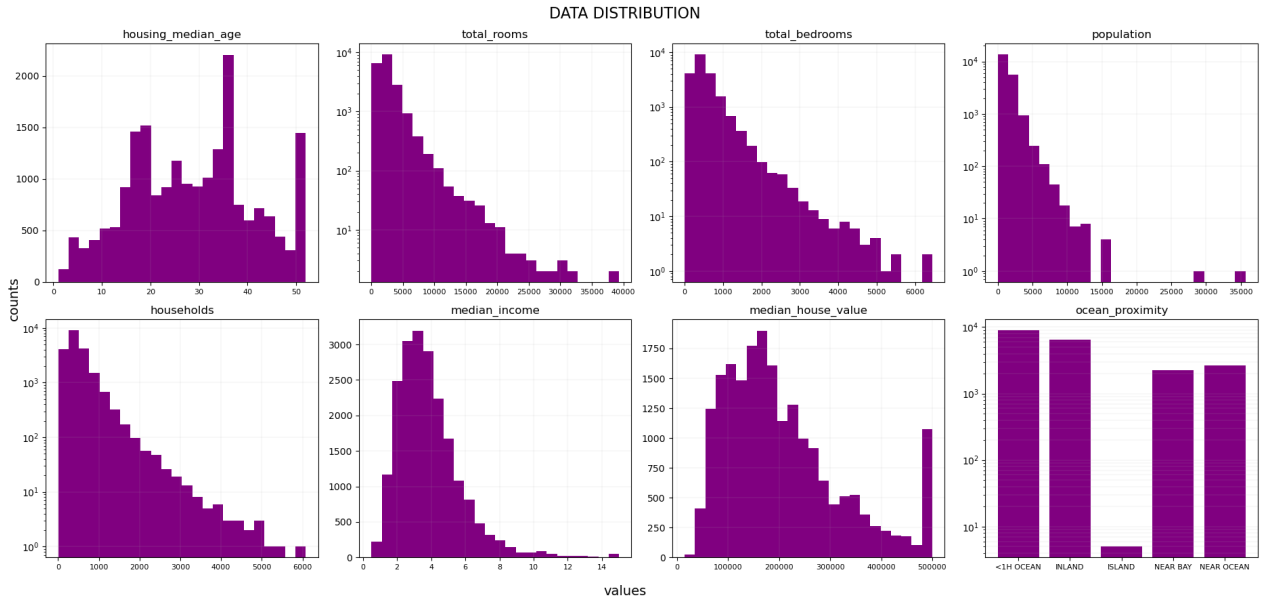

Figure 2: Distribution of data (excluding the geographical features).

The dataset's distribution, as depicted in the histograms (see Figure 2), reveals several key insights. The housing_median_age exhibits a multimodal distribution with distinct peaks around 18, 35, and 50 years, indicating periods of heightened construction activity. The total_rooms, total_bedrooms, population, and households display right-skewed distributions, with most values concentrated in the lower range and a few extreme high values. Logarithmic scaling is applied to enhance visualization. The median income follows a roughly normal distribution with a slight right skew, where most households earn between 2 to 8 income units. The median_house_value presents a bimodal distribution, with a pronounced peak at approximately 500,000, suggesting a potential price cap. The ocean_proximity feature, represented as a bar chart (due to one-hot-encoding there are 5 features in the place of ocean_proximity, this is their combined bar chart), shows that "<1H OCEAN" and "INLAND" are the most common categories, while "ISLAND" remains rare.

2.3 Data Normalization

In most cases, the distribution of the features is not normal (see Figure 2), with many exhibiting skewed or multimodal patterns. Due to this, we opted for normalization instead of standardization, as normalization is better suited for handling non-normally distributed data by scaling values within a fixed range, preserving the original distribution without assuming a specific shape.

The geographical features were not normalized and will not be included in the analysis, as the ocean_proximity features (<1H OCEAN, INLAND, ISLAND, NEAR BAY, and NEAR OCEAN) are more interpretable and already encapsulate some of the information contained in the geographical features.

Examining Figure 3, we observe that housing_median_age tends to be higher near the bay and on the islands, while newer houses are primarily built inland. The chosen normalization method has made it more challenging to extract insights about total_rooms, total_bedrooms, population, and households from the box plot, as these features have a wide range of values with most data points concentrated toward the lower end. Min-max scaling is sensitive to outliers, which affects the interpretation of these features. Additionally, median_house_value appears to be higher on the islands, likely due to their rarity, while the lowest values are found inland.Similarly, median_income is lower in inland areas, indicating a socioeconomic disparity compared to coastal regions.

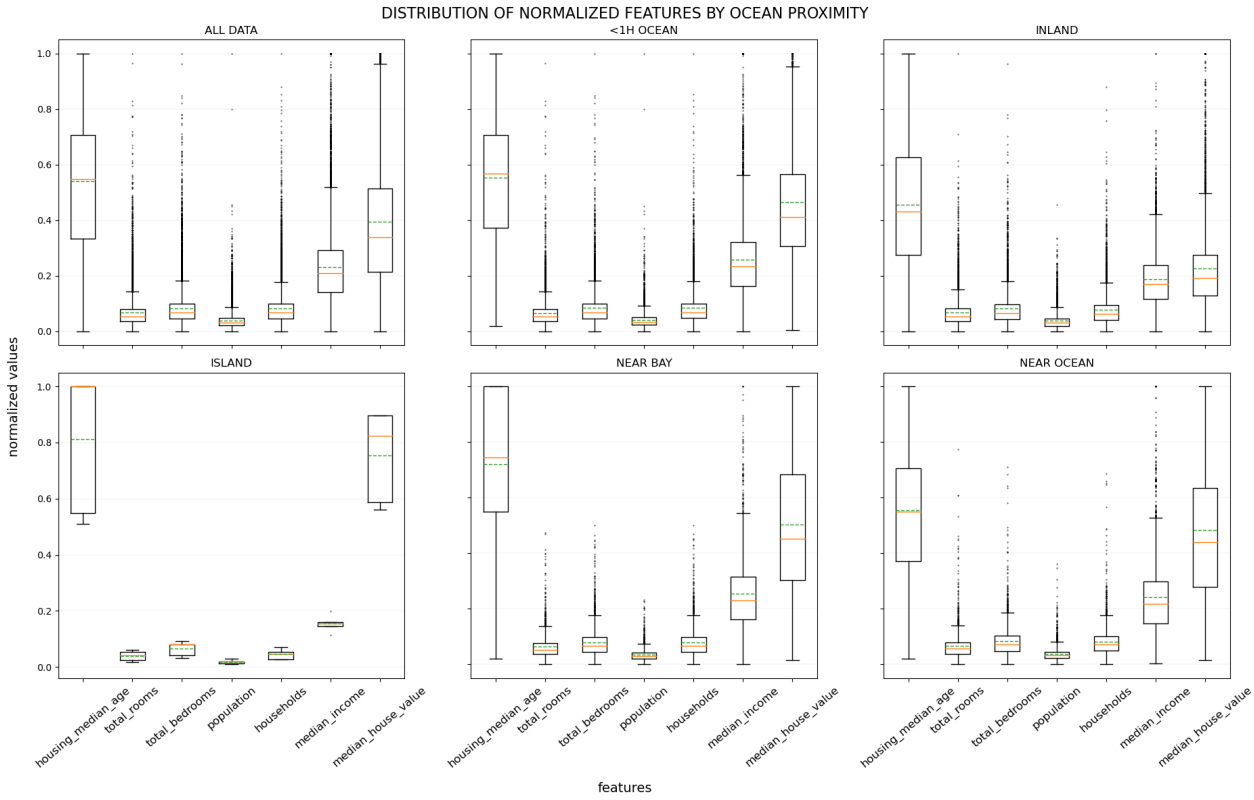

Figure 3: Box plot illustrating the distribution of housing statistics, categorized by ocean proximity.

2.4 Correlations

In this section, we examine the correlations between features. Before conducting the analysis, we anticipate strong correlations among total_rooms, total_bedrooms, population, and households, as these variables are inherently related to the size and occupancy of housing units. For this analysis, we used the Pearson correlation coefficient to measure the strength of relationships between features. Additionally, we calculated the p-values to identify statistically significant correlations, allowing us to filter out those that may have occurred by chance. The results are visualized in Figures 4, 5.

In Figure 5, the correlations that provide no additional information, including the diagonal and the symmetric elements of the matrix, have been removed. Additionally, the statistically non-significant correlations, as determined by their p-values, have been excluded. This ensures that only meaningful, significant correlations are presented.

Upon examining all the data, a weak negative correlation is observed between housing_median_age and total_rooms, total_bedrooms, population, and households. This correlation can be explained by the fact that newer housing developments, such as apartment buildings and multi-unit dwellings, typically have higher densities. These structures are designed to house more people within a smaller footprint compared to older, single-family homes. As a result, newer apartment buildings may have more rooms in total and accommodate more households due to their vertical construction, while older homes tend to have fewer rooms and are often associated with lower population and household numbers. As expected, very strong correlations are found between total_rooms, total_bedrooms, population, and households. Additionally, a strong correlation exists between median_income and median_house_value. This pattern is observed in each individual category, with the exception of the ISLAND, where a significant number of records are lacking to establish meaningful correlations.

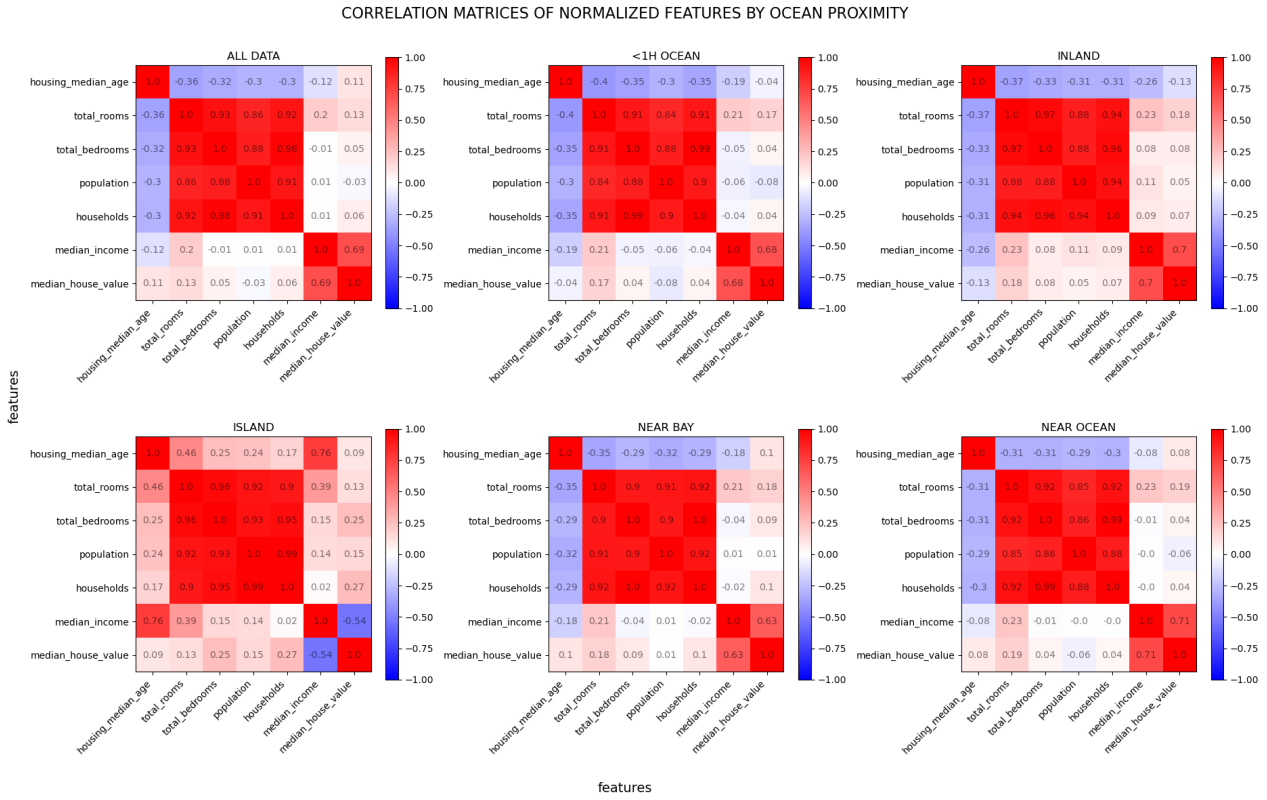

Figure 4: Correlation heatmap illustrating the relationships between housing statistics, categorized by ocean proximity.

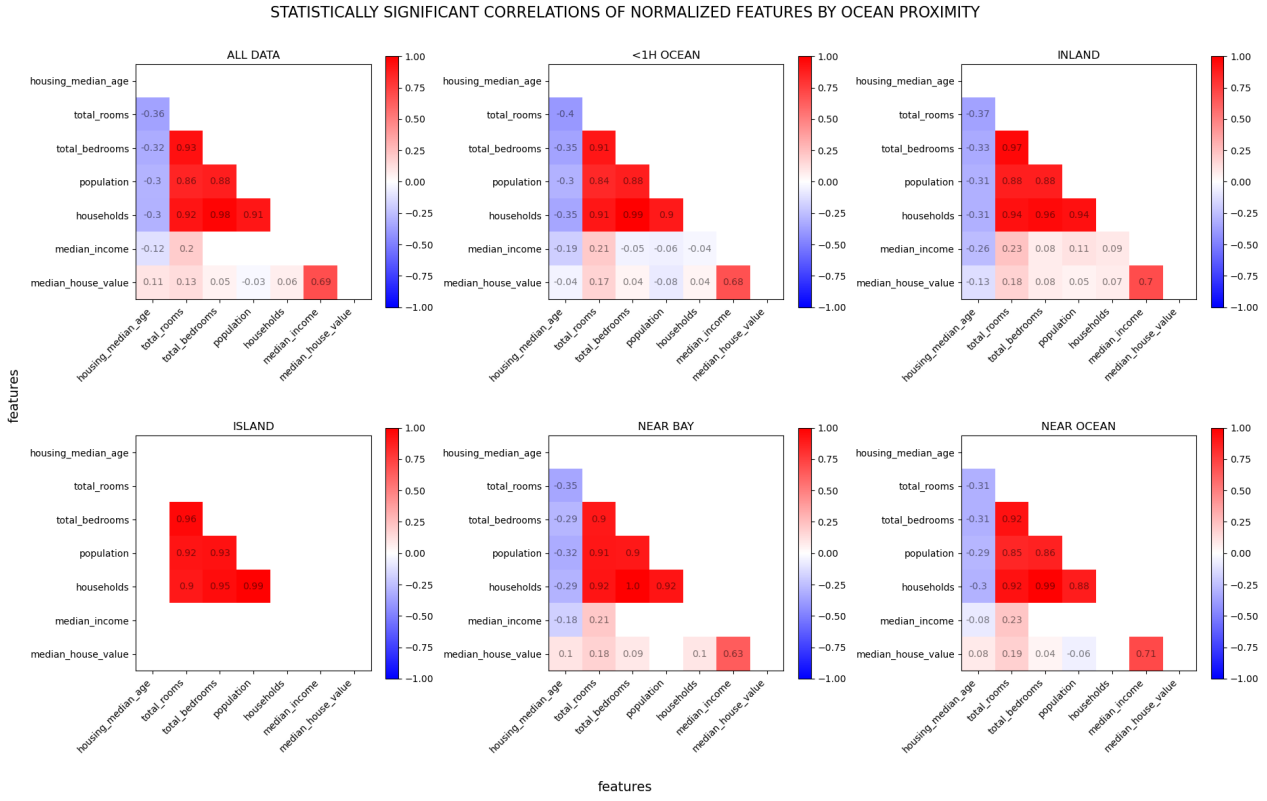

Figure 5: Correlation heatmap illustrating the statistically significant correlations between housing statistics, categorized by ocean proximity.

3 Alternative Analysis

3.1 Feature Engineering and Visualizations

The categorical data and missing values have been handled in a manner similar to the approach described in Section 2.1. In Section 2.4, very strong correlations were observed between total_rooms, total_bedrooms, population, and households. For the alternative analysis, the features total_rooms, total_bedrooms, and population will be replaced with new features: rooms_per_household, bedrooms_per_person, and people_per_household.

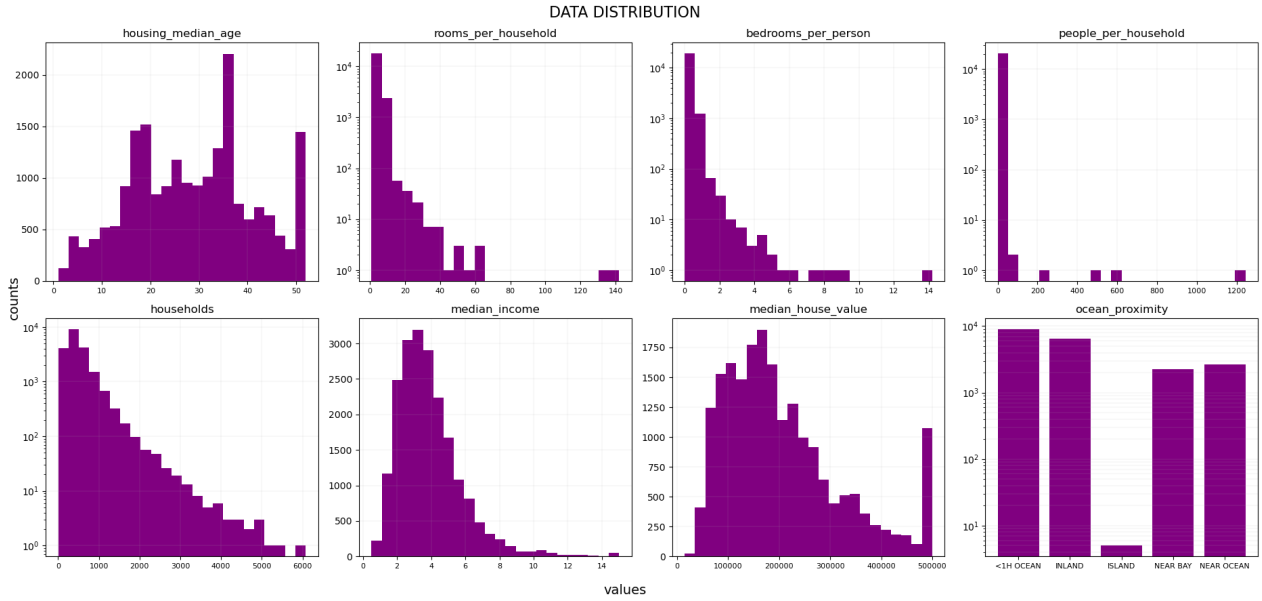

Figure 6: Distribution of data (excluding the geographical features).

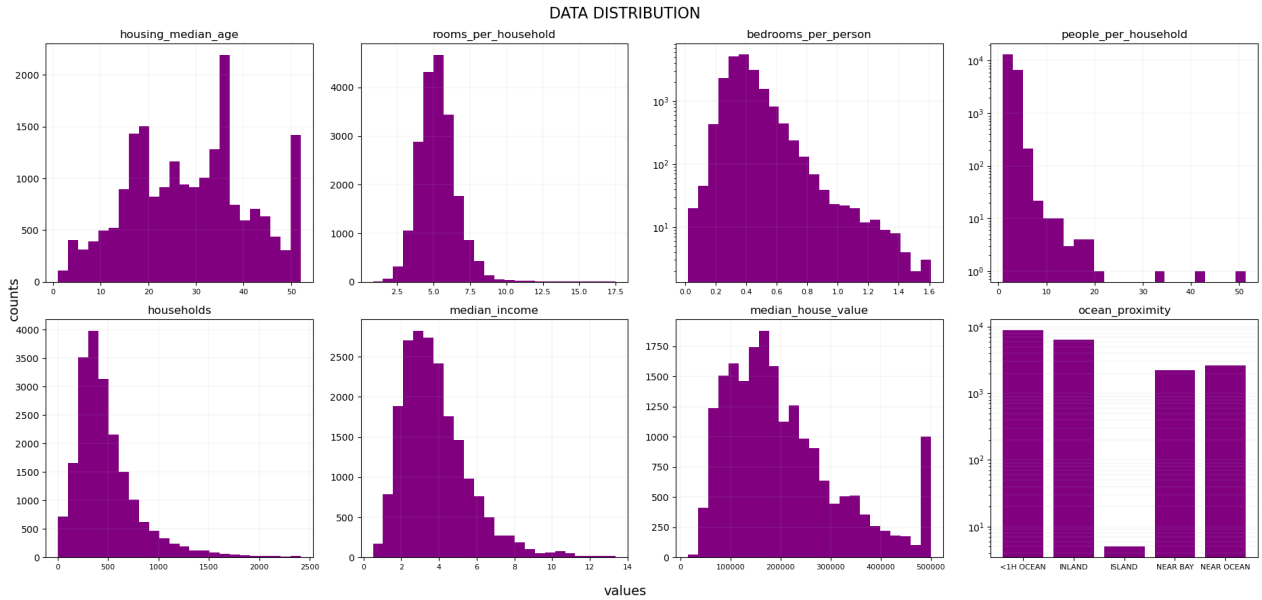

Figure 7: Distribution of data after removing outliers (excluding the geographical features).

Examining Figure 6, we observe that the rooms_per_household, bedrooms_per_person, and people_per_household features display right-skewed distributions, with most values concentrated in the lower range and a few extreme high values. These extreme values are considered outliers. Before removing outliers, we applied a

logarithmic transformation to the households feature, as it was highly right-skewed, ensuring that important values were not mistakenly removed. Outliers were then identified and removed based on their z-scores, with values exceeding a z-score of 5 being excluded, reducing the dataset from 20,433 entries to 20,157. Figure 7 displays the data distribution after applying the logarithmic transformations and removing the outliers as described above.

3.2 Data Normalization

The features were normalized in a manner similar to the approach described in Section 2.3.

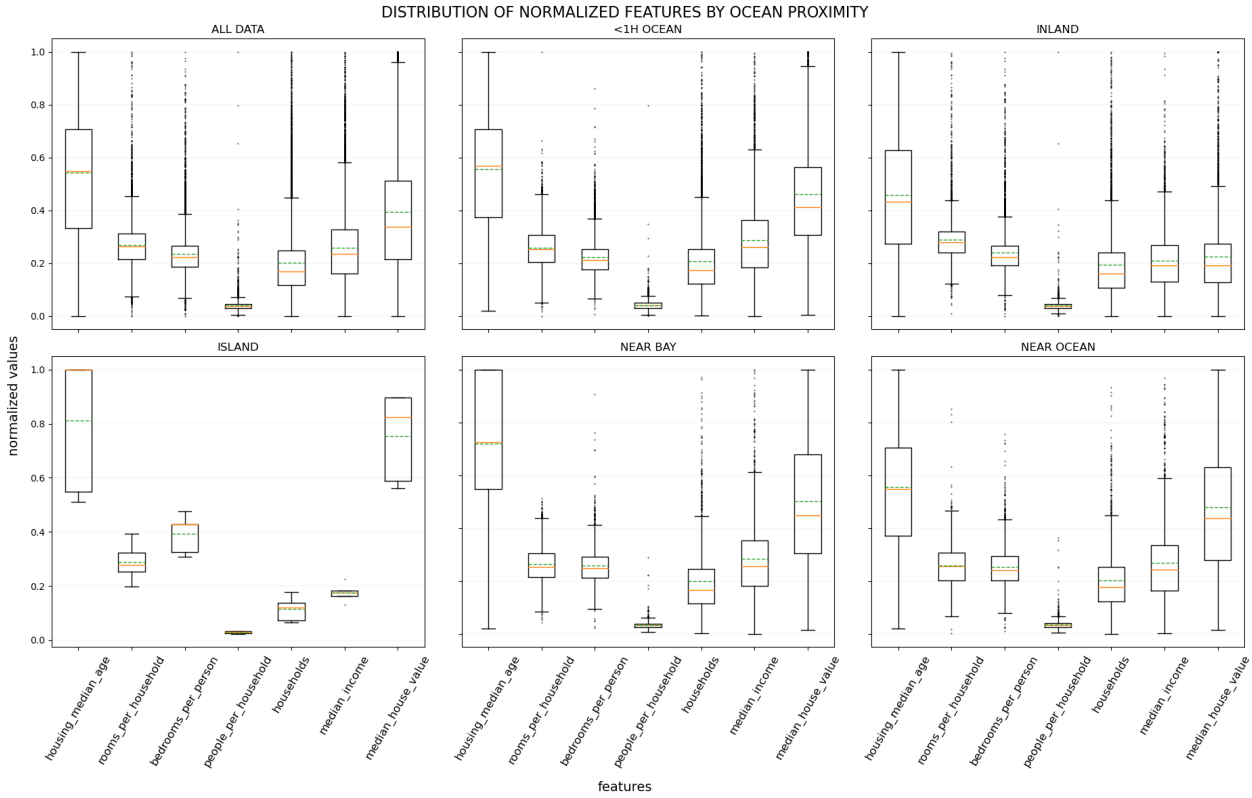

Figure 8: Box plot illustrating the distribution of housing statistics, categorized by ocean proximity.

The highest mean for rooms_per_household and bedrooms_per_person is observed on the islands, while the lowest mean appears inland. This suggests that island homes tend to be larger, with fewer housing units and lower population density, resulting in more rooms and bedrooms per household. For people_per_household, the highest mean is found inland, while the lowest mean is on the islands.

3.3 Correlations

The process described in Section 2.4 is repeated, resulting in the generation of Figures 9 and 10.

Upon examining ALL DATA in Figure 10, we observe a weak negative correlation between housing_median_age and both rooms_per_household and households. This suggests that older houses tend to have fewer rooms and accommodate fewer households. Additionally, there is a strong correlation between rooms_per_household and median_income, as well as a weaker correlation with median_house_value. A strong positive correlation is also observed between median_housing_value and median_income. These correlations imply that higher-value homes typically have more rooms and are located in areas with higher income levels. Lastly, there is a strong negative correlation between people_per_household and bedrooms_per_person, indicating that as the number of people per household increases, the number of bedrooms per person decreases.

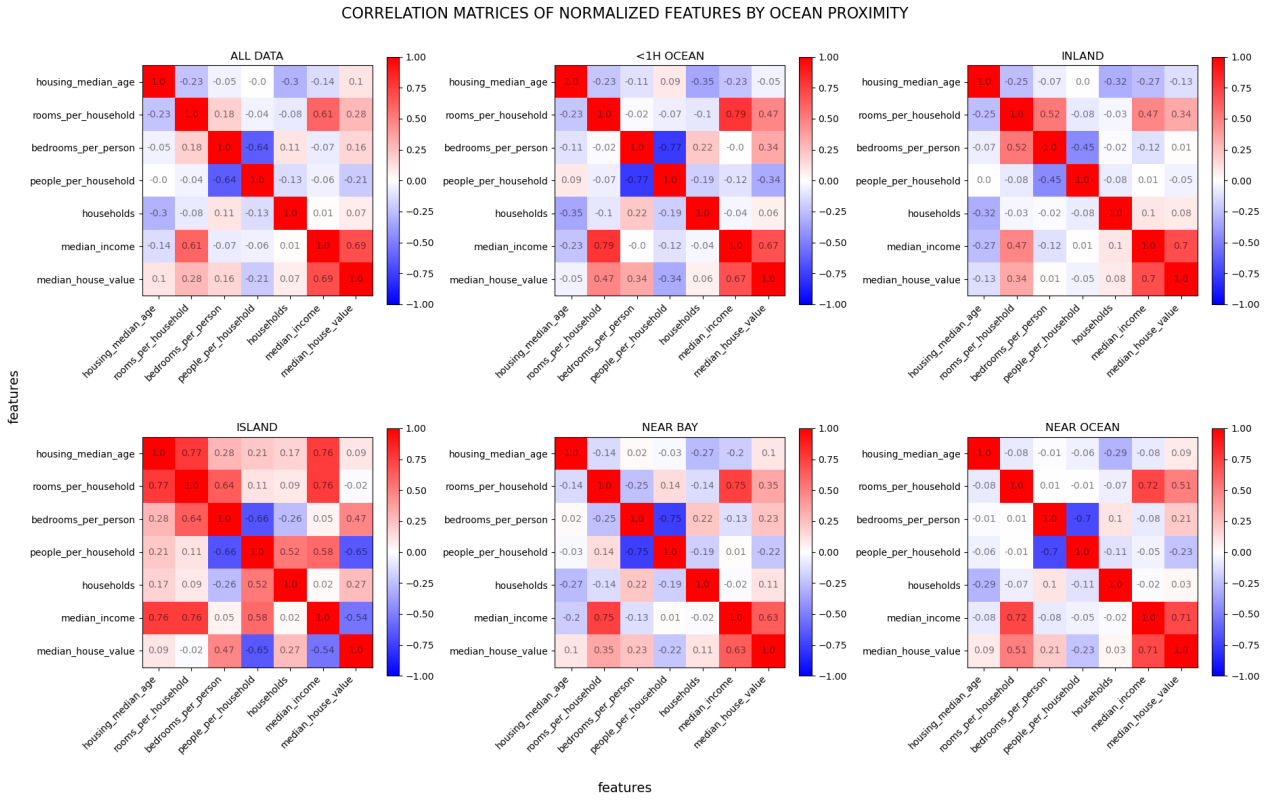

Figure 9: Correlation heatmap illustrating the relationships between housing statistics, categorized by ocean proximity.

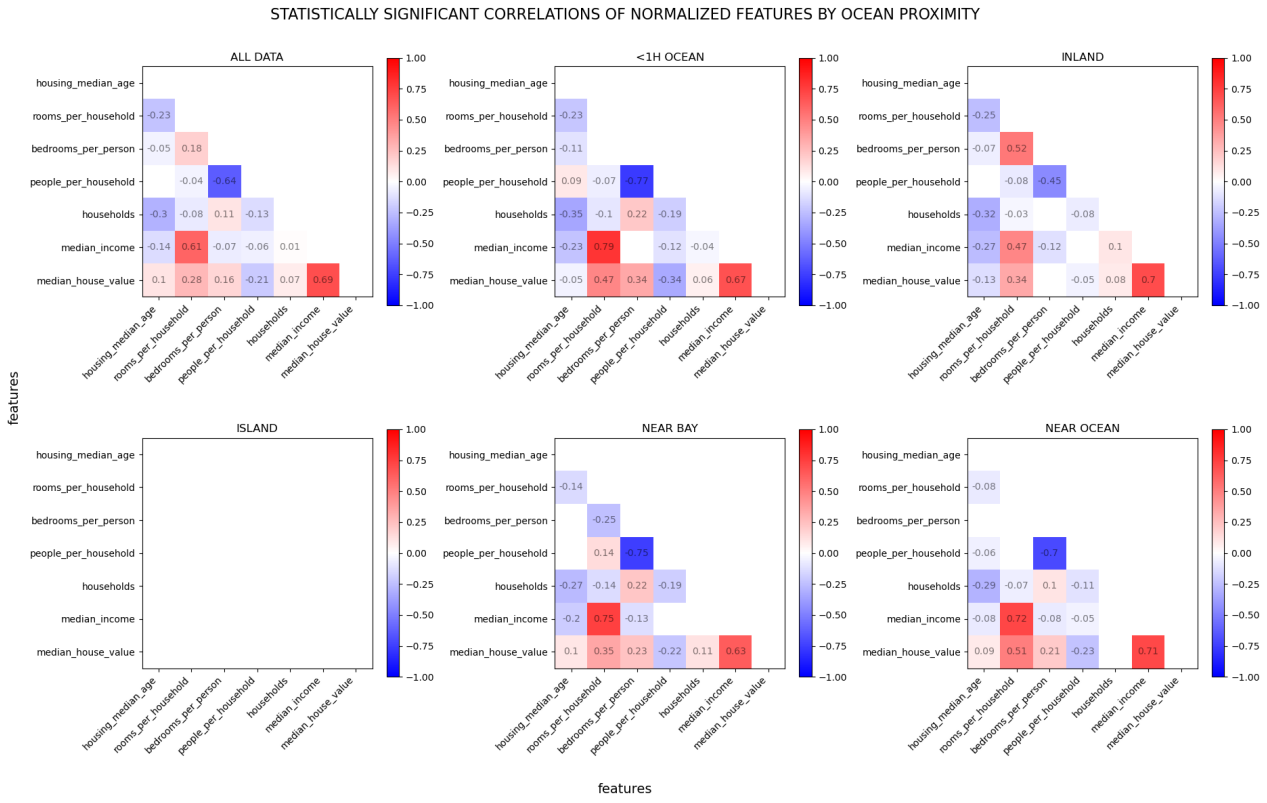

Figure 10: Correlation heatmap illustrating the statistically significant correlations between housing statistics, categorized by ocean proximity.

For the <1H OCEAN category, we observe a stronger negative correlation between median_income and housing_median_age compared to ALL DATA, as well as a stronger positive correlation with rooms_per_household, highlighting their relationship. The correlation between median_house_value and rooms_per_household is moderate, making it stronger than in ALL DATA. The negative correlation between people_per_household and bedrooms_per_person is more pronounced in this category. The remaining correlations closely mirror those observed in the ALL DATA category.

For the INLAND category, there is a weak negative correlation between median_income and housing_median_age. The negative correlation between median_house_value and housing_median_age is also weak. The correlation between bedrooms_per_person and rooms_per_household is moderate and significantly stronger than in ALL DATA. Meanwhile, the negative correlation between people_per_household and bedrooms_per_person is moderate but weaker than in ALL DATA. The other correlations are largely consistent with those in the ALL DATA category.

There are no statistically significant correlations for the ISLAND category due to the limited sample size, so we move on to the NEAR BAY category. Here, the negative correlation between rooms_per_household and housing_median_age is very weak, even weaker than in ALL DATA. Additionally, there is a weak negative correlation between housing_median_age and both households and median_income. A weak negative correlation between bedrooms_per_person and rooms_per_household suggests the presence of shared housing spaces or smaller one-space homes. There is also a weak correlation between households and bedrooms_per_person, as well as between median_house_value and bedrooms_per_person. The rest of the correlations closely align with those in the ALL DATA category.

For the NEAR OCEAN category, similar to NEAR BAY, the negative correlation between rooms_per_household and housing_median_age is very weak. There is a moderate positive correlation between median_house_value and rooms_per_household, as well as a weak positive correlation between median_house_value and bedrooms_per_person. The remaining correlations are similar to those found in the ALL DATA category.

4 Summary

The dataset was processed using one-hot encoding for the ocean_proximity feature (see Section 2.1), which created five new binary columns. This encoding method transformed the categorical variable into multiple binary attributes, making it suitable for machine learning models. In addition, missing values in the total_bedrooms feature were removed. Given the large size of the overall dataset, this removal had a negligible impact on the statistical properties and did not significantly affect the analysis.

Histograms revealed that many features exhibited non-normal distributions (see Section 2.2). To address this, min-max normalization was initially applied instead of standardization (see Section 2.3), as it scales the data to a range of [0, 1]. However, due to the sensitivity of min-max normalization to outliers, which can complicate the interpretation of certain features, some additional steps were taken. In an alternative approach (see Section 3.2), we applied logarithmic scaling to certain features to reduce skewness and improve interpretability. Additionally, some outliers were removed to mitigate their impact on the analysis.

Pearson correlation analysis confirmed strong correlations between total_rooms, total_bedrooms, population, and households (see Section 2.4), which was expected. However, these strong correlations did not provide additional meaningful information, as they essentially capture overlapping data regarding the number of households and population. To address this redundancy and extract more valuable insights, in an alternative approach(see Section 3.3), we combined these features into new variables — rooms_per_household, bedrooms_per_person, and people_per_household — while retaining the original households feature for reference.

The results indicate that housing characteristics, location, and economic factors significantly influence median house value. Coastal regions, particularly near the bay and islands, tend to have higher-income populations

and more expensive homes, whereas inland areas generally feature lower incomes and more affordable housing. Older homes are concentrated along the coast, while newer developments are more common inland, where there is a stronger negative correlation between housing age and value. Among all factors, median income emerges as the strongest predictor of median house value, highlighting the close relationship between economic status and housing affordability.