Stage 1

I have chosen a dataset named “Seattle House Sales Prices” which consists of numerical values only. This dataset is public and available by this link *https://www.kaggle.com/sameersmahajan/seattle-house-sales-prices.* This dataset has following columns: id, date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, zipcode, lat, long, sqft_living15, sqft_lot15.

Each column name gives a proper understanding of what it represents, except sqft_living15 and sqft_lot15. I have no idea what those columns might mean. However, I suppose that it means the same values as sqft_living and sqft_lot but in year of 2015. The Official Kaggle page has no description regarding those columns, that is why I dropped them. Also, I have noticed that Bathrooms column can have floating numbers. It sounded wrong and weird to me, but this link helped me to understand this logic *https://www.brickunderground.com/blog/2011/02/the_175_bath_apartment*

There are 2 columns that evaluate condition and total grade of houses. Condition is counted on 5 scale bases while grade has 12 scale grading. I can understand rating from 0 to 5 but grading from 0 to 12 sounds unusual to me. Number 12 is the highest number met in the “Grade” column, so I think it is the maximal value. Anyway, I have not applied any changes to these columns.

For my research I have considered just 8 of them (price, bedrooms, bathrooms, sqft_living, floors, condition, grade, yr_built). In my opinion, these columns are the best to be researched.

Original file has 21613 rows of data. I have decided to take only first 3000 of them for the training set. I have declared 8 variables for training set with prefix d, which means that this variable points to training set.

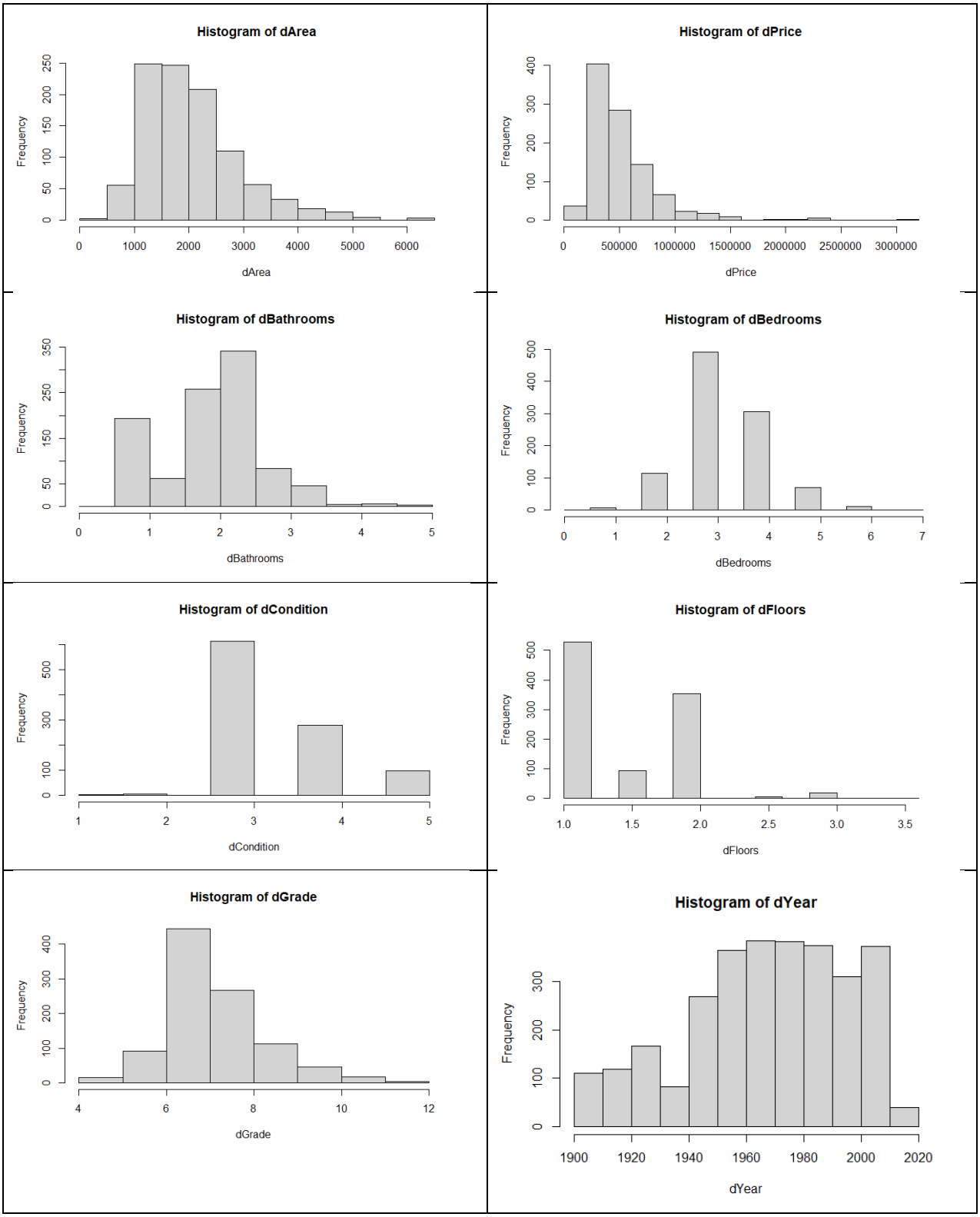
```
dPrice <- main$"price"
dArea <- main$"sqft_living"
dBedrooms <- main$"bedrooms"
dBathrooms <- main$"bathrooms"
dFloors <- main$"floors"
dCondition <- main$"condition"
dGrade <- main$"grade"
dYear <- main$"yr_built"
```

To find the number of missing values I used following commands. In my case, I had no NULL values from the beginning

```
sum(is.na(dPrice))
sum(is.na(dArea))
sum(is.na(dBedrooms))
sum(is.na(dBathrooms))
sum(is.na(dFloors))
sum(is.na(dCondition))
sum(is.na(dGrade))
```

After that I created histogram graphs of all my 8 variables.

Histogram of Variables



By looking at histogram graphs, I can say the following about my data

Area – Most houses have living Area in range from 1000 sqft to 2000 sqft
Price – Most frequent prices are below 500.000$
Bathrooms – People mostly have 2.5 bathrooms at home
Bedrooms – 3 Bedrooms are commonly used at Seattle houses
Floors – 1 floored Houses are common thing
Condition – Most houses have appropriate condition, which is middle number 3
Grade – 7 is the most frequent grade for Seattle houses. That is more than half, not bad
Year – Most houses were built in years from 1970 to 1990 and in 2010

Stage 2

Here are central tendency measures I have calculated with my variables.

	MEAN	*MEDIAN*	*MODE*	*GM MEAN*
Area	2052.95	1900	1560	1889.451
Price	516922.72	442750	550000	452024.6
Bathrooms	2.07	2	2.5	1.925442
Bedrooms	3.39	3	3	3.269665
Floors	1.44	1	1	1.353453
Condition	3.46	3	3	3.398386
Grade	7.61	7	7	7.526904
Year	1967.91	1970	1977	1967.711

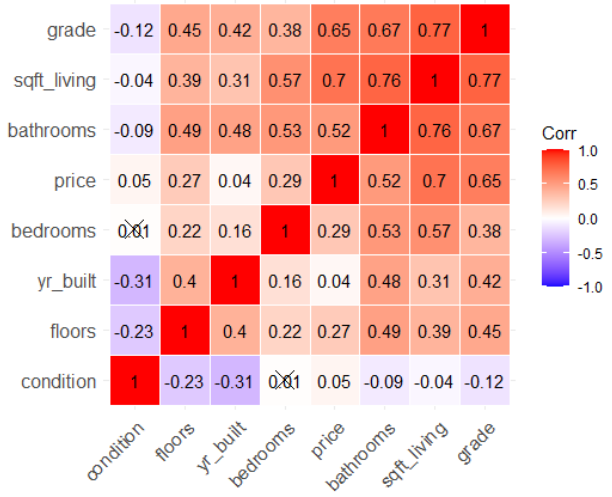
By looking at central tendency measures and comparing them with histogram graphs I can confirm that the information corresponds each other. Mode values are mostly similar to my observations I made on Stage 1 with Histograms

	Range	*Interquartile Range*	*Variance*	*Standard Deviation*
Area	6970	1060	751449.2	866.86
Price	2917500	317375	91966540178	303259.86
Bathrooms	4.5	1	0.5480549	0.74
Bedrooms	6	1	0.7709461	0.88
Floors	2	1	0.2573793	0.51
Condition	4	1	0.4627718	0.68
Grade	7	1	1.270155	1.13
Year	115	40	775.9381	27.86

By looking at variability measures I can confirm that Range values match with Distribution graphs. Also, by calculating Ranges from (Mean - SD) to (Mean + SD) I can say that these ranges look similar to original distribution graphs

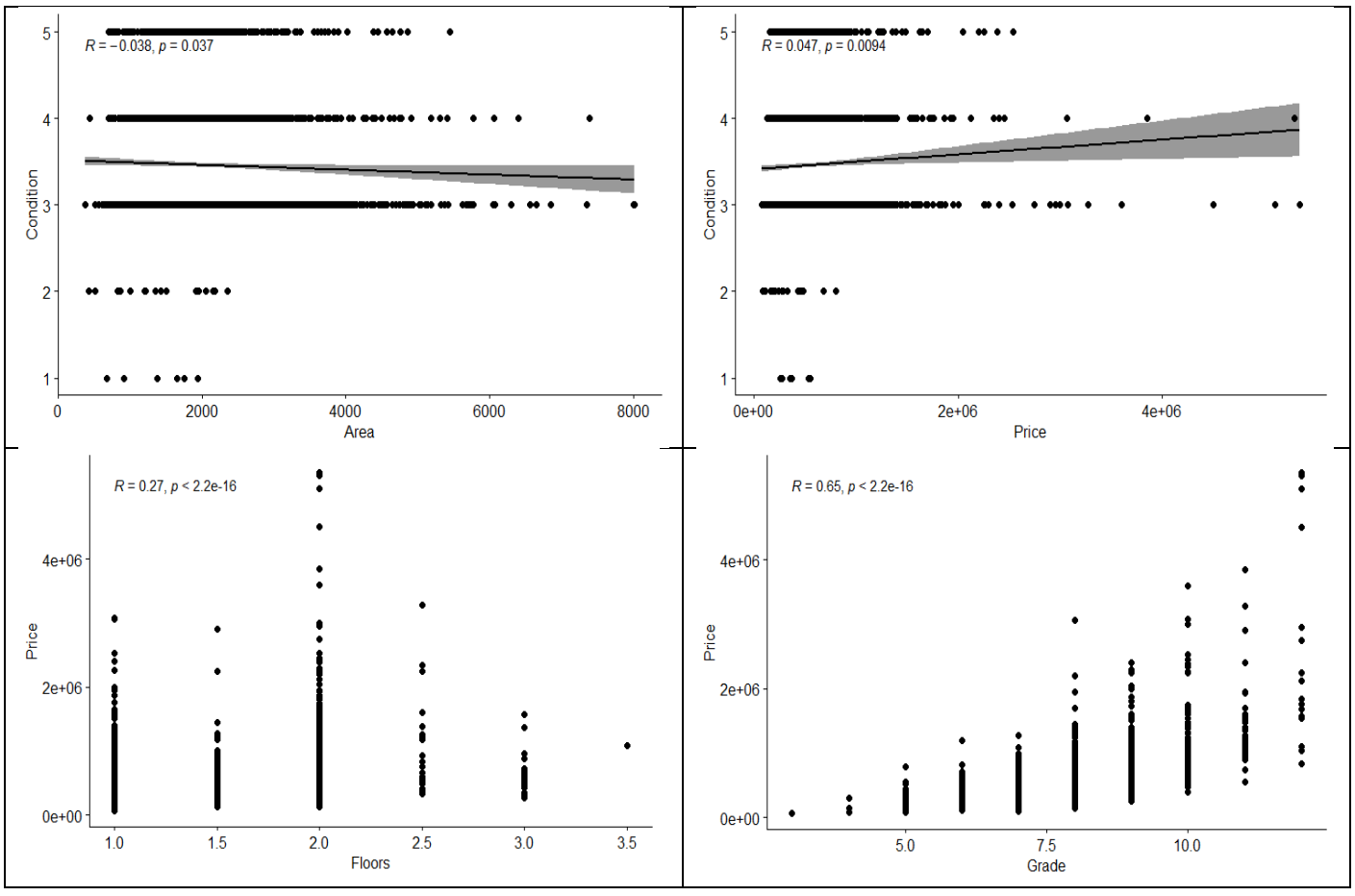
Stage 3

To find linear relationship between my values I had to compute correlation matrix to find the correlation value between my variables.

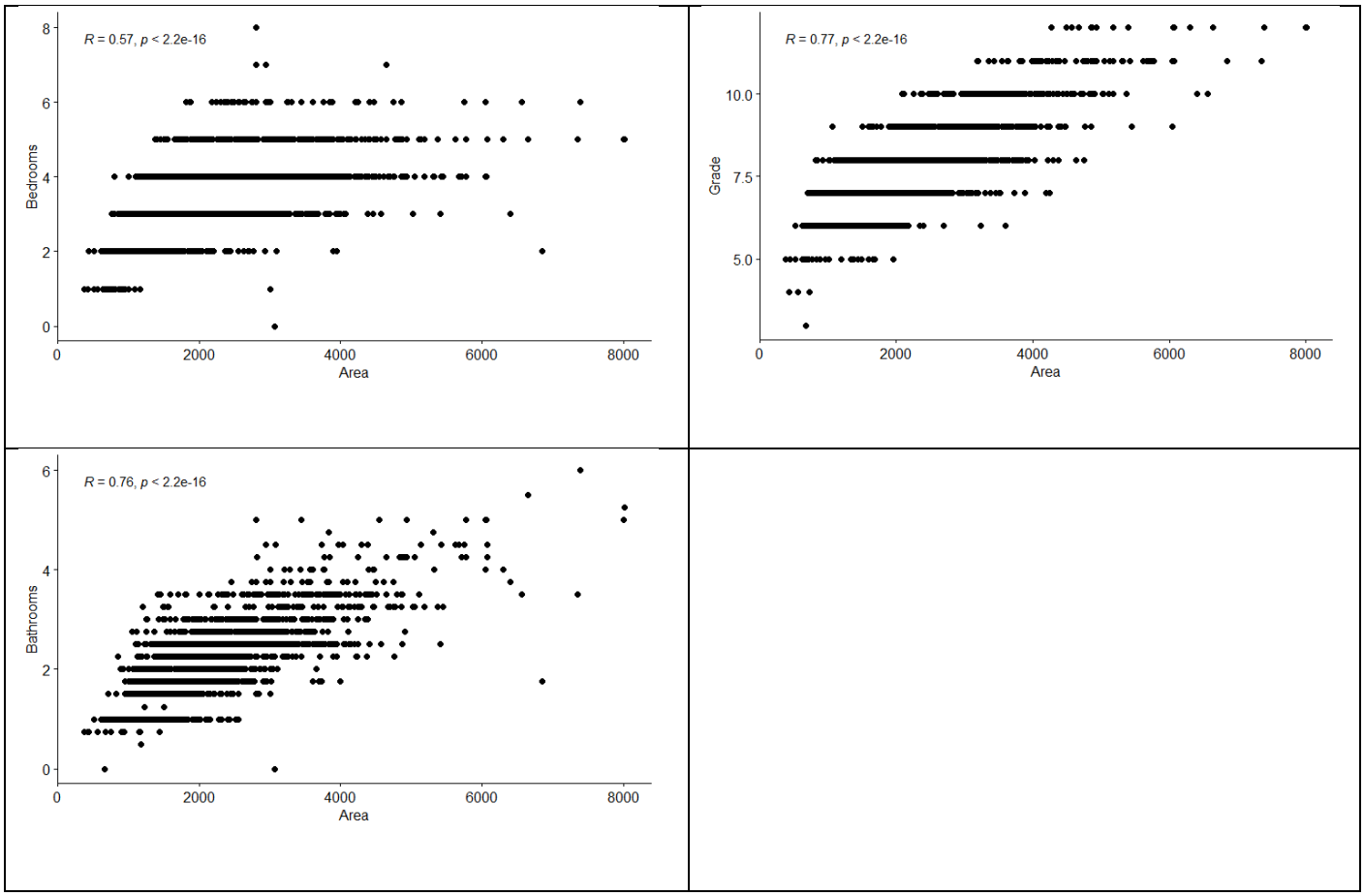


By looking at this correlation graph, I have concluded that Area-Price, Area-Bathrooms, Area-Bedrooms, Area-Grade and Grade-Price have high correlation values. Although they have high correlation values, only Area-Price has a logical connection between each other. Here, my Independent variable is “Area” (sqft_living) because I can see that price is calculated based on this variable.

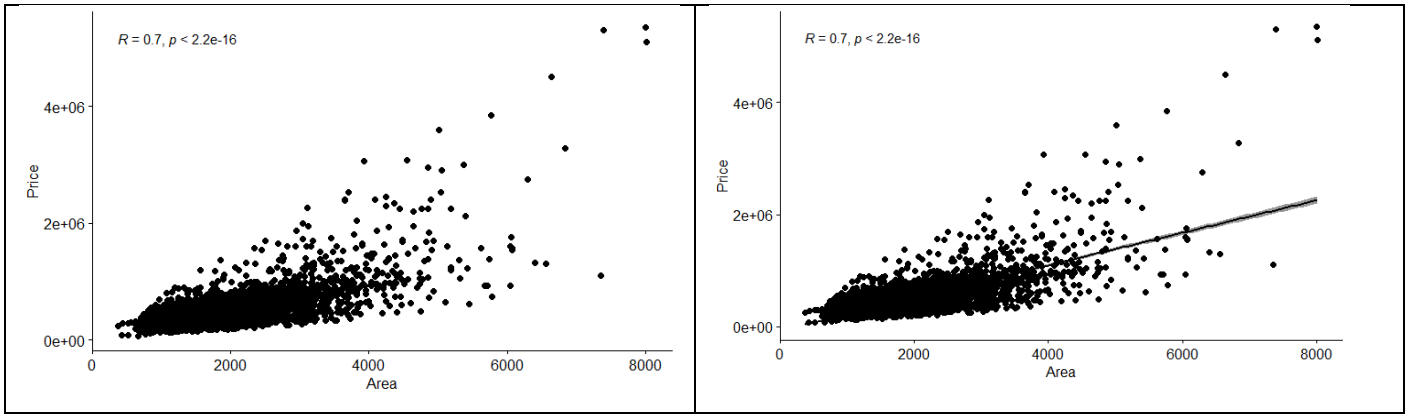
Here are connections that I thought they might be logically connected but they have really weak correlation or no connection at all, so I won’t use them.



Following scatter plots have a good correlation but the graphs show no connection at all



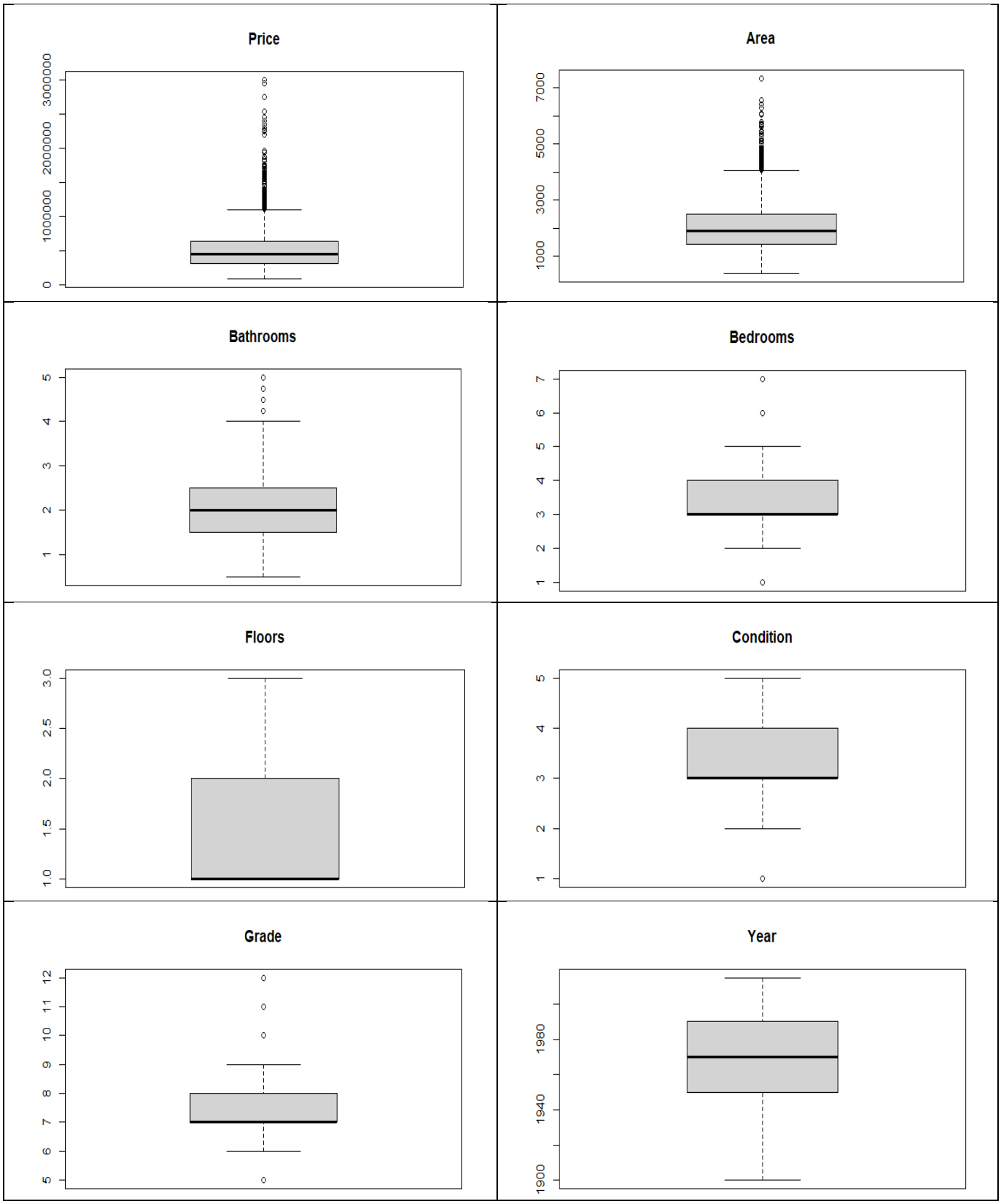
Here is the “dependent variable vs independent variable” scatter plot with and without regression line.



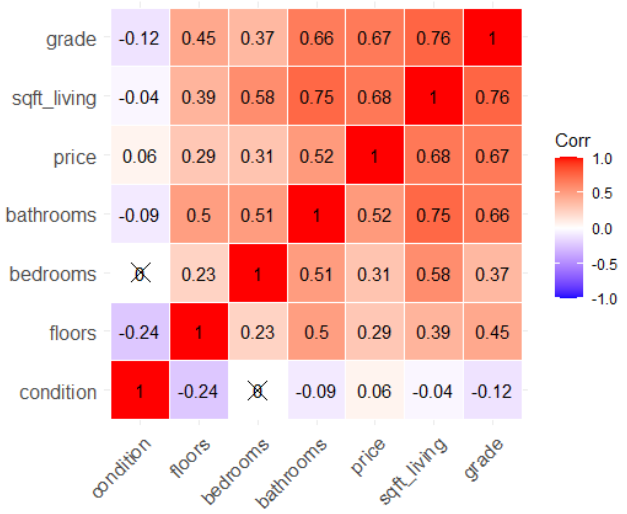
By looking at above linear models I can see that they have some outliers that I would like to remove.

Box plots can be used to identify outliers. We cannot decide upon the outliers on distribution plots alone. Box plots just help us to know more about outliers in features. The dots with no filling outside the box depict the data-points that pose as outliers.

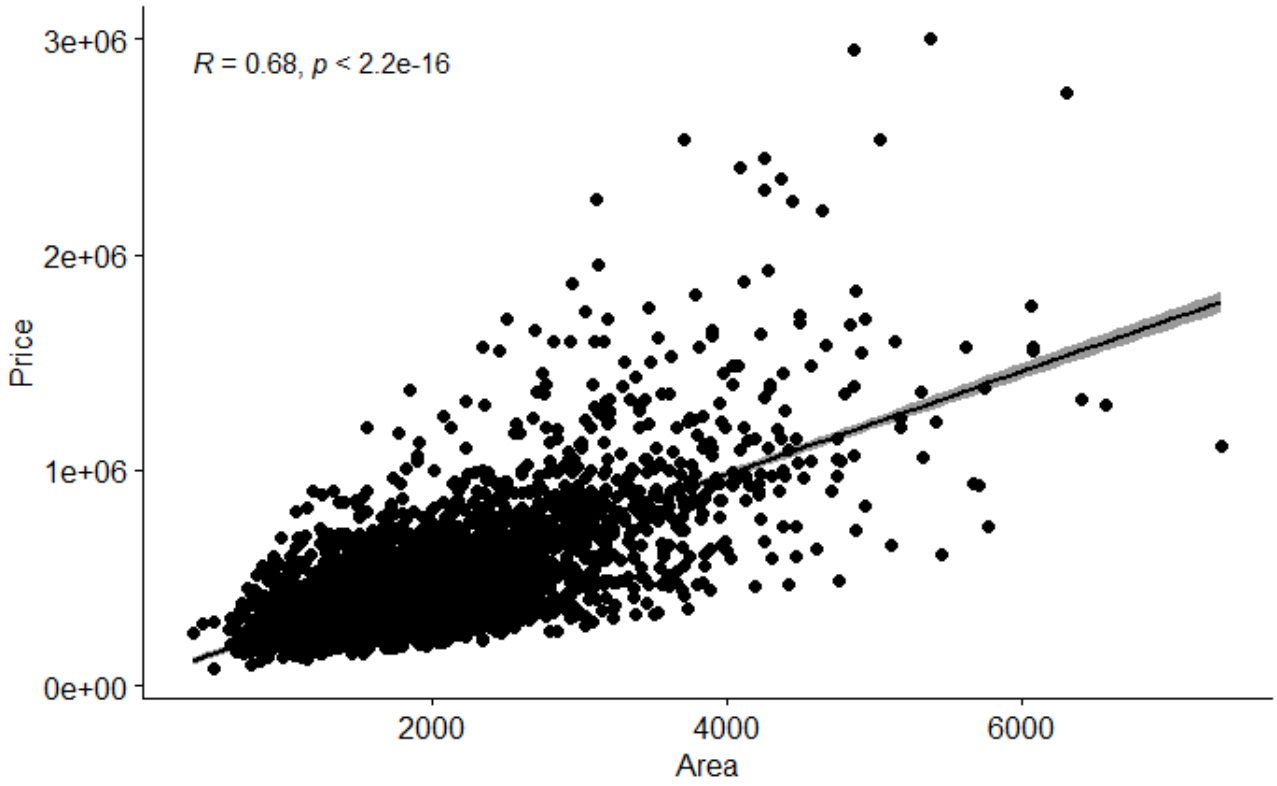
Boxplots



Correlation matrix after removing some, but not all, outliers. Interesting thing is that after removing some outliers my correlation values have slightly decreased



Here is my Area-Price “dependent variable vs independent variable” scatter plot with regression line with less outliers:



Stage 4

I have chosen VikingLotto for Probability theory part. This lottery is played in Estonia and has following main rule. There are 6 main numbers and 1 bonus number. You can choose 6 main numbers from range 1 to 48 without repetitions and 1 bonus number within range 1 to 8.

This is the formula I have used to calculate the probability:

Only main numbers	Main numbers + bonus number
$C_n^k = \frac{n!}{(n-k)!\,k!}$ Where n is total numbers in pool (1-48) and k is number of matches (3-6)	$C_n^k = \frac{n!}{(n-k)!\,k!} * \frac{8!}{(8-1)!\,1!}$ Where n is total numbers in pool (1-48) and k is number of matches (3-6)

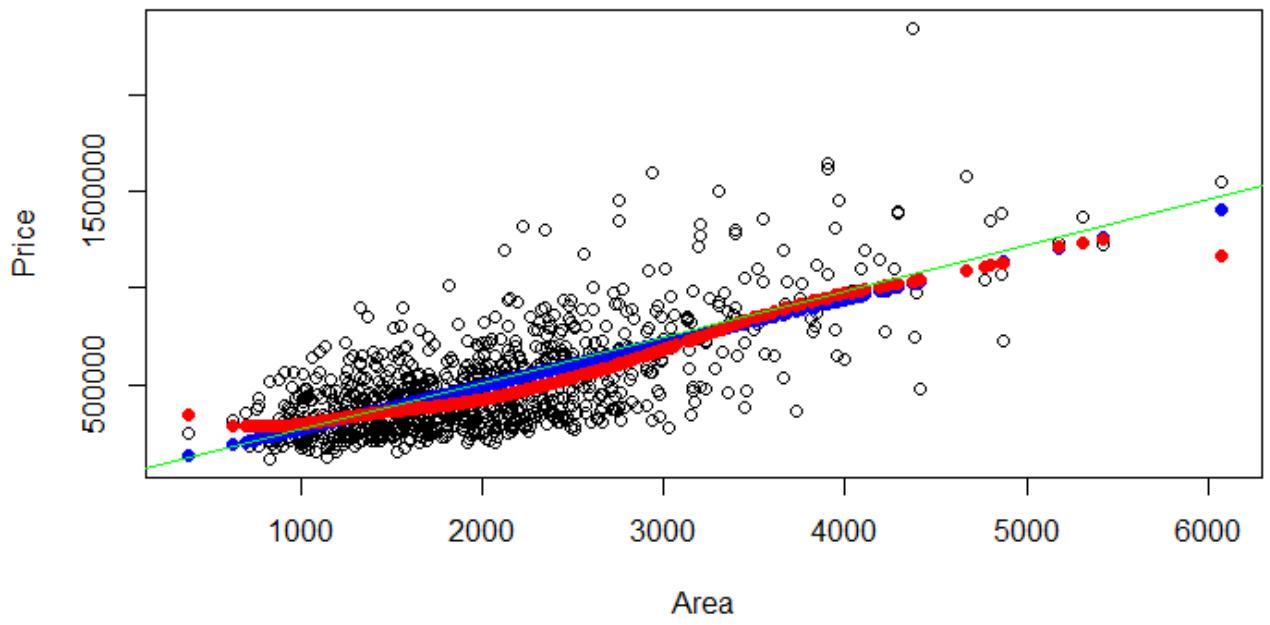
Price level	*Number of matches*	*Probability*
I (Jackpot)	6 main numbers + 1 bonus number	1:98 172 096
II	6 main numbers	1:12 271 512
III	5 main numbers + 1 bonus number	1:13 698 432
IV	5 main numbers	1:1 712 304
V	4 main numbers + 1 bonus number	1:1 556 640
VI	4 main numbers	1:194 580
VII	3 main numbers + 1 bonus number	1:138 368
VIII	3 main numbers	1:17 296

On formula with bonus number, we can just calculate main number part probability and multiply it by 8, because we have only one choose option on bonus part.

As it can be seen, even winning the lowest price level with 3 numbers match is challenging. There is only one conclusion. Do not play lotteries

Stage 5

I have divided my training and testing set with proportion 70/30. My testing data set consists of 30% of randomly chosen values from training set with less outliers. I have 892 items in my testing set. I have done predictions of linear model prediction and SVM (support vector machine) prediction. Here I am trying to predict the "Price" value based on "Area". The results of predictions can be seen on graphs below where blue colored dotted line means LM prediction, red colored dotted line means SVM prediction, green line means regression line of training set and black dots represent testing set values.



By using MAPE function I have calculated mean difference in percentage between original testing set values and predicted set values.

	Area & Price
lm	34.18%
svm	31.14%

By looking at error values, I can say that prediction worked out quite well. None of predictions had error more than 50%. I can see that SVM predictions have less error values in percentage comparing to LM row