California Housing Dataset Analysis and Modeling Report

1 Introduction

This report presents a comprehensive analysis of the California Housing dataset, which contains information about housing in various districts across California. The goal is to understand the factors influencing housing prices and develop predictive models to estimate median house values. Through data preprocessing, exploratory analysis, feature engineering, and model development, we aim to create an accurate prediction system for California's housing market.

2 Dataset Overview

The dataset consists of 20,640 entries, with each entry representing a block group (a geographical unit) in California. It contains 10 features:

- **Geographical features**: longitude, latitude
- **Housing features**: housing_median_age, total_rooms, total_bedrooms
- **Demographic features**: population, households, median_income
- **Categorical feature**: ocean_proximity (NEAR BAY, ¡1H OCEAN, INLAND, NEAR OCEAN, ISLAND)
- **Target variable**: median_house_value

The data provides a comprehensive view of housing characteristics across different California regions, allowing for analysis of geographical, demographic, and economic factors affecting housing prices.

3 Exploratory Data Analysis (EDA)

Initial examination revealed:

- The dataset has 20,640 rows and 10 columns.

- Most features are numerical (9 columns) with one categorical feature (ocean_proximity).
- The ocean_proximity feature has 5 categories with the following distribution:
  - ¡1H OCEAN: 9,136 entries
  - INLAND: 6,551 entries
  - NEAR OCEAN: 2,658 entries
  - NEAR BAY: 2,290 entries
  - ISLAND: 5 entries

Correlation analysis showed significant relationships:

- Strong positive correlations between:
  - total_rooms & total_bedrooms (0.927)
  - households & total_bedrooms (0.974)
  - population & households (0.907)
  - median_income & median_house_value (0.688)
- Strong negative correlations between:
  - longitude & latitude (-0.925)
  - median_house_value & ocean_proximity_INLAND (-0.485)

3.1 Distribution of Total Bedrooms

To examine the distribution of the total_bedrooms feature and detect potential outliers, we generated a box plot (Figure 1).

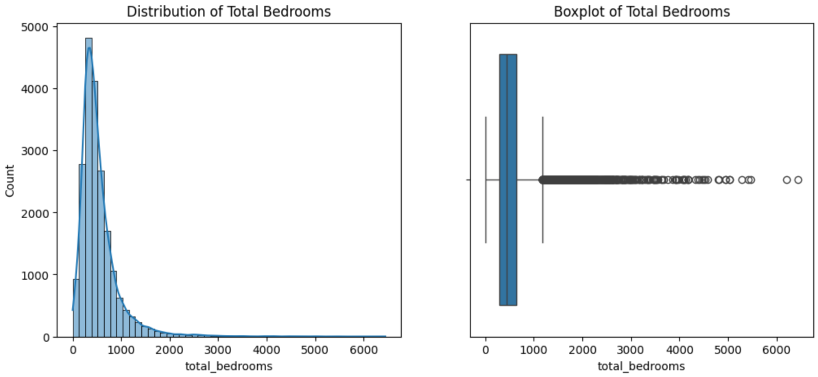
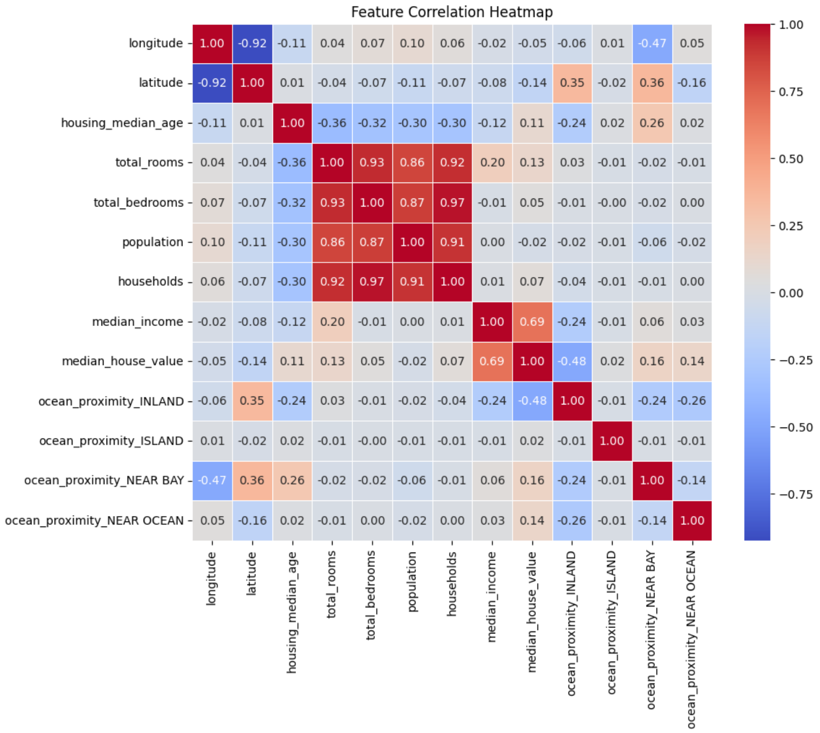


Figure 1: Box Plot of Total Bedrooms Distribution

3.2 Correlation Matrix

To visualize feature relationships, we plotted a correlation matrix (Figure 2).



	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity_INLAND	ocean_proximity_ISLAND	ocean_proximity_NEAR BAY	ocean_proximity_NEAR OCEAN
longitude	1.00	-0.92	-0.11	0.04	0.07	0.10	0.06	-0.02	-0.05	-0.06	0.01	-0.47	0.05
latitude	-0.92	1.00	0.01	-0.04	-0.07	-0.11	-0.07	-0.08	-0.14	0.35	-0.02	0.36	-0.16
housing_median_age	-0.11	0.01	1.00	-0.36	-0.32	-0.30	-0.30	-0.12	0.11	-0.24	0.02	0.26	0.02
total_rooms	0.04	-0.04	-0.36	1.00	0.93	0.86	0.92	0.20	0.13	0.03	-0.01	-0.02	-0.01
total_bedrooms	0.07	-0.07	-0.32	0.93	1.00	0.87	0.97	-0.01	0.05	-0.01	-0.00	-0.02	0.00
population	0.10	-0.11	-0.30	0.86	0.87	1.00	0.91	0.00	-0.02	-0.02	-0.01	-0.06	-0.02
households	0.06	-0.07	-0.30	0.92	0.97	0.91	1.00	0.01	0.07	-0.04	-0.01	-0.01	0.00
median_income	-0.02	-0.08	-0.12	0.20	-0.01	0.00	0.01	1.00	0.69	-0.24	-0.01	0.06	0.03
median_house_value	-0.05	-0.14	0.11	0.13	0.05	-0.02	0.07	0.69	1.00	-0.48	0.02	0.16	0.14
ocean_proximity_INLAND	-0.06	0.35	-0.24	0.03	-0.01	-0.02	-0.04	-0.24	-0.48	1.00	-0.01	-0.24	-0.26
ocean_proximity_ISLAND	0.01	-0.02	0.02	-0.01	-0.00	-0.01	-0.01	-0.01	0.02	-0.01	1.00	-0.01	-0.01
ocean_proximity_NEAR BAY	-0.47	0.36	0.26	-0.02	-0.02	-0.06	-0.01	0.06	0.16	-0.24	-0.01	1.00	-0.14
ocean_proximity_NEAR OCEAN	0.05	-0.16	0.02	-0.01	0.00	-0.02	0.00	0.03	0.14	-0.26	-0.01	-0.14	1.00

Figure 2: Correlation Matrix of Numerical Features

4 Handling Missing Values

The dataset contained 207 missing values in the ‘total_bedrooms‘ column. Analysis of this feature showed:

- Mean: 537.87
- Median: 435.00
- Standard deviation: 421.39
- Range: 1 to 6,445

Due to the presence of outliers (as observed in the box plot), the median was chosen as the imputation method to avoid skewing the data. After imputation, the dataset had no remaining missing values.

5 Categorical Data Encoding

The ‘ocean_proximity‘ categorical feature was one-hot encoded using the pandas ‘get_dummies‘ function with ‘drop_first=True‘ to avoid the dummy variable trap. This resulted in four binary columns:

- ocean_proximity_INLAND
- ocean_proximity_ISLAND
- ocean_proximity_NEAR BAY
- ocean_proximity_NEAR OCEAN

The reference category (¡1H OCEAN) was excluded from the encoding.

6 Feature Scaling

Standardization (Z-score scaling) was applied to all numerical features to ensure they have similar scales. This prevents features with larger numerical ranges from dominating the model training process.

7 Feature Engineering

To improve model performance, several new features were created:

- **rooms_per_household**: Captures housing density and living space availability.
- **bedrooms_per_room**: Represents the proportion of bedrooms to total rooms.
- **population_per_household**: Indicates average household size.
- **log_median_income**: Log transformation of income to handle skewness.
- **log_population_per_household**: Log transformation to normalize distribution.

8 Model Training and Evaluation

The dataset was split into training (80%) and testing (20%) sets. Multiple models were trained and evaluated:

The Random Forest model significantly outperformed all others, explaining approximately 80.7% of the variance in house prices.

Model	MSE	R^2 Score
Linear Regression	5.8945e+09	0.5502
Polynomial Regression	5.4167e+09	0.5866
Ridge Regression	5.8946e+09	0.5502
Lasso Regression	5.8945e+09	0.5502
Random Forest	2.5349e+09	0.8066

Table 1: Model Performance Metrics

9 Conclusion and Insights

- Random Forest was the best-performing model, demonstrating the non-linear nature of housing price relationships.
- Median income and location are the most significant factors influencing house prices.
- Coastal properties tend to have higher values than inland properties.
- Further improvements could include hyperparameter tuning and external data integration.