Topic: Housing Price Prediction System

Table of content:

1. Introduction

The goal of this project is to predict housing prices based on multiple factors such as location, income, house attributes and ocean proximity. Combining machine learning techniques and implementing them with various models helps identify a relationship between the factors and the housing price. This helps investors and buyers in making informed decisions.

The problem this project addresses is the challenge of estimating housing prices in a dynamic environment considering various factors. Traditional methods sometimes fail to predict the prices since the data can be complex and many factors can affect it. This project uses data-driven techniques to overcome the limitations faced by traditional ways and ensures a precise and reliable prediction.

The interest behind this project derives from the growing need for data-driven decisions in the real estate industry. Accurate price prediction can help organizations to identify trends and put a proper valuation on the properties.

2. Dataset Description

The dataset used in this project has data on the median house prices for California districts derived from the 1990 census. The dataset is sourced from kaggle.com website. And it's a publicly available dataset.

Source:https://www.kaggle.com/datasets/camnugent/california-housing-prices

Features: The dataset contains the following features.

Feature	Description
longitude	Geographic coordinate specifying the east-west position of a house.
latitude	Geographic coordinate specifying the north-south position of a house.
housing_median_age	Median age of houses in a particular block.
total_rooms	Total number of rooms in a block.
total_bedrooms	Total number of bedrooms in a block (with missing values handled via imputation).
population	Total population residing in a block.
households	Total number of households in a block.
median_income	Median income of residents in a block (scaled for privacy).
median_house_value	Median house price in a block (target variable).
ocean_proximity	Categorical variable indicating the block's proximity to the ocean.

Target Variable: The target of the project is to predict the median_house_value feature which represents the median price of houses in a block.

Data Characteristics:

Size:
There are 10 columns and 20641 rows of data.

Type:
Mixture of numerical and categorical data.

Missing Values:
Present in total_bedrooms column.

Categorical Variables:
ocean_proximity column contains categorical values.

Number of features:
The dataset contains 10 features.

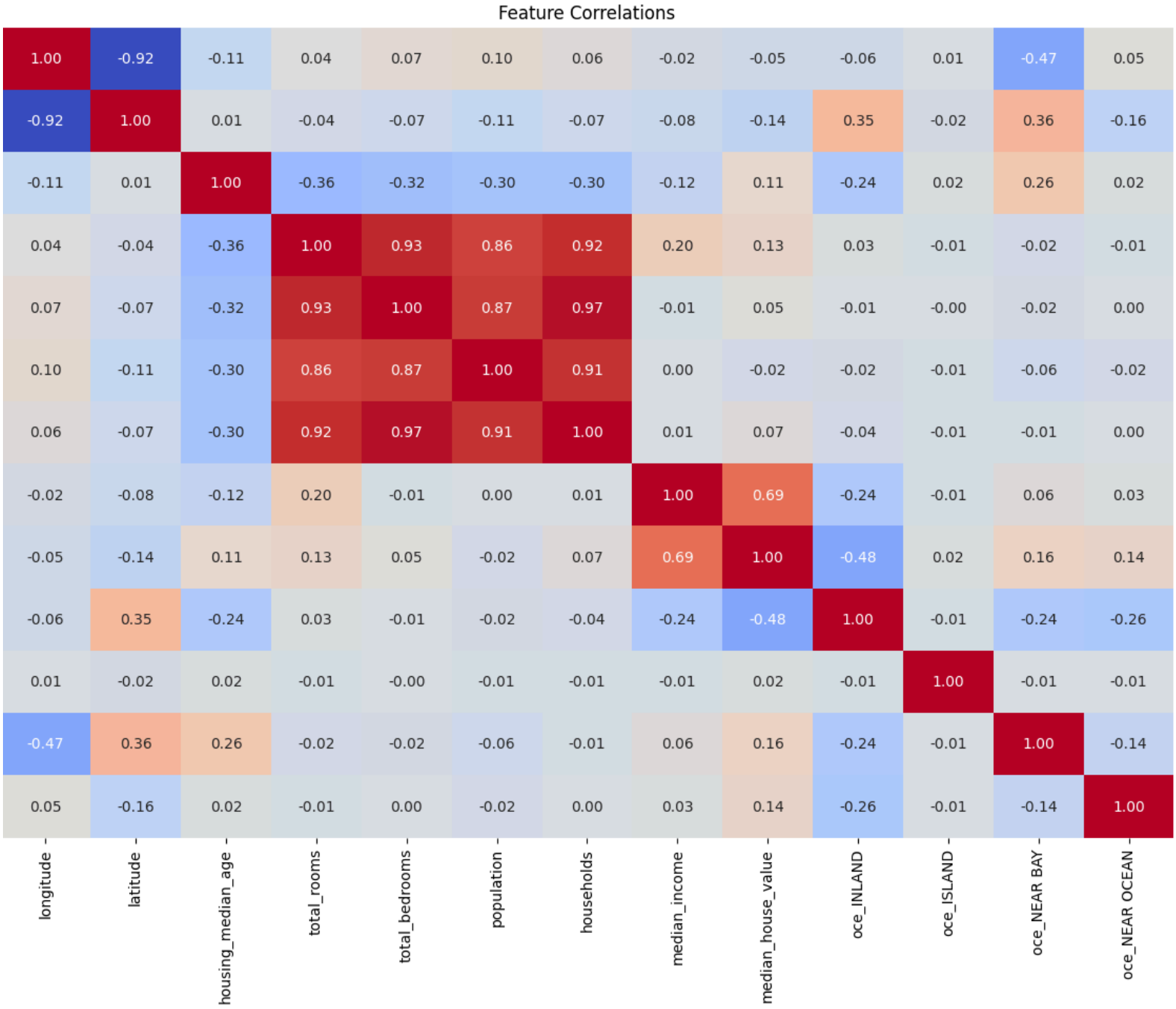
Type of Problem:
This is a regression based problem because the target variable median_house_value is continuous and it is a numeric type of data. The median house value represents house prices which is this project's goal to predict.

Types of Features:

i. Quantitative:
longitude, latitude, housing_median_age, total_rooms, total_bedrooms, population, households, median_income, median_house_value.

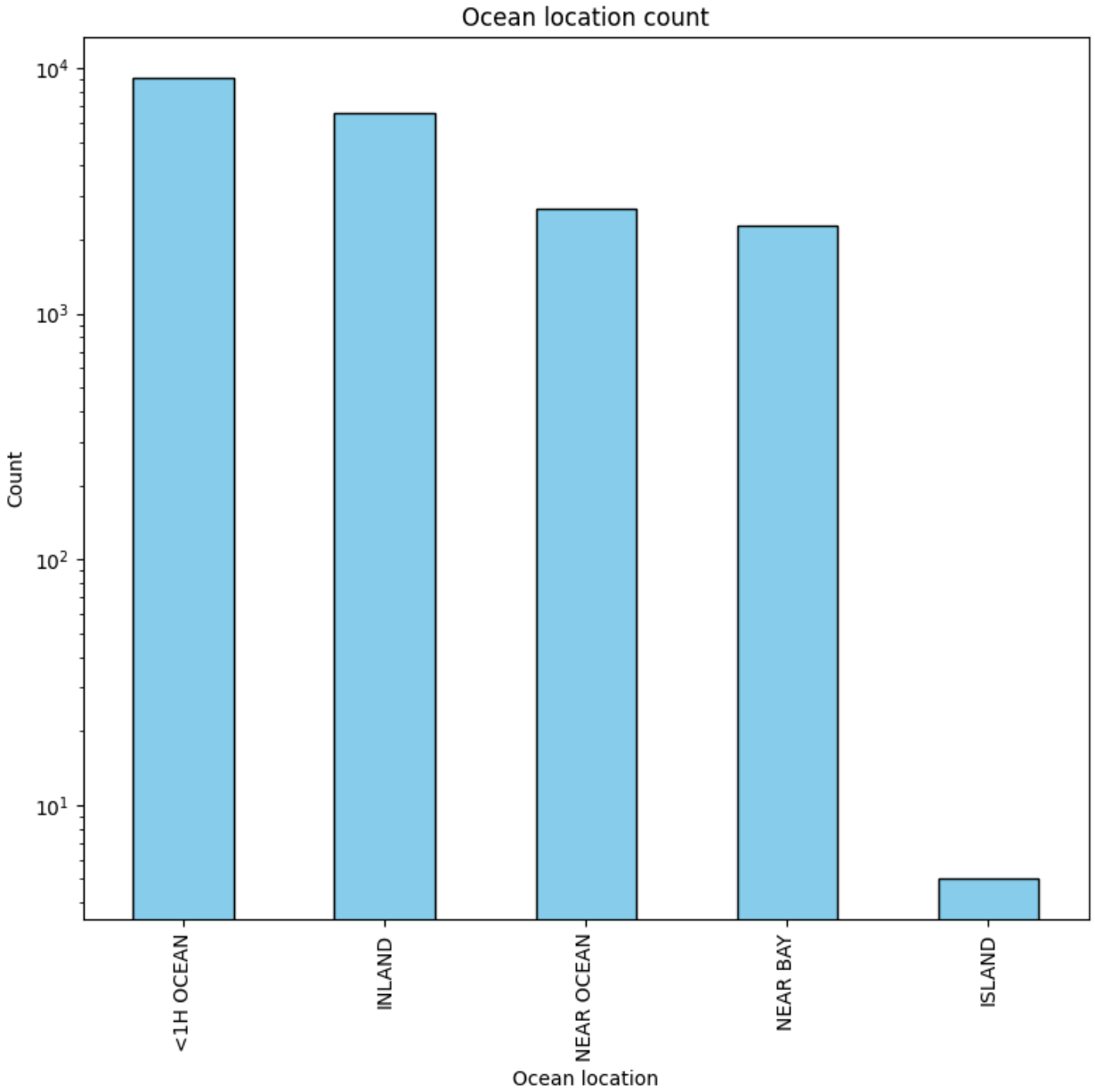
ii. Categorical:
ocean_proximity.

Correlation of Features:

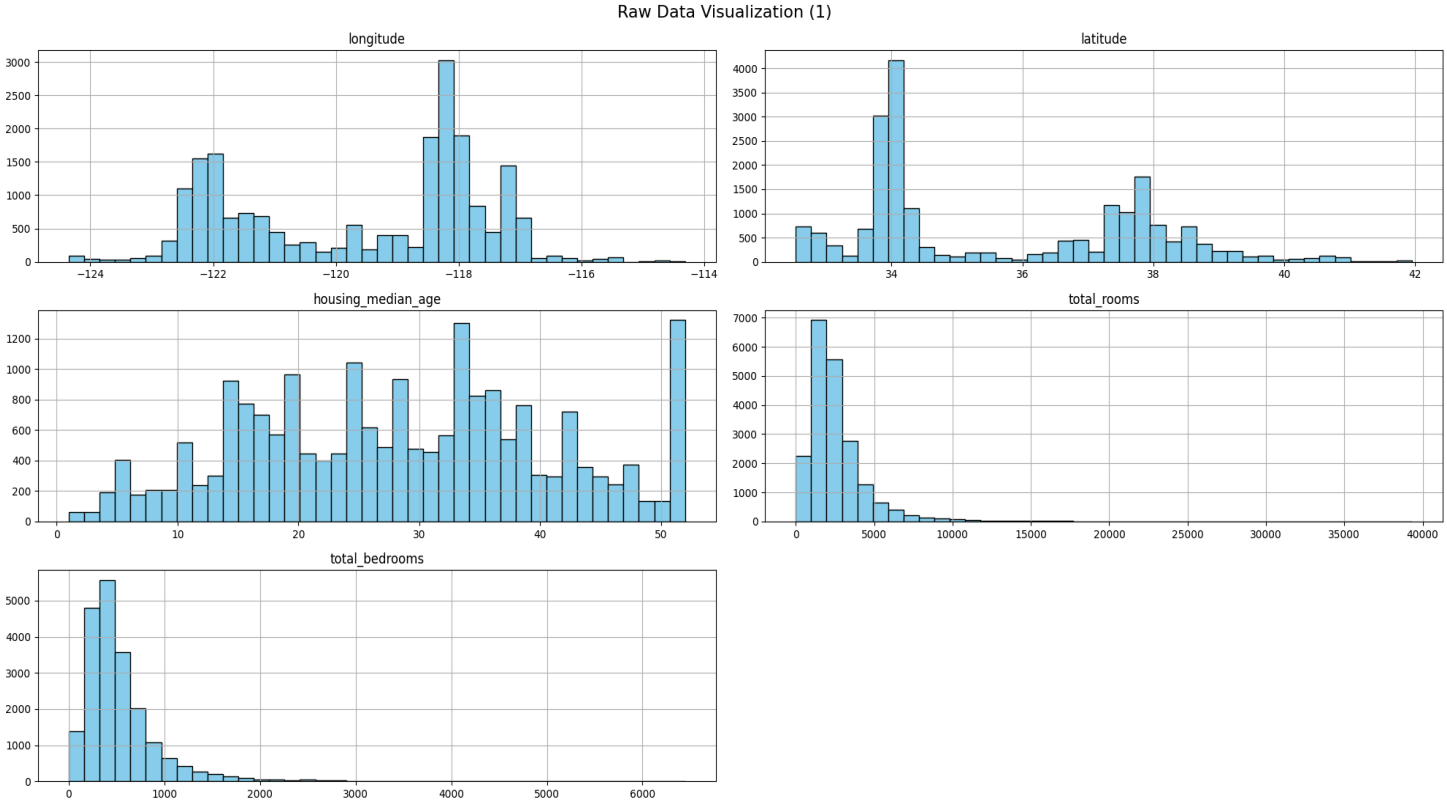


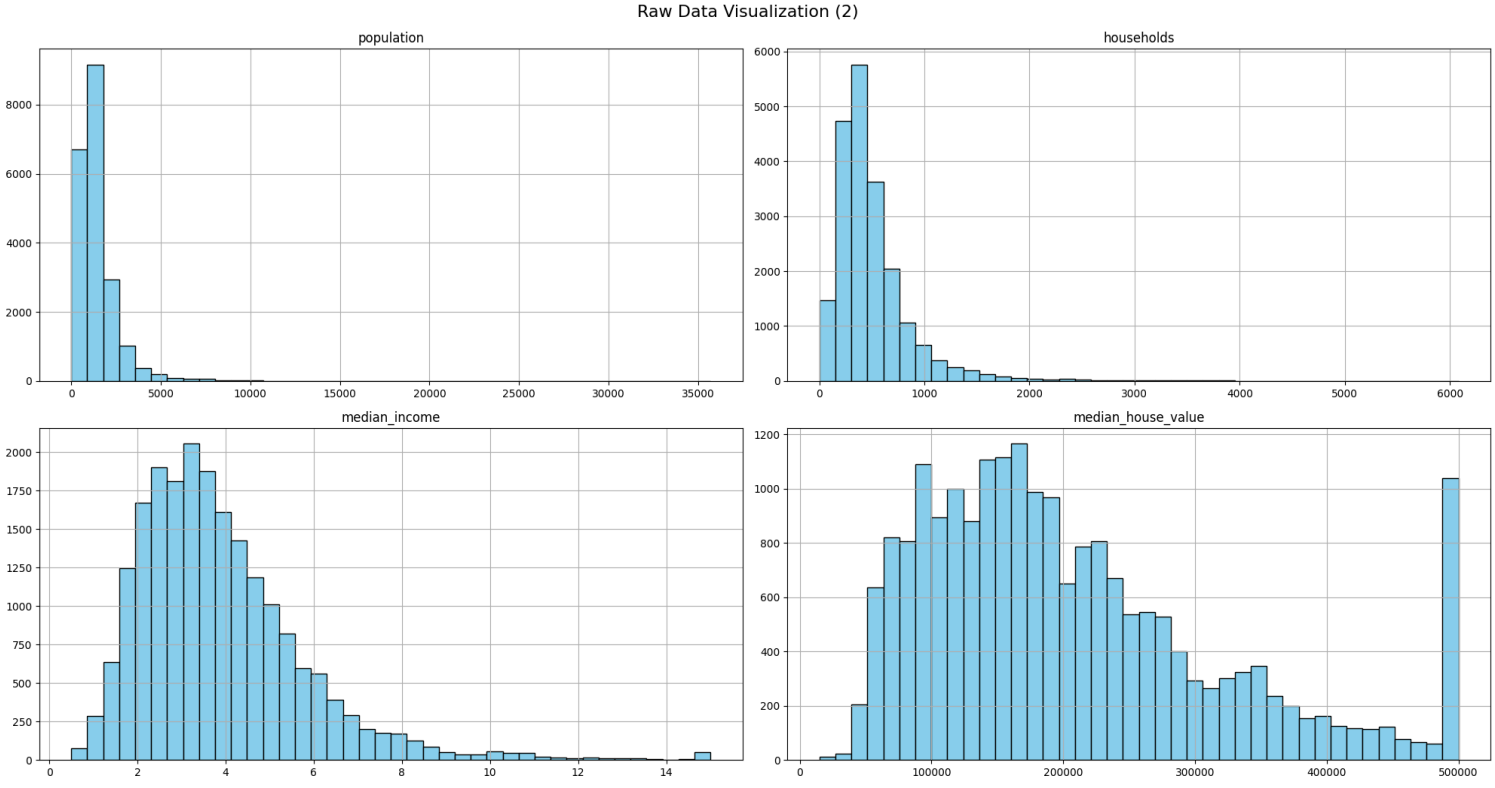
Imbalanced Dataset

For categorical features like ocean_proximity, The representation of unique categories having an equal number of instances or not has been done using a bar chart in log scale.



3. Data Visualization





4. **Dataset Pre-Processing**:

In the dataset, we had null values in total_bedrooms. Missing values in this column were imputed using the median and after imputation no null or missing values were there in the dataset.

Categorical variable ocean_proximity has been encoded using one-hot encoding in the dataset. one-hot encoding has been done on the ocean_proximity column to convert its categorical data into multiple binary (0/1) columns, each for their unique category like Inland, <1H Ocean. This transformation ensures that the data is numerical and it is necessary for the machine learning models to factor in this category for more accurate prediction. This ensures that the model interprets the categorical data in a structured and meaningful manner. Here are the first five rows on how the one-hot encoding makes columns.

<1H OCEAN	INLAND	ISLAND	NEAR_BAY	NEAR_OCEAN
False	False	False	True	False
False	False	False	True	False
False	False	False	True	False
False	False	False	True	False
False	False	False	True	False

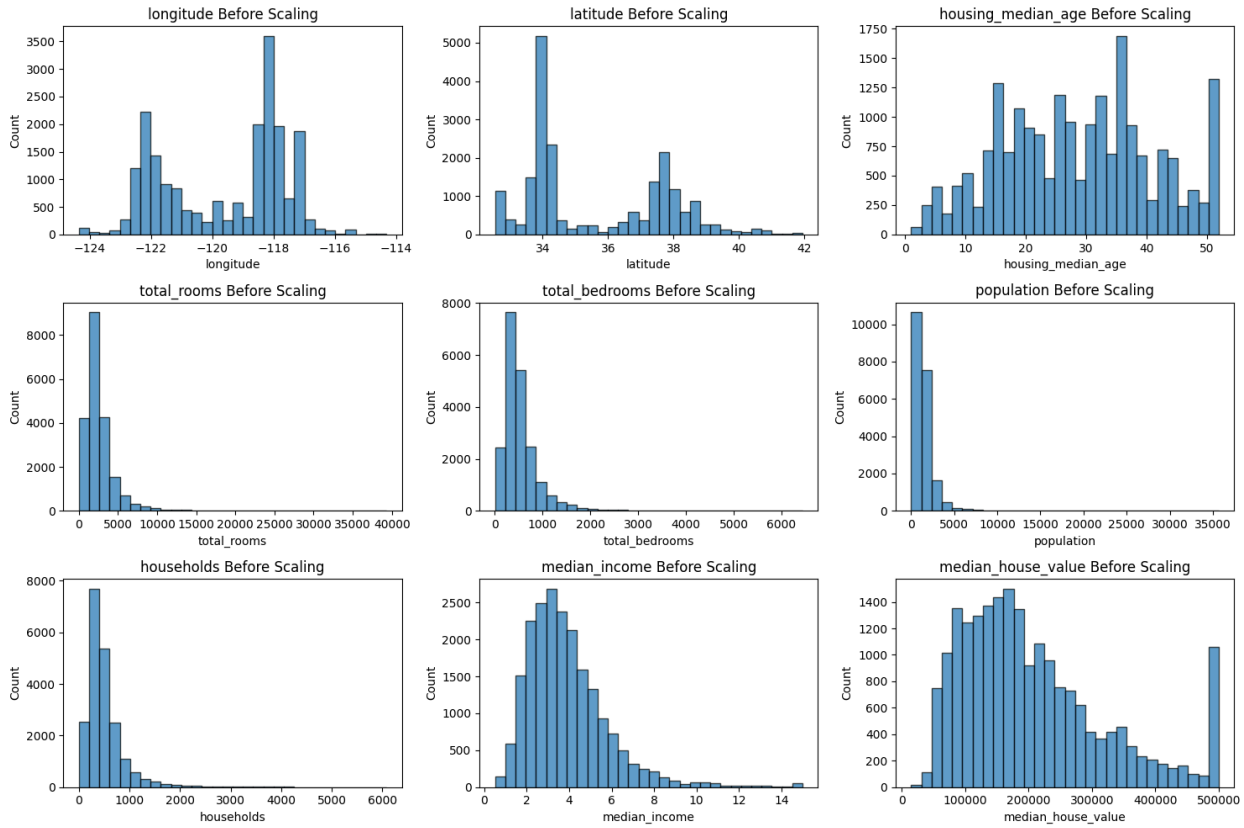
5. Feature Scaling:

Feature scaling has been done to standardize and normalize the input features to ensure all the features have comparable scale. The formula used for this is

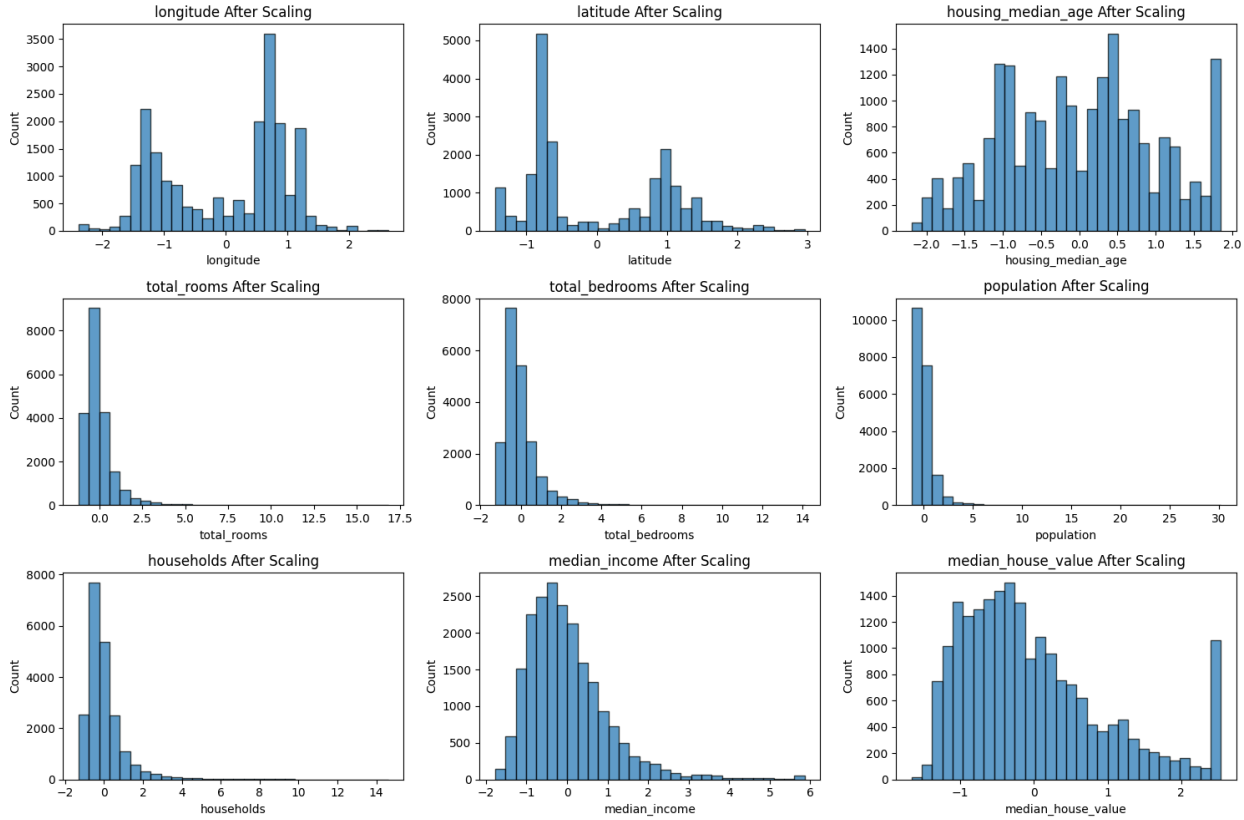
$$z = x - \mu / \sigma$$

Here, x is the original feature from the dataset, μ is the mean of the feature and σ is the standard deviation of the feature.

Plots of the feature before scaling:



Plots of the feature after scaling:



6. Dataset Splitting

The dataset has been split, and 70% has been given for training and 30% has been assigned for testing the dataset.

7. Model Training and Testing

The models used in this project are:

- Decision Tree
- Linear Regression
- SVR Analysis

8. Comparison Analysis

Implementing the models on the dataset, we got the following statistics for each model:

A. Decision Tree

Result:
Coefficient of Determination (R^2): Testing: 0.6337039516467126, Training: 1.0
Training MSE: 4.859656585697901e-32
Testing MSE: 0.3610665363282769 5
Mean Absolute Percentage Error (MAPE): 0.3814424737524185
Cross-Validation Score: -0.3820203106733014 4

B. Linear Regression

Result (Linear Regression)
Coefficient of Determination (R^2): Testing: 0.6393611711434395, Training: 0.6470480227253683
Training MSE: 0.3551111393738062 3
Testing MSE: 0.355490083461491 2
Mean Absolute Percentage Error (MAPE): 0.4339793296167354 7
Cross-Validation Score: -0.3577179441308399

Result (LASSO Regression)
Coefficient of Determination (R^2): Testing: -3.9309721393543384e-06, Training: 0.0
Training MSE: 1.0061174387400997
Testing MSE: 0.985727138728307 3

Mean Absolute Percentage Error (MAPE): 0.7853181385120985
Cross-Validation Score: -1.0063472314048727

Result (Ridge Regression)
Coefficient of Determination (R^2): Testing: 0.6393006195303357, Training: 0.6470327158530575
Training MSE: 0.3551265398849707
Testing MSE: 0.3555497705951929
Mean Absolute Percentage Error (MAPE): 0.434032840814229 27
Cross-Validation Score: -0.357748620052376 84

C. Support vector regression (SVR) analysis

Result
SVR Coefficient of Determination (R^2): Testing: 0.7563448795795759, Training: 0.7735179566321192
Training MSE: 0.227867533393916 44
Testing MSE: 0.240176520561314 97
Mean Absolute Percentage Error (MAPE): Training: 1.9126891527728367, Testing: 1.2013576648002484
Cross-Validation Score: 0.6686961243555268

Conclusion

Based on the analysis using Decision Tree, Linear Regression and Support Vector Regression(SVR), the best performing model in terms of performance is Support Vector Regression (SVR). SVR showed the highest Coefficient of Determination (R^2) for both

testing (0.7563) and training (0.7735),which indicates that SVR explains significant portion of the variance in the dataset.

Linear Regression and Ridge Regression with R^2 values close to 0.64 suggests they also performed well for testing and training dataset.

Decision Tree had a high training R^2 (1.0) but the testing performance R^2 value of 0.633 indicates overfitting. The MSE and cross-validation scores were also quite low which suggests that it has limited generalization to unseen data.

LASSO Regression performed worst with a very low R^2 and high Mean Absolute Percentage Error (MAPE) resulting in over-regularization.

Through decision trees and linear regression the project identified key features like median_income, ocean_INLAND, and longitude as the most significant predictors for the model.

Model Evaluation: SVR not only demonstrated superior performance but also showed good generalization capability in both training and testing phase of the dataset. The cross validation score of 0.668 confirms that it's the most consistent. Decision tree overfitted to training data while Ridge regression provided a good balance between bias and variance.

Overall, Support Vector Regression stands out as the most reliable model for predicting the target variable with its high accuracy and lowest error rates.