Louis ATALLAH

Project Report

Project Overview

The “housing” dataset provides information about Californian houses grouped by neighbourhood. It contains many features such as the median income of the inhabitants, number of households, number of rooms, the median house value, and so on…

The goal of this study is to determine whether the median house value can be predicted and if it can be what are the most significant features to take in account to estimate the price.

Dataset & Variables

Each observation in the dataset is a bundle of households for a specific area in California. It includes 20,640 entries and 10 variables. For each neighbourhood the following data is available:

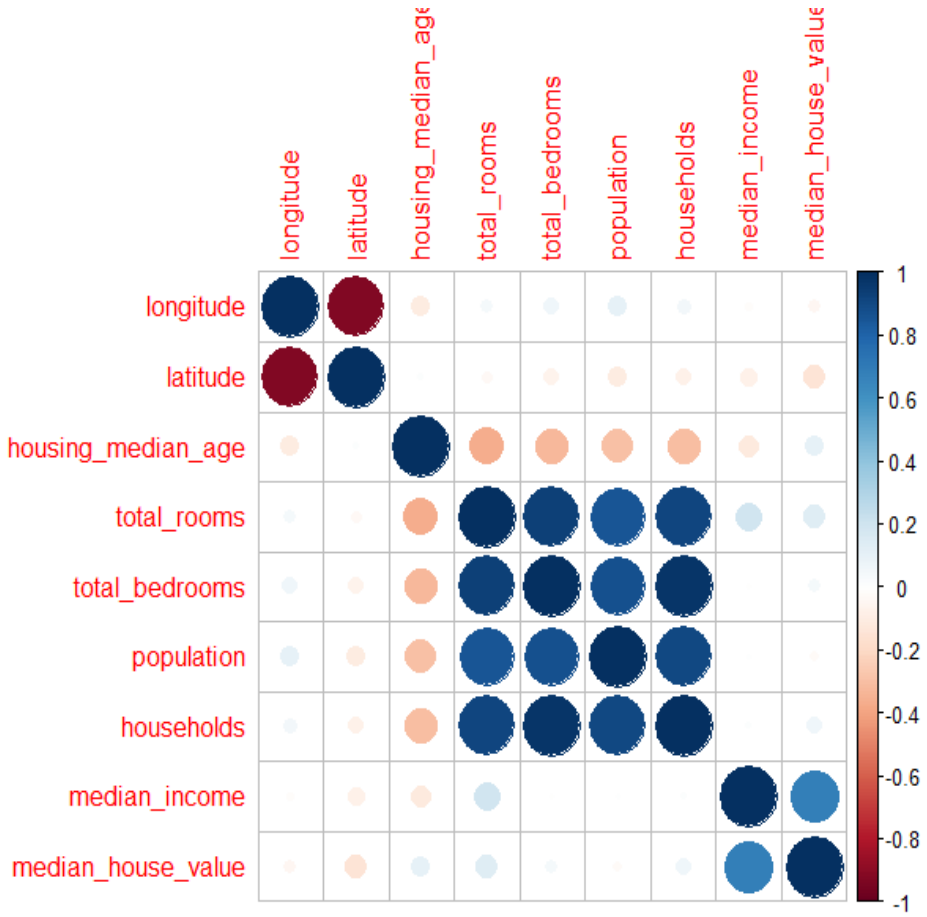
- longitude,
- latitude,
- housing median age,
- number of rooms/bedrooms,
- inhabitants of the area,
- number of households,
- median income,
- median house value
- ocean proximity.

All those features are numeric except the ocean proximity which originally is a character variable.

EDA

Corrplot Matrix

First, let's see the correlation matrix including all variables of the study:



After visualizing the corrplot matrix, we can observe that many numeric features are correlated to each other.

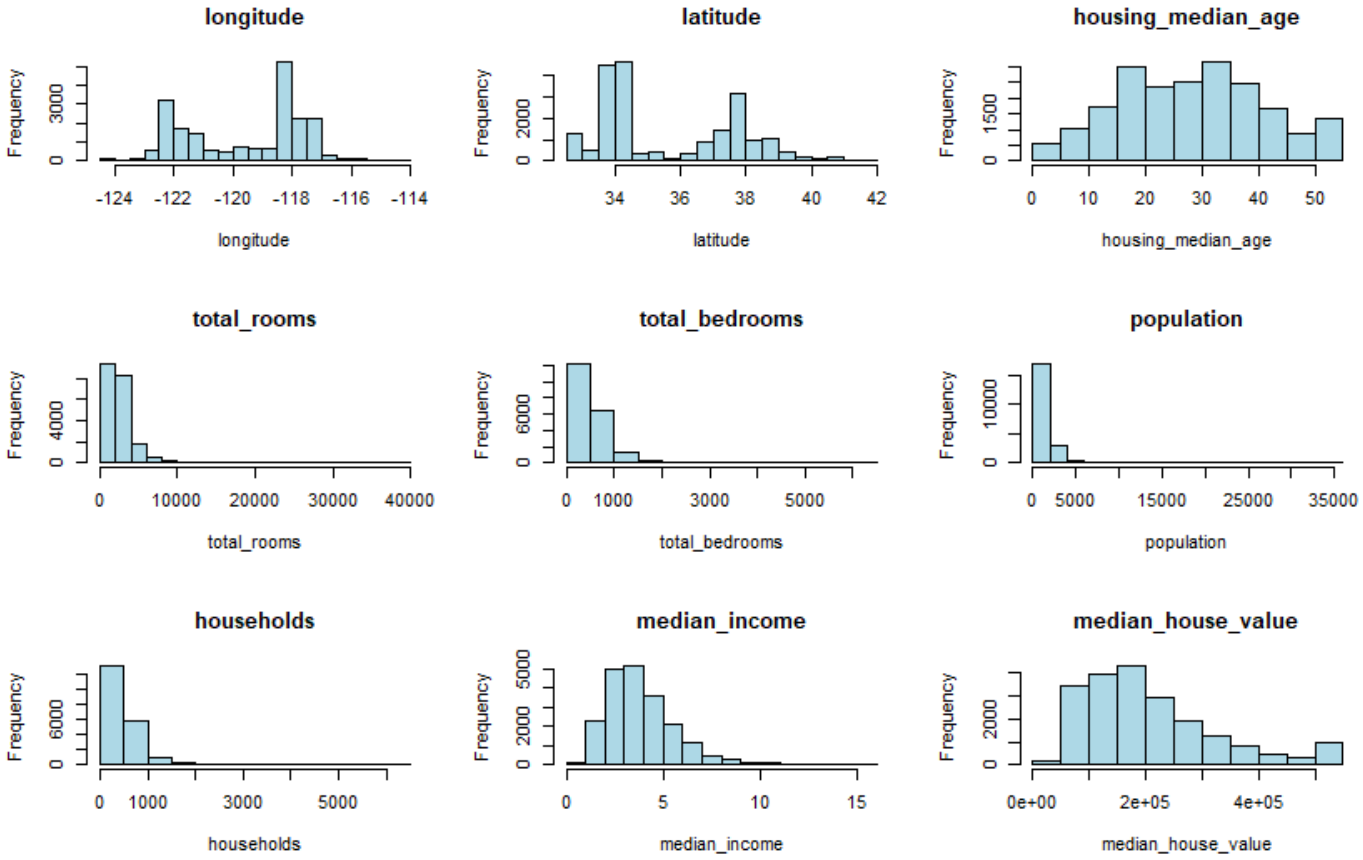
We notice a strong positive pairwise correlation (close to 1) between the number of rooms (total_rooms & total_bedrooms), the population and the households. Hence, for example if we take total_rooms and the population: the more bedrooms the houses of the area have, the more people live in the area.

We can see a positive correlation as well between the median house value and the median income (correlation coefficient of 0.6).

While the longitude and the latitude are strongly negatively correlated, which means the higher is the latitude, the lower the longitude is.

Histograms for each numeric variable

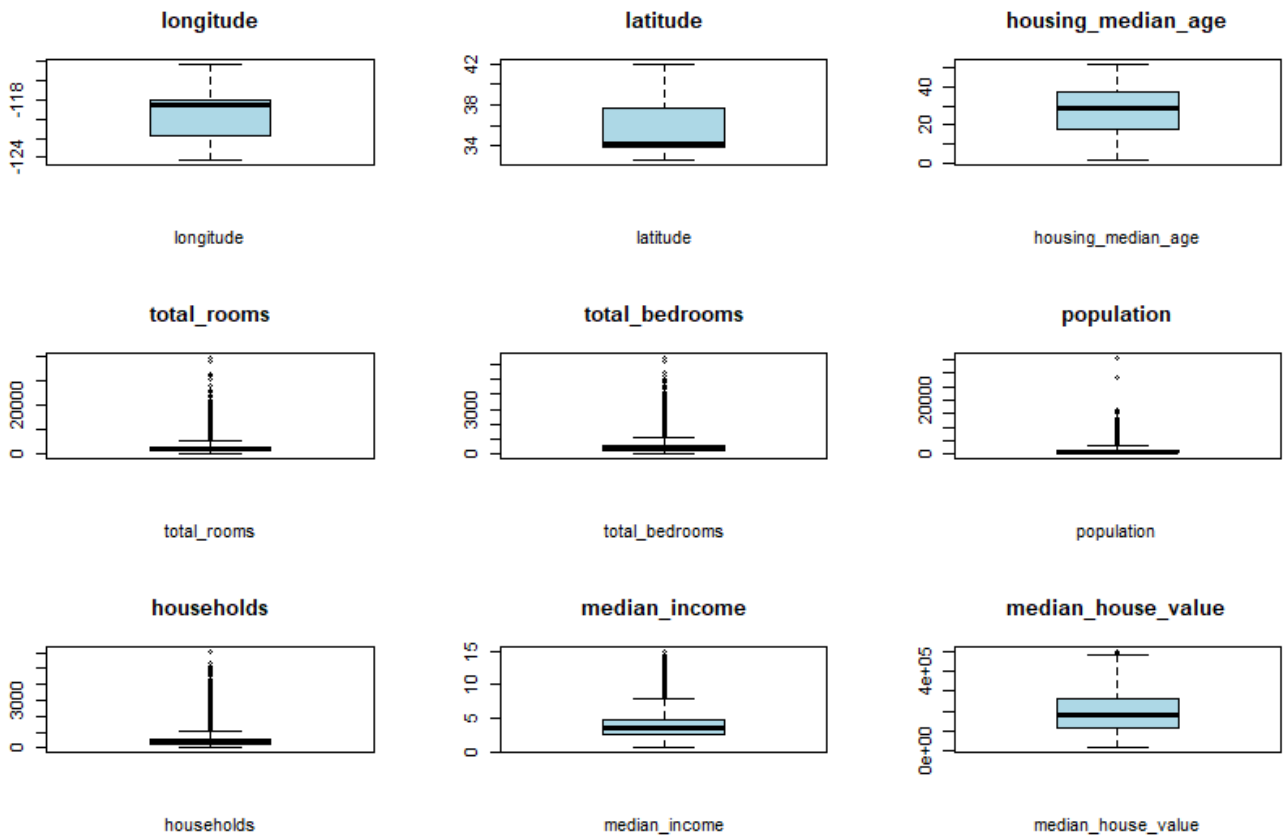
By implementing a for loop, I have displayed all histograms for the 9 numeric features.



Histograms of many various numeric variables are shown above. As we can see, the housing median age histogram is relatively flat which means that the dataset contains houses from every age class. The number of rooms is mainly concentrated between 0 and 1000 for an area, the same as the number of households. The population doesn't exceed 5000 per neighbourhood. The median income and the median house value are decreasing as we move onto the x axis.

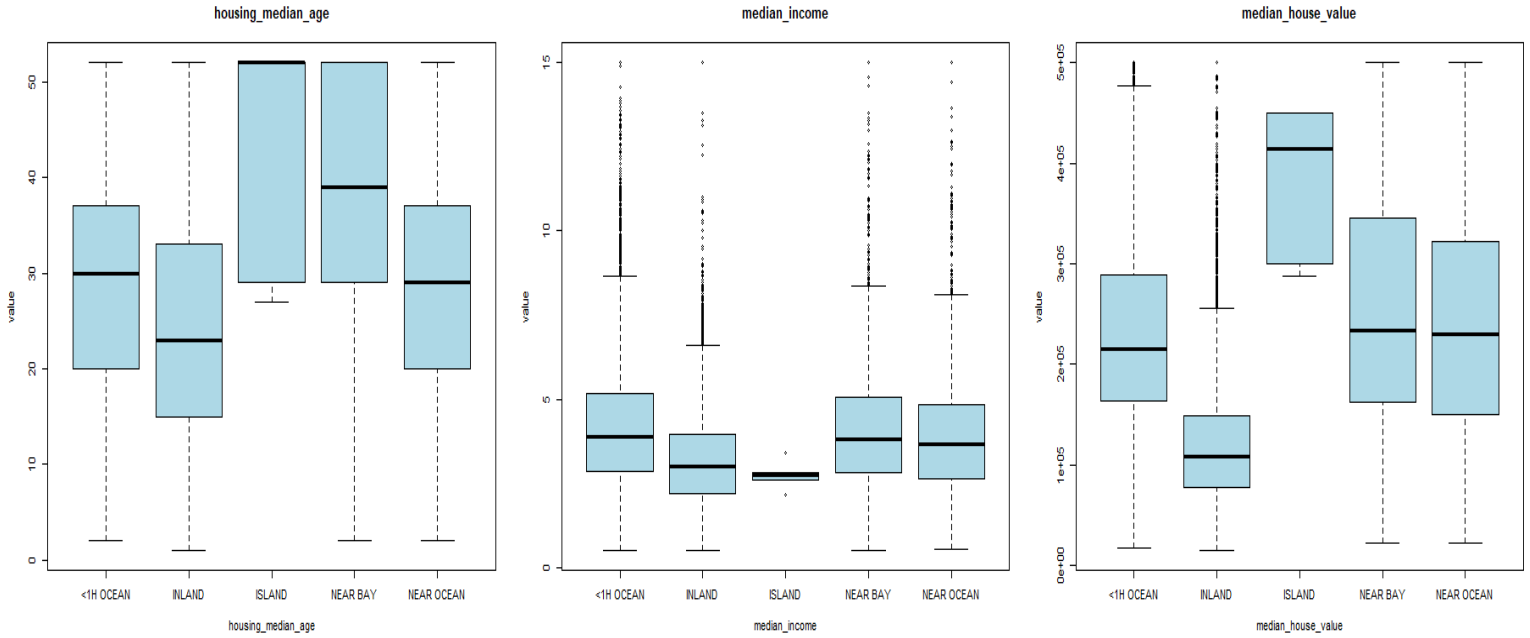
Box plots

For the box plots I have proceeded to the same way as I did for the histograms:



The latitude and longitude box plots show diversity concerning the neighbourhood location. The housing median age does not include outliers either, it shows that the study includes from very recent houses to housings aged of almost 60 years old. The median house value varies a lot as well, with houses from less than $100,000 to almost $500,000. For the following features: total bedrooms, total rooms, population, households; these contain a lot of outliers, whereas their first and third quartiles are close. It means that some observations gather more than the average tendency of other observations from the same variable.

Below, the 3 box plots display the house median age, the median income and the median house value according to their location from the ocean by using the ocean proximity feature:



From the box plots we can infer that the oldest houses are located on the island parts with a median above 60, while the more recent ones are in the inland part with a median of 23. The island areas gather mainly recent houses due to the low variance.

However, if the island part has both the oldest houses and the most valuable houses with a median greater than $400,000, it is the lowest median income of the 5 groups. This observation might lead us to infer that the housing median age is related to the median house value. Indeed, if we compare the housing median age box plot and the median house value box plot, we can recognize a similar tendency. The cheapest area remains the inland part with the newest houses.

Data Munging

First, to manipulate the ocean_proximity variable I had to convert it into a categorical one. Then, the dataset contained some missing values for the total_bedrooms feature that I replaced by the median value. Once my data cleaning done, I created 5 categorical variables related to the area situation (near the bay, ocean, island, inland, greater than 1 hour from the ocean); hence, I removed the ocean_proximity column. I did another feature modification about the number of rooms by computing the mean number of rooms and bedrooms per household. After adding these two numeric features I removed the total_rooms and total_bedrooms as well. To end up my data munging process I have scaled every numeric feature except the responsive variable which is the median house value.

My new cleaned dataset contains 14 variables with 20,640 entries.

Statistical Model

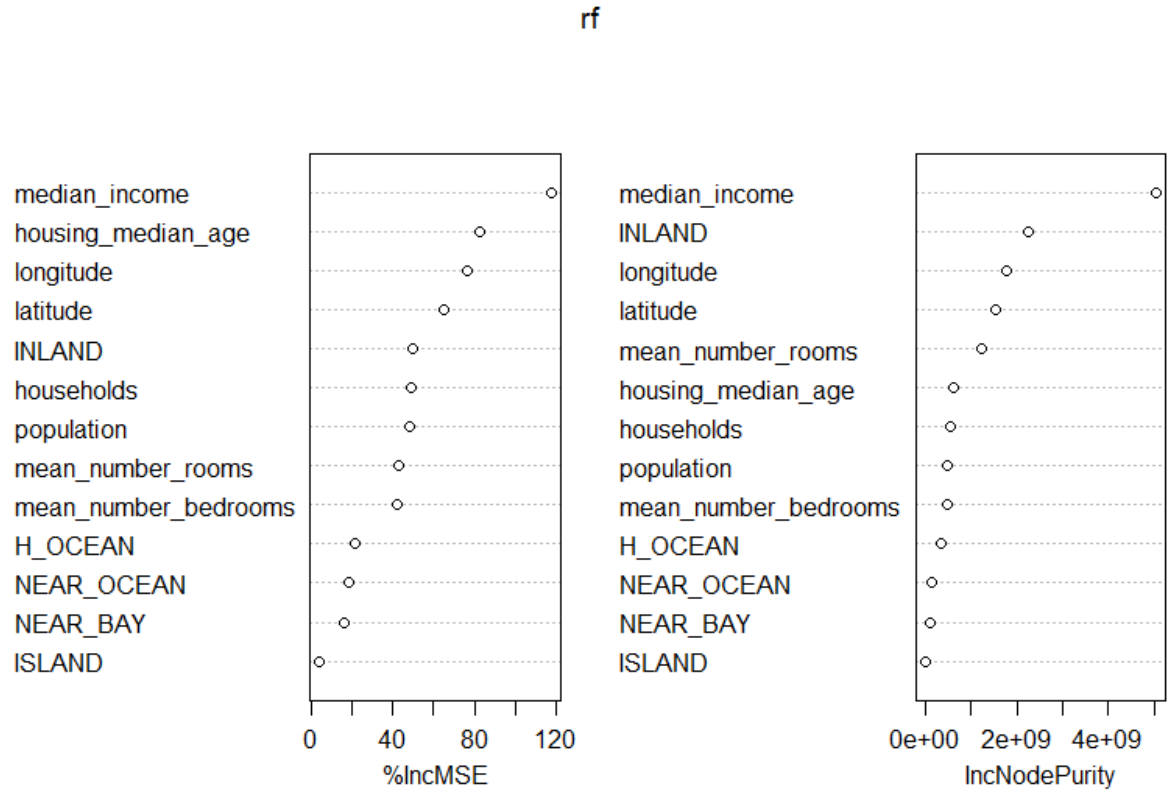
For this part, I have prepared the data to be fitted to a random forest model. The data has been split into a 70-30 percentage for the train and test sets respectively. I have set a random seed of 123 to preserve the samples. Once I obtained my two sets, I have separated both to compute the x part with all the features and the y part including only the responsive variable (the one to be predicted) the median house value. After doing these steps, I applied my random forest model of 500 trees to fit the data for the training part.

Metrics & Results

The random forest model gives a RMSE of 431.0554 which is not so bad regarding a median house value prediction of hundreds of thousands of dollars.

On the test set, the RMSE is even a bit better with a value of 426.5007. Between the rmse on the train and the test sets there is only less than 5 points of difference with a slight improvement for the test set. This means that the model is not overfitting and confirm good predictions.

The graph below displays the importance of each variable to make predictions. The median income is the most important feature, while ISLAND (whether the neighbourhood is in the inland part) is the less.



Let's try to compute the model again but without the 4 least important features, which are H_OCEAN, NEAR_OCEAN, NEAR_BAY, ISLAND.

The results are not better than the first model. Based on the RMSE, the first model with every feature is a bit more performant. It has a RMSE on the test set of 426.5007 against 427.1518 for the second lightened model.

Business Answer

After analysing the models, it seems that the median income and the housing median age are the most significant features with the geographical coordinates and the Inland features to predict the median house value. A lightened model has been tested to see if it is more performant than the basic one; however, it turns out that it is not the case thus, it is preferable to keep every feature to make predictions about the median house price.