UNIVERSITÀ
DEGLI STUDI
DI PADOVA

California House Price Prediction

Statistical Learning Final Project

Betti Gianmarco – 2097050

Marinelli Andrea – 2091700

Rinaldi Giorgia – 2092226

Outline

- Objectives
- Data Collection
- Exploratory Data Analysis (EDA)
- Model Data & Analysis
- Model Evaluation
- Conclusion

Objectives

- The objective of this project is to predict median house values in California
- By exploring the data and building a regression model, we aim to achieve accurate predictions and valuable insights for the housing market

Data collection

Dataset: California House Price

Source: *https://www.kaggle.com/datasets/shibumohapatra/house-price*

Description: California House Price dataset is a publicly available dataset based on the 1990 U.S. Census Bureau. It contains information about various factors that influence the median house prices in California. Each observation represent a block group, that is a small geographical unit whose population typically range from 600 to 3000 individuals.

Dimension: 20640 observations - 10 columns

UNIVERSITÀ DEGLI STUDI DI PADOVA

Features

longitude (signed numeric - float) : Longitude value for the block in California, USA [°]

latitude (numeric - float) : Latitude value for the block in California, USA [°]

housing_median_age (numeric - int) : Median age of the house in the block [years]

total_rooms (numeric - int) : Total number of rooms (excluding bedrooms) in all houses in the block

total_bedrooms (numeric - float) : Total number of bedrooms in all houses in the block

Features

population (numeric - int) : Total number of population in the block

households (numeric - int) : Total number of households in the block

median_income (numeric - float) : Median of the total household income of all the houses in the block [10k $]

ocean_proximity (numeric - categorical) : Type of the landscape of the block [Unique Values : 'NEAR BAY', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND']

median_house_value (numeric - int) : Median of the household prices of all the houses in the block [$]

UNIVERSITÀ DEGLI STUDI DI PADOVA

Data Preparation & Cleaning

- Checked for duplicated rows (there weren't any)
- Detected and Removed missing values (for numerical variables)
- Checked unique values for categorical variable (**ocean_proximity**)

Data Preparation & Cleaning

Exploratory Data Analysis (EDA)

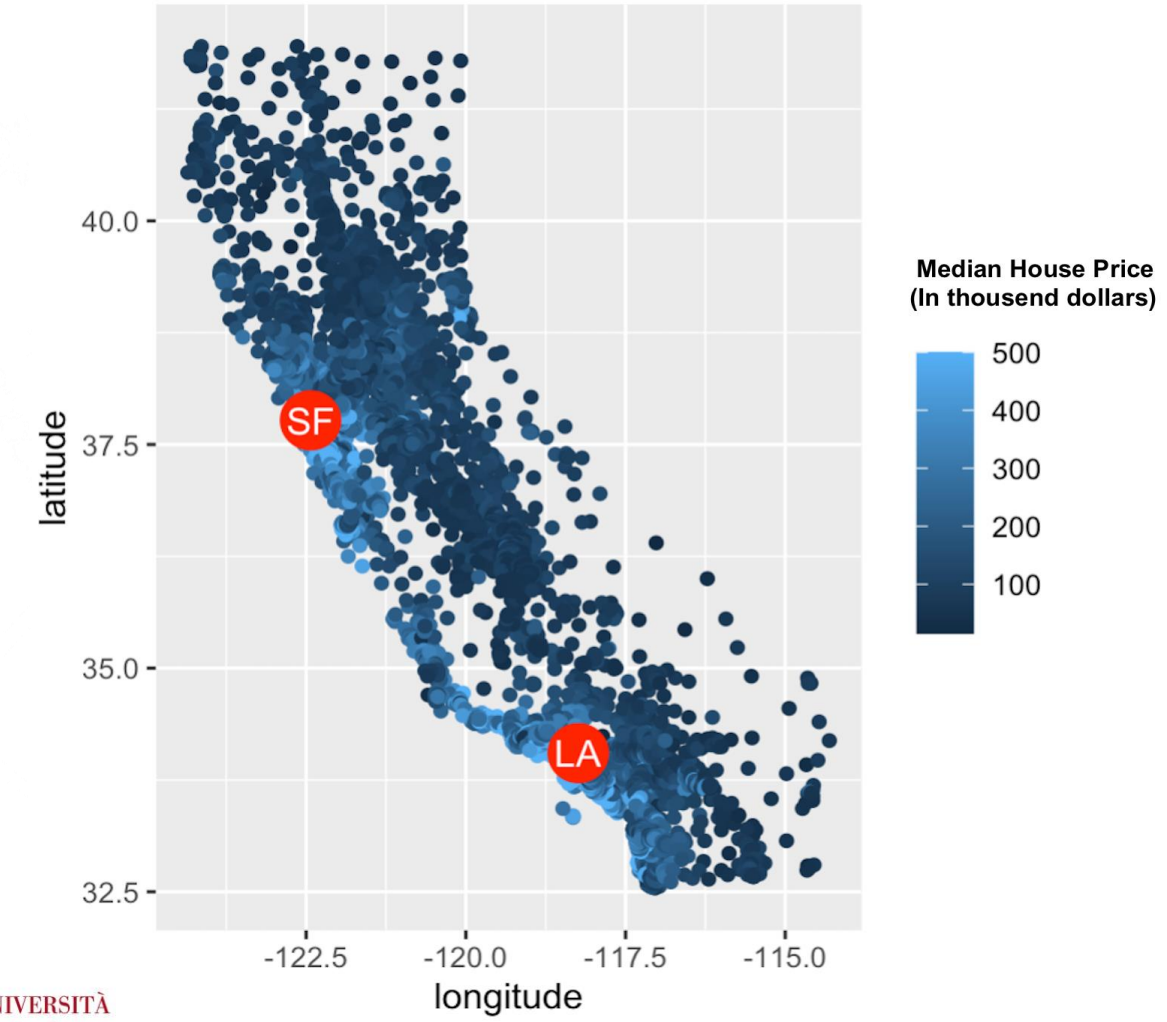
We can resume our EDA in these main phases:

- Creation of new features
- Analysing features distribution
- Detecting and treating outliers
- Checking feature correlation

Creation of New Features

Areas in close proximity to Los Angeles and San Francisco tend to exhibit higher values of the target variable, "median house value."

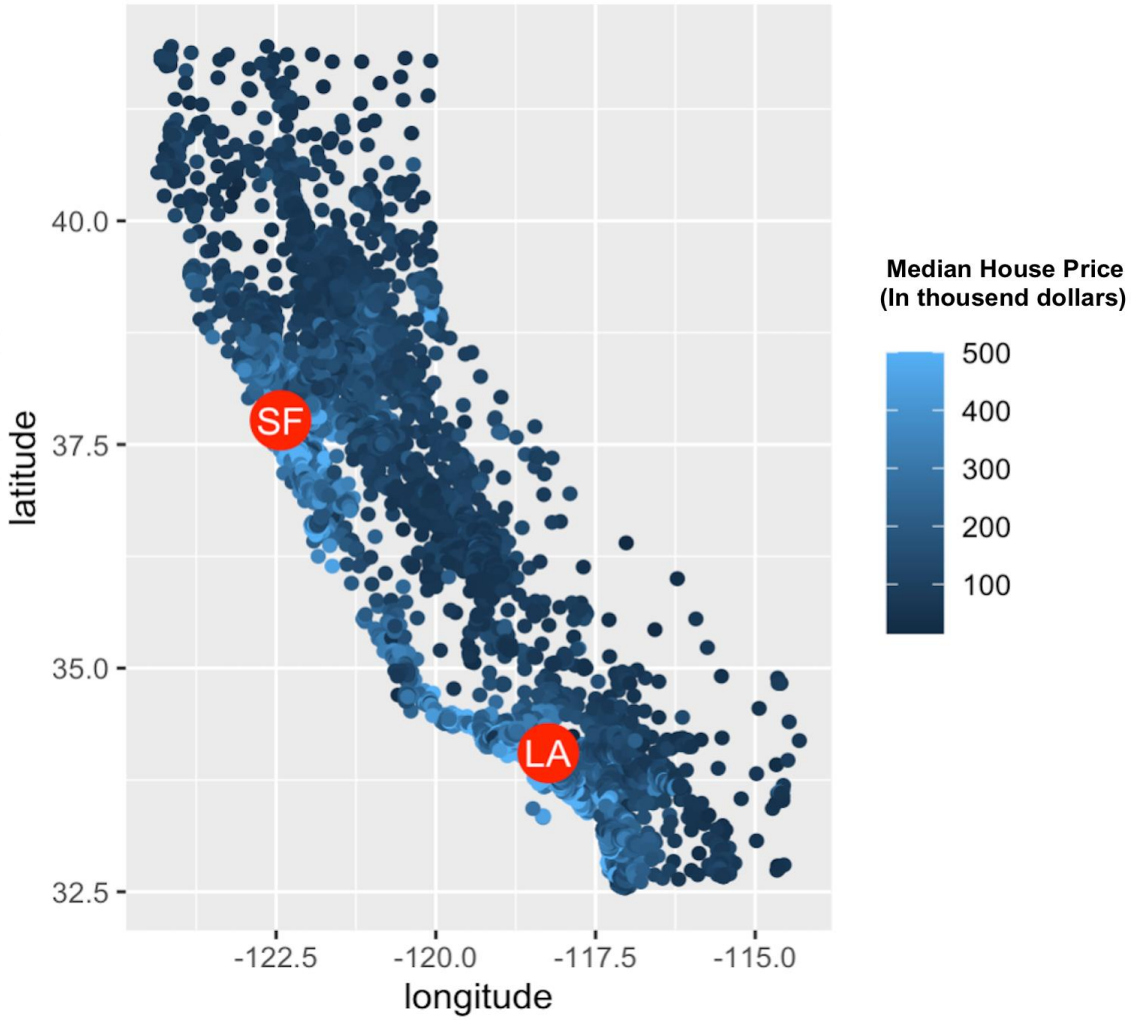
We have decided to combine "latitude" and "longitude"
to create a new numerical variable called "distance".



Creation of New Features

Let distSF (x) and distLA(x) be the euclidean distance of block x coordinates respectively from San Francisco and Los Angeles coordinates. The variable distance is defined as:

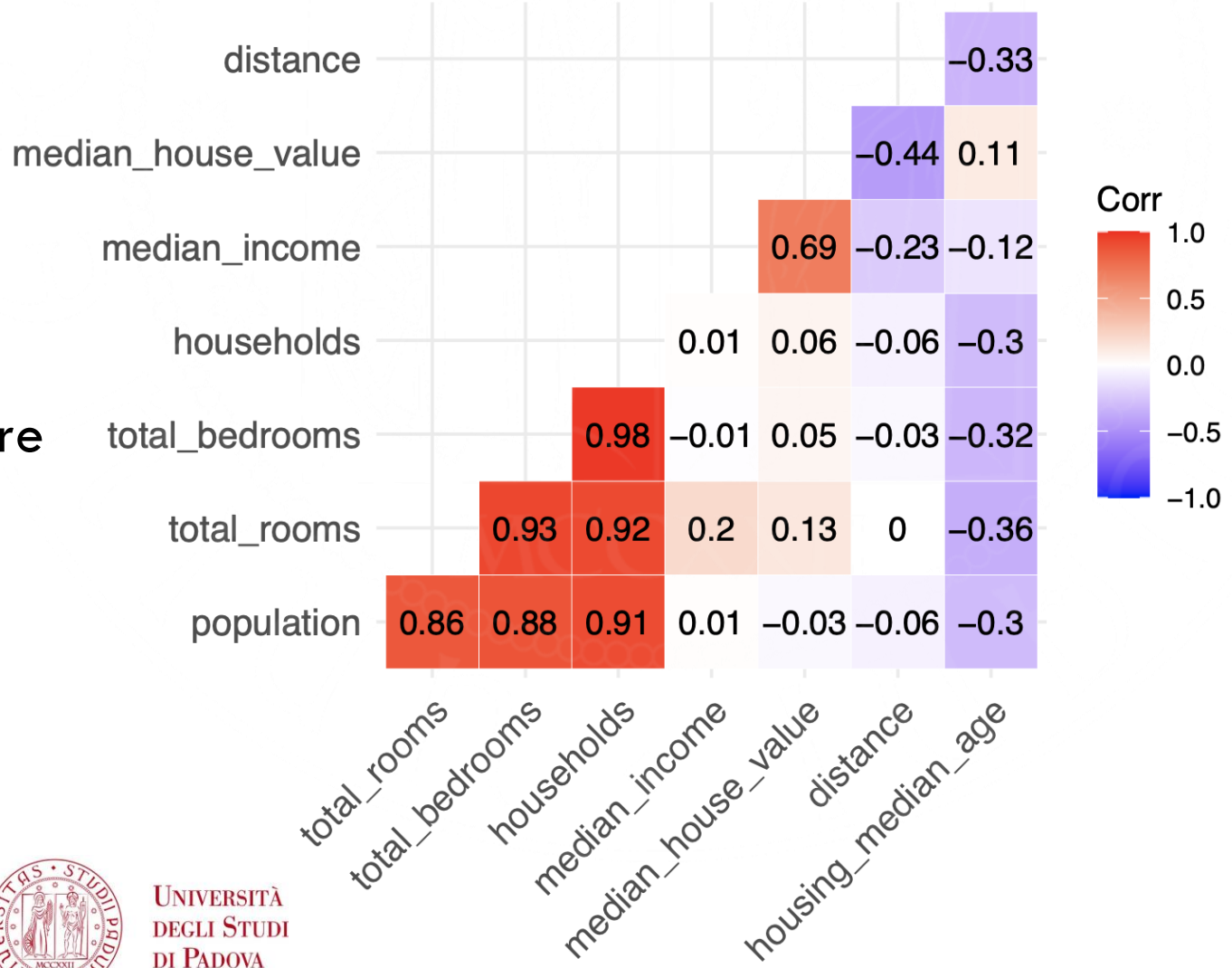
$$Distance = min\{dist_{SF}; dist_{LA}\}$$



Correlation Matrix

Looking at the correlation matrix we can inspect the correlation among variables in our current dataframe.

We can notice that some variables are high correlated.



Creation of New Features

We noticed that the variable "households" has high correlation with "total rooms" and "total bedrooms". Therefore we decide to make them interact creating:

$$rooms\_ph = \frac{total\_rooms}{households}$$

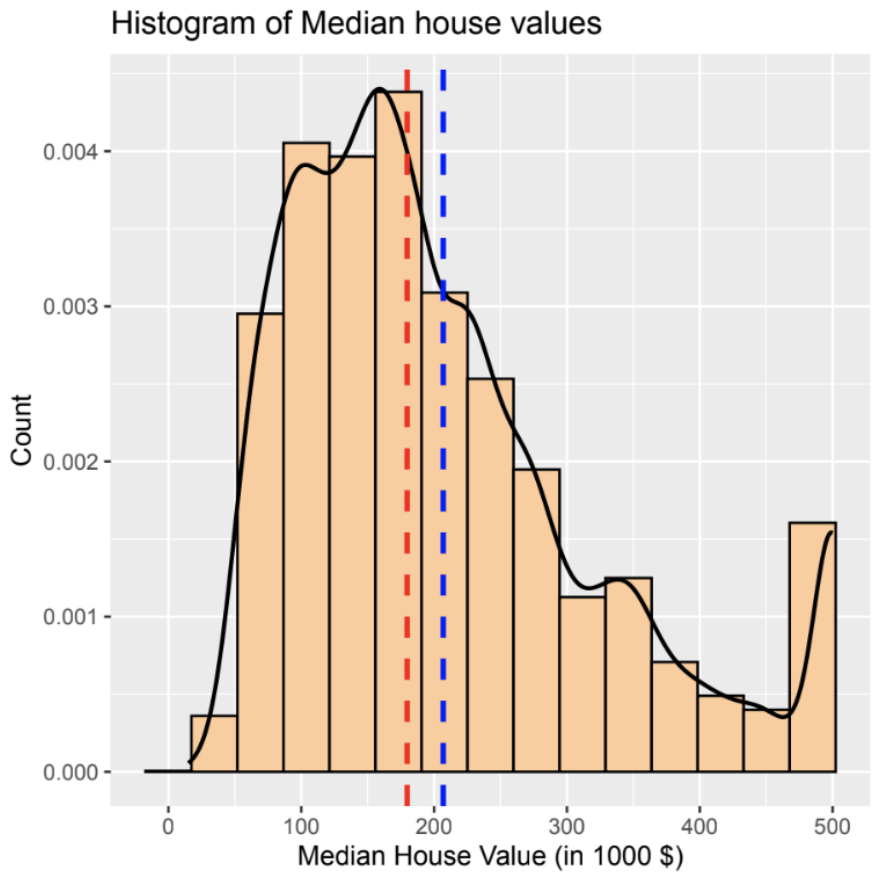
$$bedrooms\_ph = \frac{total\_bedrooms}{households}$$

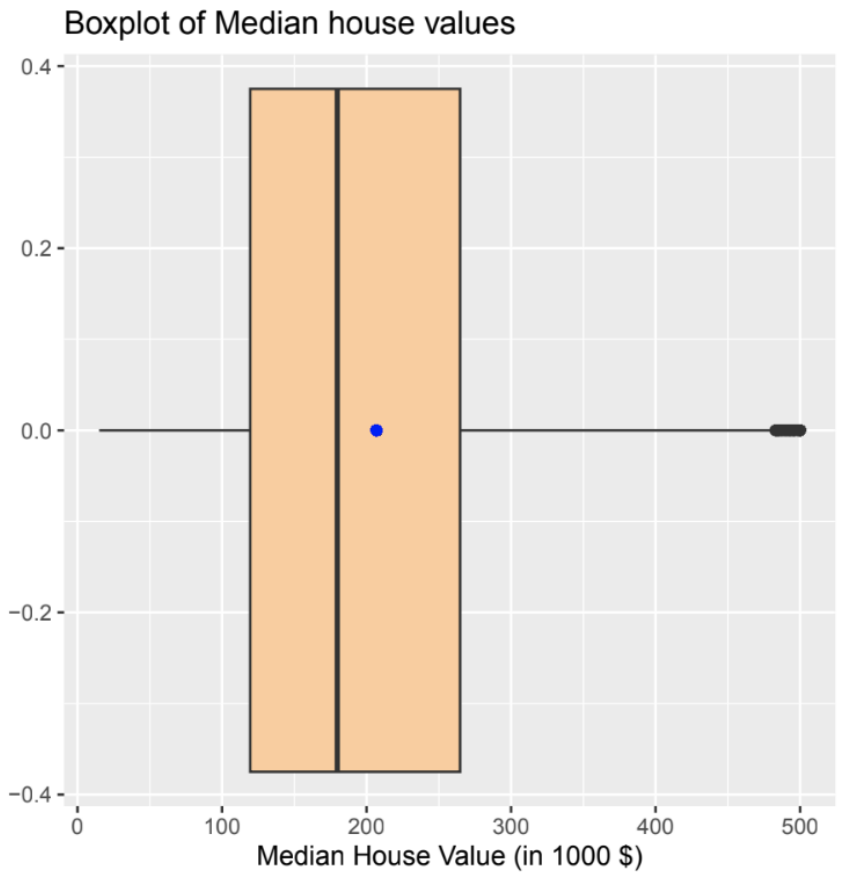
Correlation Matrix

The correlation matrix above shows an improvement. We are going to detect it again at the end of the Exploratory Data Analysis, right after the analysis of eventual outliers.

Analysing Features Distribution

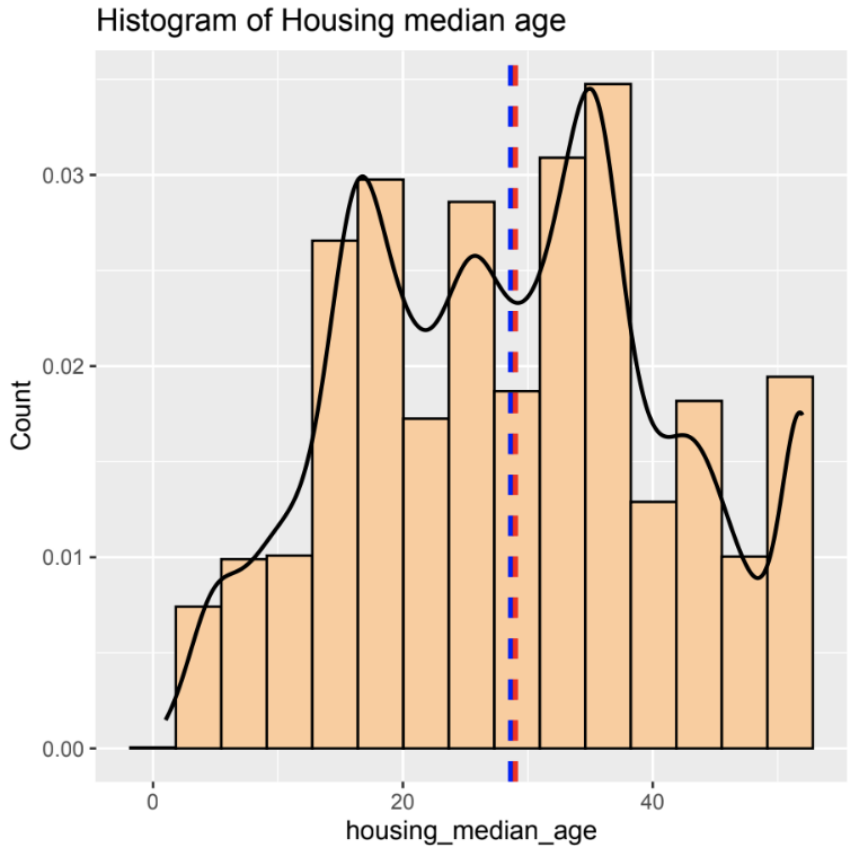
Median House Value

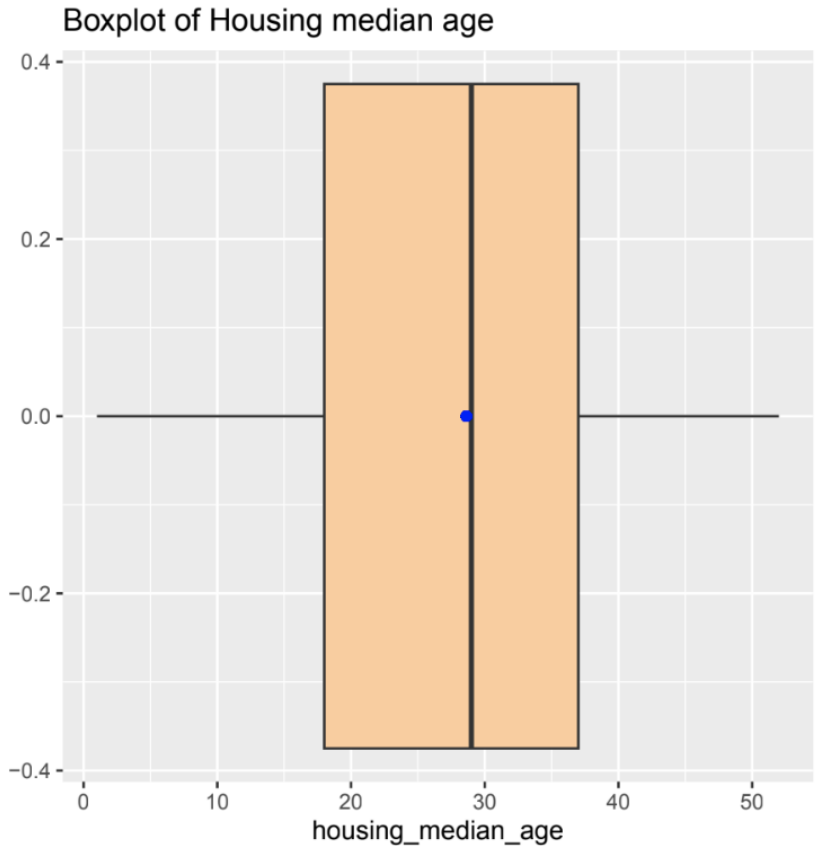




Analysing Features Distribution

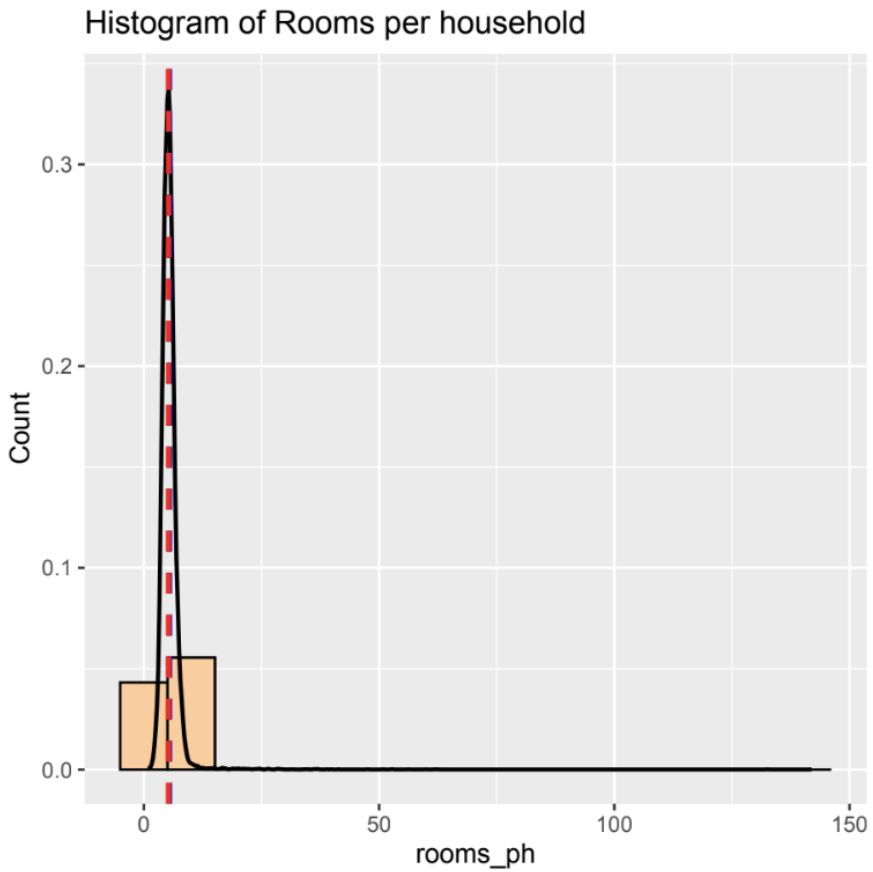
Housing Median Age

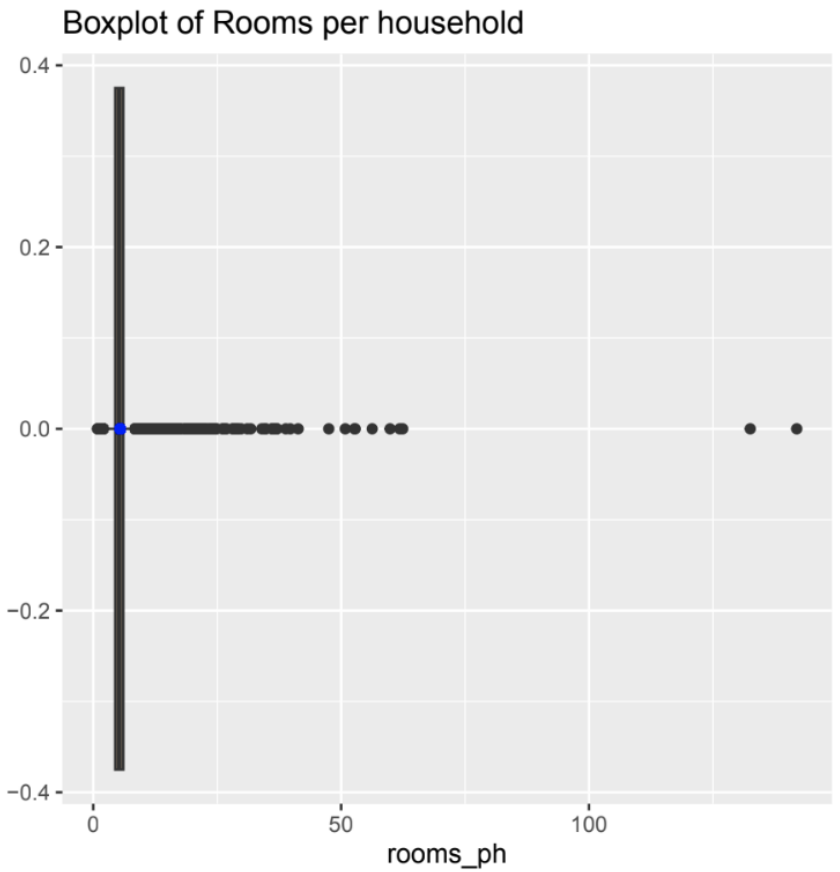




Analysing Features Distribution

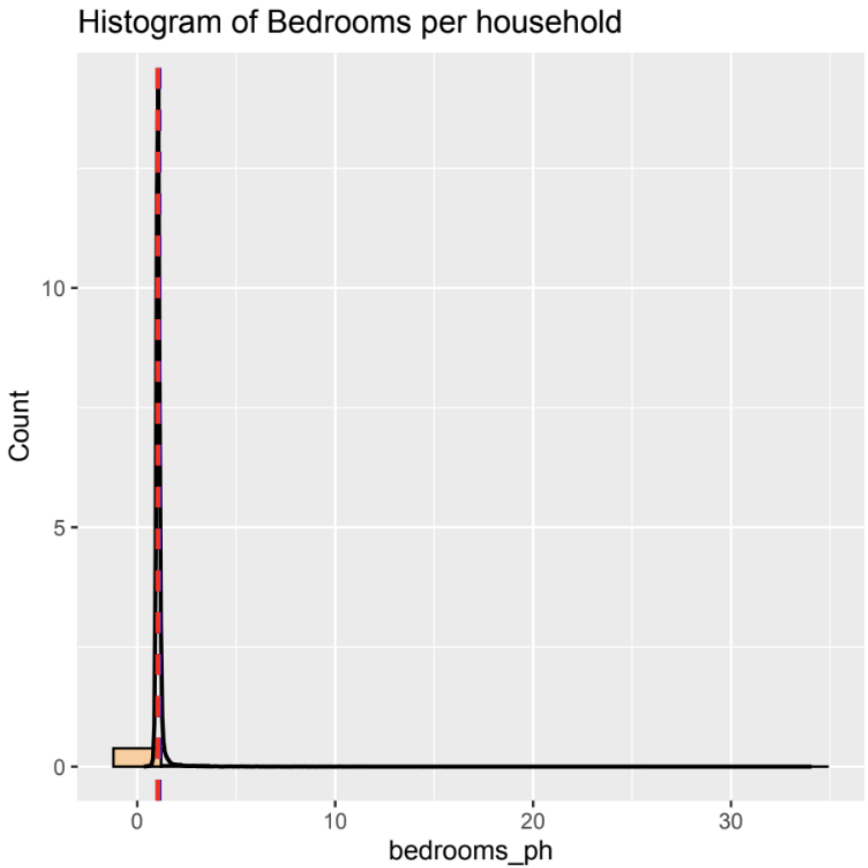
Room per Household

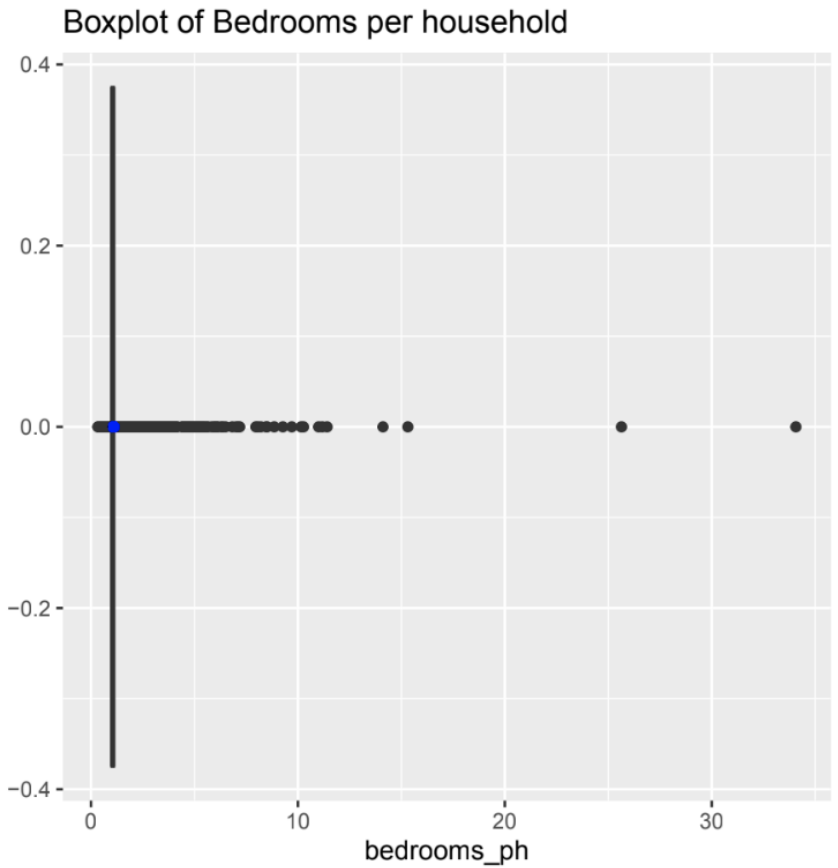




Analysing Features Distribution

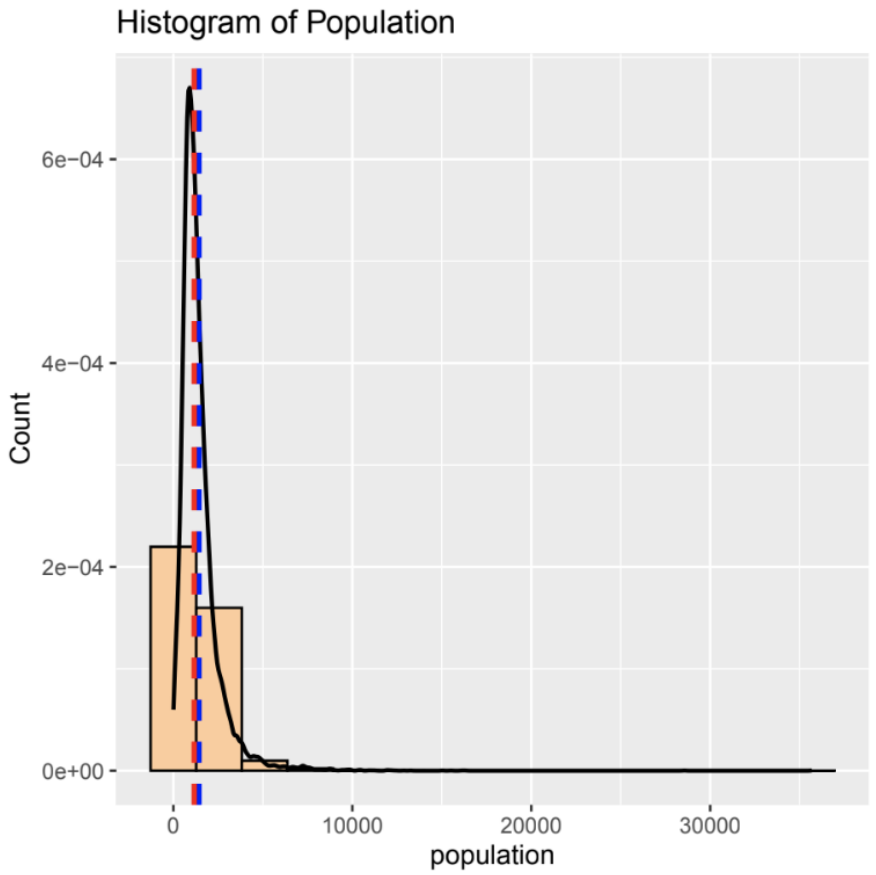
Bedrooms per Households

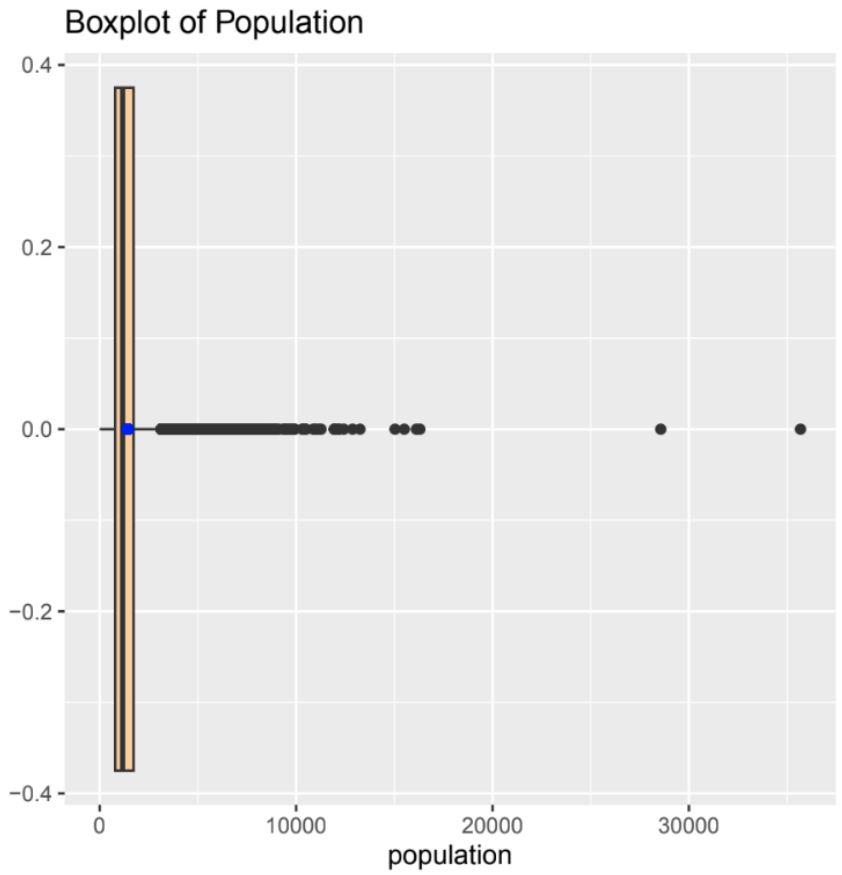




Analysing Features Distribution

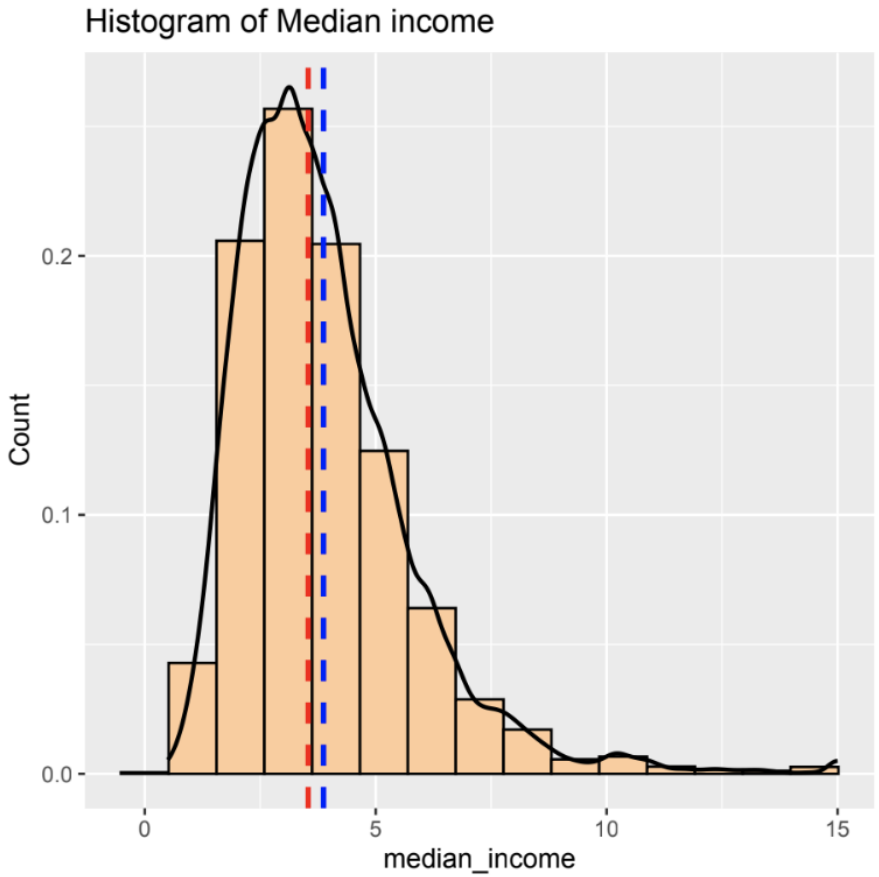
Population

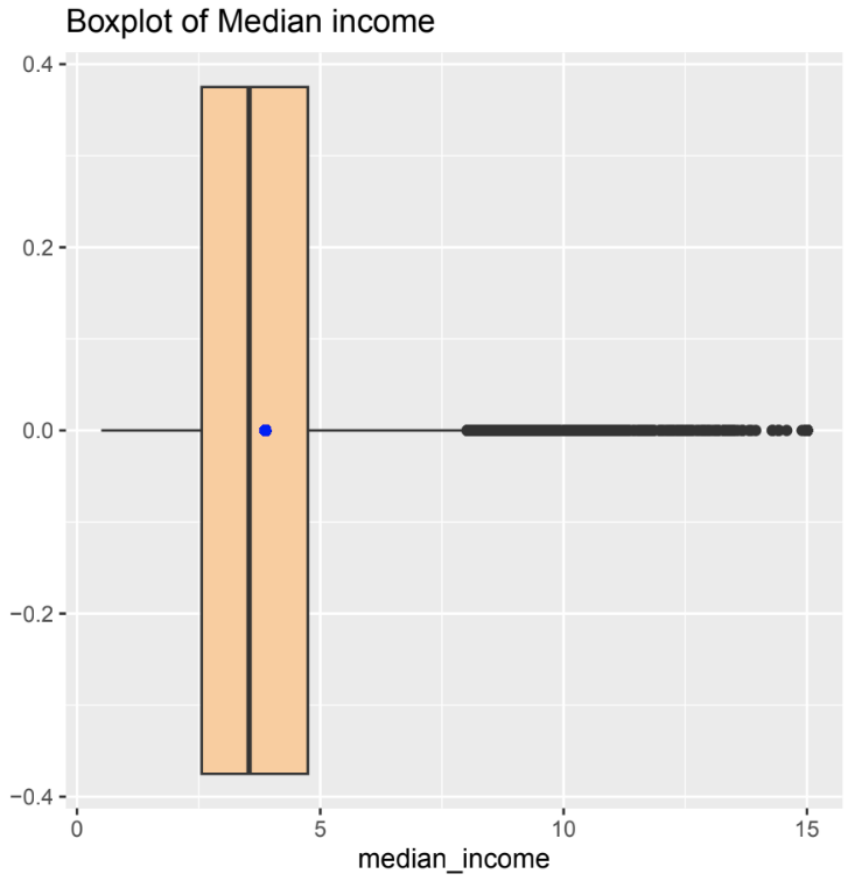




Analysing Features Distribution

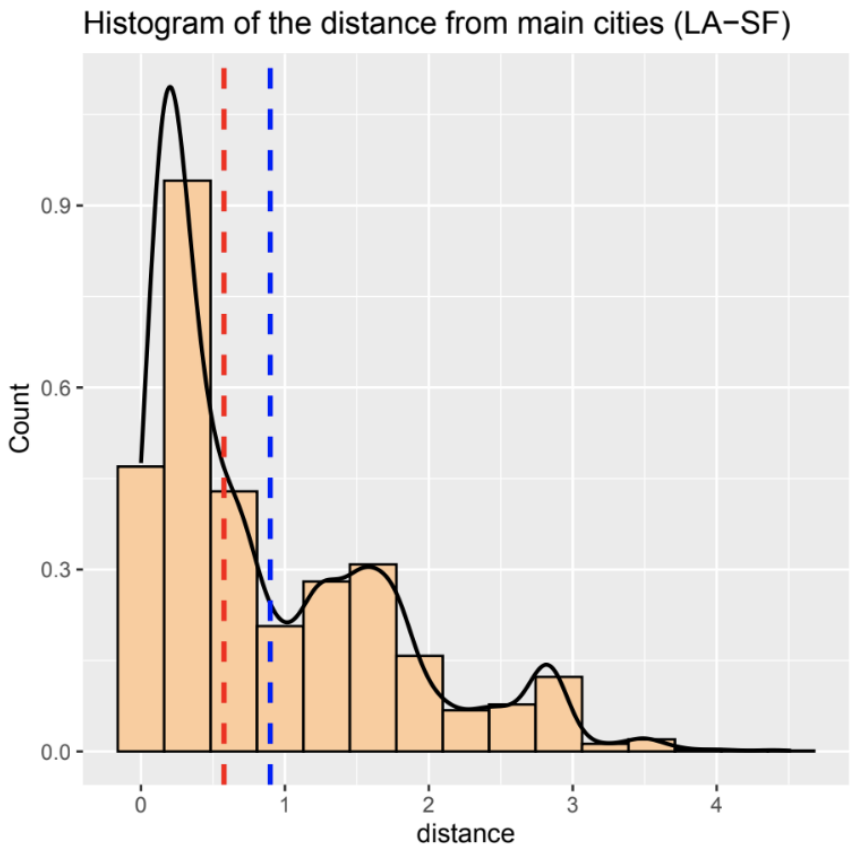
Median Income





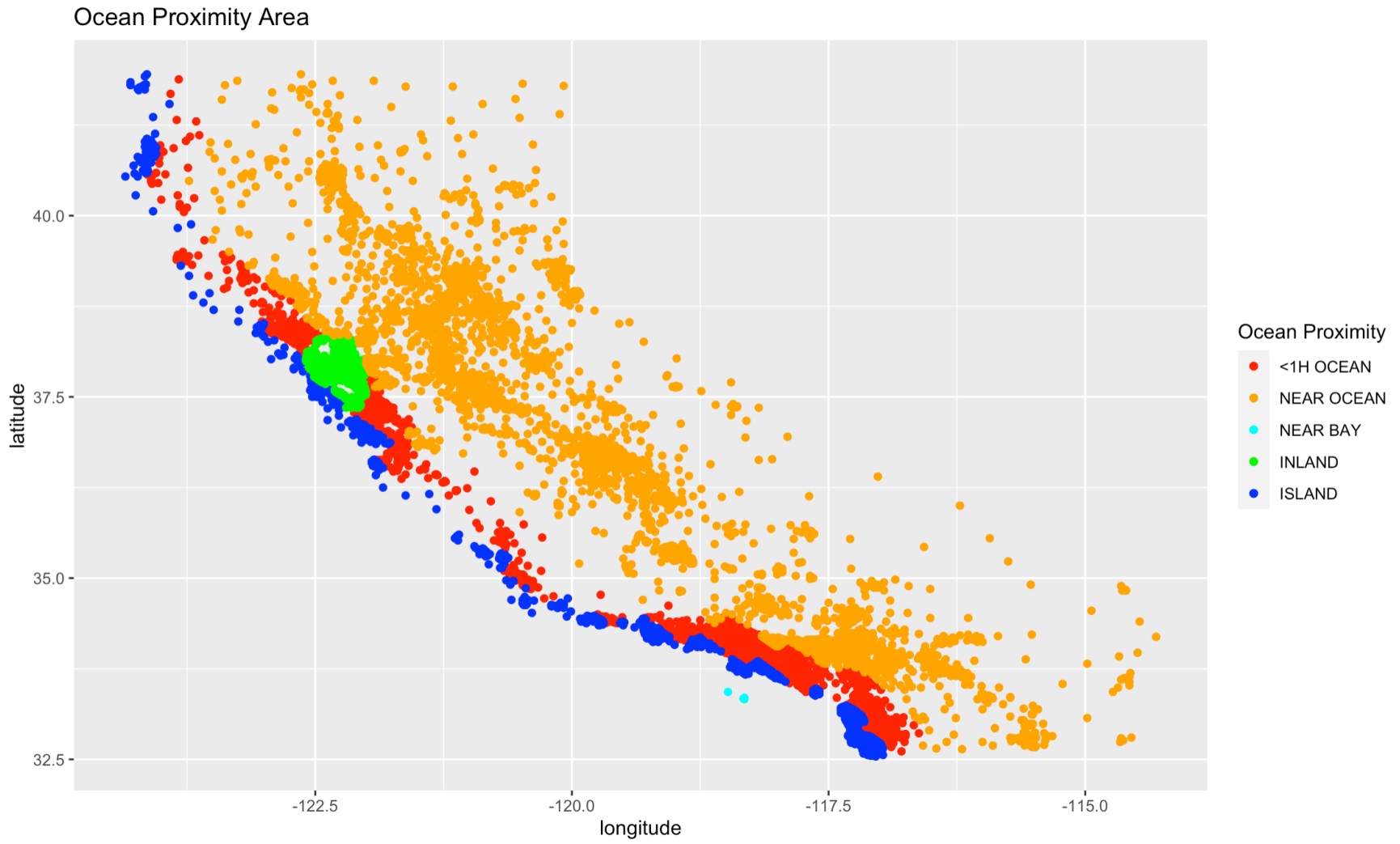
Analysing Features Distribution

Distance



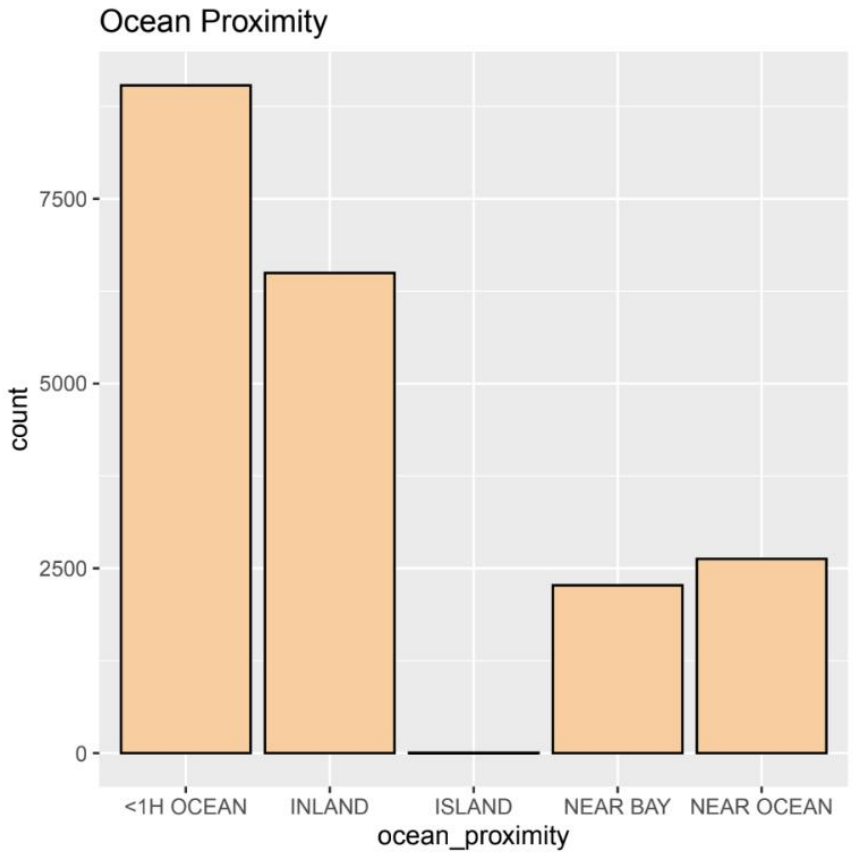
Analysing Features Distribution

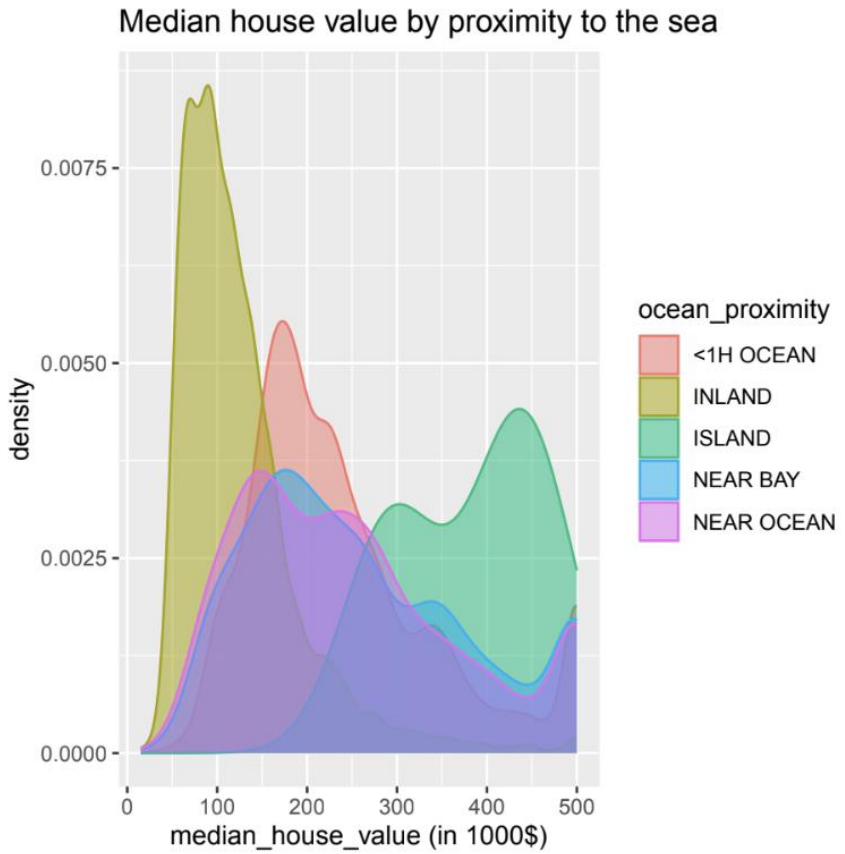
Ocean Proximity



Analysing Features Distribution

Ocean Proximity





Outliers

Previous charts have highlighted the presence of outliers. We've treated outliers by using the Interquartile Range (IQR) rule, where data points below

$$Q_1 - 1,5\ IQR$$

or above

$$Q_3 + 1,5\ IQR$$

are considered outliers and subsequently removed from the dataset.

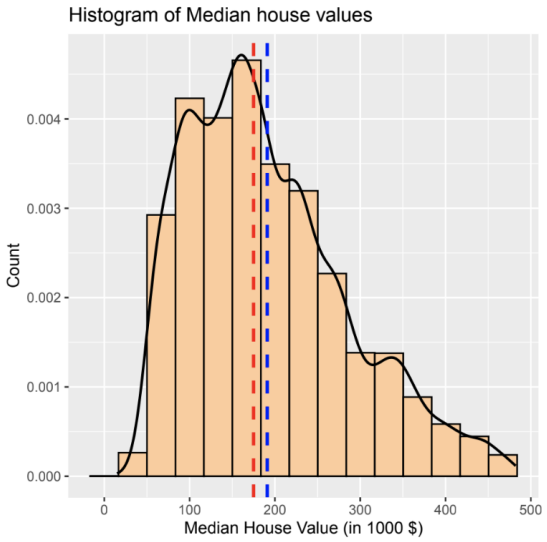
Correlation Matrix

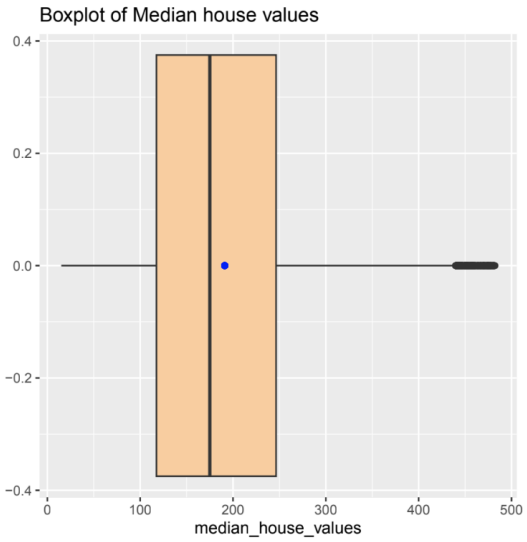
Now the cleaned dataframe has 16665 observations and 8 variables.

Now we check again correlation matrix in order to see if correlation among variables improves deleting outliers.

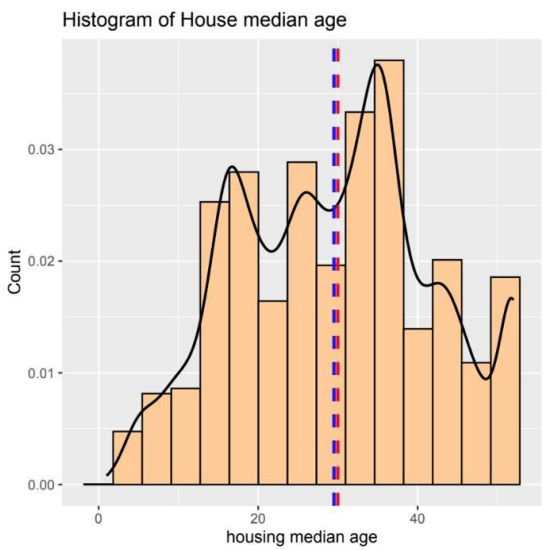
Analysing Features Distribution

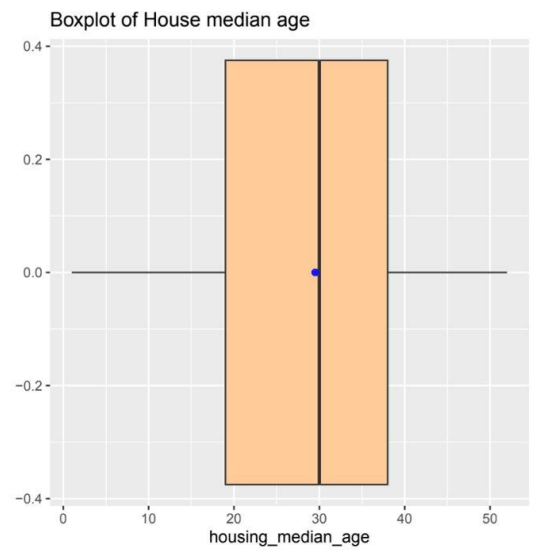
Median House Value





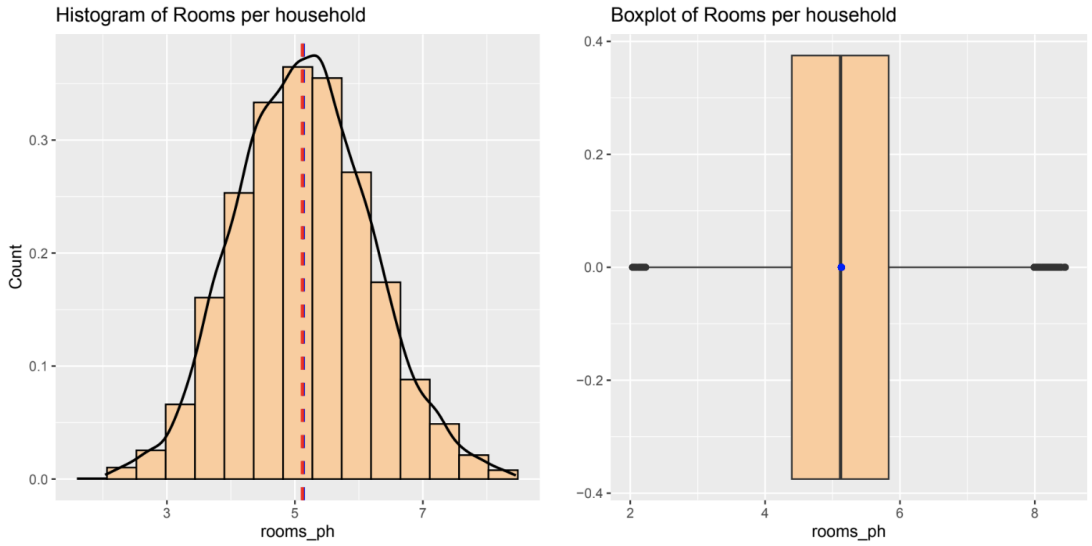
Housing Median Age



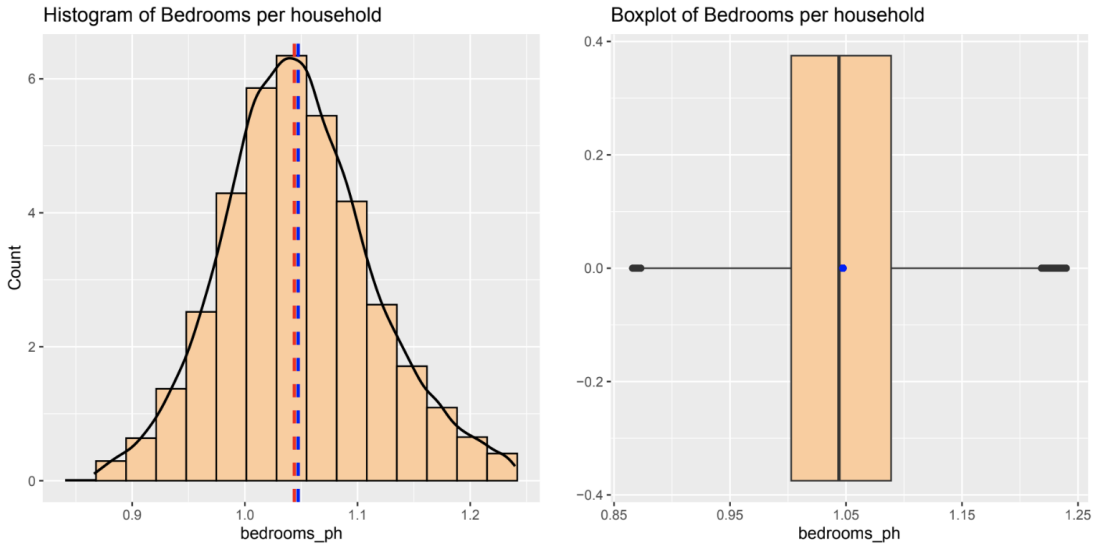


Analysing Features Distribution

Room per Household

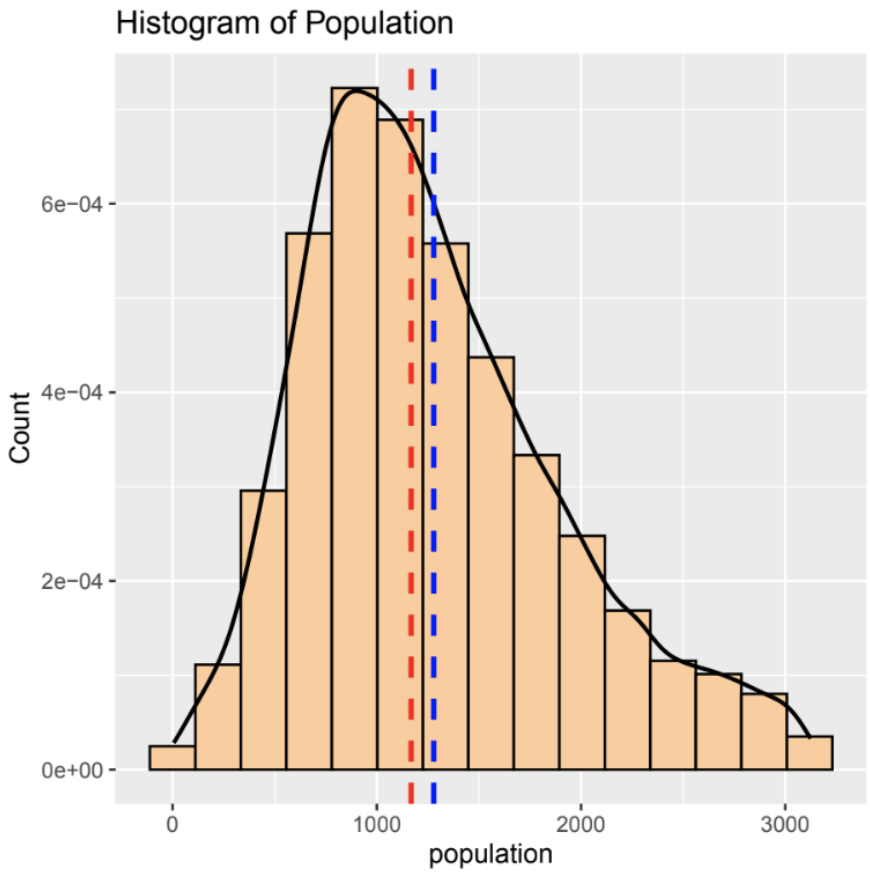


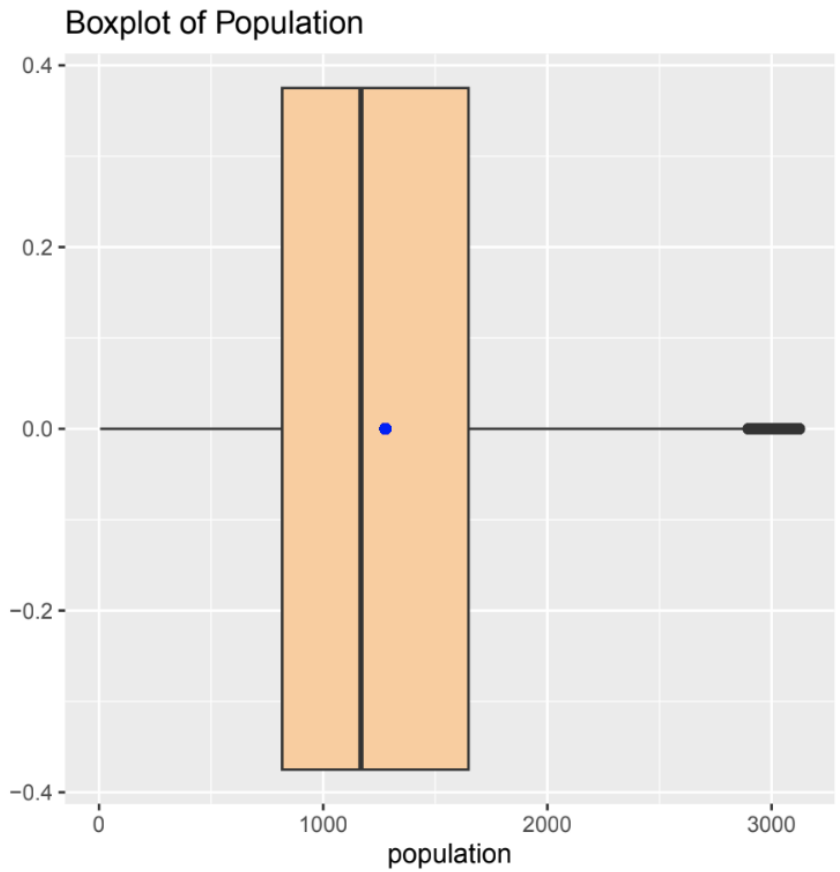
Bedrooms per Households



Analysing Features Distribution

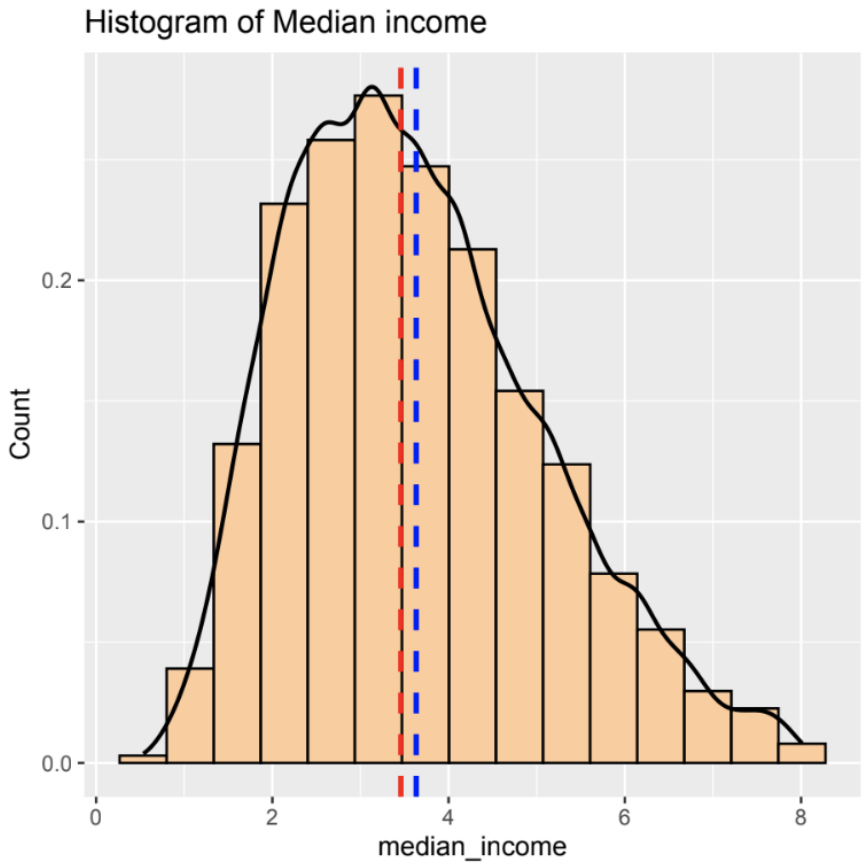
Population

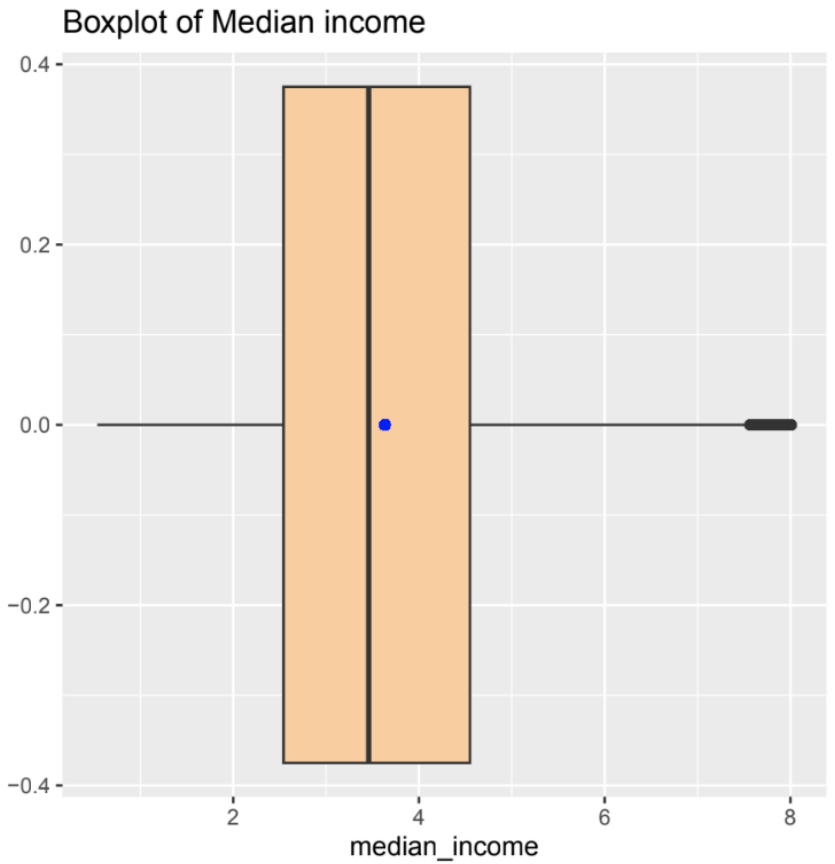




Analysing Features Distribution

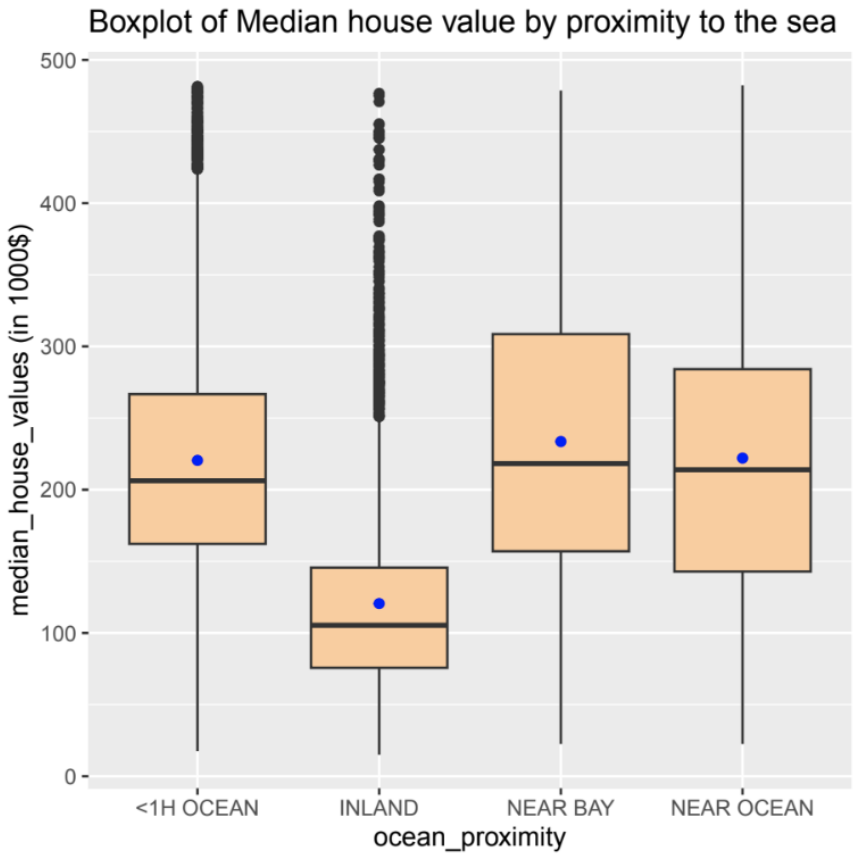
Median Income

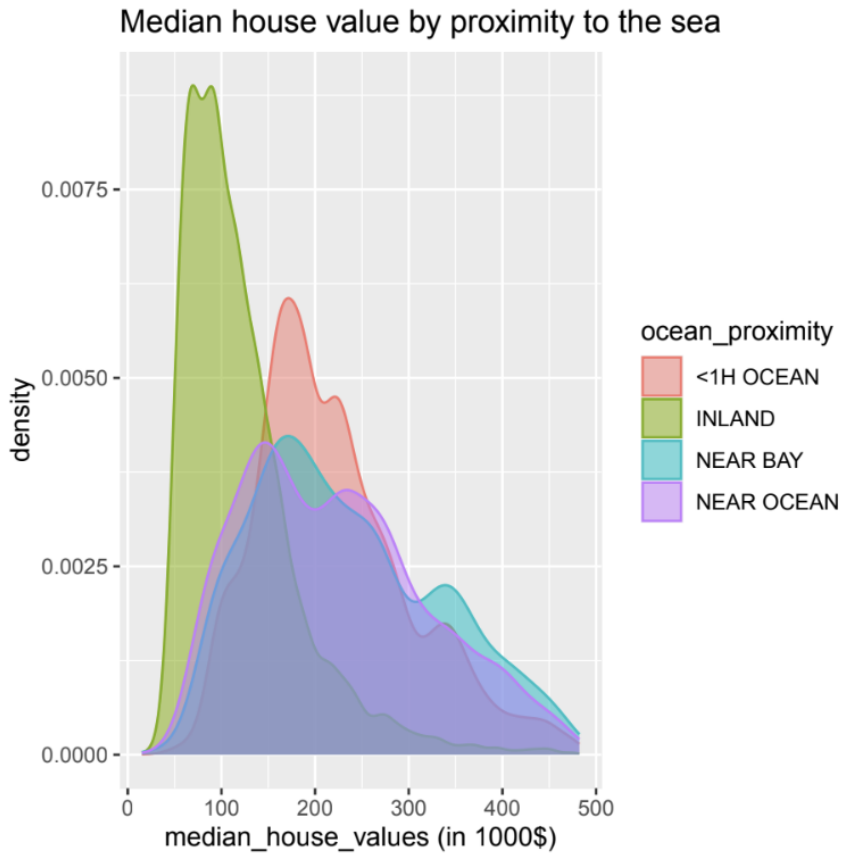




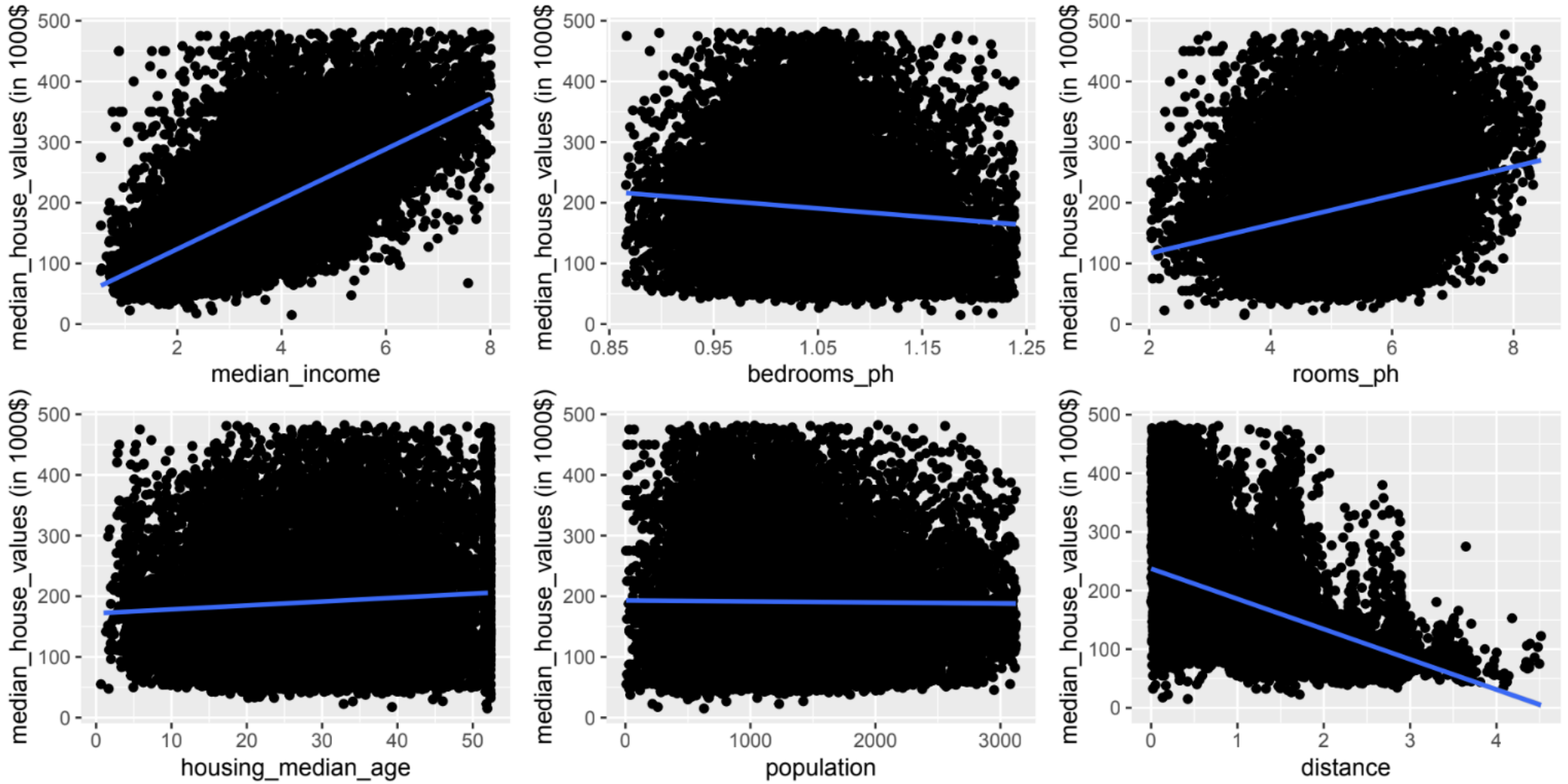
Analysing Features Distribution

Median House Value VS Proximity to the Sea





Median House Values vs num. Features



Model Data & Analysis

Multiple Linear Regression Model

Utilizing a multiple linear regression model, we analyze the impact of various features on the "median house value" response variable.

Model Preparation

Prior to computation, data modification is necessary to accommodate the numerical and categorical variables. Dummy variables are created for each label of the categorical variable "ocean proximity" with "NEAR BAY" serving as the reference category.

Multiple Linear Regression Model

Model Data & Analysis

```
##
## Call:
## lm(formula = median_house_value ~ ., data = dataset_with_dummy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -353.60  -39.40   -8.79   28.26  398.42
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -4.459e+01  8.779e+00  -5.079 3.83e-07 ***
## distance           -2.139e+01  7.985e-01 -26.795  < 2e-16 ***
## bedrooms_ph         1.479e+02  7.656e+00  19.323  < 2e-16 ***
## rooms_ph           -1.114e+01  7.696e-01 -14.474  < 2e-16 ***
## housing_median_age  5.500e-01  4.592e-02  11.977  < 2e-16 ***
## population         -2.979e-03  7.852e-04  -3.794 0.000149 ***
## median_income       4.395e+01  6.033e-01  72.851  < 2e-16 ***
## INLAND             -4.627e+01  1.922e+00 -24.072  < 2e-16 ***
## Min_1H_OCEAN       -7.700e+00  1.596e+00  -4.823 1.42e-06 ***
## NEAR_OCEAN          2.214e+01  2.060e+00  10.746  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.05 on 16655 degrees of freedom
## Multiple R-squared:  0.5908, Adjusted R-squared:  0.5906
## F-statistic:  2672 on 9 and 16655 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression Model

Assessing the presence of multicollinearity

```
vif(model)
```

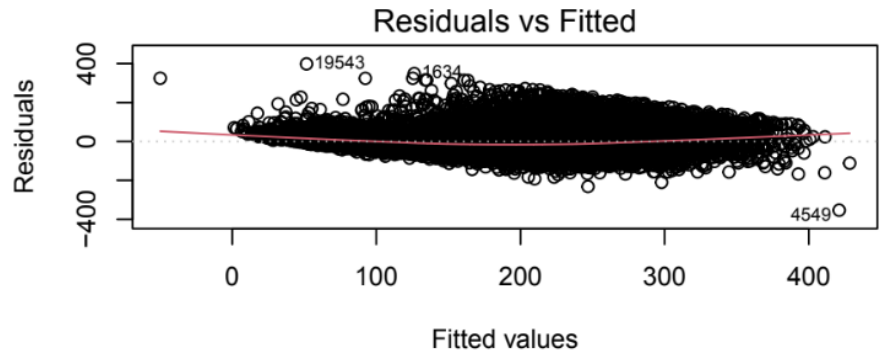
```
##           distance        bedrooms_ph          rooms_ph housing_median_age
##           2.049730           1.207103           3.008486           1.436901
##         population      median_income            INLAND       Min_1H_OCEAN
##           1.128676           3.492837           3.661619           2.913252
##         NEAR_OCEAN
##           2.185327
```

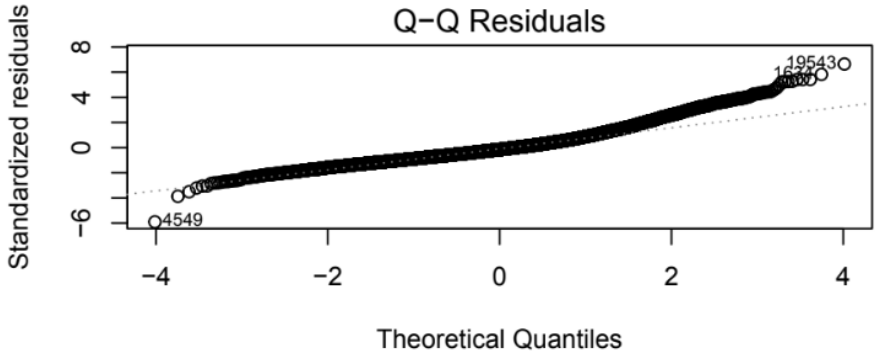
Typically, a VIF (Variance Inflation Factor) value above 5 or 10 is considered a threshold for concern, indicating moderate to high multicollinearity. In such cases, it may be necessary to address multicollinearity by removing one or more correlated variables. However, none of the variables in this case show a concerning VIF value.

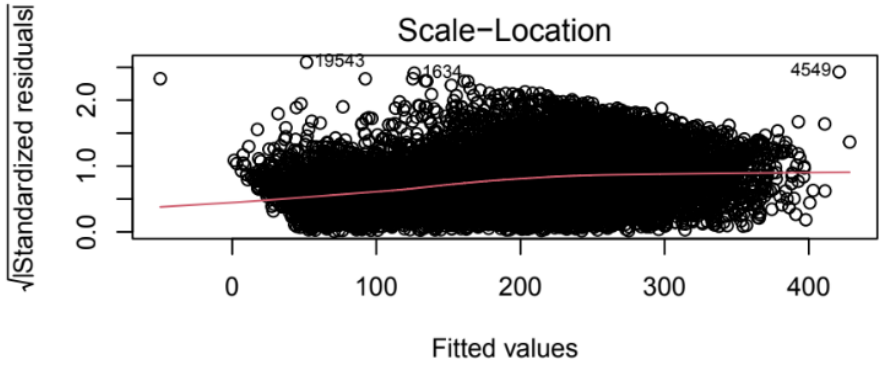
Multiple Linear Regression Model

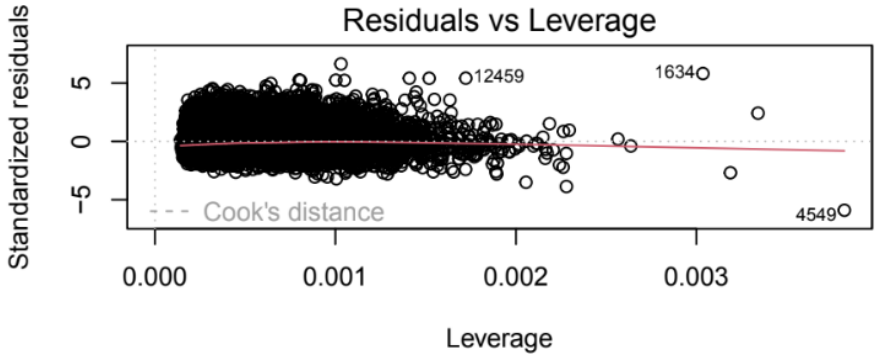
Model Data & Analysis

Checking Model Assumptions: Residuals and QQ plots









Stepwise variables selection

Stepwise approach

Since the model including all independent variables seems not to perform so well, we try a stepwise approach to select variables using backward selection. at each step, the least significant variable is removed based on a predefined criterion: AIC for a first attempt, then BIC.

Stepwise variables selection

Stepwise using AIC

```
##
## Call:
## lm(formula = median_house_value ~ distance + bedrooms_ph + rooms_ph +
##     housing_median_age + population + median_income + INLAND +
##     Min_1H_OCEAN + NEAR_OCEAN, data = dataset_with_dummy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -353.60  -39.40   -8.79   28.26  398.42
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -4.459e+01  8.779e+00  -5.079 3.83e-07 ***
## distance           -2.139e+01  7.985e-01 -26.795  < 2e-16 ***
## bedrooms_ph         1.479e+02  7.656e+00  19.323  < 2e-16 ***
## rooms_ph           -1.114e+01  7.696e-01 -14.474  < 2e-16 ***
## housing_median_age  5.500e-01  4.592e-02  11.977  < 2e-16 ***
## population         -2.979e-03  7.852e-04  -3.794 0.000149 ***
## median_income       4.395e+01  6.033e-01  72.851  < 2e-16 ***
## INLAND             -4.627e+01  1.922e+00 -24.072  < 2e-16 ***
## Min_1H_OCEAN       -7.700e+00  1.596e+00  -4.823 1.42e-06 ***
## NEAR_OCEAN          2.214e+01  2.060e+00  10.746  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.05 on 16655 degrees of freedom
## Multiple R-squared:  0.5908, Adjusted R-squared:  0.5906
## F-statistic:  2672 on 9 and 16655 DF,  p-value: < 2.2e-16
```

Stepwise using BIC

```
##
## Call:
## lm(formula = median_house_value ~ distance + bedrooms_ph + rooms_ph +
##     housing_median_age + population + median_income + INLAND +
##     Min_1H_OCEAN + NEAR_OCEAN, data = dataset_with_dummy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -353.60  -39.40   -8.79   28.26  398.42
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -4.459e+01  8.779e+00  -5.079 3.83e-07 ***
## distance           -2.139e+01  7.985e-01 -26.795  < 2e-16 ***
## bedrooms_ph         1.479e+02  7.656e+00  19.323  < 2e-16 ***
## rooms_ph           -1.114e+01  7.696e-01 -14.474  < 2e-16 ***
## housing_median_age  5.500e-01  4.592e-02  11.977  < 2e-16 ***
## population         -2.979e-03  7.852e-04  -3.794 0.000149 ***
## median_income       4.395e+01  6.033e-01  72.851  < 2e-16 ***
## INLAND             -4.627e+01  1.922e+00 -24.072  < 2e-16 ***
## Min_1H_OCEAN       -7.700e+00  1.596e+00  -4.823 1.42e-06 ***
## NEAR_OCEAN          2.214e+01  2.060e+00  10.746  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.05 on 16655 degrees of freedom
## Multiple R-squared:  0.5908, Adjusted R-squared:  0.5906
## F-statistic:  2672 on 9 and 16655 DF,  p-value: < 2.2e-16
```

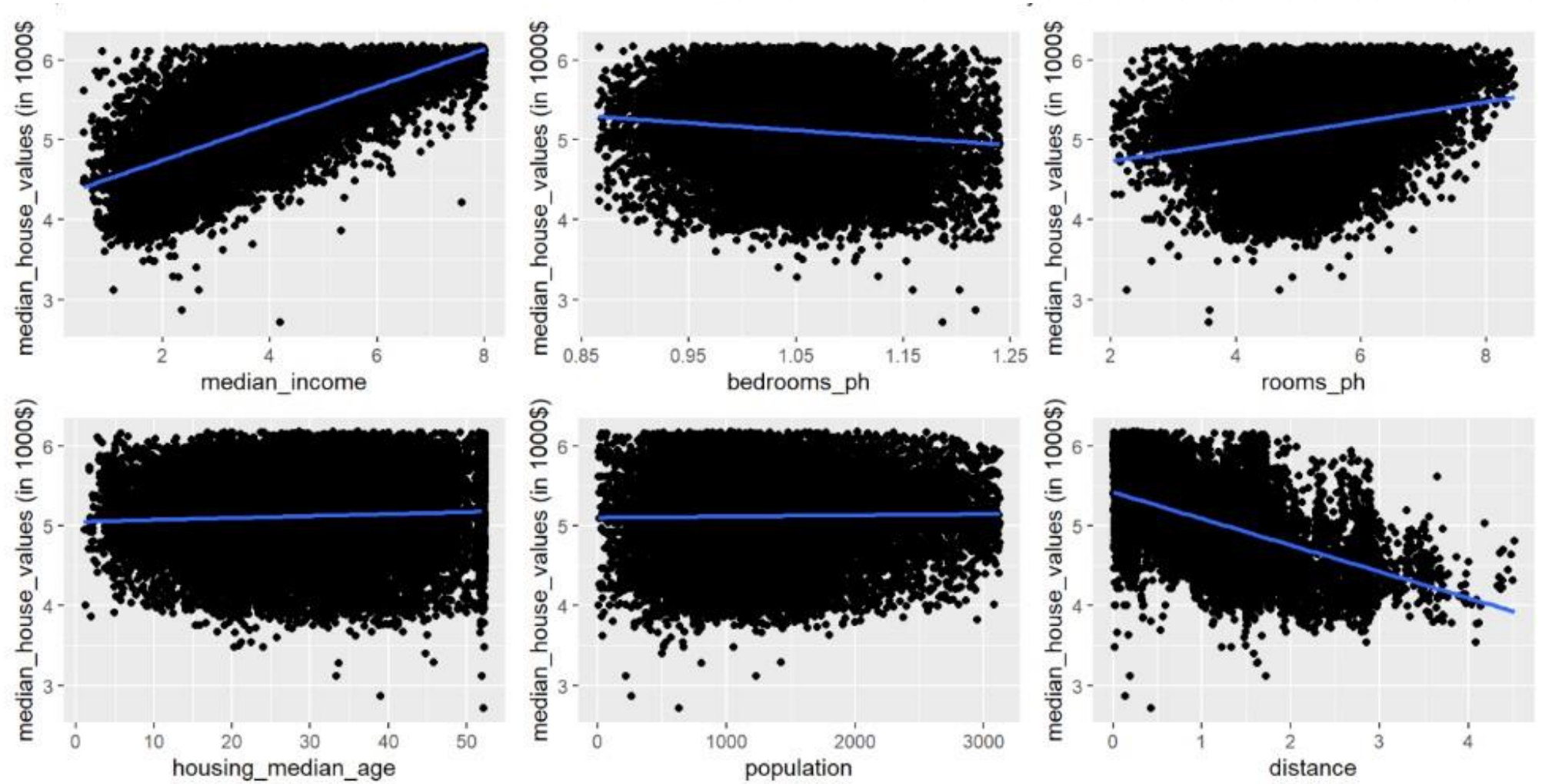
Logarithmic Transformation

The outputs suggest to maintain all features in the model, but AIC and BIC are quite high.

Since we are not satisfied with results obtained, we try to transform the response variable using logarithmic transformation.

Using log-transformation of Y, relations between variables seem to get better, as we can see in the following slides.

Logarithmic Transformation



Logarithmic Transformation


```
##
## Call:
## lm(formula = log(median_house_value) ~ ., data = dataset_with_dummy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54484 -0.20042 -0.02433  0.18142  1.90254
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.213e+00  4.418e-02  95.353  < 2e-16 ***
## distance           -1.748e-01  4.018e-03 -43.491  < 2e-16 ***
## bedrooms_ph         5.442e-01  3.853e-02  14.125  < 2e-16 ***
## rooms_ph           -3.799e-02  3.873e-03  -9.808  < 2e-16 ***
## housing_median_age  4.825e-05  2.311e-04   0.209  0.83460
## population         -1.134e-05  3.952e-06  -2.868  0.00413 **
## median_income       2.154e-01  3.036e-03  70.931  < 2e-16 ***
## INLAND             -2.944e-01  9.674e-03 -30.433  < 2e-16 ***
## Min_1H_OCEAN       -8.252e-03  8.034e-03  -1.027  0.30435
## NEAR_OCEAN          1.600e-01  1.037e-02  15.432  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3022 on 16655 degrees of freedom
## Multiple R-squared:  0.6638, Adjusted R-squared:  0.6636
## F-statistic:  3654 on 9 and 16655 DF,  p-value: < 2.2e-16
```

Logarithmic Transformation


```
R2_log <- summary(model_log)$r.squared
R2_log
```

```
## [1] 0.6638152
```

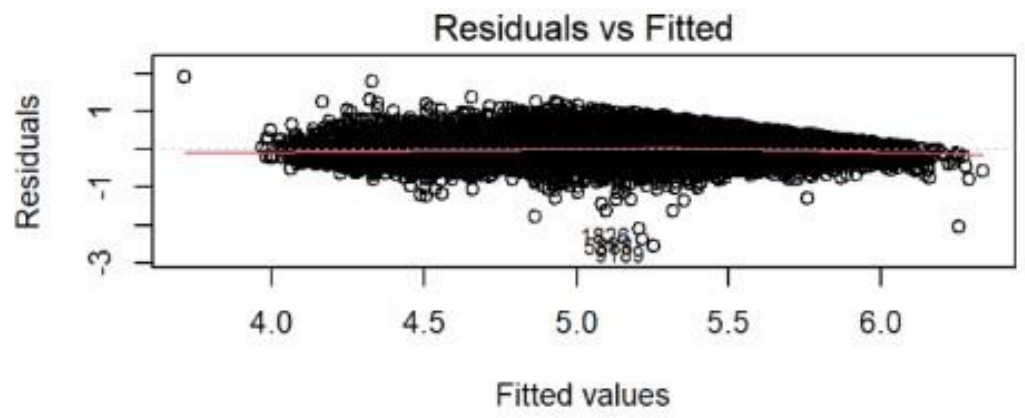
```
adjusted_R2_log <- summary(model_log)$adj.r.squared
adjusted_R2_log
```

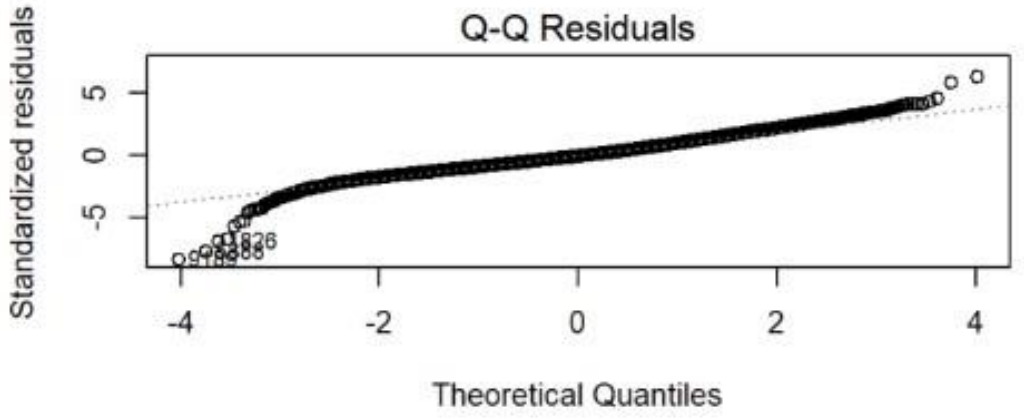
```
## [1] 0.6636335
```

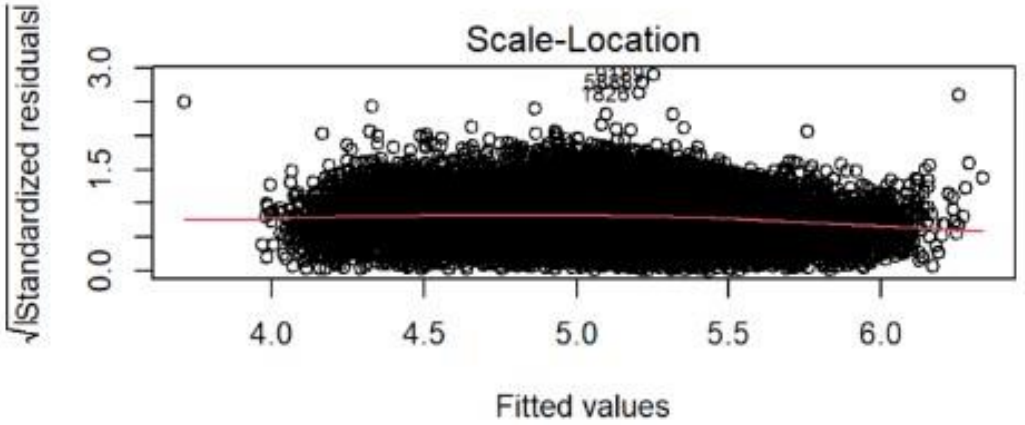
```
RSE_log <- summary(model_log)$sigma
RSE_log
```

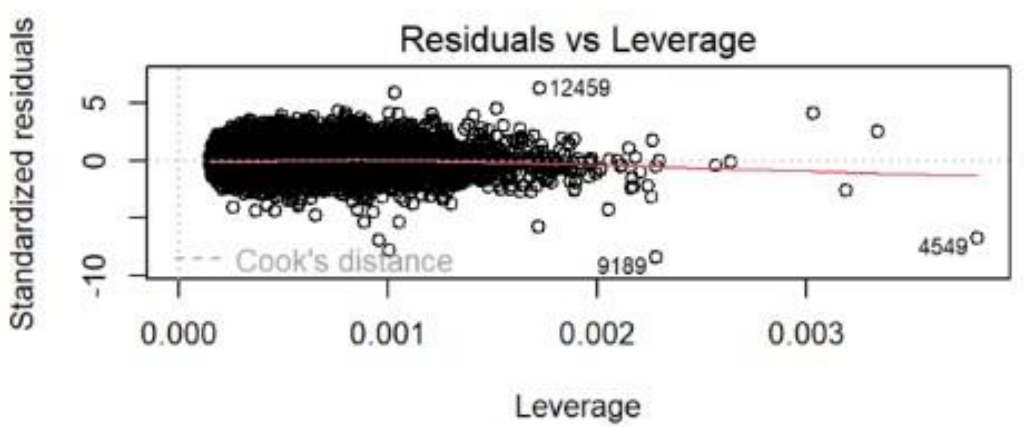
```
## [1] 0.3022116
```

Checking Assumption (for Log Model)









Stepwise (Log Model)

Both AIC and BIC stepwise backward selection suggest to remove variables "housing median age" and "<1H OCEAN" variables.

W.r.t. previous model:

- AIC passes from 183797.5 to 7421.612;
- BIC passes from 183882.5 to 7506.544.

In the following slides we print the summary of these 2 models.

Stepwise (for Log Model)

Stepwise using AIC

```
##
## Call:
## lm(formula = log(median_house_value) ~ distance + bedrooms_ph +
##     rooms_ph + population + median_income + INLAND + NEAR_OCEAN,
##     data = dataset_with_dummy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54347 -0.19999 -0.02458  0.18085  1.90206
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.210e+00  4.023e-02 104.641   <2e-16 ***
## distance      -1.756e-01  3.761e-03 -46.685   <2e-16 ***
## bedrooms_ph    5.431e-01  3.817e-02  14.228   <2e-16 ***
## rooms_ph      -3.754e-02  3.839e-03  -9.779   <2e-16 ***
## population    -1.194e-05  3.759e-06  -3.175   0.0015 **
## median_income  2.149e-01  2.912e-03  73.794   <2e-16 ***
## INLAND        -2.877e-01  6.874e-03 -41.852   <2e-16 ***
## NEAR_OCEAN     1.671e-01  7.874e-03  21.220   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3022 on 16657 degrees of freedom
## Multiple R-squared:  0.6638, Adjusted R-squared:  0.6636
## F-statistic:  4698 on 7 and 16657 DF,  p-value: < 2.2e-16
```

Stepwise using BIC

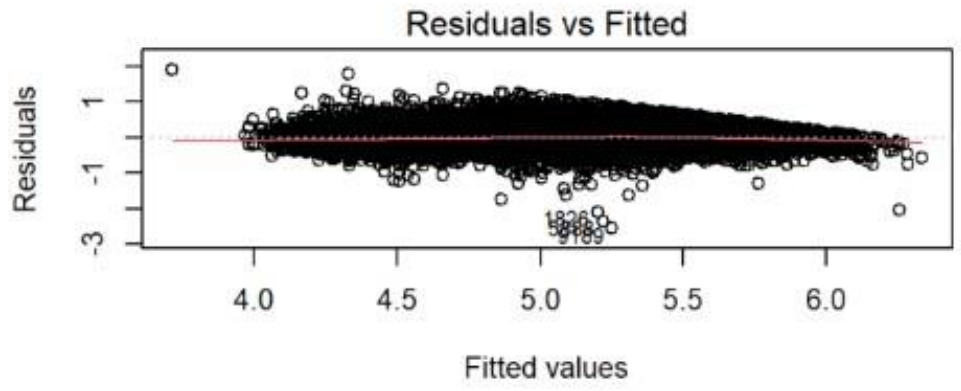
```
##
## Call:
## lm(formula = log(median_house_value) ~ distance + bedrooms_ph +
##     rooms_ph + population + median_income + INLAND + NEAR_OCEAN,
##     data = dataset_with_dummy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54347 -0.19999 -0.02458  0.18085  1.90206
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.210e+00  4.023e-02 104.641   <2e-16 ***
## distance      -1.756e-01  3.761e-03 -46.685   <2e-16 ***
## bedrooms_ph    5.431e-01  3.817e-02  14.228   <2e-16 ***
## rooms_ph      -3.754e-02  3.839e-03  -9.779   <2e-16 ***
## population    -1.194e-05  3.759e-06  -3.175   0.0015 **
## median_income  2.149e-01  2.912e-03  73.794   <2e-16 ***
## INLAND        -2.877e-01  6.874e-03 -41.852   <2e-16 ***
## NEAR_OCEAN     1.671e-01  7.874e-03  21.220   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3022 on 16657 degrees of freedom
## Multiple R-squared:  0.6638, Adjusted R-squared:  0.6636
## F-statistic:  4698 on 7 and 16657 DF,  p-value: < 2.2e-16
```

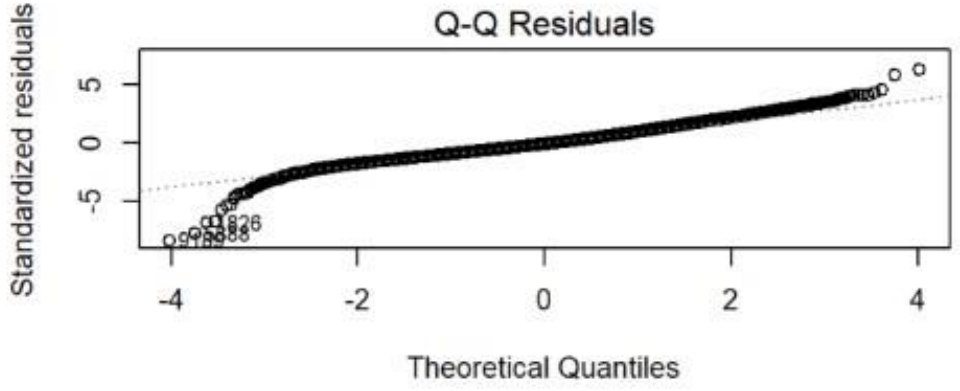
New Model

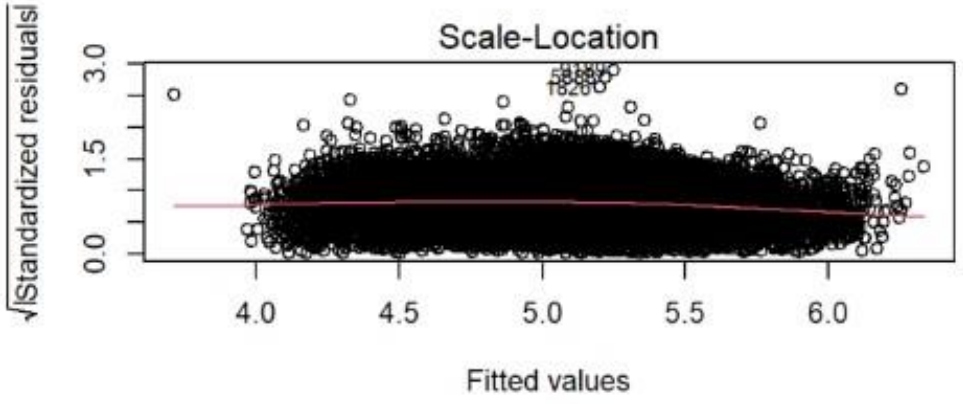
Given that we update our model considering all variables but “housing median age” and “<1H OCEAN” variables.

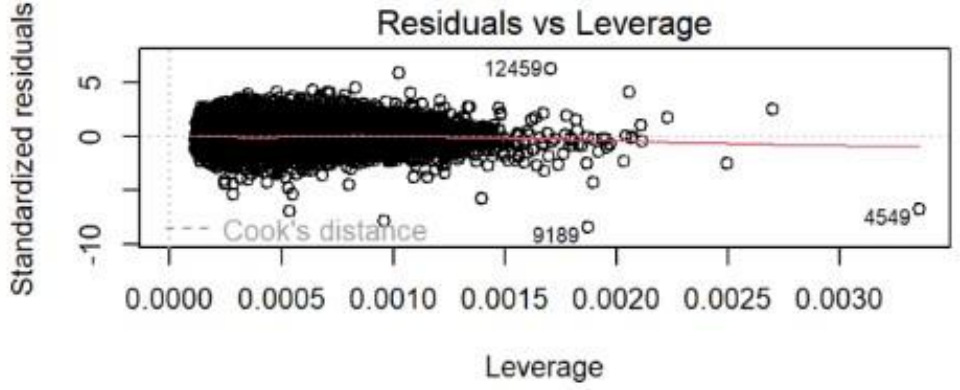
```
##
## Call:
## lm(formula = log(median_house_value) ~ . - Min_1H_OCEAN - housing_median_age,
##     data = dataset_with_dummy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54347 -0.19999 -0.02458  0.18085  1.90206
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.210e+00  4.023e-02 104.641   <2e-16 ***
## distance      -1.756e-01  3.761e-03 -46.685   <2e-16 ***
## bedrooms_ph    5.431e-01  3.817e-02  14.228   <2e-16 ***
## rooms_ph      -3.754e-02  3.839e-03  -9.779   <2e-16 ***
## population    -1.194e-05  3.759e-06  -3.175   0.0015 **
## median_income  2.149e-01  2.912e-03  73.794   <2e-16 ***
## INLAND        -2.877e-01  6.874e-03 -41.852   <2e-16 ***
## NEAR_OCEAN     1.671e-01  7.874e-03  21.220   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3022 on 16657 degrees of freedom
## Multiple R-squared:  0.6638, Adjusted R-squared:  0.6636
## F-statistic:  4698 on 7 and 16657 DF,  p-value: < 2.2e-16
```

Checking Model Assumption









Stepwise (for Log Model)

Stepwise using AIC

```
##
## Call:
## lm(formula = log(median_house_value) ~ (distance + bedrooms_ph +
##     rooms_ph + housing_median_age + population + median_income +
##     INLAND + Min_1H_OCEAN + NEAR_OCEAN) - Min_1H_OCEAN - housing_median_age,
##     data = dataset_with_dummy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54347 -0.19999 -0.02458  0.18085  1.90206
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.210e+00  4.023e-02 104.641   <2e-16 ***
## distance      -1.756e-01  3.761e-03 -46.685   <2e-16 ***
## bedrooms_ph    5.431e-01  3.817e-02  14.228   <2e-16 ***
## rooms_ph      -3.754e-02  3.839e-03  -9.779   <2e-16 ***
## population    -1.194e-05  3.759e-06  -3.175   0.0015 **
## median_income  2.149e-01  2.912e-03  73.794   <2e-16 ***
## INLAND        -2.877e-01  6.874e-03 -41.852   <2e-16 ***
## NEAR_OCEAN     1.671e-01  7.874e-03  21.220   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3022 on 16657 degrees of freedom
## Multiple R-squared:  0.6638, Adjusted R-squared:  0.6636
## F-statistic:  4698 on 7 and 16657 DF,  p-value: < 2.2e-16
```

Stepwise using BIC

```
##
## Call:
## lm(formula = log(median_house_value) ~ (distance + bedrooms_ph +
##     rooms_ph + housing_median_age + population + median_income +
##     INLAND + Min_1H_OCEAN + NEAR_OCEAN) - Min_1H_OCEAN - housing_median_age,
##     data = dataset_with_dummy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54347 -0.19999 -0.02458  0.18085  1.90206
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.210e+00  4.023e-02 104.641   <2e-16 ***
## distance      -1.756e-01  3.761e-03 -46.685   <2e-16 ***
## bedrooms_ph    5.431e-01  3.817e-02  14.228   <2e-16 ***
## rooms_ph      -3.754e-02  3.839e-03  -9.779   <2e-16 ***
## population    -1.194e-05  3.759e-06  -3.175   0.0015 **
## median_income  2.149e-01  2.912e-03  73.794   <2e-16 ***
## INLAND        -2.877e-01  6.874e-03 -41.852   <2e-16 ***
## NEAR_OCEAN     1.671e-01  7.874e-03  21.220   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3022 on 16657 degrees of freedom
## Multiple R-squared:  0.6638, Adjusted R-squared:  0.6636
## F-statistic:  4698 on 7 and 16657 DF,  p-value: < 2.2e-16
```

Comparisons

The two models are similar but the model without “housing median age” and “<1H OCEAN” variables shows slightly better results, expecially in AIC and BIC values.

In fact, AIC and BIC decrease after the variables removal.

- AIC: 7418.803
- BIC: 7488.293

We have now reached the best model for our response variable “median house value”.

ANOVA

ANOVA is used to determine if there are significant differences between models in terms of how well they fit the data. We use ANOVA to compare model_log and model_log_2.

Those outputs show that there aren't significant differences among the two models, since p-value (0.5513) is greater than alpha (0.05)

ANOVA

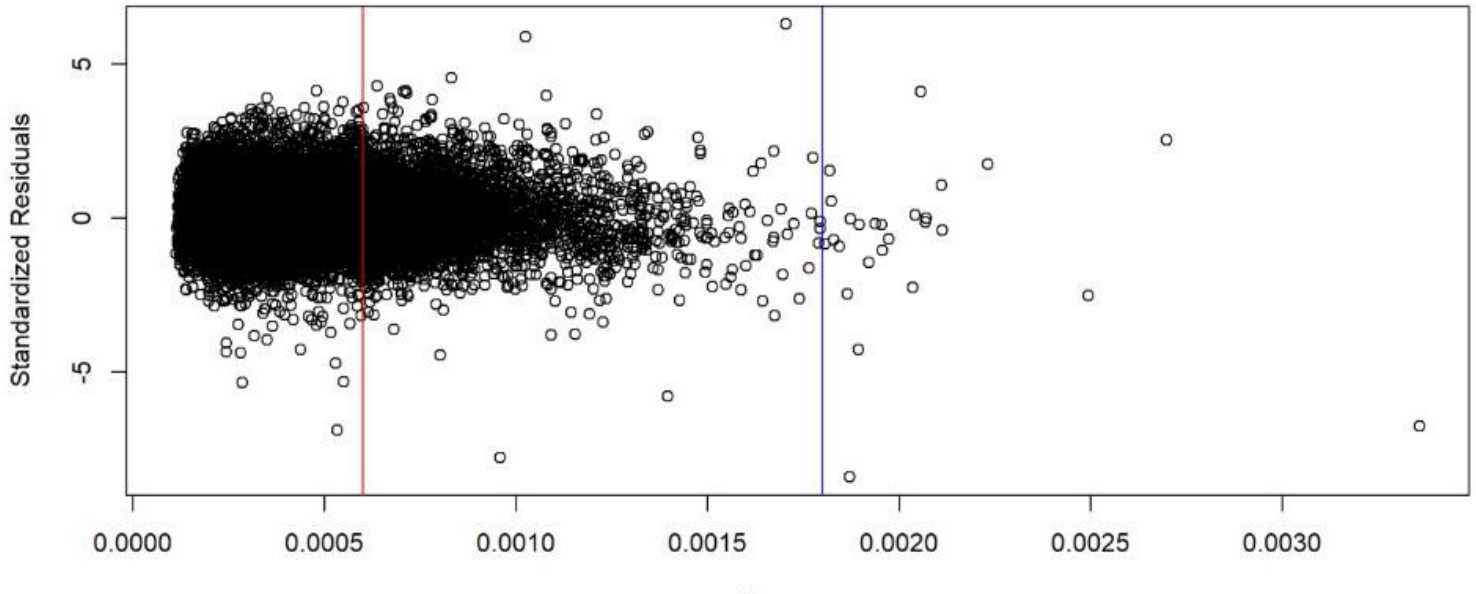

```
anova(model_log, model_log_2)
```

```
## Analysis of Variance Table
##
## Model 1: log(median_house_value) ~ distance + bedrooms_ph + rooms_ph +
##     housing_median_age + population + median_income + INLAND +
##     Min_1H_OCEAN + NEAR_OCEAN
## Model 2: log(median_house_value) ~ (distance + bedrooms_ph + rooms_ph +
##     housing_median_age + population + median_income + INLAND +
##     Min_1H_OCEAN + NEAR_OCEAN) - Min_1H_OCEAN - housing_median_age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1  16655 1521.1
## 2  16657 1521.2 -2  -0.10877 0.5954 0.5513
```

High Leverage Point

We check for high leverage points using as criterion the rule of thumb that there is high leverage when:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} > 3\,\frac{(p+1)}{n}$$



Since by definition high leverage points are data which have a strong impact on the model, we decide to delete them in order to check if the model without high leverage could be better.

Model (Without High Leverage Point)

The model without high leverage points seems to produce a slight improovement.

```
##
## Call:
## lm(formula = log(median_house_value) ~ . - housing_median_age -
##     Min_1H_OCEAN, data = data_w_hlp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36284 -0.20031 -0.02406  0.18059  1.91115
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.195e+00  4.013e-02 104.545  < 2e-16 ***
## distance      -1.746e-01  3.779e-03 -46.201  < 2e-16 ***
## bedrooms_ph    5.669e-01  3.811e-02  14.875  < 2e-16 ***
## rooms_ph      -4.201e-02  3.849e-03 -10.916  < 2e-16 ***
## population    -1.316e-05  3.746e-06  -3.514 0.000443 ***
## median_income  2.184e-01  2.921e-03  74.775  < 2e-16 ***
## INLAND        -2.848e-01  6.874e-03 -41.432  < 2e-16 ***
## NEAR_OCEAN     1.678e-01  7.845e-03  21.385  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3007 on 16631 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6662
## F-statistic:  4745 on 7 and 16631 DF,  p-value: < 2.2e-16
```

Summary

	R2	adjusted_R2	RSE	AIC	BIC
model	0,590806759	0,59058564	60,05008157	183798	183882
model_log	0,663815213	0,663633546	0,302211596	7421,61	7506,54
model_log_2	0,663791175	0,663649885	0,302204256	7418,8	7488,29
model_w_hlp	0,666329578	0,666189136	0,300711267	7242,43	7311,91

Cross-Validation

In order to evaluate the model performance, we split the data in training and test sets.

We decide to use the 80% of data randomly selected as training set and the other 20% as test set.

Then we build the linear model with training data and make predictions.

```
# Split the data into train and test and compute MSE

# setting seed to generate a reproducible random sampling
set.seed(569)

# Define training subset
n_train = floor(dim(data_w_hlp)[1]*0.8) # 13243 samples for test(80 % of the data)
i_train <- sample( 1:n, size = n_train, replace = FALSE) # indexes of training samples

# linear model with training data
model_train <- lm(log(median_house_value) ~ . -Min_1H_OCEAN -housing_median_age, data = data_
w_hlp , subset = i_train)

# Prediction based on fitted model
y_pred <- predict(model_train, newdata = data_w_hlp[-i_train, ])

# mean squared error on test data
MSE <- mean((log(data_w_hlp$median_house_value[-i_train])-y_pred)^2)
MSE
```

```
## [1] 0.09299337
```

```
# MSE su log(y) che implica MSE=exp(MSE) su median_house_value/1000
MSE_c = exp(MSE) * 1000
MSE_c
```

```
## [1] 1097.454
```

K-fold Cross-Validation

To have an improved evaluation of the model's ability to generalize, we make use of K-fold Cross-Validation. Initially, we use 10 folds, and then we increase it to 20 folds to observe any differences.

10 Folds

```
##       Error
## 0.09040445
```

```
error_10 = exp(mean.cv.errors_10) * 1000
error_10
```

```
##     Error
## 1094.617
```

20 Folds

```
##       Error
## 0.09037392
```

```
# mean.cv.errors_20 e mean.cv.errors_10 errori su log(y)
error_20 = exp(mean.cv.errors_20) * 1000
error_20
```

```
##     Error
## 1094.583
```

Both cross-validations produce very similar errors, it indicates that the increased subdivision into more folds has not significantly affected the variability of the model's performance.

Ridge Regression

We randomly select a 80% of values and assign them to “train” vector. Then, we creat e the test vector using all data but those on train vector: these elements are used for testing the trained model.

On the training data we perform 10-fold cross validation using RIDGE regression and select the best value of regularization parameter (lambda).

Then we compute RIDGE regression model coefficients at the optimal lambda value.

```
ridge_model <- cv.glmnet(X[train, ], y[train], alpha = 0, nfold=10) # alpha = 0 regressione R
IDGE
lambda_min <- ridge_model$lambda.min
lambda_min
```

```
## [1] 0.03346832
```

```
coefficient_ridge <- coef(ridge_model, s = "lambda.min")
coefficient_ridge
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                              s1
## (Intercept)     4.334212e+00
## distance       -1.750514e-01
## bedrooms_ph     3.795451e-01
## rooms_ph       -8.211852e-03
## population     -6.410539e-06
## median_income   1.851018e-01
## INLAND         -2.993836e-01
## NEAR_OCEAN      1.636651e-01
```

Ridge Regression (Model Evaluation)

We have predicted test set values comparing them with the real ones, and then we computed statistic metrics.

```
# Residual Standard Error
RSE <- sqrt(RSS/(n_R - p_R - 1))
RSE
```

```
## [1] 0.3017948
```

```
# R Squared statistic
R2_R <- 1 - RSS/TSS
R2_R
```

```
## [1] 0.6385339
```

```
# adjusted R square
adjR2_R <- 1 - (1-R2_R)*((n_R-1)/(n_R-p_R-1))
adjR2_R
```

```
## [1] 0.6377717
```

```
MSE_R <- mean((predictions - y_test)^2)
MSE_R
```

```
## [1] 0.09086114
```

```
MSE_R_c = exp(MSE_R) * 1000
MSE_R_c
```

```
## [1] 1095.117
```

UNIVERSITÀ DEGLI STUDI DI PADOVA

ANOVA

In this part our aim it to apply ANOVA to evaluate if there are significant differences between means of different groups of data. In order to compute ANOVA, a foundamental assumption is homoscedasticity between groups. To verify this assumption we use Bartlett test.

```
# checking homoscedasticity between categorical variables
bartlett.test(log(median_house_value) ~ ocean_proximity , data = dataset_last)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  log(median_house_value) by ocean_proximity
## Bartlett's K-squared = 327.2, df = 3, p-value < 2.2e-16
```

The result shows that p-value is less than alpha = 0.05. Therefore the null hypotesis is not verified. It means that ANOVA cannot be applied.

Conclusion

The results show that the best model for making predictions regarding the "median house value" variable is the model with the logarithmic transformation of Y, which excludes two variables (housing_median_age and MIN_1H_OCEAN) and does not consider the High Leverage point. Indeed, the application of ridge regression did not bring significant improvements to it, which demonstrates its goodness. This says that from all the numeric variables the only one that does not influence the estimation of median house value is housing_median_age.

This model explains approximately 66.62% of the data and has a RSE of 0.300711267.

Moreover, also both AIC and BIC have a good value: AIC: 7242.433, BIC: 7311.908. The computed MSE is equal to 0.09299337 for log(Y), and MSE is equal to 1097.454