Project Report: Predictive Modeling for Median House Prices in California

Project Overview: Goals and Hypothesis

The goal of this project is to develop a predictive model for estimating median house prices in various districts in California. The business question we aim to address is: "What factors influence median house prices, and can we build a model that accurately predicts these prices based on relevant features?"

Our hypothesis is that factors such as median income, housing median age, geographical location (longitude and latitude), and other demographic attributes significantly impact median house prices. By constructing a predictive model, we aim to provide valuable insights into the California housing market.

Dataset and Variables

Our dataset comprises data collected during the 1990 Census for various block groups across California. Each block group typically consists of approximately 1425.5 individuals living in a geographically compact area. The dataset contains 20,640 observations and nine related factors:

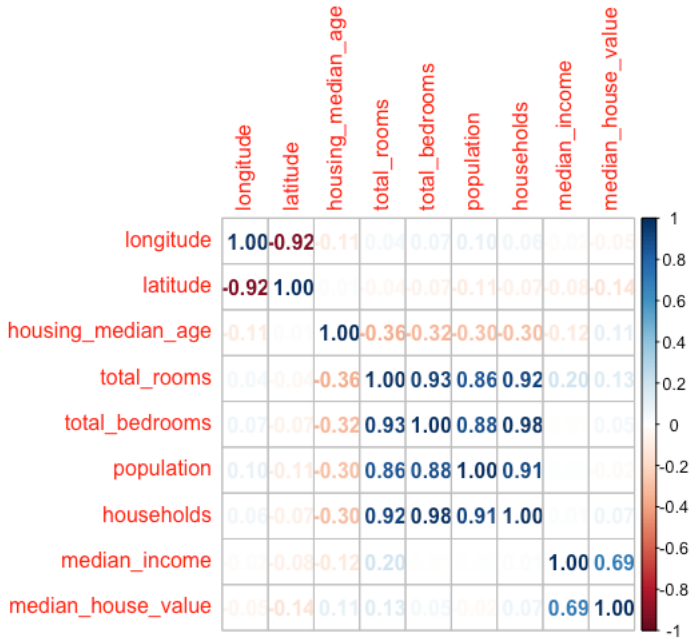
1. median_house_value: The median value of housing units within the block group (the dependent variable).
2. latitude: The geographical latitude of the block group.
3. longitude: The geographical longitude of the block group.
4. housing_median_age: The median age of housing units within the block group.
5. total_rooms: The total number of rooms in housing units within the block group.
6. total_bedrooms: The total number of bedrooms in housing units within the block group.
7. population: The total population residing within the block group.
8. households: The total number of households within the block group.
9. median_income: The median income of residents within the block group.
10. ocean_proximity: A categorical variable indicating the proximity of the block group to the ocean.

Exploratory Data Analysis (EDA) and Data Visualization

Our correlation analysis of the numeric variables uncovers notable relationships within the dataset. We observe strong positive correlations between several variables, including total rooms and total bedrooms, as well as population and total rooms. These correlations imply that larger properties typically contain more bedrooms and can accommodate larger populations. In contrast, a strong negative correlation exists between total bedrooms and median age,

suggesting that older housing units tend to have fewer bedrooms. Additionally, we find a positive correlation between median income and median house value, indicating that regions with higher income levels also tend to have higher home values. The correlation plot below visually depicts these key relationships uncovered in the data.

In summary, the correlation analysis reveals meaningful connections between property attributes like size and bedrooms, demographics like population and age, and economic factors like income and home values. These insights enhance our understanding of the housing market dynamics reflected in the data. The correlation plot effectively summarizes the most salient relationships for further investigation.

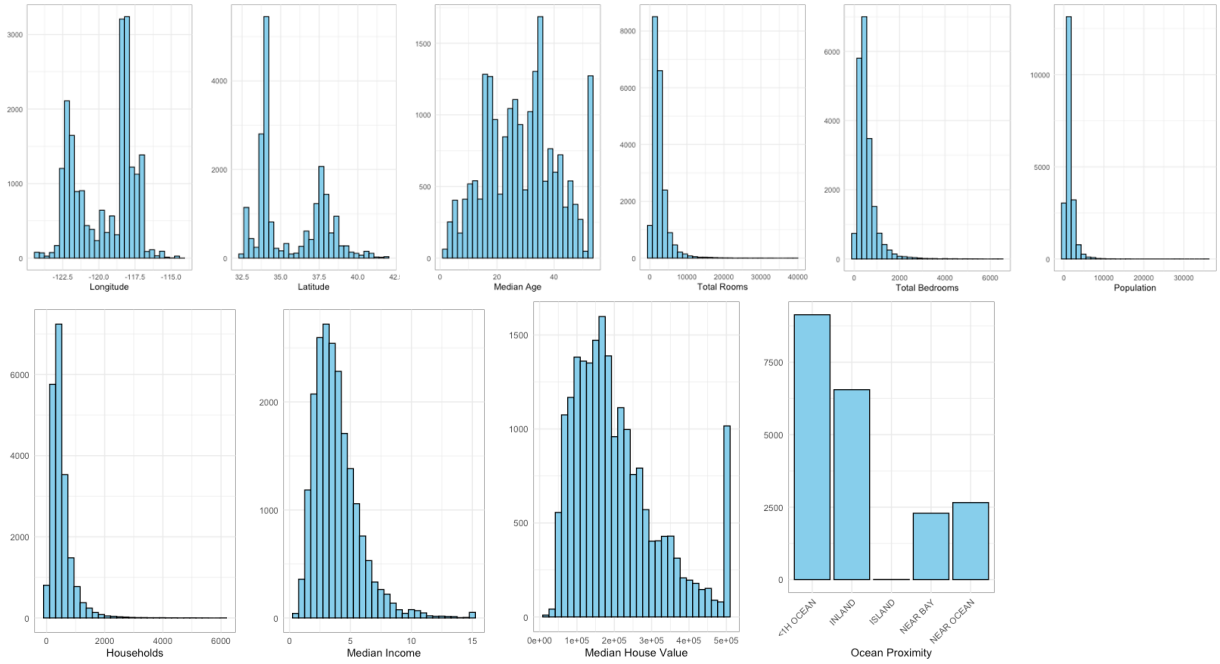


The histograms reveal valuable insights into the distribution patterns of the numeric variables. The longitude histogram displays a bimodal shape with prominent peaks around -112.5 and -117.5. This suggests the data contains two distinct geographic clusters concentrated around those longitude values. Similarly, the latitude histogram shows dual peaks at approximately 33.5 and 37.5, indicating two geographic regions clustered around those latitudes.

In contrast, variables such as median age, total rooms, total bedrooms, population, and households exhibit single-peaked, bell-shaped histograms, indicative of normal distributions. The age, size, and demographic variables lack obvious clustering effects.

In summary, the spatial attributes of longitude and latitude display clear evidence of non-normal distributions and geographic clustering within the data. Meanwhile, the histograms for other numeric variables like age and households conform more closely to normal distribution

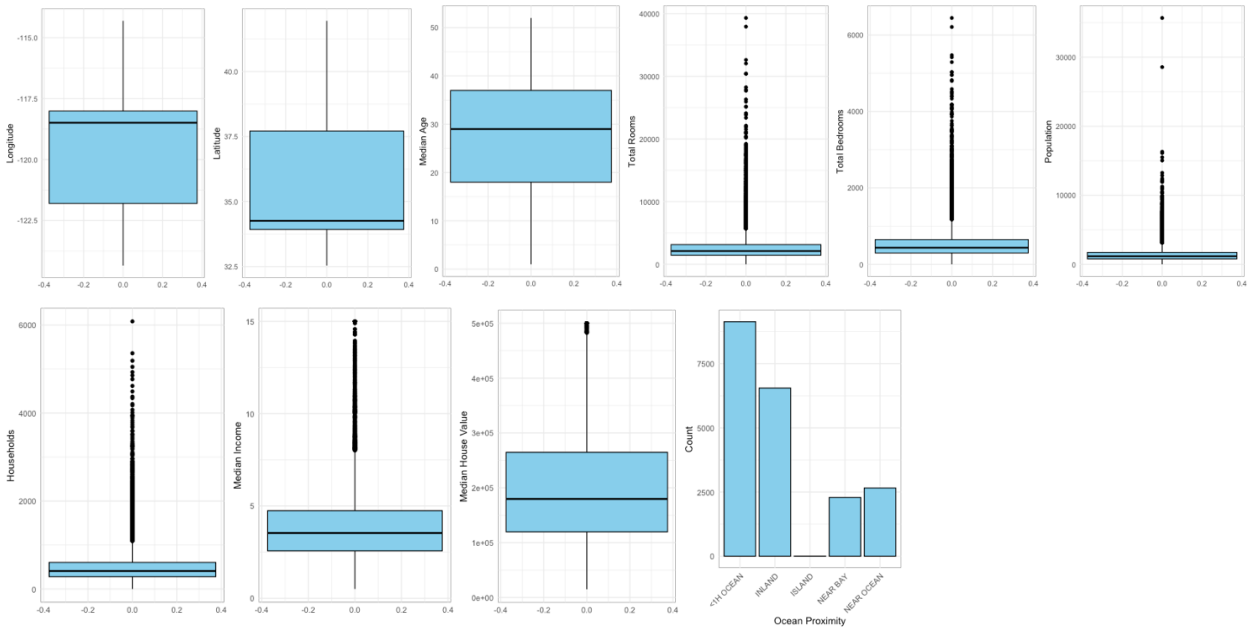
expectations. Analyzing these distribution patterns provides perspective on the underlying data structure and relationships. The histograms enable quick visual identification of clustering effects and anomalies for additional investigation.

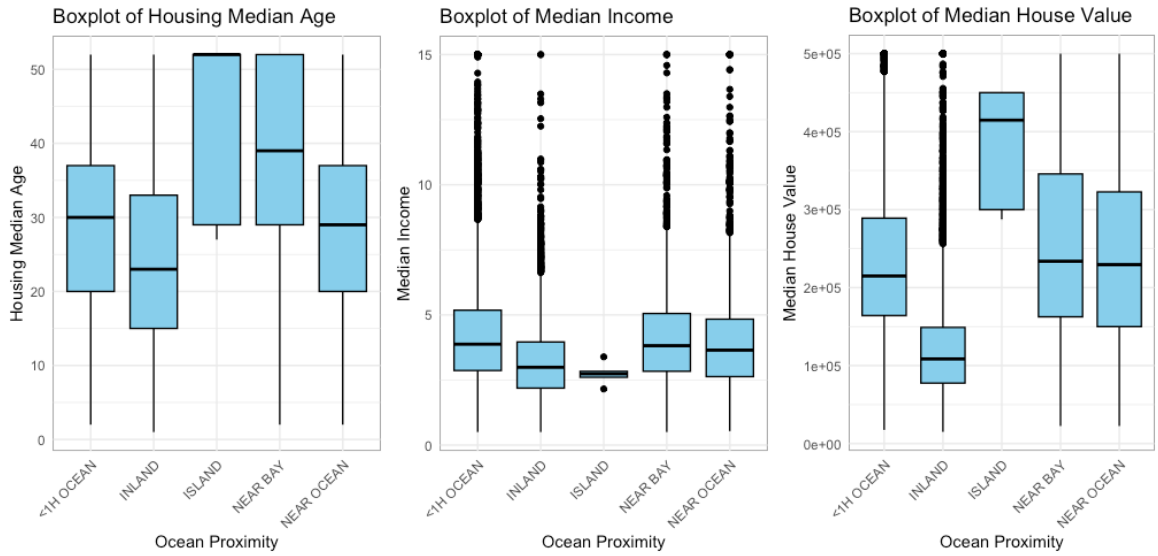


The latitude boxplots reveal an unusual distribution, with the median line positioned at the bottom of the plot. This reflects the bimodal shape observed in the latitude histogram, indicating two clusters in the data. In contrast, the longitude boxplot shows the median line at the top, corresponding to the dual-peak histogram resulting from clusters around -112.5 and -117.5.

For variables like median age, total rooms, total bedrooms, population and households that exhibited normal histograms, the median lines sit in the middle of the boxplots as expected. However, numerous outliers are present across these boxplots, signaling potential anomalies or extreme values worthy of further investigation.

In summary, the boxplots provide visual confirmation of the distribution patterns found in the histograms. The latitude and longitude plots clearly display the effects of bimodality and dual-clustering respectively. The centralized median lines in the other boxplots match the normal histogram shapes. But the prevalence of outliers suggests underlying data anomalies that could skew analyses if not properly addressed. Analyzing boxplots and histograms in conjunction facilitates deeper understanding of the data distributions.





Data Preprocessing

Our data preparation process involved several steps to refine the dataset's quality and structure:

1. We filled 207 missing values in the total bedrooms column using median imputation to ensure data completeness.

2. The ocean proximity categorical variable was transformed into five binary columns representing NEAR BAY, <1H OCEAN, INLAND, NEAR OCEAN, and ISLAND through one-hot encoding.
3. To provide better representations of housing attributes, two new variables were constructed: mean bedrooms and mean rooms, calculated by dividing total bedrooms and total rooms by households.
4. For improved scale consistency and modeling stability, numerical variables were standardized to a common scale, excluding median house value and the binary categorical columns.

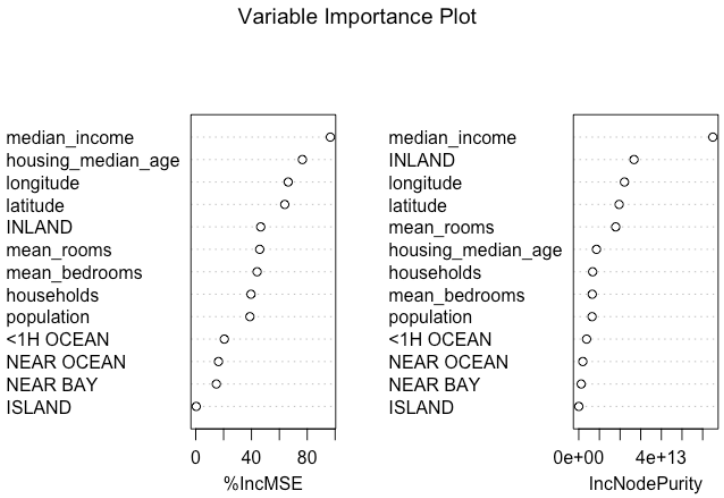
In summary, through imputation, encoding, feature engineering, and scaling, we processed the raw data into a refined dataset more suitable for analysis and modeling, while preserving the completeness, integrity and distributional characteristics of the original data.

Modeling Approach

We used the Random Forest algorithm for regression to predict median house values. The feature vector (predictors) included all the variables except for the target variable, "median_house_value." The response variable was "median_house_value." The model was trained on 70% of the data, validated on the remaining 30% test set.

Description of the Performance Metric Results for the Model

The performance of the Random Forest model was evaluated using the Root Mean Squared Error (RMSE) metric. The RMSE for the trained model was approximately 48,879, while the RMSE for the test set was approximately 50,654. These values suggest that the model has some predictive power, but it may not be highly accurate in predicting house values.



Business Answer

Our random forest model demonstrates moderate ability to predict median housing prices, with test set errors around $50,000. This provides a useful starting point but further refinement could improve accuracy. The model results reveal location attributes like longitude, latitude, and ocean proximity are by far the most important factors driving median values. Property characteristics like total rooms and bedrooms also contribute but are secondary to location. Demographic factors like median income have a smaller but measurable effect on median prices.

Given these insights, we recommend further developing the model by incorporating additional location attributes and economic factors that likely influence prices, such as proximity to key employers and amenities. While the current model requires more precision before deployment, it provides valuable insights on the relationships between location, property features, demographics, and housing prices. These business insights are useful for understanding the California housing market even if the model needs refinement.

In summary, the random forest model shows initial promise as the foundation for a housing price estimation tool if further tuned. The most important next steps are adding more granular location data and testing other algorithms. But the model and feature importance results already deliver actionable business insights into the housing market that can inform strategy.