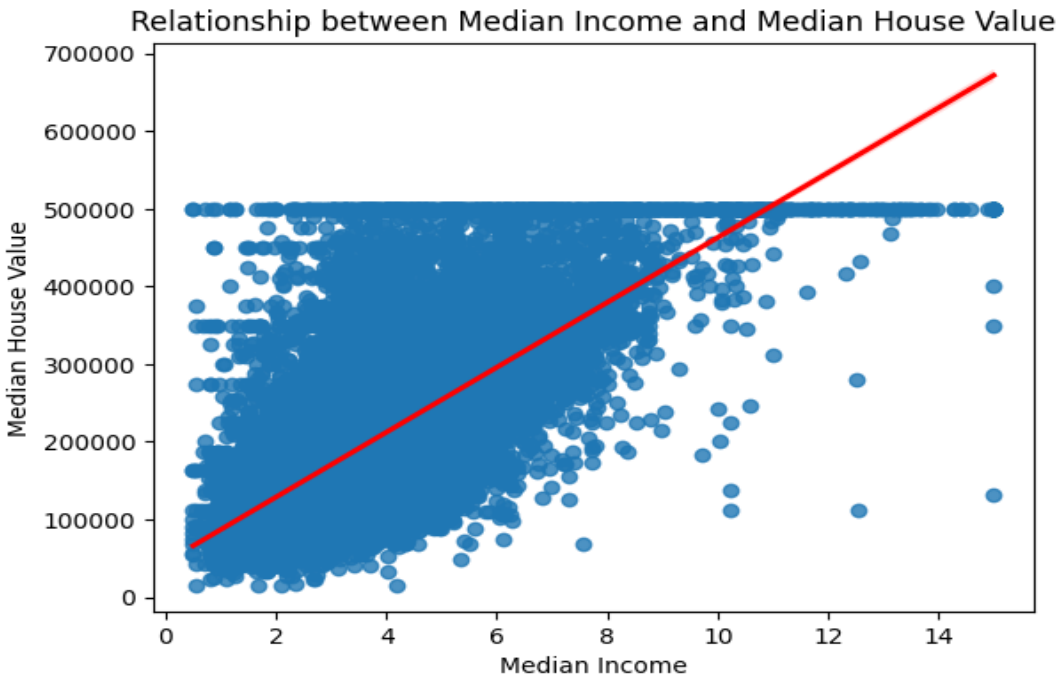
Predicting California Housing Prices

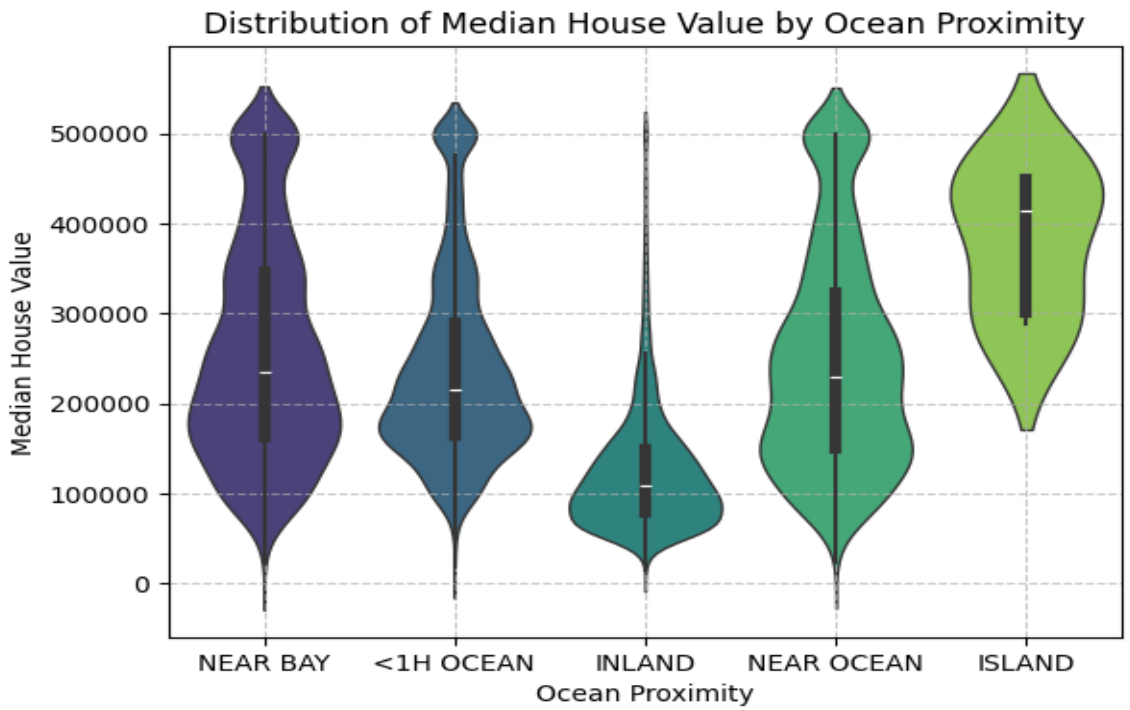
Introduction

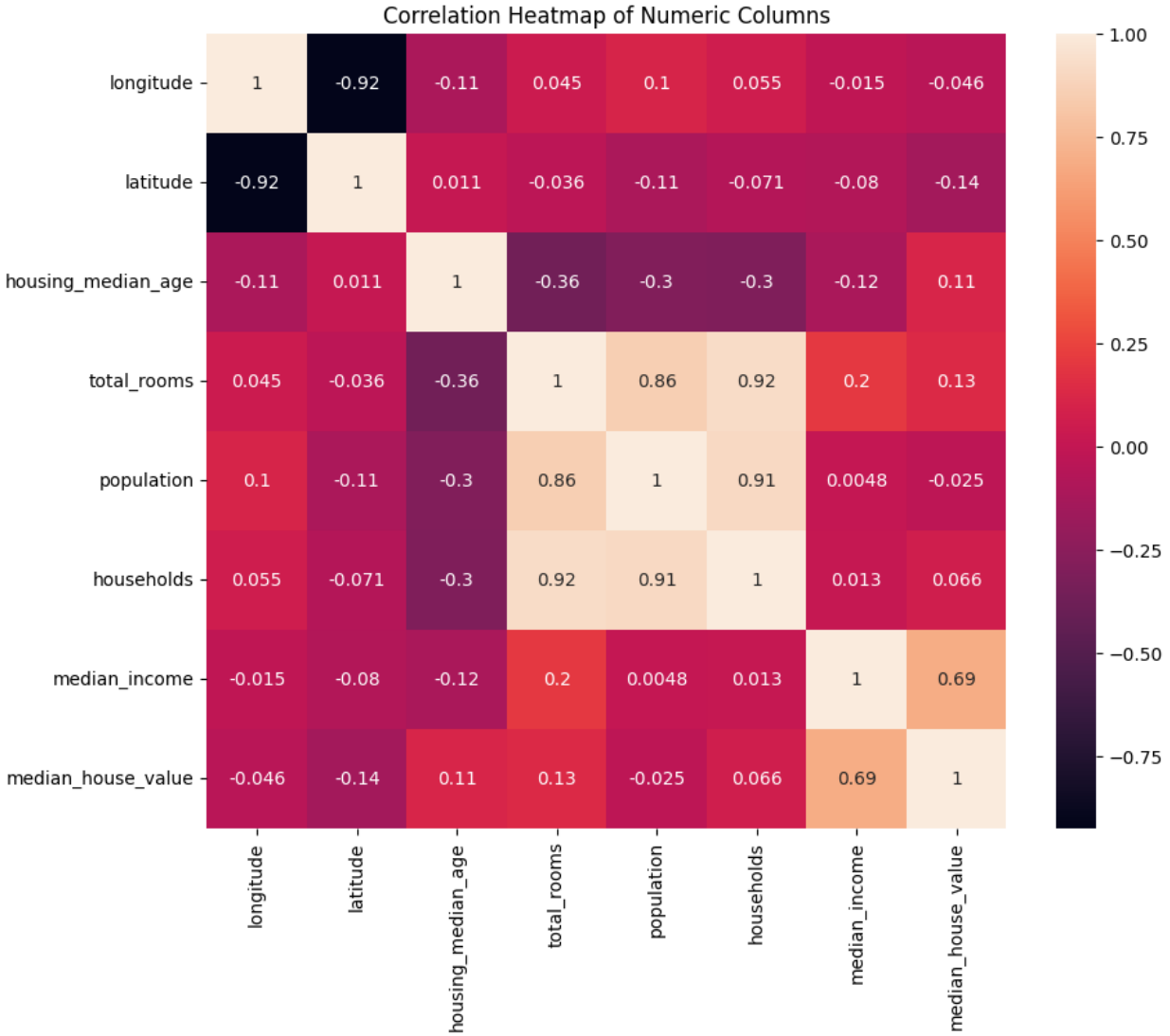
The aim of this project is to predict housing prices in California based on various features such as location, housing characteristics, and demographic data. The dataset used for this project contains information collected during the 1990 census. The features include longitude, latitude, housing median age, total rooms, total bedrooms, population, households, median income, and ocean proximity.

Data Exploration and Preprocessing

- Data Loading: The dataset was loaded into a Pandas DataFrame.
- Initial Exploration: Initial exploration of the data was performed using methods like head(), describe(), info(), and isnull().sum() to understand its structure, summary statistics, and missing values.
- Handling Missing Values: Missing values in the 'total_bedrooms' column were filled with the mean value of that column.
- Data Visualization: Various visualizations such as histograms, correlation heatmap, regression plots, and violin plots were created to understand the relationships between different features and the target variable.




Distribution of Median House Value by Ocean Proximity
Median House Value
500000
400000
300000
200000
100000
0
NEAR BAY
<1H OCEAN
INLAND
NEAR OCEAN
ISLAND
Ocean Proximity


Correlation Heatmap of Numeric Columns
longitude 1 -0.92 -0.11 0.045 0.1 0.055 -0.015 -0.046
latitude -0.92 1 0.011 -0.036 -0.11 -0.071 -0.08 -0.14
housing_median_age -0.11 0.011 1 -0.36 -0.3 -0.3 -0.12 0.11
total_rooms 0.045 -0.036 -0.36 1 0.86 0.92 0.2 0.13
population 0.1 -0.11 -0.3 0.86 1 0.91 0.0048 -0.025
households 0.055 -0.071 -0.3 0.92 0.91 1 0.013 0.066
median_income -0.015 -0.08 -0.12 0.2 0.0048 0.013 1 0.69
median_house_value -0.046 -0.14 0.11 0.13 -0.025 0.066 0.69 1
longitude
latitude
housing_median_age
total_rooms
population
households
median_income
median_house_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Feature Engineering

- Encoding Categorical Data: The categorical feature 'ocean_proximity' was encoded using Label Encoding to convert it into a numerical format.
- Scaling Numerical Data: Numerical features were scaled using Min-Max Scaling to bring them to a common scale.

Feature Selection

- SelectKBest: SelectKBest algorithm with f_regression scoring was used for feature selection to select the top k features that are most relevant for predicting the target variable.

Model Building and Evaluation

1. Linear Regression Model:
    - The Linear Regression model was trained on the training data.
    - Predictions were made on the testing data.
    - Evaluation metrics like Mean Squared Error (MSE) and R2 Score were calculated to assess the model's performance.
2. Random Forest Regression Model:
    - A Random Forest Regression model was trained on the training data.
    - Predictions were made on the testing data.
    - Evaluation metrics (MSE and R2 Score) were calculated to evaluate the model's performance.

Conclusion

Both Linear Regression and Random Forest Regression models were implemented to predict housing prices in California. The Random Forest Regression model outperformed the Linear Regression model in terms of predictive accuracy, as evidenced by lower Mean Squared Error and higher R2 Score. However, further optimization and tuning of models could potentially improve their performance. Additionally, more sophisticated feature engineering techniques and exploring other regression algorithms could be beneficial for enhancing the predictive capabilities of the models.