Project Report: Prediction of House Values in California

1. Goals and Hypothesis

This project explores the factors that influence the house value in California by using a supervised machine learning technique. The response variable (`median_house_value`) belongs to the California Housing data set which was published in 1997 and is based on census data collected in 1990. The project aims to understand the relationship between the collected variables such as income, house size, age, or proximity to the ocean, and the house value from the data set. The goal is to predict house values in the future based on the insights gained, as the hypothesis is that the house value is dependent on one or more of the collected variables.

2. Description of Data Set and all Variables

The California Housing data set is stored in the data frame `housing` and consists of the following variables which all refer to a census block group. A census block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data.

Table 1: Variables of the California Housing data set

Name	*Class of the Variable*	*Description (refers to a census block group)*
`longitude`	numeric	coordinate that describes the geographic location
`latitude`	numeric	coordinate that describes the geographic location
`housing_median_age`	numeric	median age of the house
`total_rooms`	numeric	total number of rooms
`total_bedrooms`	numeric	total number of bedrooms
`population`	numeric	population
`households`	numeric	total number of households
`median_income`	numeric	median income of households
`median_house_value`	numeric	median house value
`ocean_proximity`	factor	distance to ocean measured in five levels

3. Discussion of EDA and Data Visualization

The Exploratory Data Analysis (EDA) shows that `housing_median_age`, `median_income`, and `median_house_value` are higher in `NEAR BAY` than in `INLAND`. Moreover, `total_rooms`, `total_bedrooms`, `population`, `households`, and `median_income` have outliers as the maximum values are very high and their probability distributions are right-skewed. `total_bedrooms` also consists of many NA values, 207 were counted. The results found can be seen in the following histograms and boxplots of all numeric variables.

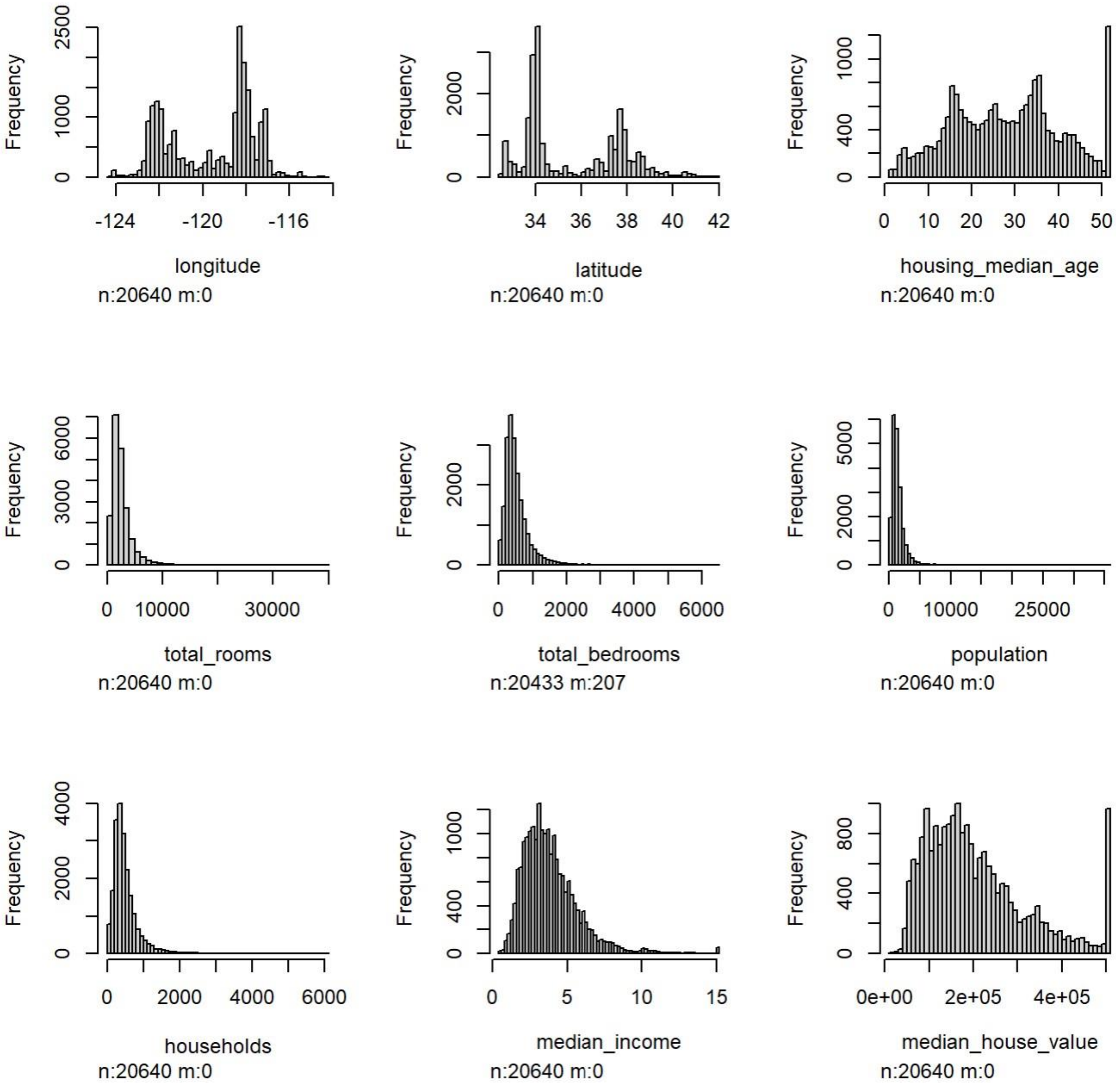


Figure 1: Histograms of all numeric variables

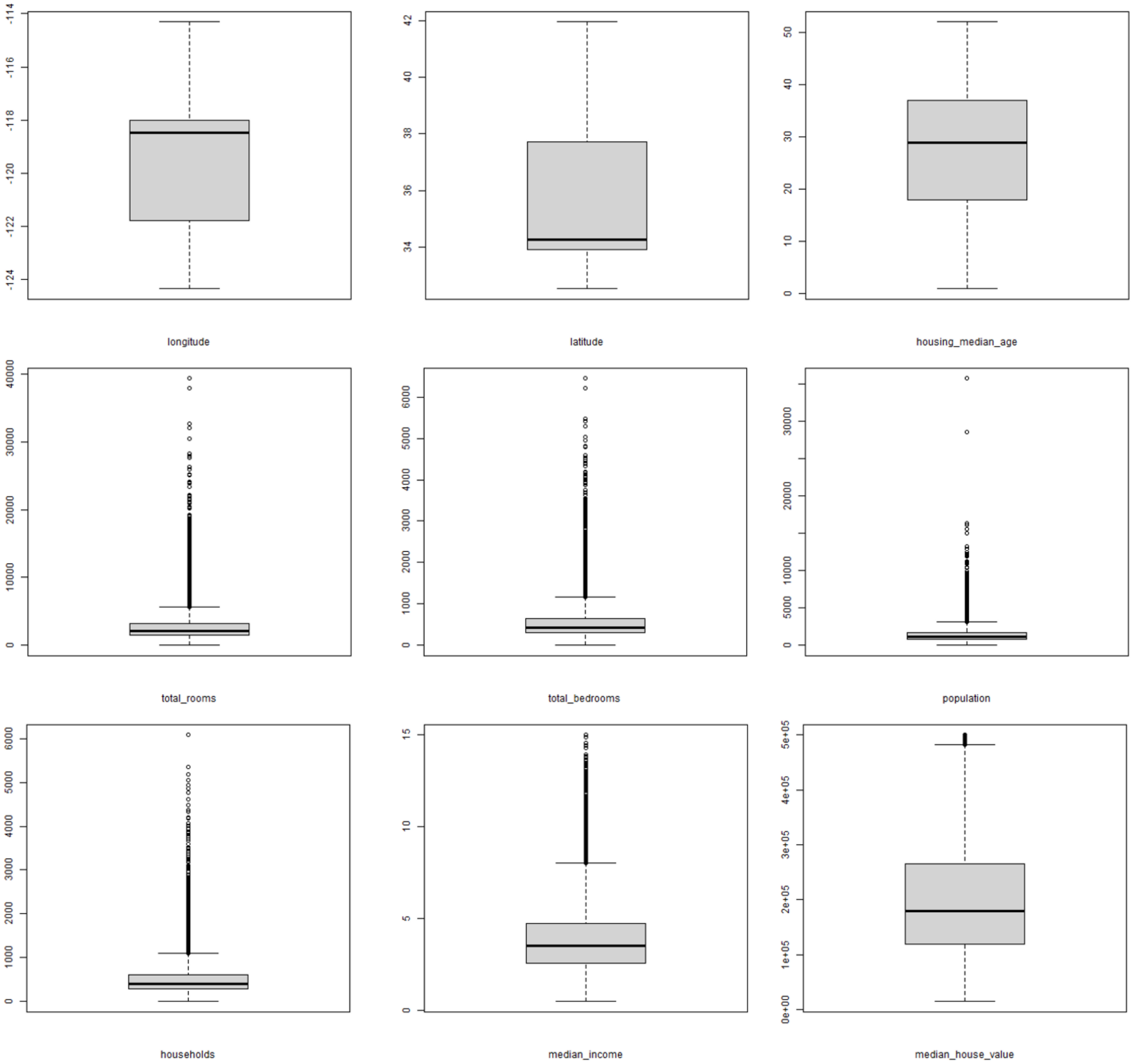


Figure 2: Boxplots of all numeric variables

Furthermore, in the following boxplots distributed by ocean proximity, it can be seen that the sample for the `ocean_proximity` level `ISLAND` is very small, with only five observations. Therefore, the observations of the `ISLAND` level have the least variance in all three plots displayed. The age of the houses of the `ISLAND` level is the highest, followed by `NEAR BAY`. The income of the `ISLAND` level is the lowest, while the house value of this level is the highest. The house value of the `INLAND` level is the lowest overall. Therefore, the variable `ocean_proximity` could be a possible feature variable for the `median_house_value`.

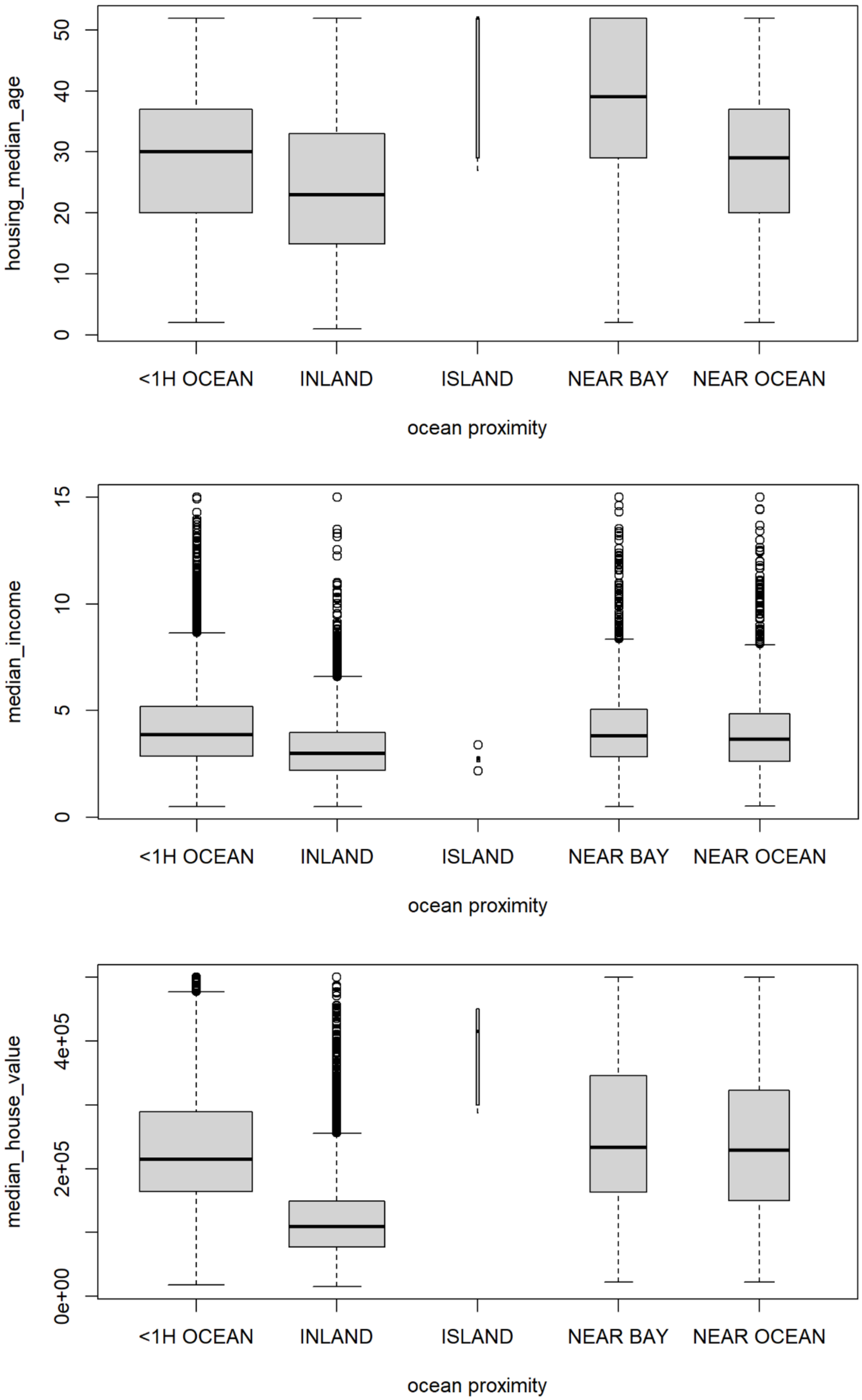


Figure 3: Boxplots by ocean proximity of housing_median_age, median_income, and median_house_value

The correlation analysis below shows that `longitude` and `latitude` are not dependent on the other variables. There is a slight negative dependence between `housing_median_age` and the other variables, except for the two mentioned above. `total_rooms`, `total_bedrooms`, `households` and `population` have a strong dependence. The house value only correlates strongly with income, which could indicate a possible feature variable for the `median_house_value`.

Table 2: Pearson coefficients of correlation between all numeric variables (indicates p < .05; ** indicates p < .01)*

Variable	*M*	*SD*	1	2	3	4	5	6	7	8
1. longitude	-119.57	2.00								
2. latitude	35.63	2.14	-.92** [-.93, -.92]							
3. housing_median_age	28.64	12.59	-.11** [-.12, -.09]	.01 [-.00, .02]						
4. total_rooms	2635.76	2181.62	.04** [.03, .06]	-.04** [-.05, -.02]	-.36** [-.37, -.35]					
5. total_bedrooms	537.87	421.39	.07** [.06, .08]	-.07** [-.08, -.05]	-.32** [-.33, -.31]	.93** [.93, .93]				
6. population	1425.48	1132.46	.10** [.09, .11]	-.11** [-.12, -.10]	-.30** [-.31, -.28]	.86** [.85, .86]	.88** [.87, .88]			
7. households	499.54	382.33	.06** [.04, .07]	-.07** [-.08, -.06]	-.30** [-.32, -.29]	.92** [.92, .92]	.98** [.98, .98]	.91** [.90, .91]		
8. median_income	3.87	1.90	-.02* [-.03, -.00]	-.08** [-.09, -.07]	-.12** [-.13, -.11]	.20** [.18, .21]	-.01 [-.02, .01]	.00 [-.01, .02]	.01 [-.00, .03]	
9. median_house_value	206855.82	115395.62	-.05** [-.06, -.03]	-.14** [-.16, -.13]	.11** [.09, .12]	.13** [.12, .15]	.05** [.04, .06]	-.02** [-.04, -.01]	.07** [.05, .08]	.69** [.68, .70]

4. Description of Data Pipeline (Data Munging Steps)

Before the feature variables can be finally selected, data munging steps are necessary to clean up the existing `housing` data frame. Therefore, four different data munging steps are performed, each resulting in a new data frame that is serially numbered, e.g., `housing_1`.

1. Imputation of missing values (NA's) in `total_bedrooms` by using the median and the `impute()` function to obtain a better model.
2. Transformation of the factor variable `ocean_proximity` into a binary categorical variable for better handling the variable in supervised machine learning.
3. Creation of two new variables based on `total_bedrooms` and `total_rooms` by dividing the values by households to get the `mean_number_bedrooms` and `mean_number_rooms`.
4. Feature scaling of each numeric variable except `median_house_value` (response variable) and the binary categorical variables of `ocean_proximity`.

The resulting data frame is stored in `cleaned_housing` with all variables moved in the order given below.

Table 3: Variables of the California Housing data set after data transformation

Name	*Class of the Variable*	*Description (refers to a census block group)*
`NEAR_BAY`	logical	binary categorical variable for ocean proximity
`<1H_OCEAN`	logical	binary categorical variable for ocean proximity
`INLAND`	logical	binary categorical variable for ocean proximity
`NEAR_OCEAN`	logical	binary categorical variable for ocean proximity
`ISLAND`	logical	binary categorical variable for ocean proximity
`longitude`	numeric	coordinate that describes the geographic location
`latitude`	numeric	coordinate that describes the geographic location
`housing_median_age`	numeric	median age of the house
`population`	numeric	population
`households`	numeric	total number of households
`median_income`	numeric	median income of households
`mean_number_rooms`	numeric	mean number of rooms
`mean_number_bedrooms`	numeric	mean number of bedrooms
`median_house_value`	numeric	median house value

5. Description of Statistical Model and Performance Metric Results

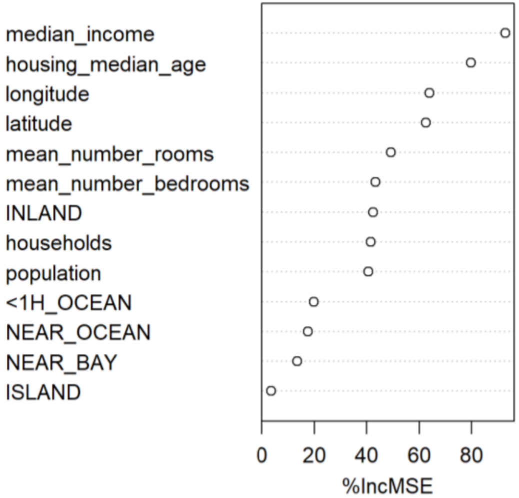
The data set cleaned_housing is divided into a training set (70%) and a test set (30%). A second exploratory analysis on the training set shows a positive trend between median_income and median_house_value, confirming the relationship found earlier.

Next, the randomForest() algorithm is applied to the training set for training and inference. The response variable is median_house_value, while all other variables act as the feature vector. The resulting trained model is stored in rf.

As residuals are a measure of how far from the regression line data points are, the Root Mean Squared Error (RMSE) is a measure of how spread out these residuals are and how concentrated the data is around the line of best fit. Lower values of RMSE indicate better fit. The calculated RMSE for the rf model is 49,566.27, while the calculated RMSE for the test set using the predicted and the actual values of the variable median_house_value is 48,700.28. Both RMSEs are close together indicating a good prediction performance and no overfitting of the model.

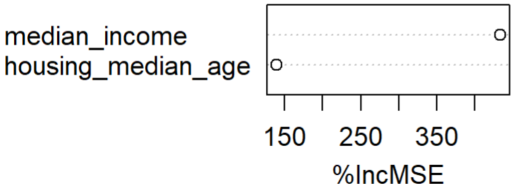
The variable importance plot below shows that the selected feature vector can be further adjusted to improve the model performance. Age and income have the greatest impact on predicting the house values. Therefore, the model is retrained using only housing_median_age and median_income as the feature vector.

Table 4: Variable importance plot of random forest model



After retraining the model, the difference between RMSE train (80,270.55) and test (80,489.76) is closer, but the absolute value is now higher. This means that the model predicts the house value about 80,000 off the actual value. The variable importance plot below confirms that income is the best predictor for the house value. Therefore, a linear regression model with income as the independent variable can also be considered.

Table 5: Variable importance plot of retrained random forest model



6. Conclusion

In this project, the `randomForest()` algorithm was used to predict the `median_house_value` of the California Housing data set. The median house value is primarily influenced by the income of the households, followed by the age of the house. Since the performance of the model is within a good range, the model can be used to predict house values in the future based on a given income, age, or other variables collected by the U.S. Census Bureau for a census block group.