PROJECT REPORT

Dec 10,2024
Celine XIAO

INTRODUCTION

This current project is a data science project which the main goal is to be able to predict the median house value, given certain information about a property.
It can help to solve some business issue faced by estate agencies: “How to set the price of a property, whether an apartment or a house, in a way that is reasonable while also maximizing profits?”.

We will work with the following dataset: ‘housing.csv’.
This data set appeared in a 1997 paper titled Sparse Spatial Autoregressions by Pace, R. Kelley and Ronald Barry, published in the *Statistics and Probability Letters* journal. The researchers built it using the 1990 California census data. It contains one row per census block group.

This dataset contains is made of 20,640 observations and 10 features.
The target feature is of course this one:

- **median_house_value (**in dollars)

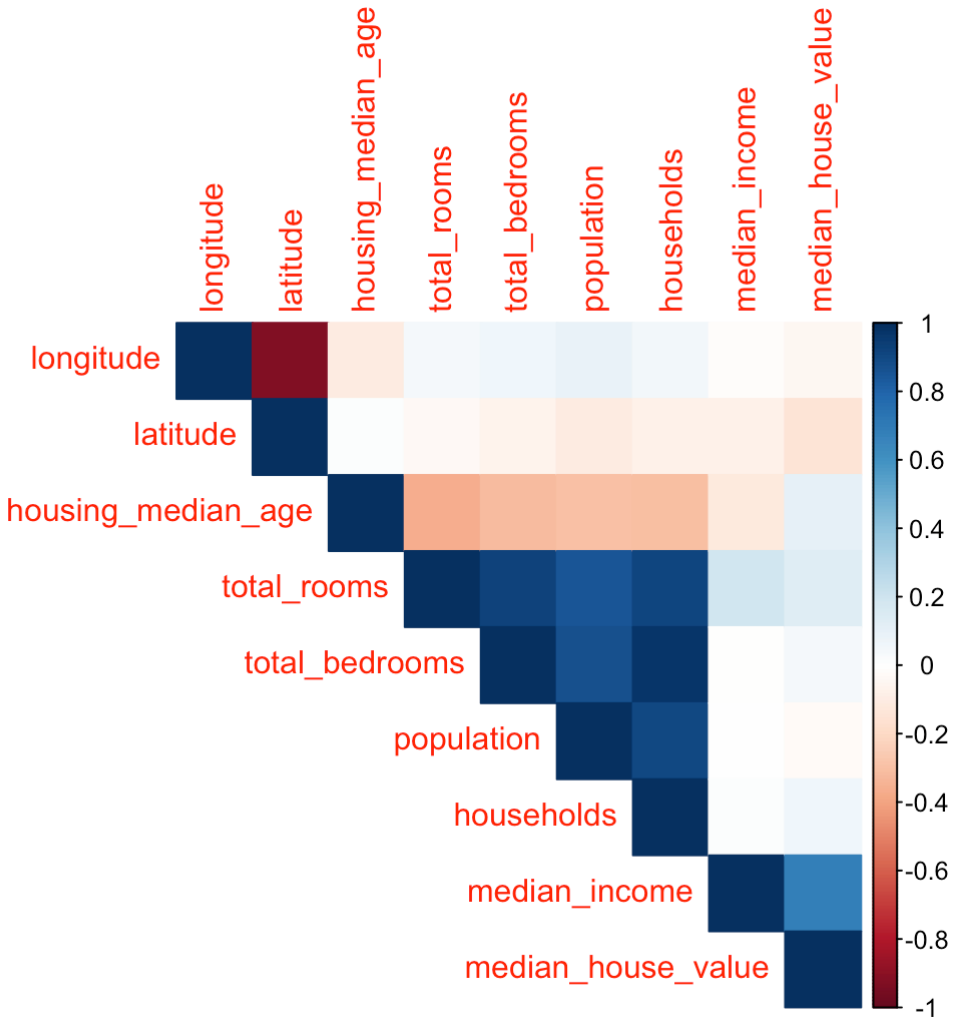
The features we will used to make the median house value prediction are the following ones:

- **longitude**: The geographical coordinate specifying the east-west position of a property. (The unit is degrees)
- **latitude**: The geographical coordinate specifying the north-south position of a property. Both indicate the position of the property.
- **housing_median_age**: The median age of houses in a census block group (in the neighborhood). The unit used is years.
- **total_rooms:** The total number of rooms in all housing units within a block.
- **total_bedrooms:** The total number of bedrooms in all housing units within a block.
- **population:** The total number of people residing in the block
- **households:** The total number of households (occupied housing units) in the block.
- **median_income:** The median income of a households in the block. (In tens of thousands of dollars)
- **ocean_proximity:** A categorical variable describing the property’s proximity to the ocean.

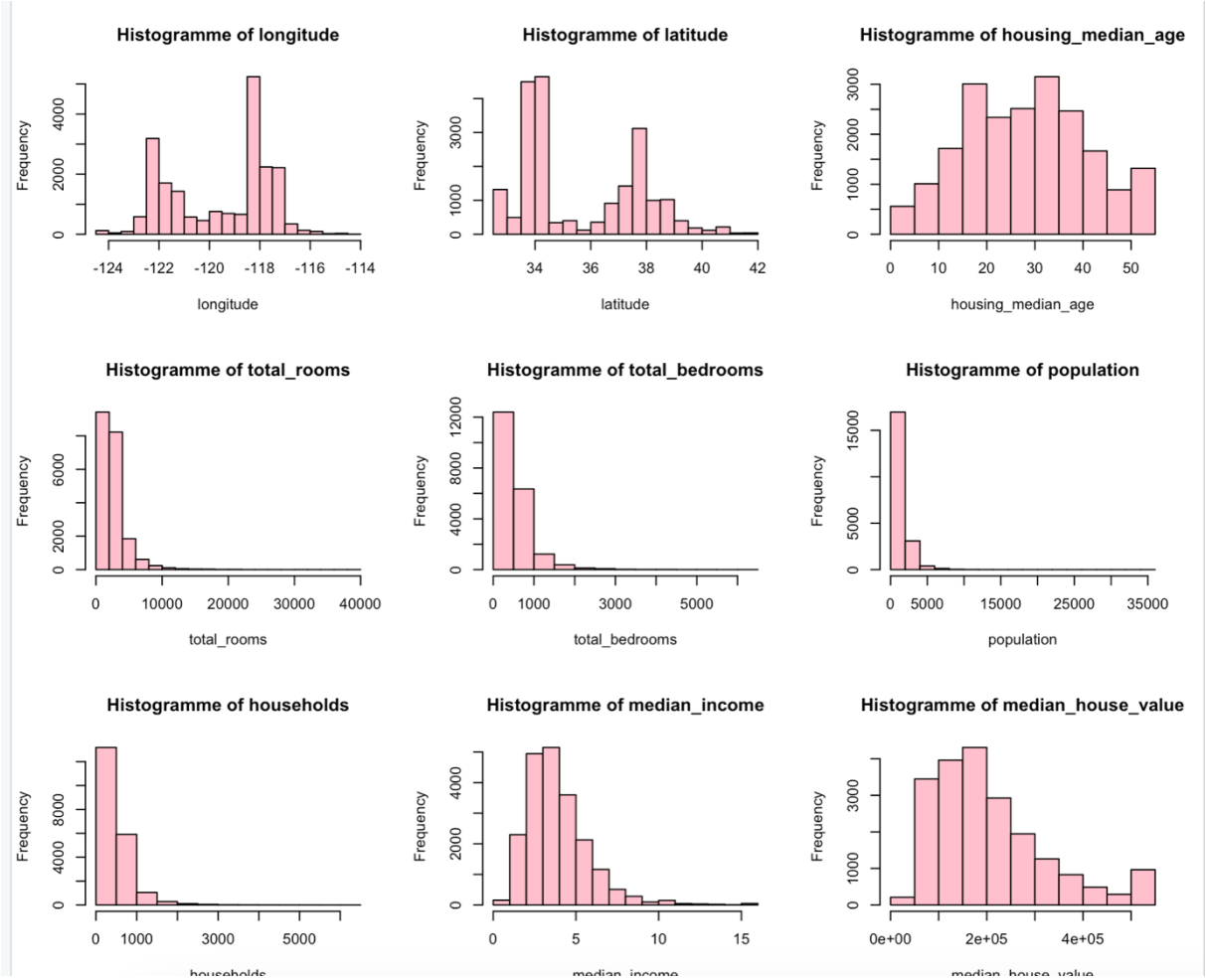
Here is a little look on the dataset:

```
#longitude latitude housing_median_age total_rooms total_bedrooms population
#1   -122.23    37.88                41         880            129        322
#2   -122.22    37.86                21        7099           1106       2401
#3   -122.24    37.85                52        1467            190        496
#4   -122.25    37.85                52        1274            235        558
#5   -122.25    37.85                52        1627            280        565
#6   -122.25    37.85                52         919            213        413
#households median_income median_house_value ocean_proximity
#1        126        8.3252            452600        NEAR BAY
#2       1138        8.3014            358500        NEAR BAY
#3        177        7.2574            352100        NEAR BAY
#4        219        5.6431            341300        NEAR BAY
#5        259        3.8462            342200        NEAR BAY
#6        193        4.0368            269700        NEAR BAY
```

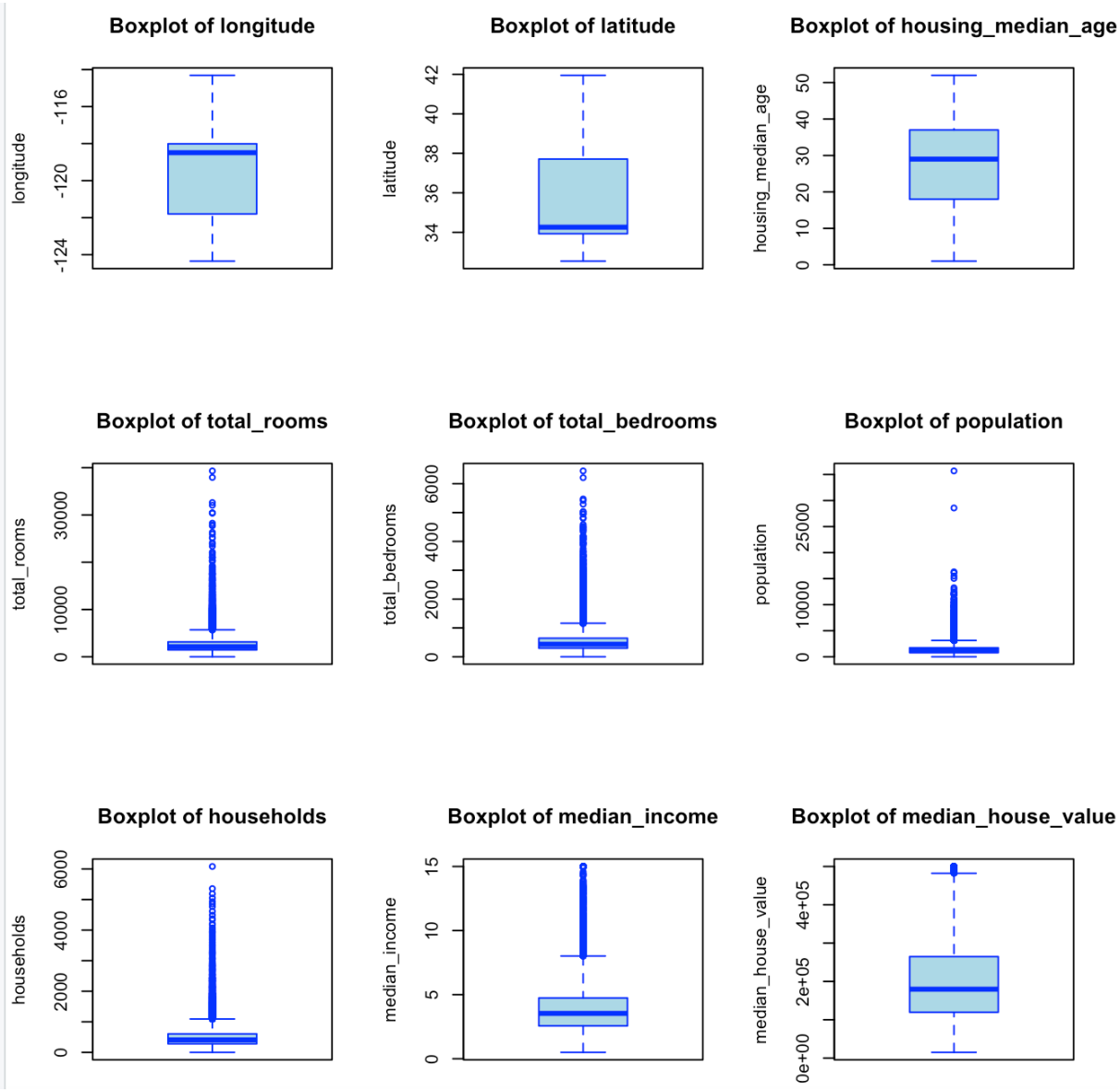
Let’s dive into the study of our features and examine the correlations between them.



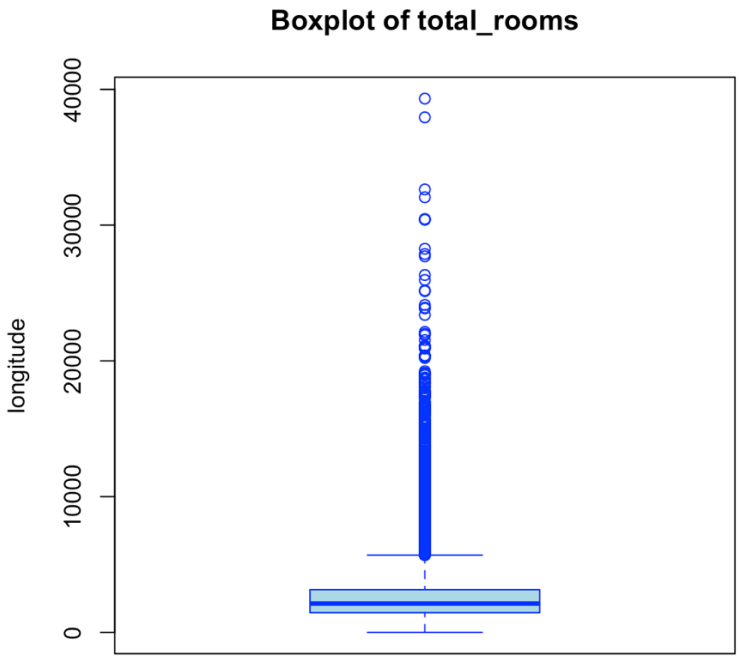
By examining this correlation map, we can observe that the darkest colors represent the strongest correlations between features. First, we notice that only one feature, **median_income**, is highly correlated with our target value. This feature will contribute the most to our model’s predictions. Additionally, we observe that **total_rooms**, **total_bedrooms**, **population**, and **households** are strongly correlated with each other, which makes sense given their inherent relationships.

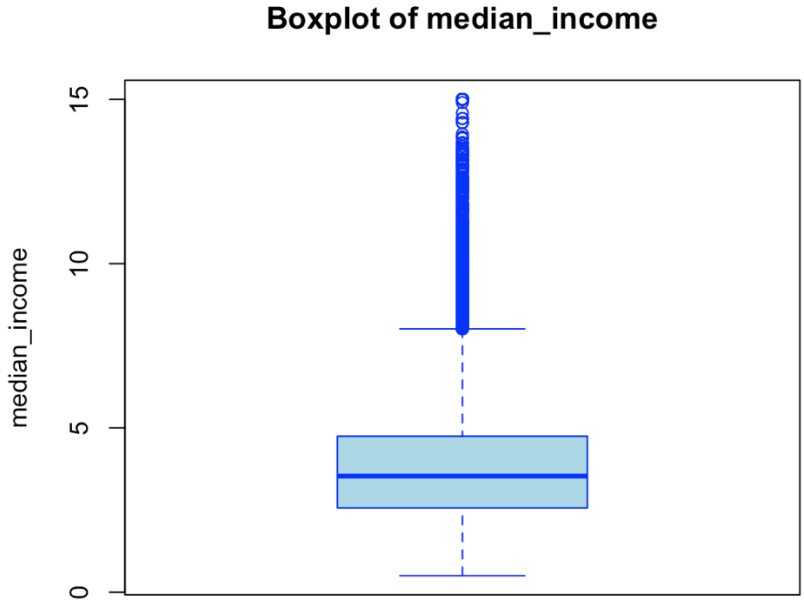


Here is look on the histogram of different features,
We can see for some feature such as the median income, housing median age. They follow a normal distribution.

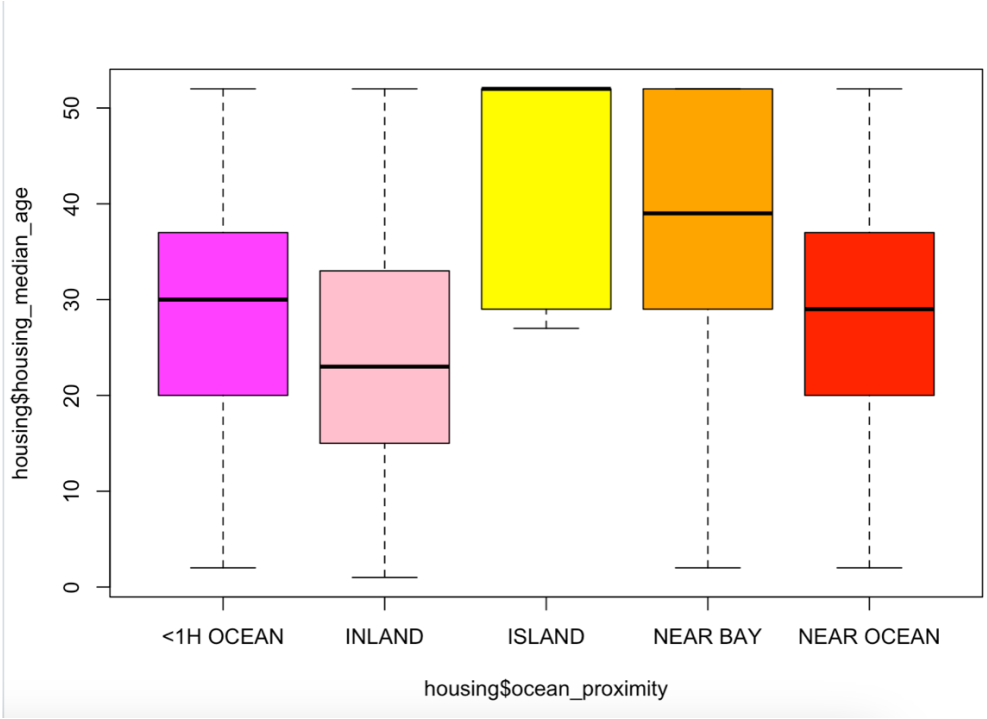


Immediately, by examining the boxplots of these different features, some appear to have unusual distributions.
The following features: **median_income**, **total_bedrooms**, **population**, and **total_rooms** show a significant number of values outside the whiskers, indicating potential outliers. But the number of points outside the whiskers seems a little too significant to be considered potential outliers.

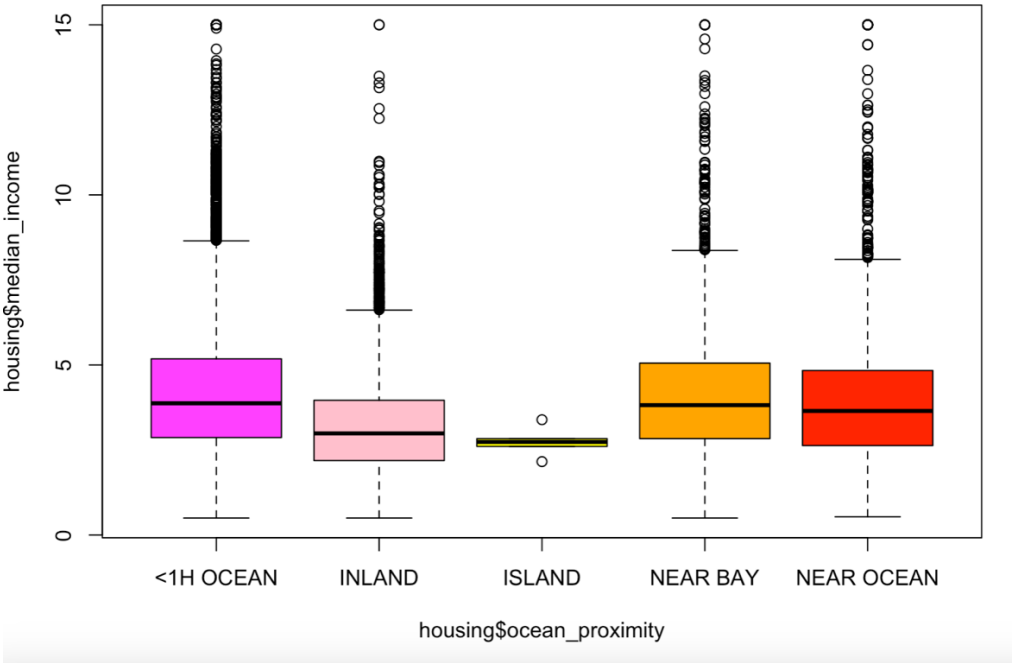




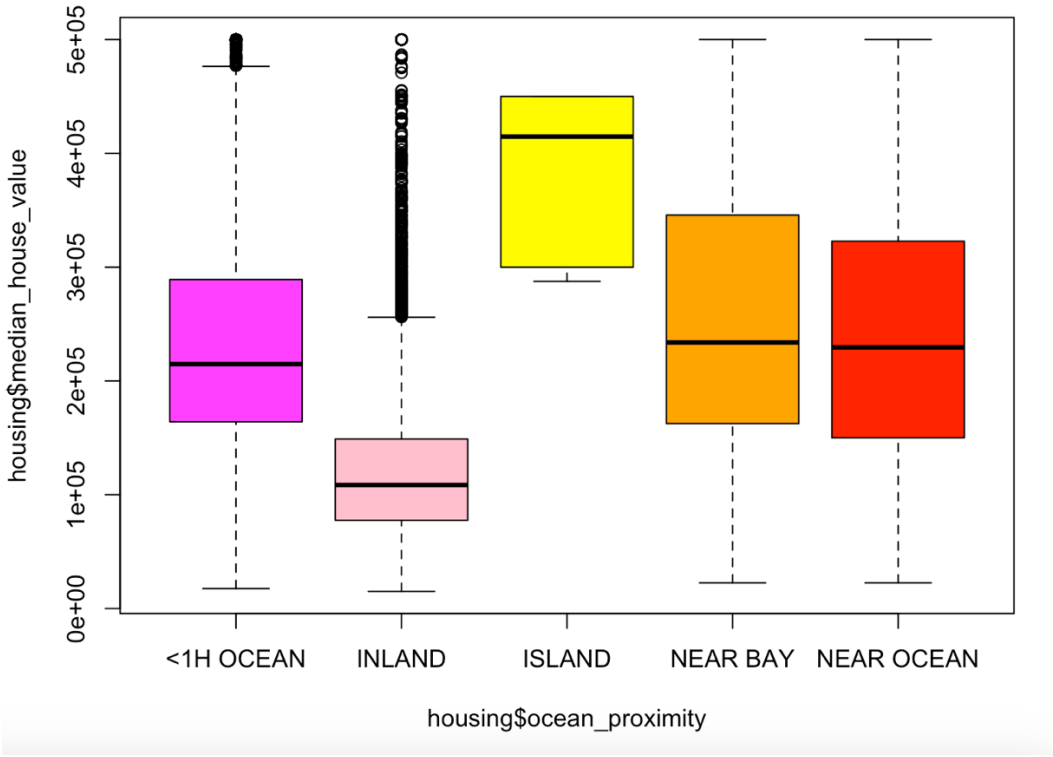
Here is the boxplot for the feature (housing_median_age , median_income and median_home_value)with respect to the factor variable ocean_proximity.



Basically, we observe that the oldest house is in ISLAND or Nearby the ocean.



Here We have a can deduce that the house that are located less than 1h or near from the ocean belongs to the household who tend to have a highest median income.



Here, we observe that the most expensive houses are mostly located on islands, followed by those near the ocean (within 1 hour). However, the boxplot for **INLAND** shows a significant number of outliers. This could indicate specific neighborhoods where house prices are potentially overpriced.

DATA MUNGING STEP

We examined the dataset and made a few modifications.
First, we transformed **Ocean_proximity**, a categorical variable, into four separate features using one-hot encoding.
Next, we addressed the **total_bedrooms** feature, which contained missing values (NAs). We replaced these missing values with the median of the column.
After cleaning the dataset, we now have a well-prepared dataset that looks like this:

```
#> head(housing)
#longitude latitude housing_median_age population households median_income median_house_value ocean_proximity_<1H OCEAN
#1   -122.23    37.88                41        322        126        8.3252            452600                        0
#2   -122.22    37.86                21       2401       1138        8.3014            358500                        0
#3   -122.24    37.85                52        496        177        7.2574            352100                        0
#4   -122.25    37.85                52        558        219        5.6431            341300                        0
#5   -122.25    37.85                52        565        259        3.8462            342200                        0
#6   -122.25    37.85                52        413        193        4.0368            269700                        0
#ocean_proximity_INLAND ocean_proximity_ISLAND ocean_proximity_NEAR BAY ocean_proximity_NEAR OCEAN mean_bedrooms mean_rooms
#1                      0                     0                        1                         0     1.0238095   6.984127
#2                      0                     0                        1                         0     0.9718805   6.238137
#3                      0                     0                        1                         0     1.0734463   8.288136
#4                      0                     0                        1                         0     1.0730594   5.817352
#5                      0                     0                        1                         0     1.0810811   6.281853
#6                      0                     0                        1                         0     1.1036269   4.761658
```

We normalized the features, except for **median_house_value**.
Afterward, we split the dataset into training and testing sets: 70% of the data was used for training (**x_train**, **y_train**), and the remaining 30% was used for testing (**x_test**, **y_test**).
The **x_train** and **x_test** datasets contain the features we will use to train our model. Here are the feature names:

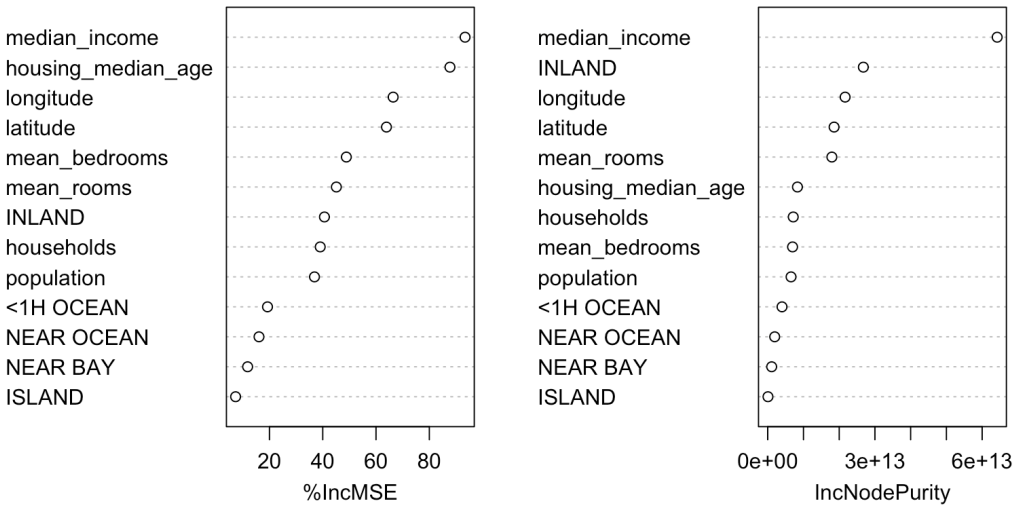
#[1] "longitude" "latitude" "housing_median_age" "population"
#[5] "households" "median_income" "<1H OCEAN" "INLAND"
#[9] "ISLAND" "NEAR BAY" "NEAR OCEAN" "mean_bedrooms"
#[13] "mean_rooms"

And y_train and y_test is our target variable: median_house_value.

Statistical Model

We used a Random Forest model for this project. Random Forest is an ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Specifically, this Random Forest model consists of 500 trees(ntree=500), which ensures robust and stable predictions by averaging the outputs of these trees.

We analyzed feature importance using two metrics provided by the Random Forest model:



%Inc MSE (Percentage Increase in Mean Squared Error):
This metric shows how much the prediction error increases when a particular feature is excluded. Features with a high %Inc MSE contribute significantly to the model's predictive accuracy.

IncNodePurity (Increase in Node Purity):
This metric measures how much a feature improves the homogeneity of the target variable within each tree node. Here again, median_income emerged as the most impactful feature.

We observe that on the first graph (%Inc MSE Increasing in mean squarred error) the feature 'median_income' and 'housing_median_age' contribute the most to the model predictive accuracy. The features 'longitude' and 'latitude' also impact the model's accuracy which is logical since the location of a house significantly influence its price.

By examining "IncNodePurity," we observe that the feature median_income has the greatest impact on node purity. This variable is highly significant for the tree.

Description of the Performance Metric Results

The model's performance was evaluated using Root Mean Squared Error (RMSE):
Initial Model (All Features, 500 Trees):

- Training RMSE: **49,396.36**
- Testing RMSE: **49,372.00**

Reduced Model (Top 4 Features, 500 Trees):

- Training RMSE: **51,537.84**
- Testing RMSE: **50,735.52**

The initial model, which utilized all 13 features, provides better predictions, as expected due to the larger amount of information available for training. However, the simplified model, using only the top 4 features, still performs reasonably well. The RMSE for the simplified model increases by only **$1,000–$2,000**, which is not a significant difference in this context. This demonstrates that the reduced model is both efficient and effective, offering competitive performance while being more interpretable and computationally efficient.

We conclude that the initial model with all features performed slightly better, with lower RMSE values. However, the reduced model, using only the top four features, still performed reasonably well. Both models generalized effectively, as the training and testing RMSE values were similar, indicating the model is not overfitting.

Business answer

This analysis highlights the key factors that influence house prices. Among these, median income emerged as the strongest predictor, emphasizing its importance in determining property values. Other significant features, such as housing median age and geographic factors like longitude and latitude, also reflect trends seen in real estate markets.

Insights for Stakeholders:

Median income acts as a solid indicator of housing demand, making it a crucial factor for setting property prices and identifying high-value areas.

Geographic factors like proximity to desirable locations play a key role in determining house prices. Properties close to the ocean or in affluent neighborhoods tend to be priced higher.

The simplified model, based on just four key features, shows that reliable predictions can still be made. This makes the model easier to interpret and implement, without losing much accuracy.

With an error margin of about $49,000–$51,000, the model provides valuable insights for real estate businesses, investors, and policymakers to make informed decisions about property pricing, investment opportunities, and urban development.