ICY0006 Probability Theory and Statistics

Final Project Work

December 2020

Almaz Kydyrmin

Tallinn University of Technology

Introduction

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'.

The main goal of this research is to provide a model which could give as most accurate price as possible for every flat of Tallinn by taking in account most important properties.

Project structure:

'train.csv' – data set for model training.

'test.csv' – data set for testing.

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

'final_project.pdf' – project description.

'probability.pdf' – probability assignment.

Dataset generation method:

To make the data set closer to real life, I used Python3 random library which I consider as a seller's own opinion. In fact, the same flat seems to be costing differently for different people. However, it is not the right way to make prices totally random, as there are many factors which impact on the final price. The first one which comes into mind is area. It is obvious that the larger flat area is the more expensive it will be. In addition to area, distance from city center plays big role. There are more work opportunities as well as touristic places in city center. There are also such factors as floor, number of rooms, number of floors in building and number of windows which I find as secondary factors.

I made ranges of price for different distances from city center and let Python3 choose the price from this range randomly. Flat area is also chosen randomly, but as most of flats in our city are not huge, I also took this into account. Since we do not have many skyscrapers, the maximum number of floors is 12. I looked at prices in local popular 'Okidoki' platform and set prices as following:

Distance from city center in meters	*Price per meter squared in euros*
<500	*2500 - 3500*
500 - 2000	*2000 - 3200*
2000 - 5000	*1700 - 2700*
>5000	*1300 - 2000*

Dataset description

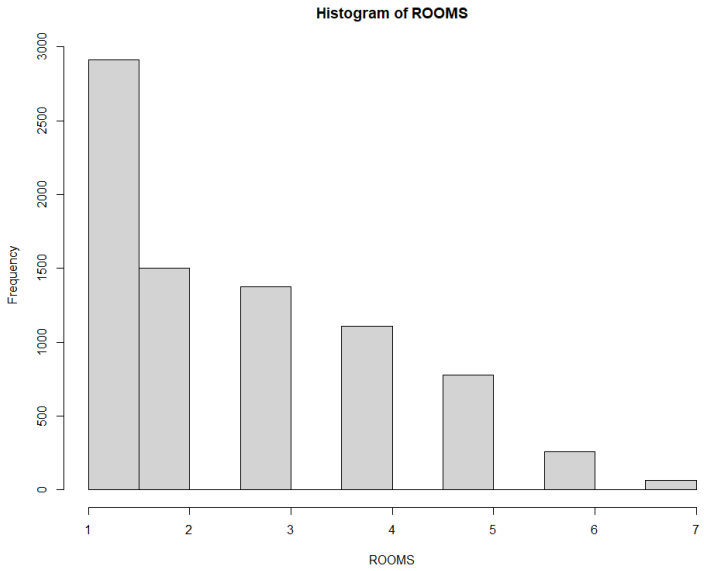
There are seven variables in my dataset:

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.*

Four variables are completely independent: 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.

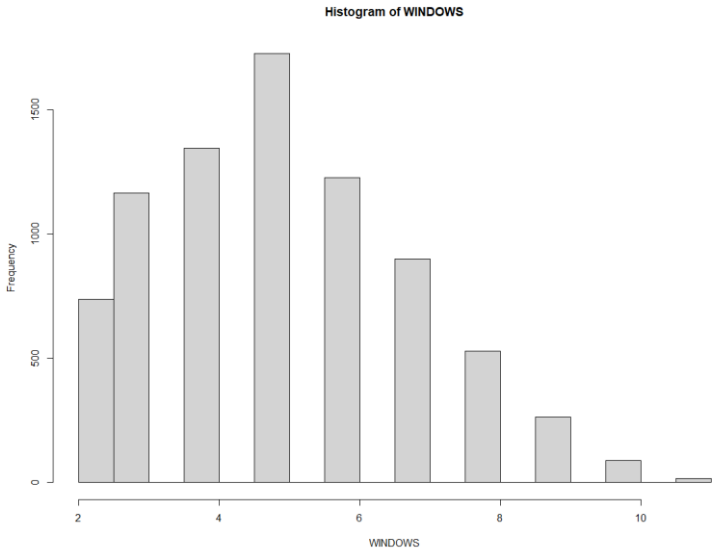
Data visualization

1. Number of rooms



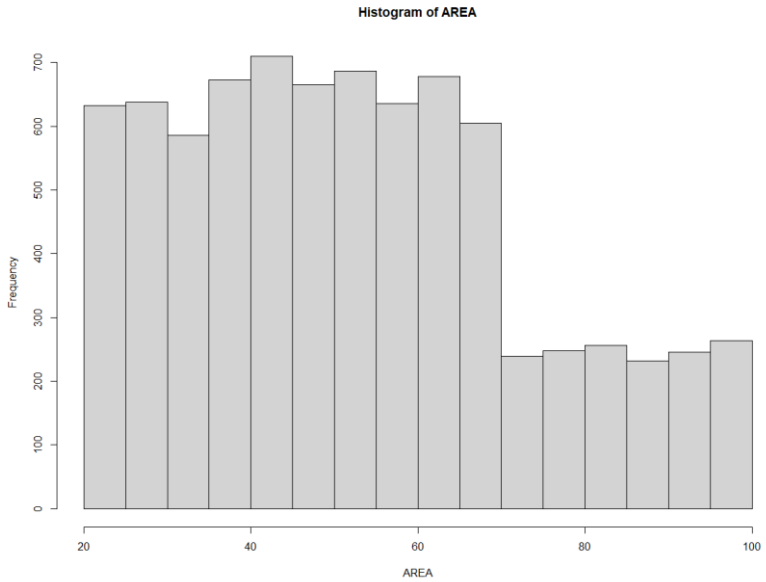
This histogram implies that that there is at least one room in each flat and that the maximum number of rooms is 7. The most popular among all variants is single room flat. Number of houses with specific number of rooms is in inverse ratio with number of rooms. The more there are rooms in a flat a customer wants to buy, the less there are choices.

2. Number of windows



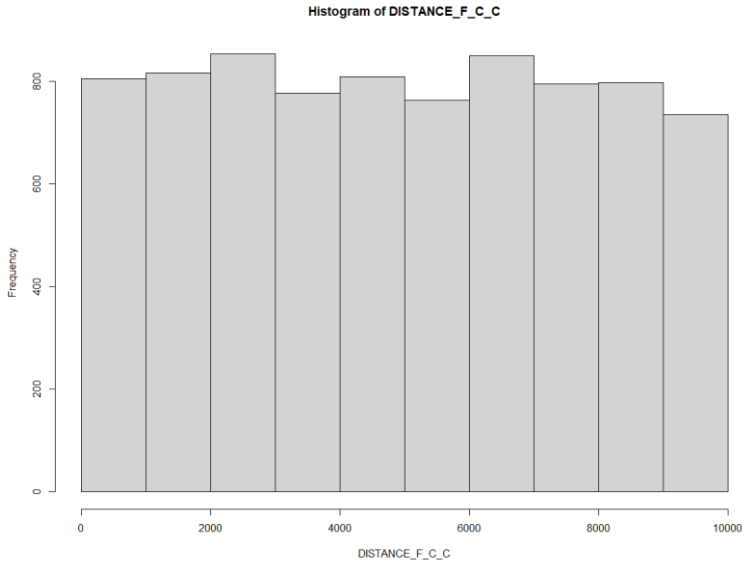
This histogram shows that the range of possible number of windows is between 2 and 11. The most popular number is 5 with more than 1500 rooms having such number of windows. The most unpopular is number 11.

3. *Area*



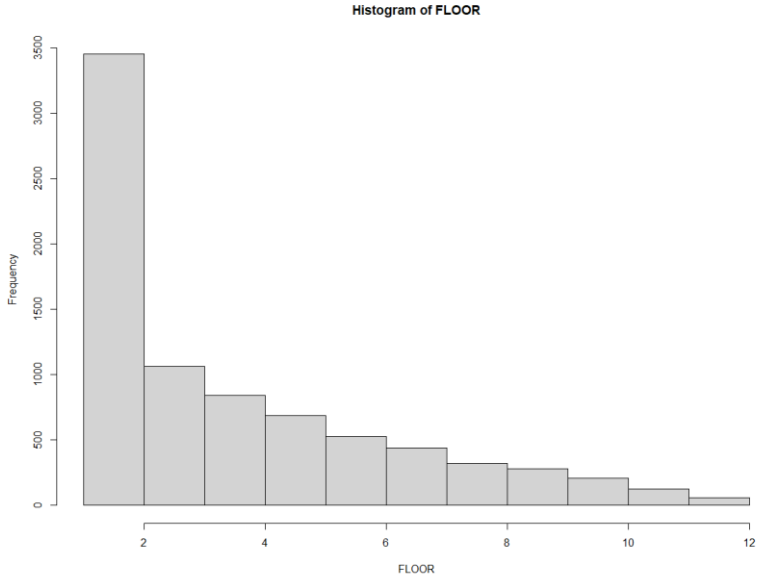
As I already mentioned in the very beginning flats with small area are more used in Tallinn. It is obvious from this graph and relation is approximately 2:1 (small:huge area). There is no distinct winner in this histogram, it goes a bit up and down from 20 to 70 after what there is significant drop till the end.

4. *Distance from city center*



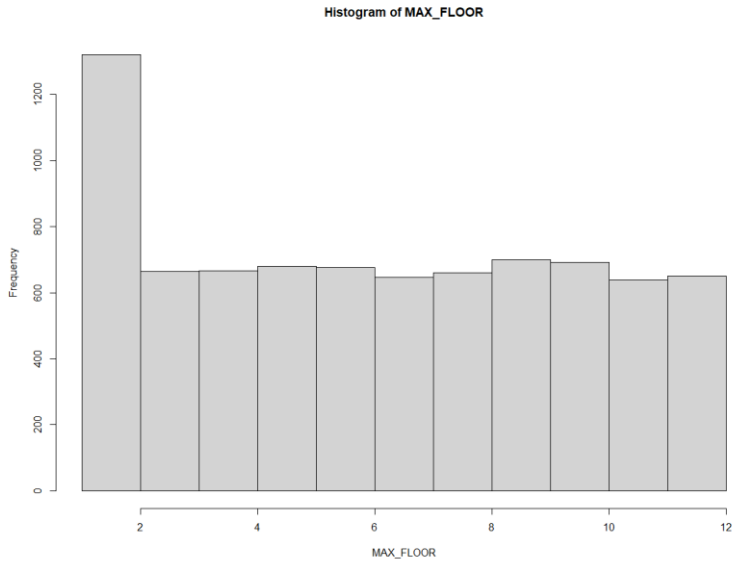
One of the most static histograms among all of them. In every column there is about 800 flats.

5. *Floor number*



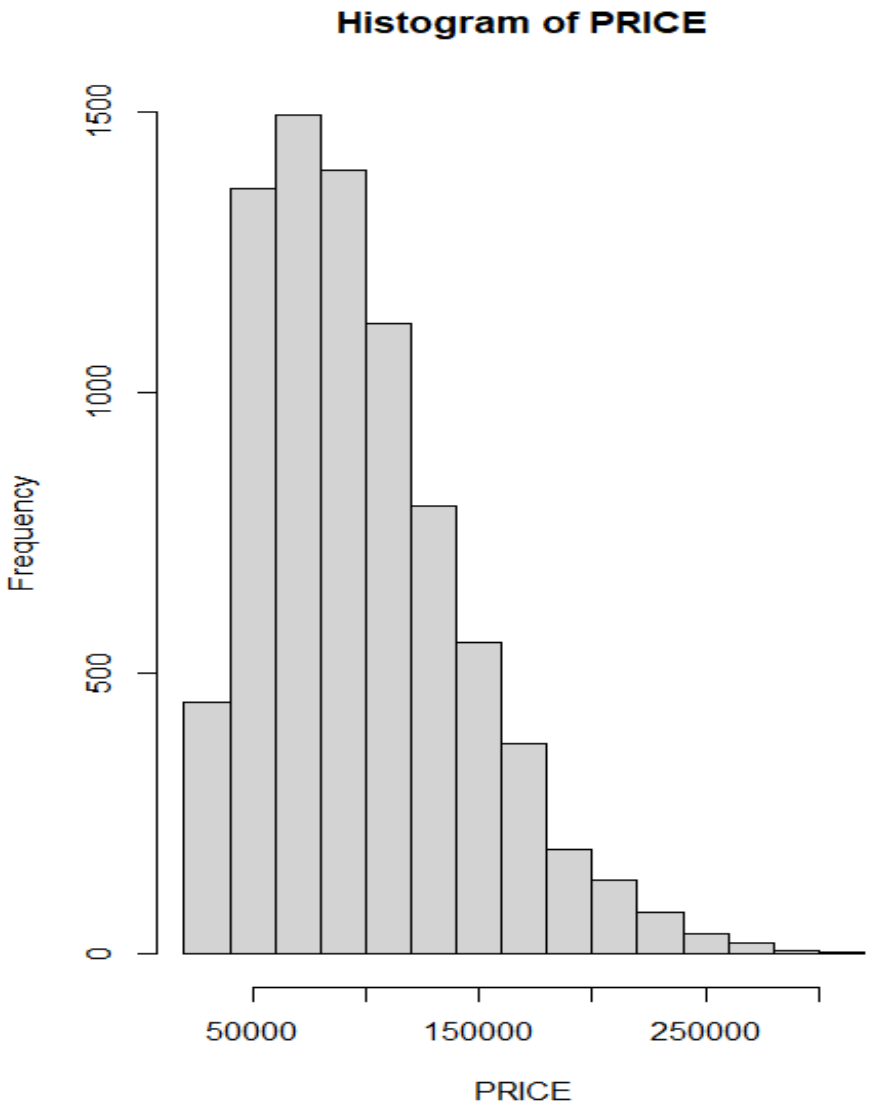
There is a noticeable decline with each increase in the floor number. The absolute winner is 1, with around 3500 flats. The range is between 1 and 12.

6. *Maximum floor number*



The most popular is one. Maximum number is 12.

7. Price



The most interesting histogram. Also, I would mention that most frequent price lies between 50000 and 100000. The cheapest flat variant costs a bit less than 30k, whilst the most expensive one is about 300k euro.

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:

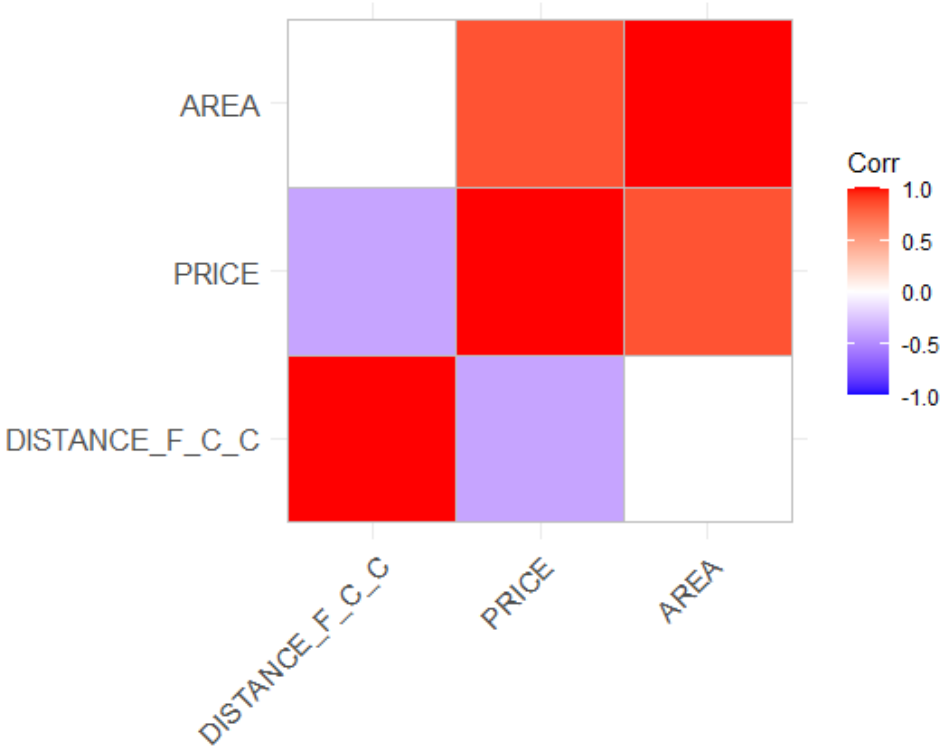
I could predict those values in first stage, by looking at the histograms, but here are exact values of mean, mode, median, range, interquartile range, variance, and standard deviation. The average area is 52.6 meters squared, and the average price is 97580.

Correlations

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

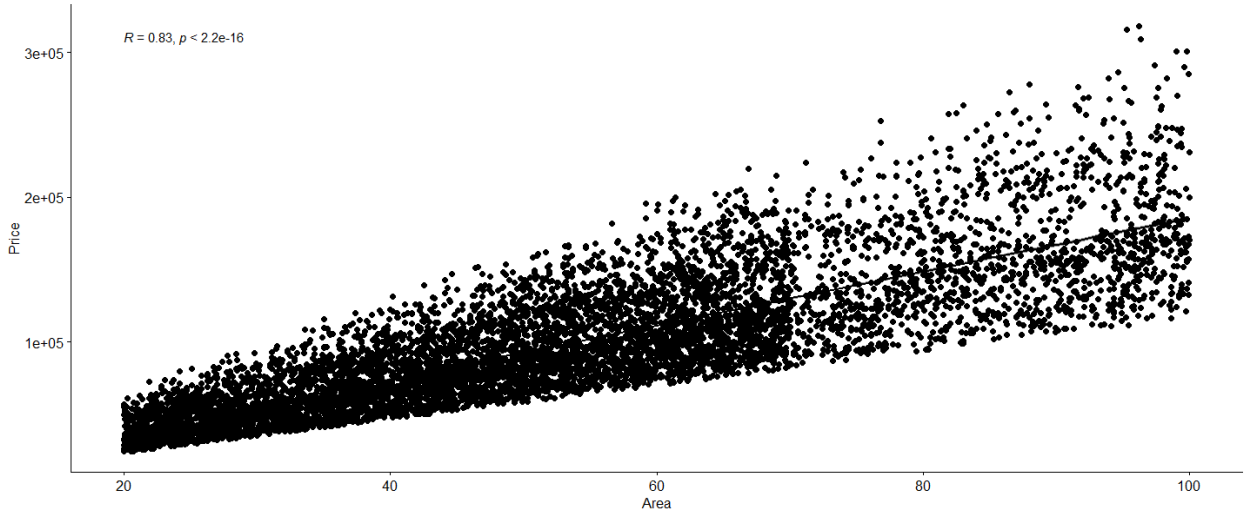
Correlation coefficient between price and area: ***0.8315173***

Correlation coefficient between price and distance from city center: ***0.393264680***

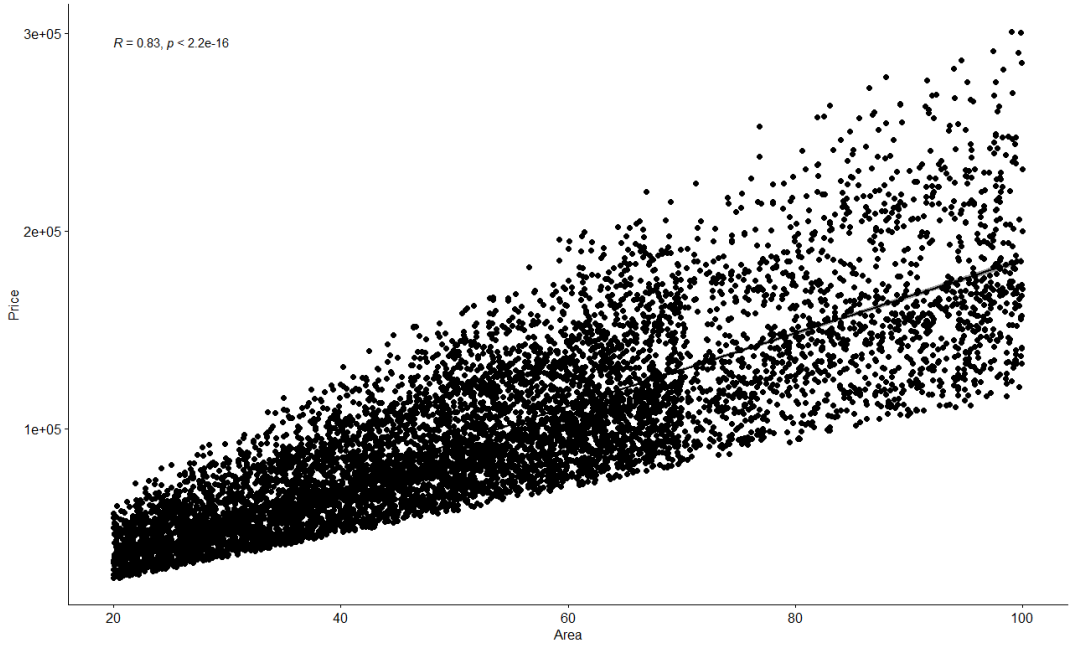


I was able to find some outliers which impacted on the correlation. After removing them, I got a bit better result.

Before:



After:



Results and Conclusion

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 the results to 'result.csv'. Using excel I easily counted uncertainty between predicted value and the actual price.

"

```
model <-glm(PRICE ~ AREA + DISTANCE_F_C_C, data=data)

prediction <- predict(model, test)

write.csv(prediction,"predictions.csv", row.names = FALSE)
```

"

The only step left is to find average error in my model. I used agrimetsoft [https://agrimetsoft.com] to find it out. The result pleased me: ***The Mean Absolute Error is 13899.176 euros.***

As I already mentioned in the very beginning two different people may sell the same flat with totally different prices. It is called human factor and I do not see a way to deal with it.

This project helped me to understand the basics of Data Science. I got a bit different result in comparison with what I expected, but it was nice experience which I believe I will be able to use during my career path.

The end!

Almaz Kydyrmin.