ICY0006 Probability Theory and Statistics

Final Project Work

Almaz Kydyrmin

This dataset was generated by using a Python3 script, which is in the project directory. The topic of my project is house pricing in Tallinn, Estonia. There are 8000 rows of training data and 2000 of test data. There are no missing values in both data sets. Both Python3 data generator and R script for calculations are located in this repository: ‘generator.py’, ‘main.r’.

‘train.csv’ – data set for model training.

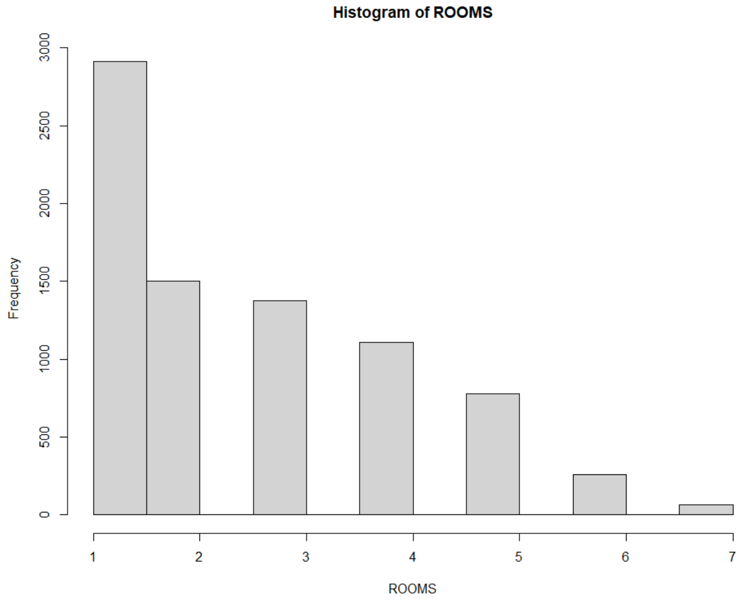
‘test.csv’ – data set for testing.

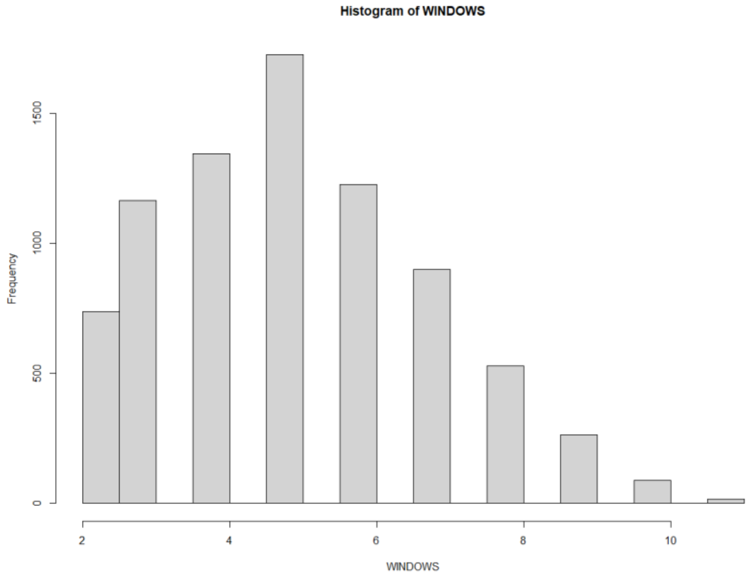
‘result.csv. – results after applying data model.

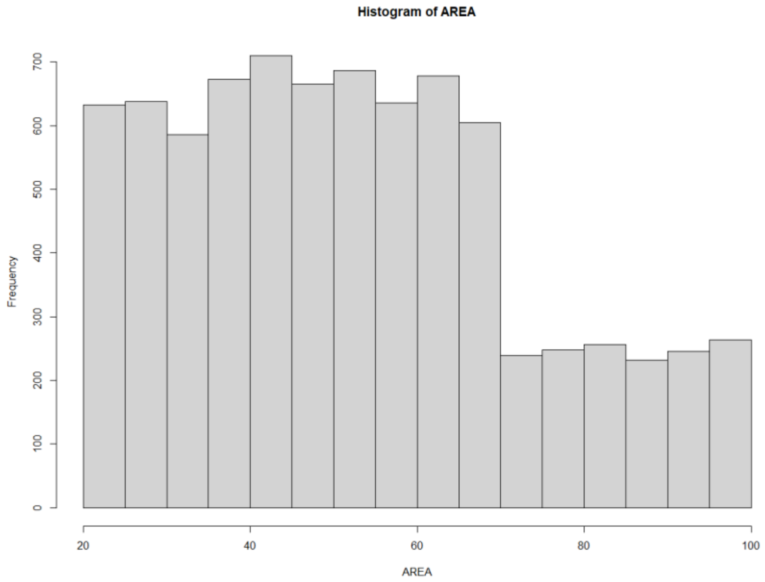
Dataset description:

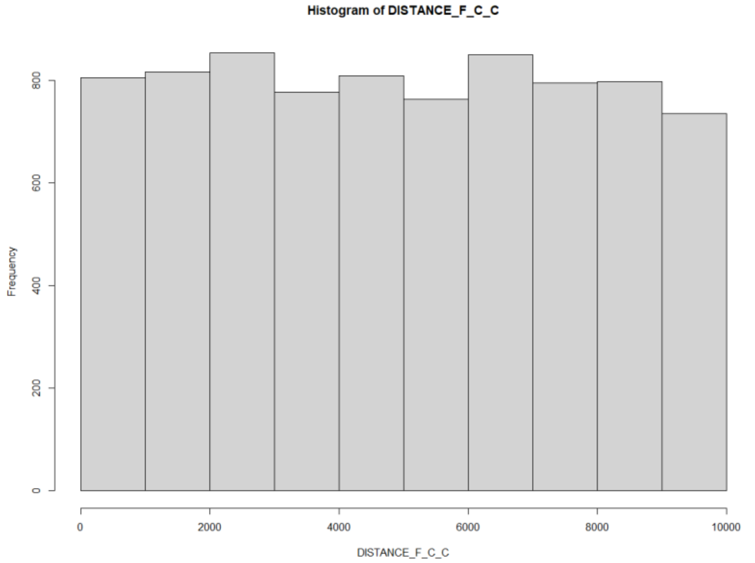
1. AREA: Flat area in meters squared
2. DISTANCE_F_C_C: Distance from city centre in meters
3. MAX_FLOOR: Number of floors in the building
4. FLOOR: Flat’s floor
5. ROOMS: Number of rooms
6. WINDOWS: Number of windows
7. PRICE: Flat price in euros

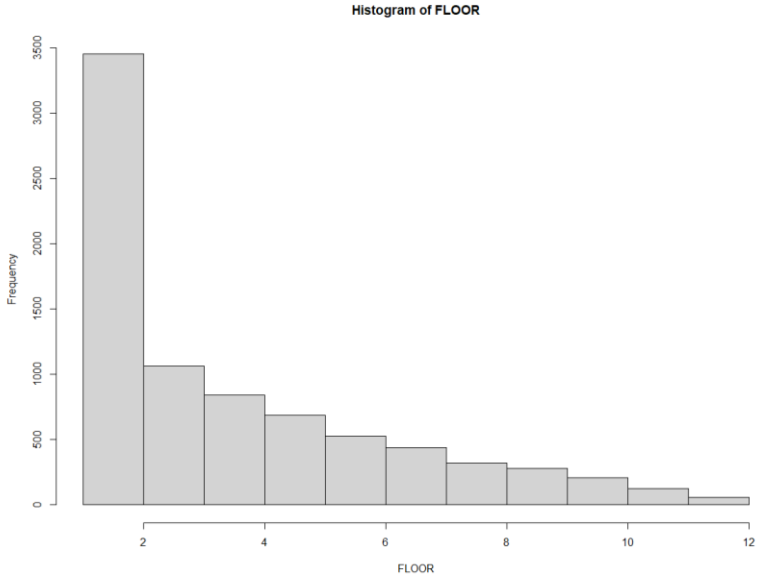
Data visualization:

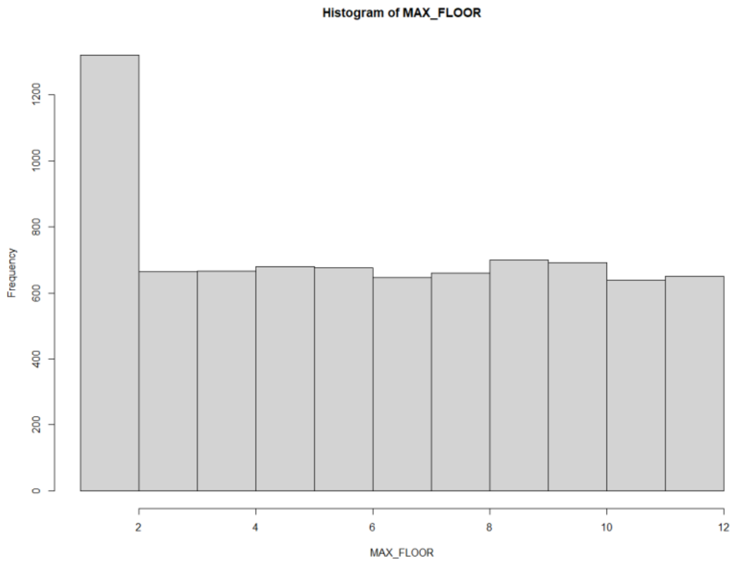


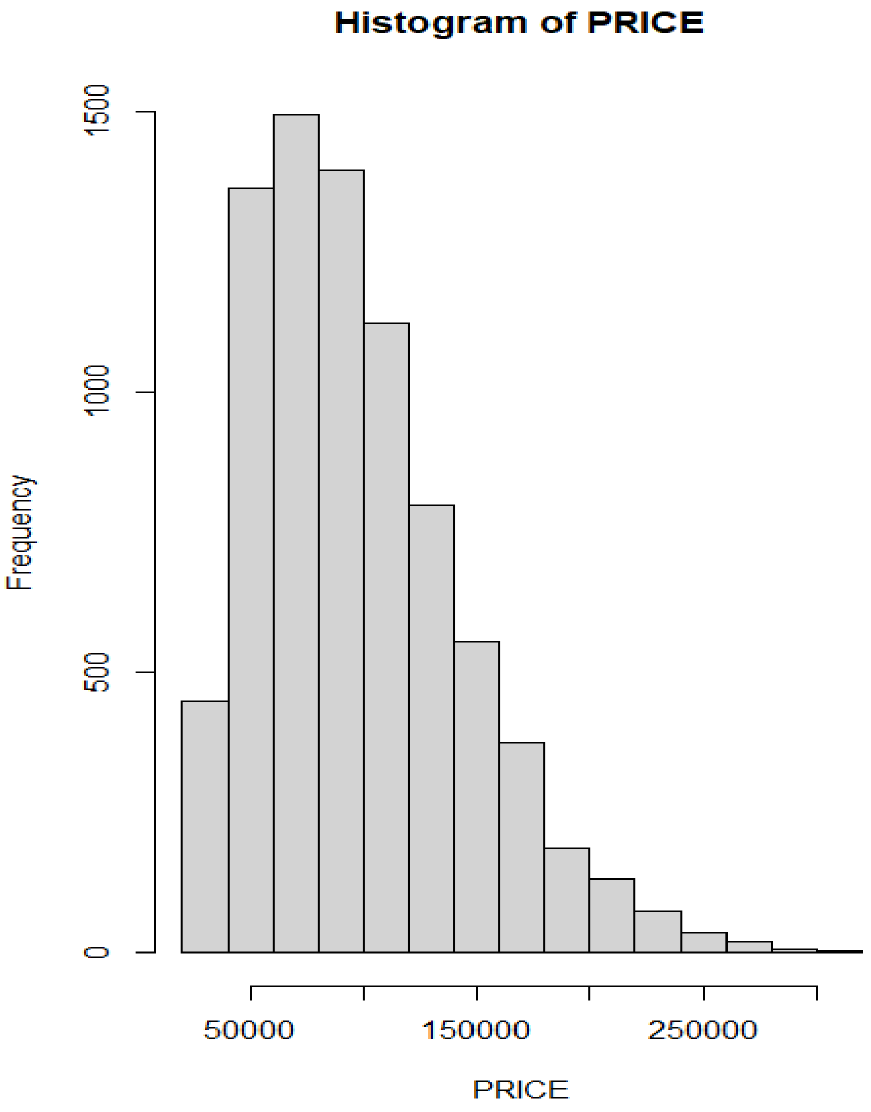



Histogram of AREA
Frequency
0
100
200
300
400
500
600
700
20
40
60
80
100
AREA


Histogram of DISTANCE_F_C_C
Frequency
0
200
400
600
800
0
2000
4000
6000
8000
10000
DISTANCE_F_C_C


Histogram of FLOOR
Frequency
0
500
1000
1500
2000
2500
3000
3500
2
4
6
8
10
12
FLOOR


Histogram of MAX_FLOOR
Frequency
0
200
400
600
800
1000
1200
2
4
6
8
10
12
MAX_FLOOR


Histogram of PRICE
Frequency
0
500
1000
1500
50000
150000
250000
PRICE

Description:

There are four totally independent variables: area, floor, total number of floors and distance from city center. Number of rooms and windows depends on area and the only totally dependent variable is price.

By looking at these histograms I can say that there are more flats with a smaller number of rooms. Also, I would mention that most frequent price lies between 50000 and 100000.

Central tendency and variability measures:

Mean:

AREA: 52.61148

DISTANCE_F_C_C: 4993.585

FLOOR: 3.782669

MAX_FLOOR: 6.497687

WINDOWS: 5.03964

ROOMS: 2.54608

PRICE: 97580.39

Mode:

PRICE: 49050

AREA: 34.9

DISTANCE_F_C_C: 8400

FLOOR: 1

MAX_FLOOR: 9

WINDOWS: 5

ROOMS: 1

Median:

PRICE: 89100

AREA: 50.7

DISTANCE_F_C_C: 5000

FLOOR: 3

MAX_FLOOR: 6

WINDOWS: 5

ROOMS: 2

Range:

PRICE: 24480 300800

AREA: 20 100

DISTANCE_F_C_C: 100 10000

FLOOR: 1 12

MAX_FLOOR: 1 12

WINDOWS: 2 11

ROOMS: 1 7

Variety:

PRICE: 2095964655

AREA: 421.5608

DISTANCE_F_C_C: 8195073

FLOOR: 7.600085

MAX_FLOOR: 11.81706

WINDOWS: 3.65113

ROOMS: 2.385727

Standard Deviation:

PRICE: 45781.71

AREA: 20.53195

DISTANCE_F_C_C: 2862.704

FLOOR: 2.756825

MAX_FLOOR: 3.437595

WINDOWS: 1.910793

ROOMS: 1.54458

Interquartile Range:

PRICE: 61380

AREA: 29.7

DISTANCE_F_C_C: 5000

FLOOR: 4

MAX_FLOOR: 5

WINDOWS: 2

ROOMS: 3

Description:

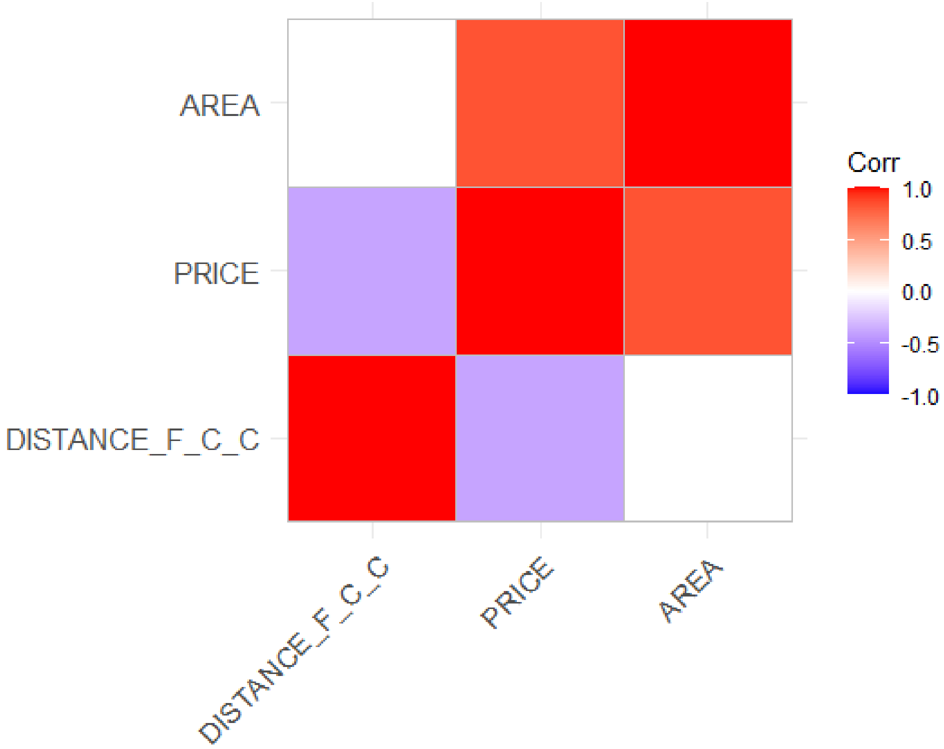
I could predict those values in first stage, by looking at data visualization, but here are exact values of mean, mode, median, range, interquartile range, variance, and standard deviation. The average area is 52.6 meters squared, and the average price is 97580.

Correlations:

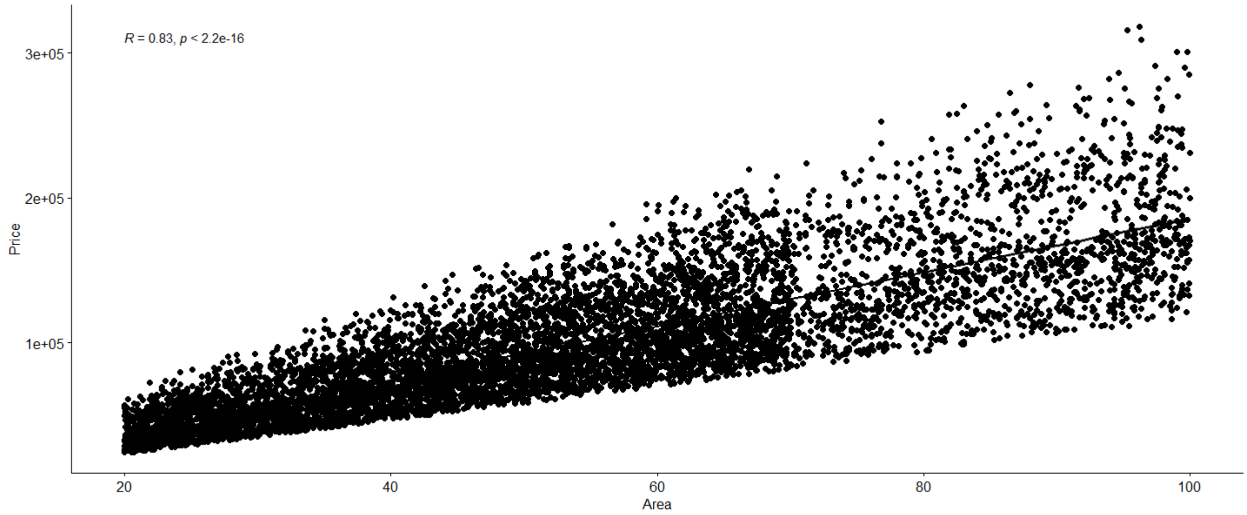
Using R, I found out that there is a definite relationship between area and price as well as between distance from city center and price.

Correlation coefficient between price and area: 0.8315173

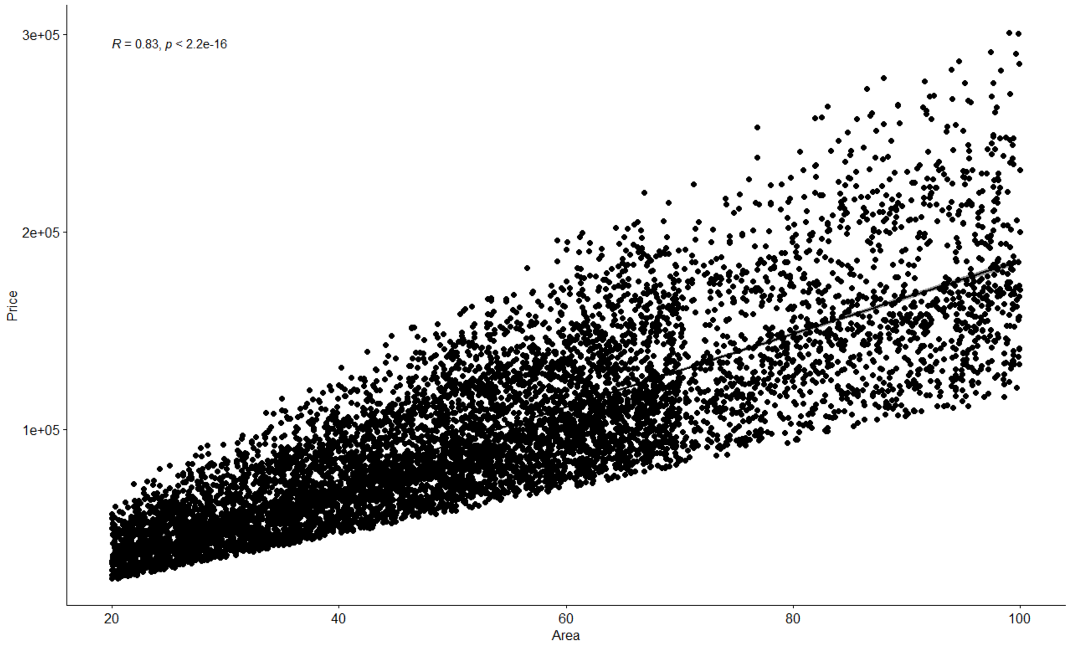
Correlation coefficient between price and distance from city center: 0.393264680



Correlation image between price and area:



I found out that there are some outliers and removed them.



Using peace of code below I created a model and applied it for test data. The result I saved to a temporary csv file and then added to ‘result.csv’. Using excel I easily counted uncertainty between predicted value and the actual price.

The Mean Absolute Error is 13899.176 euro.

My model can predict price for every flat of Tallinn with a not big error.