Report on learning practices on the course “Methods and models for multivariate data analysis”

Performed by:

Bezbochina Alexandra, J41332c

Sahatova Kseniya, J41332c

Shelestova Margorita, J41322c

Ospishev Semen, J41322c

Saint-Petersburg

2021

Table of contents

[1. ANALYSIS OF UNIVARIATE RANDOM VARIABLES 3](#_Toc69981480)

[1.1 Substantiation of chosen subsample 3](#_Toc69981481)

[1.2. Plotting a non-parametric estimation of PDF in form of a histogram and using kernel density function (or probability law in case of discrete RV) 4](#_Toc69981482)

[1.3. Order statistics estimation and its representation as “box with whiskers” plot 5](#_Toc69981483)

[1.4. Selection of theoretical distributions that best reflect empirical data 7](#_Toc69981484)

[1.5. Estimation of random variable distribution parameters using maximum likelihood technique and LS methods 8](#_Toc69981485)

[1.6. Validation of empirical and theoretical distributions using quantile biplots 11](#_Toc69981486)

[1.7. Statistical tests (2 at least). 13](#_Toc69981487)

[1.8 Brief conclusions on 1 laboratory work 14](#_Toc69981488)

[2. ANALYSIS OF MULTIVARIATE RANDOM VARIABLES 16](#_Toc69981489)

[2.1 Plotting a non-parametric estimation of PDF in form of a histogram and kernel density function for MRV. 16](#_Toc69981490)

[2.2 Estimation of multivariate mathematical expectation and variance. 17](#_Toc69981491)

[2.3 Non-parametric estimation of conditional distributions, mathematical expectations and variances. 18](#_Toc69981492)

[2.4 Estimation of pair correlation coefficients, confidence intervals for them and significance levels. 21](#_Toc69981493)

[2.5 Task formulation for regression, multivariate correlation. 23](#_Toc69981494)

[2.6 Regression model, multicollinearity and regularization (if needed). 23](#_Toc69981495)

[2.7 Quality analysis 24](#_Toc69981496)

[2.8 Brief conclusions on 2 laboratory work 24](#_Toc69981497)

[3. SAMPLING OF MULTIVARIATE RANDOM VARIABLES 24](#_Toc69981498)

[3.1 Choosing variables for sampling 24](#_Toc69981499)

[3.2 Sampling of variables from task 1 25](#_Toc69981500)

[3.3 Estimating of correlations 26](#_Toc69981501)

[3.4 Constructing of Bayesian network manually 27](#_Toc69981502)

[3.5 Constructing of Bayesian network algorithmically 28](#_Toc69981503)

[3.6. Analyzing a quality of constructed Bayesian networks 29](#_Toc69981504)

[3.7 Brief conclusions on 3 laboratory work 30](#_Toc69981505)

[4. STATIONARITY OF THE PROCESSES 31](#_Toc69981506)

[4.1 Substantiation of chosen sampling 31](#_Toc69981507)

[4.2 Stationary analysis 31](#_Toc69981508)

[4.3 Covariance or correlation function analysis. 35](#_Toc69981509)

[4.4 Noise filtration and estimation of spectral density function 38](#_Toc69981510)

[4.5 Auto-regression model 39](#_Toc69981511)

# 1. ANALYSIS OF UNIVARIATE RANDOM VARIABLES

# 1.1 Substantiation of chosen subsample

Dataset contains house sale prices for King County. It includes homes sold between May 2014 and May 2015. The shape of data is (17571,21). There are a lot of variables including categorial and continuous variables.

The main variables are:

* price - price of each home sold;
* bedrooms/ bathrooms - number of bedrooms/ bathrooms;
* sqft\_living - square footage of the apartments interior living space;
* sqft\_lot - square footage of the land space;
* floors - number of floors;
* waterfront - dummy variable for whether the apartment was overlooking the waterfront or not;
* view - index from 0 to 4 of how good the view of the property was;
* condition - index from 1 to 5 on the condition of the apartment;
* grade - index from 1 to 13, where 1-3 falls short of building construction and design;
* sqft\_above - the square footage of the interior housing space that is above ground level;
* sqft\_basement - the square footage of the interior housing space that is below ground level;
* sqft\_living15 - the square footage of interior housing living space for the nearest 15 neighbors;
* sqft\_lot15 - the square footage of the land lots of the nearest 15 neighbors.

As we can see, the data contains many variables that can be analyzed.

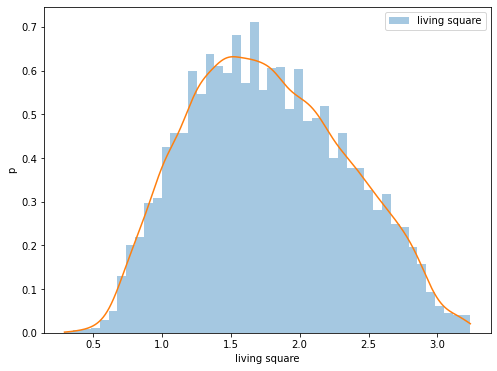
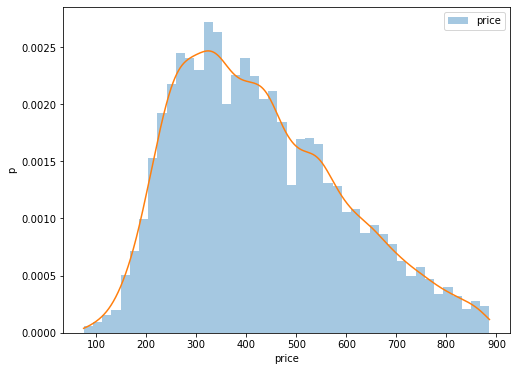
The reasons why this dataset was chosen:

1. there are continuous, categorical variables (predictors and factors);
2. there is a time series;
3. data is suitable for training regression and autorepression models.

For 1 laboratory work the following variables were selected: price (rice of each home sold), sqft\_living (square footage of the apartments interior living space), sqft\_above (the square footage of the interior housing space that is above ground level), sqft\_living15 (the square footage of interior housing living space for the nearest 15 neighbors).

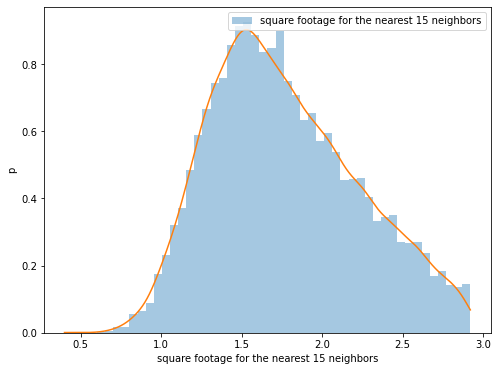
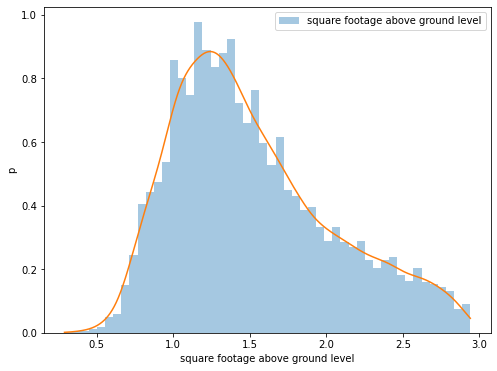
# 1.2. Plotting a non-parametric estimation of PDF in form of a histogram and using kernel density function (or probability law in case of discrete RV)

To estimate the PDF of the variables histograms were constructed and kernel density estimating was used. The results for each variable are presented below with short comments.



Picture 1.1 – Histogram kernel density estimation of PDF (the first two variables)

The distribution of the first variable (price) is similar to the gamma distribution, and the normal distribution can also be examined. The second variable (sqft\_living) is characterized normal distribution.



Picture 1.2 – Histogram kernel density estimation of PDF (the second two variables)

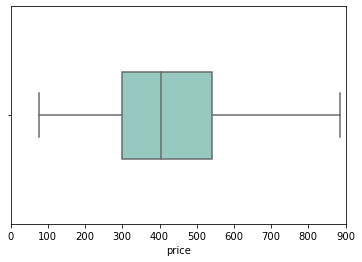
The gamma distribution is suitable for these variables (sqft\_above and sqft\_living15), but further research will be carried out for both gamma and normal distributions.

# 1.3. Order statistics estimation and its representation as “box with whiskers” plot

In this secti ordinal statistics and plot “box with whiskers” are presented for each variable. This analysis will help better select the appropriate distribution.

Boxplot is method for graphically depicting groups of numerical data through their quartiles. Outliers may be plotted as individual points. Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Quantile​​ | 0.10​​ | 0.25​​ | 0.50​​ | 0.75​​ | 0.90​​ |
| Value​​ | 235​ | 299.95​ | 403​ | 539.95​ | 667​ |



Picture 1.3 – Order statistics estimation and “box with whiskers” plot (variable – price)

The first variable (price) is characterized by slight shift to the left, there are no outliers.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Quantile​​ | 0.10​​ | 0.25​​ | 0.50​​ | 0.75​​ | 0.90​​ |
| Value​​ | 1.04​ | 1.33​ | 1.73​ | 2.18​ | 2.57​ |



Picture 1.4 – Order statistics estimation and “box with whiskers” plot (variable – sqft\_living)

The “box with whiskers” for the second variable (sqft\_living) confirms our assumptions that the variable can be distributed normally. The graph is symmetrical, there are no outliers.

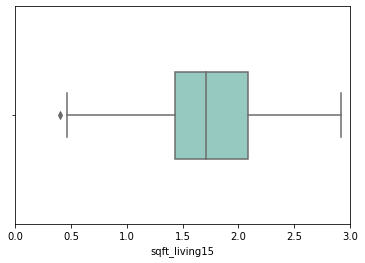
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Quantile​​ | 0.10​​ | 0.25​​ | 0.50​​ | 0.75​​ | 0.90​​ |
| Value​​ | 0.93​ | 1.13​ | 1.41​ | 1.83​ | 2.32​ |



Picture 1.5 – Order statistics estimation and “box with whiskers” plot (variable – sqft\_above)

The graph of the third variable (sqft\_above) is slightly shifted to the left, there are outliers.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Quantile​​ | 0.10​​ | 0.25​​ | 0.50​​ | 0.75​​ | 0.90​​ |
| Value​​ | 0.93​ | 1.13​ | 1.41​ | 1.83​ | 2.32​ |



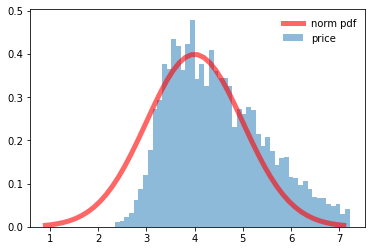
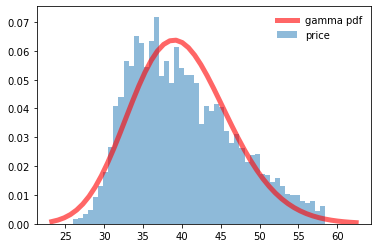
Picture 1.6 – Order statistics estimation and “box with whiskers” plot (variable – sqft\_living15)

  The plot of the third variable (sqft\_living15) is also slightly shifted to the left, there is one outlier.

# 1.4. Selection of theoretical distributions that best reflect empirical data

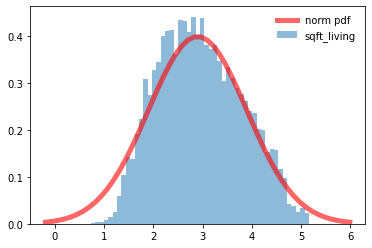
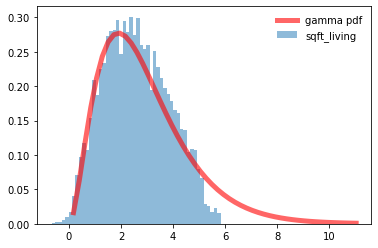
In this section, we apply two distributions to our variables and choose the one that best describes our variables.

Below you can see plots of gamma (left) and normal distribution (right) for each variable.



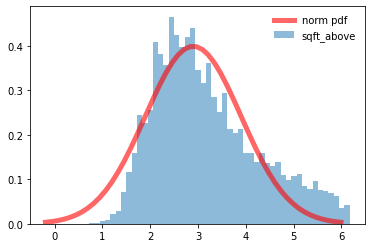
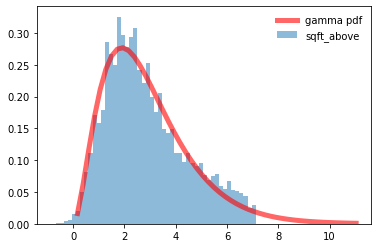
Picture 1.7 – Gamma and norm distributions (variable – price)

Gamma distribution better describes the variable price.



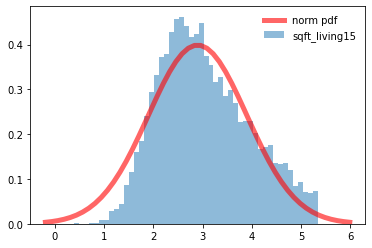
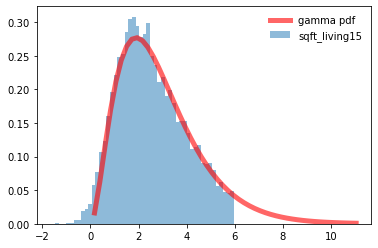
Picture 1.8 – Gamma and norm distributions (variable – sqft\_living)

Both distributions can be applied to this variable (sqft\_living), statistical tests should be performed to estimate the fitted distribution.



Picture 1.9 – Gamma and norm distributions (variable – sqft\_above)

Gamma distribution better describes the variable sqft\_above.



Picture 1.10 – Gamma and norm distributions (variable – sqft\_living15)

Gamma distribution better describes the variable sqft\_living15.

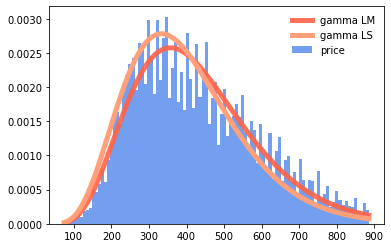
# 1.5. Estimation of random variable distribution parameters using maximum likelihood technique and LS methods

In this section, we estimate parameters for the distributions of variables. For this we used 2 methods: maximum likelihood technique and LS method.

The results of the evaluations are presented below. It should be noted that in some cases the parameters differ insignificantly.

Table 1.1 – Estimation of gamma distribution parameters using maximum likelihood technique and LS methods (variable: price)

|  |  |
| --- | --- |
| Method | Parameters of Gamma distribution |
| Maximum likelihood estimation | [5.586, 31.565, 70.999] |
| Least squares estimation | [5.601, 31.691, 65.769] |

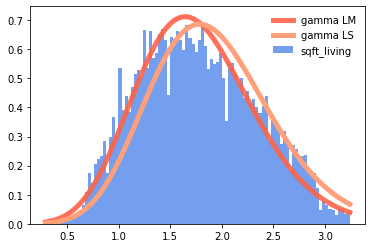


Picture 1.11 – Gamma distribution of variable obtained in two ways (variable: price)

For the first variable (price), the distribution parameters were estimated.

Table 1.2 – Estimation of gamma distribution parameters using maximum likelihood technique and LS methods (variable: sqft\_living)

|  |  |
| --- | --- |
| Method | Parameters of Gamma distribution |
| Maximum likelihood estimation | [20.954, -0.848, 0.125] |
| Least squares estimation | [20.854, -0.795, 0.130] |

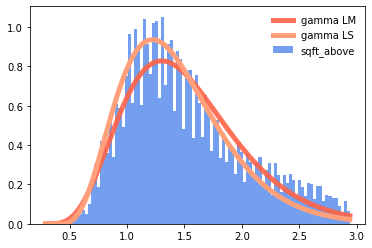


Picture 1.12 – Gamma distribution of variable obtained in two ways (variable: sqft\_living)

For the second variable (sqft\_living)), the distribution parameters were estimated.

Table 1.3 – Estimation of gamma distribution parameters using maximum likelihood technique and LS methods (variable: sqft\_above)

|  |  |
| --- | --- |
| Method | Parameters of Gamma distribution |
| Maximum likelihood estimation | [5.693, 0.276, 0.219] |
| Least squares estimation | [4.911, 0.396, 0.211] |

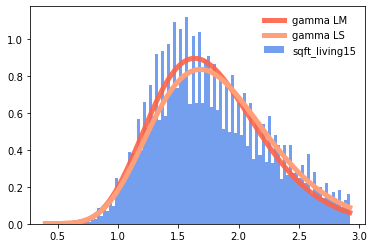


Picture 1.13 – Gamma distribution of variable obtained in two ways (variable: sqft\_above)

For the third variable (sqft\_above), the distribution parameters were estimated.

Table 1.4 – Estimation of gamma distribution parameters using maximum likelihood technique and LS methods (variable: sqft\_living15)

|  |  |
| --- | --- |
| Method | Parameters of Gamma distribution |
| Maximum likelihood estimation | [11.321, 0.217, 0.137] |
| Least squares estimation | [10.811, 0.199, 0.151] |

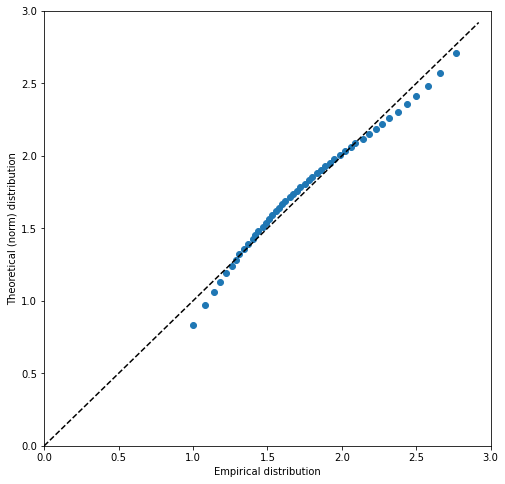
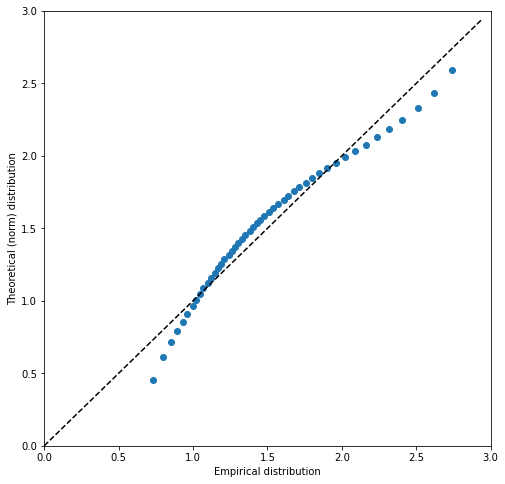
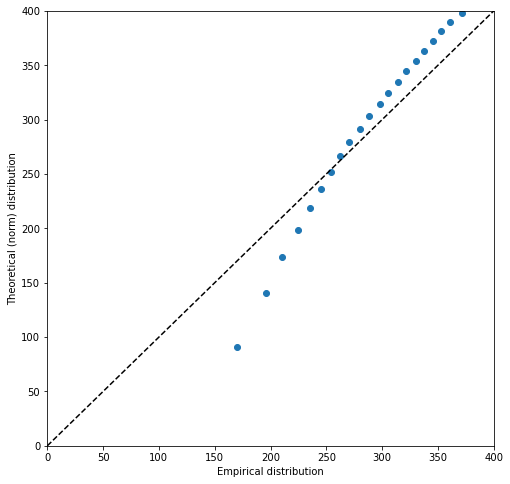


Picture 1.14 – Gamma distribution of variable obtained in two ways (variable: sqft\_living15)

For the fourth variable (sqft\_living15), the distribution parameters were estimated.

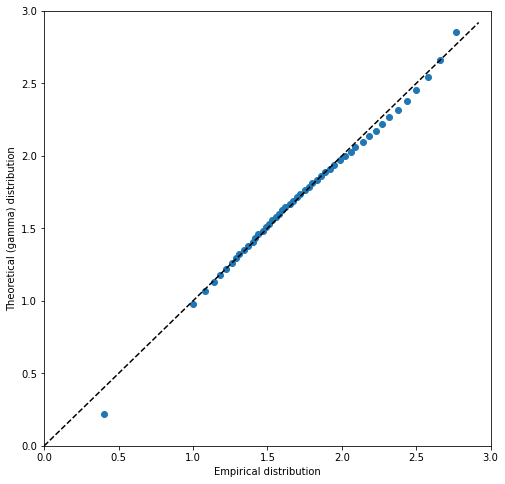
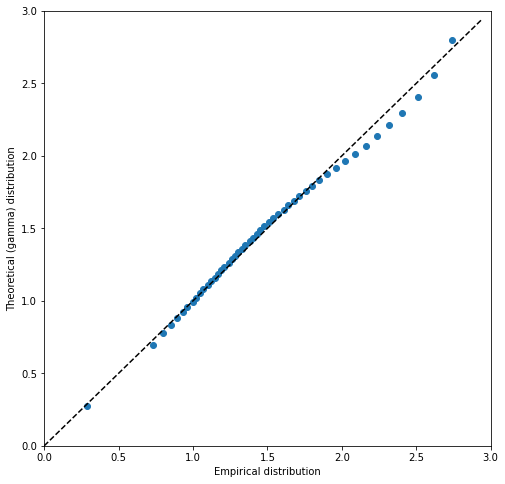
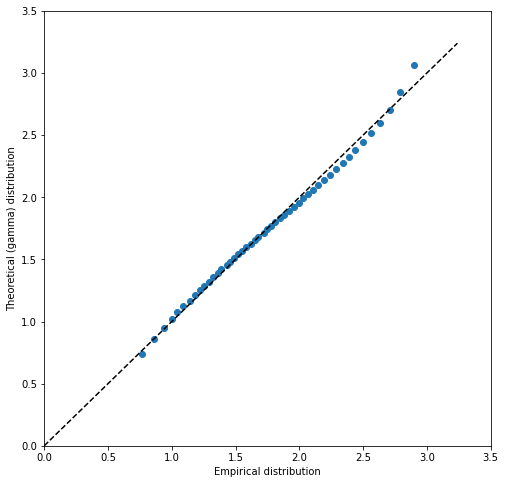
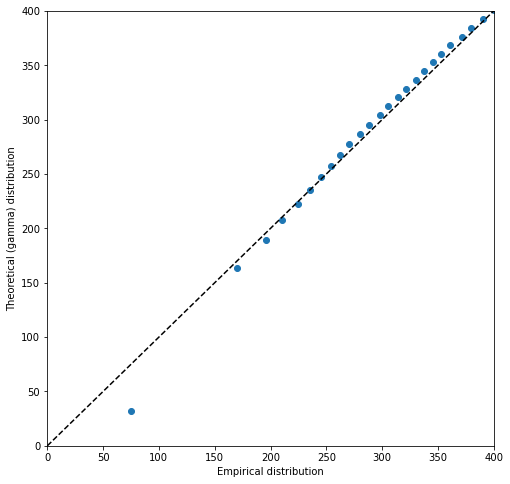
# 1.6. Validation of empirical and theoretical distributions using quantile biplots

We plotted QQ-plots for the normal distribution (all variables), and we can see that only one variable (sqft\_living) is well described by this type of distribution. Next, we will look at the gamma distribution.



Picture 1.15 – Quantile biplots of norm distribution (all variables)

QQ-plots for gamma distribution:



Picture 1.16 – Quantile biplots of gamma distribution (all variables)

As can be seen from the graphs, the distribution of the variables is as follows:

Price – gamma distribution;

Sqft\_living – gamma/normal distribution;

Sqft\_above – gamma distribution;

Sqft\_living15 – gamma distribution.

# 1.7. Statistical tests (2 at least).

Finally, we used statistical tests to determine whether our assumptions about the distribution of the variables were correct.

Kolmogorov–Smirnov test is nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution, or to compare two samples.

Cramer-von Mises test is also used for testing for goodness of fit. It is criterion used for judging the goodness of fit cumulative distribution function compared to a given empirical distribution function, or for comparing two empirical distributions.

The first variable (price) passed both tests (gamma and normal distribution). However, it should be noted that the p-value for the gamma distribution is larger than for the normal distribution.



Picture 1.17 – Results of Kolmogorov-Smirnov and Cremer-von Mises test (gamma and norm distributions, variable: price)

As already mentioned, the second variable (sqft\_living) can be described by both the gamma distribution and the normal distribution. Passed both tests:

Picture 1.18 -– Results of Kolmogorov-Smirnov and Cremer-von Mises test (gamma and norm distributions, variable: sqft\_living)

The next variable passed both tests, similarly, the p-value for the gamma distribution is greater than for the normal one.

Picture 1.19 – Results of Kolmogorov-Smirnov and Cremer-von Mises test (gamma and norm distributions, variable: sqft\_above)

We can see the same case with the last variable:

Picture 1.20 – Results of Kolmogorov-Smirnov and Cremer-von Mises test (gamma and norm distributions, variable: sqft\_living15)

Using two statistical tests to assess the correctness of the distributions, we can conclude that the fitted distributions describe the variables well. All variables (which are investigated in 1 laboratory work) can be described by a gamma distribution (also a variable sqft\_living can be described by normal distribution).

# 1.8 Brief conclusions on 1 laboratory work

Dataset contains house sale prices for King County. It includes homes sold between May 2014 and May 2015. The shape of data is (17571,21). There are a lot of variables including categorial and continuous variables.

The reasons why this dataset was chosen:

1. there are continuous, categorical variables (predictors and factors);
2. there is a time series;
3. data is suitable for training regression and autorepression models.

For 1 laboratory work the following variables were selected: price (rice of each home sold), sqft\_living (square footage of the apartments interior living space), sqft\_above (the square footage of the interior housing space that is above ground level), sqft\_living15 (the square footage of interior housing living space for the nearest 15 neighbors).

To estimate the PDF of the variables histograms were constructed and kernel density estimating was used. The distribution of the first variable (price) is similar to the gamma distribution, and the normal distribution can also be examined. The second variable (sqft\_living) is characterized normal distribution. The gamma distribution is suitable for variables sqft\_above and sqft\_living15.

Ordinal statistics and plot “box with whiskers” are presented for each variable. This analysis help better select the appropriate distribution.

Gamma distribution describes the variable price, sqft\_above and sqft\_living15. Both distributions can be applied to variable sqft\_living.

We estimated parameters for the distributions of variables. For this we used 2 methods: maximum likelihood technique and LS method. The results of the evaluations are presented.

Finally, we plotted QQ-plots and used statistical tests to determine whether our assumptions about the distribution of the variables were correct.

The first variable (price) passed both tests (gamma and normal distribution). However, it should be noted that the p-value for the gamma distribution is larger than for the normal distribution. As already mentioned, the second variable (sqft\_living) can be described by both the gamma distribution and the normal distribution. Passed both tests. The next variable passed both tests, similarly, the p-value for the gamma distribution is greater than for the normal one. We can see the same case with the last variable.

# 2. ANALYSIS OF MULTIVARIATE RANDOM VARIABLES

# 2.1 Plotting a non-parametric estimation of PDF in form of a histogram and kernel density function for MRV.

In the figure 2.1. non-parametric estimation of PDF in form of a histogram and kernel density function for two variables (price and living space) are represented. Both components have similar distribution, but we can see that point cloud is huge, there is a hint of correlation, but the spread is very large.

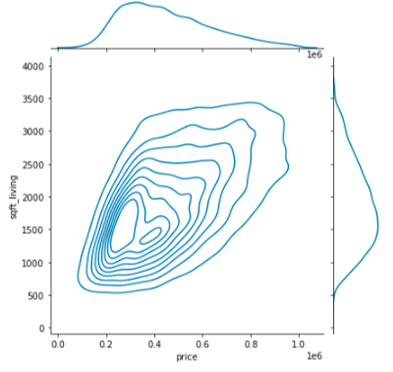
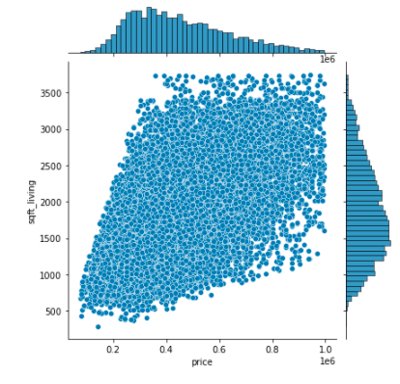
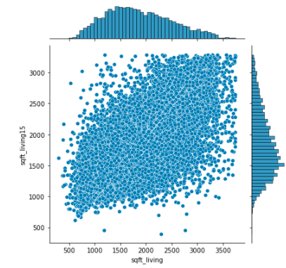
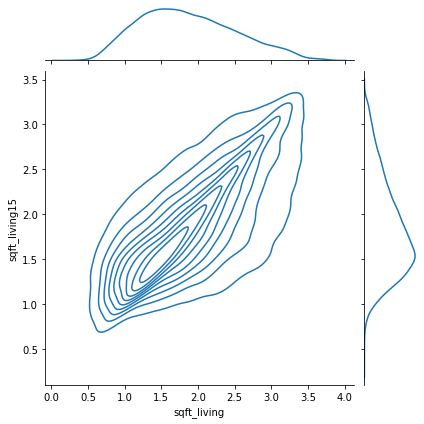
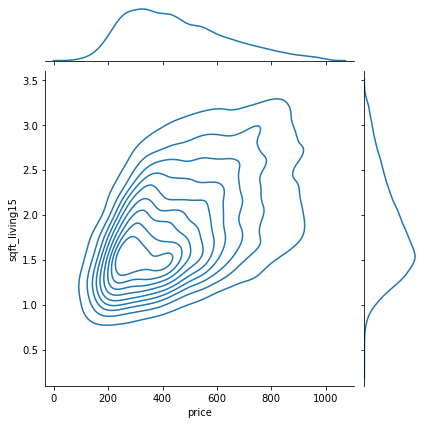
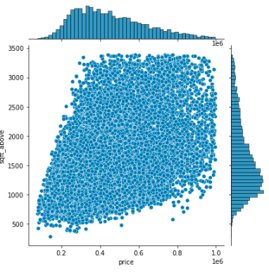


Figure 2.1 – PDF of Price of each home sold and Square living space

In the figure 2.2 there are PDF and kernel functions for other variables. All of them look rather similar.

In the dataset Price is a target variable, so we try to analyze it and see it`s relations more thoroughly.





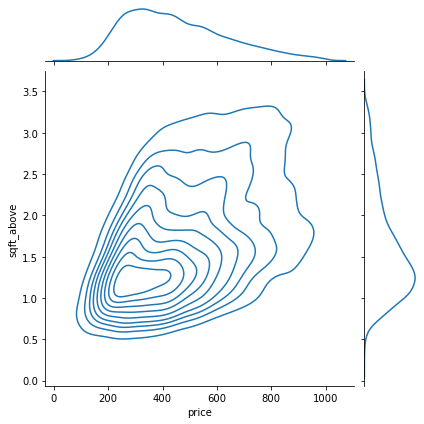
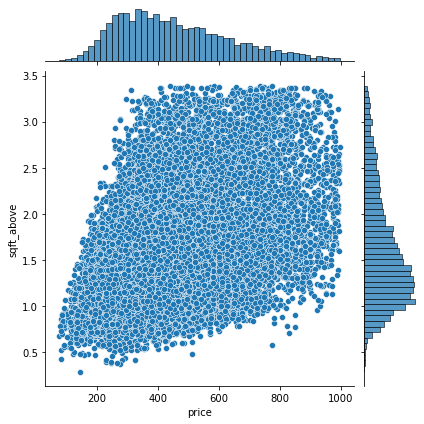


Figure 2.2 – PDF in form of a histogram and kernel density function

On the figure 2.3 there are PDFs of Price and categorial variable, such as Grade, Condition, Veiw.

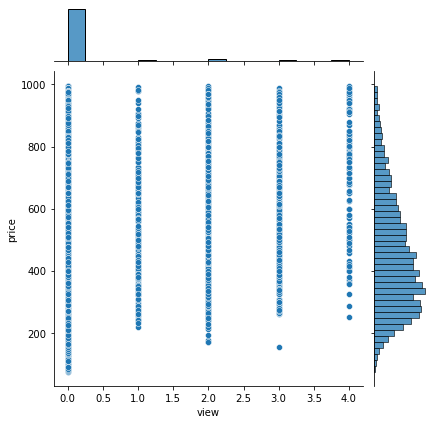
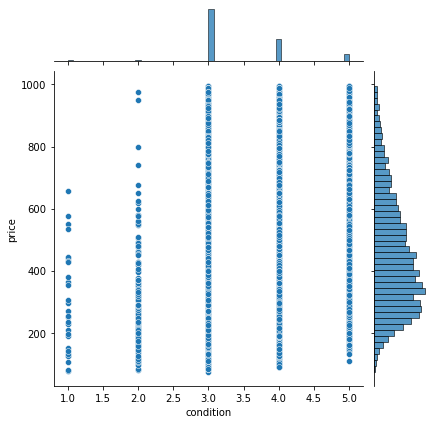
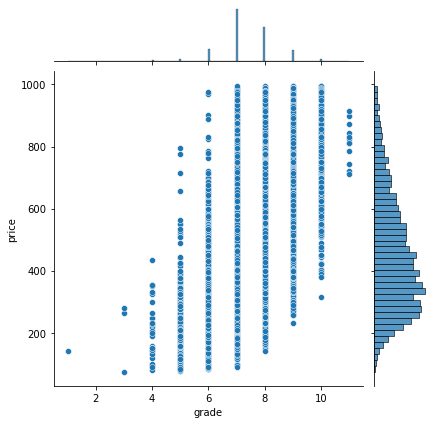
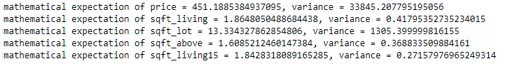


Figure 2.3 – Price and categorial variables

We can see relations on the plot Grade-Price. Here we can see clear gradation depending on the grade. The higher the grade, the bigger space and the higher the price.

# 2.2 Estimation of multivariate mathematical expectation and variance.

Mathematical expectation and variance are calculated and represented below.



Also, covariance matrix as a notion of variance to multiple dimensions is represented in the figure 2.4.

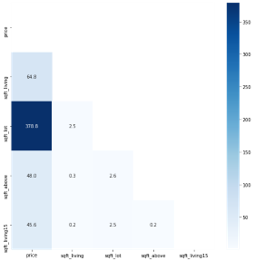


Figure 2.4 – Covariance matrix

# 2.3 Non-parametric estimation of conditional distributions, mathematical expectations and variances.

As price is target value, we are interested in clear gradation on this basis. In the figure 2.5 there are point clouds of two-dimensional variable Price-Living space in the context of the values of several categorical features. We can see number of bedrooms on the first plot correlate with living space, not with price, on the second plot there is no clear relations, on the third one clear relation with grade.

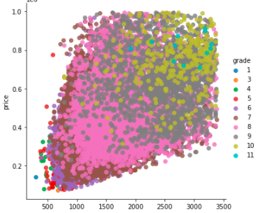
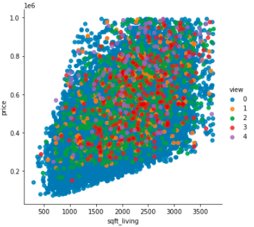
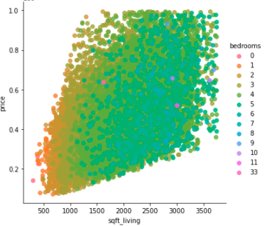


Figure 2.5 – Point clouds

The intergroup covariance is much higher than the intragroup covariance, which suggests that there is a dependence on grade and the spread between grades is much greater than within one grade. the mathematical expectation of the price increases depending on the grade.

Изображение выглядит как текст

Автоматически созданное описание

So, further we consider grade only to find conditional distribution, math expectation and variance. In the figure 2.6 there is distribution of price for different grade values.

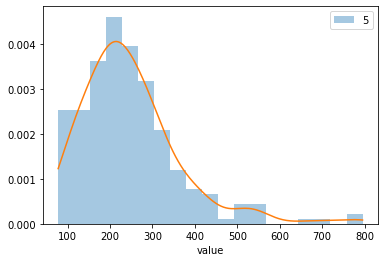
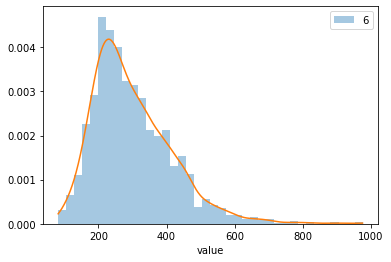
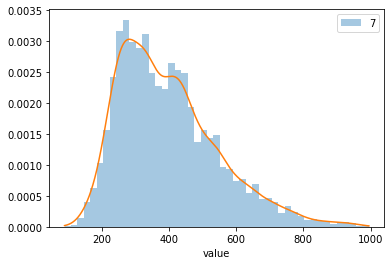
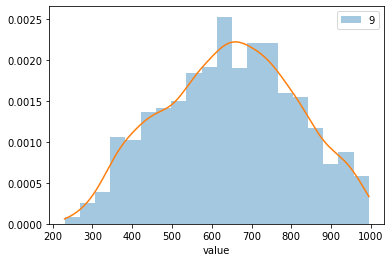
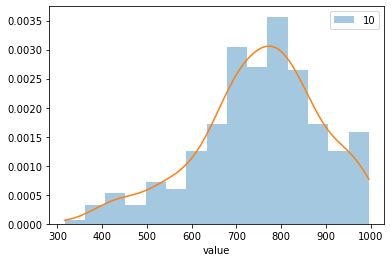
     

Figure 2.6 – Conditional distribution (Price – Grade)

On the table 2.1 and table 2.2 mathematical expectations and variance for different grade values are represented.

Table 2.1 Mathematical expectations of variables for different grade values



Table 2.2 Variance of variables for different grade values



Plots of conditional mathematical expectation and variance were built for Price variable are shown in figure 2.7.

Figure 2.7 – Conditional mathematical expectation and variance

# Estimation of pair correlation coefficients, confidence intervals for them and significance levels.

Pair correlation can be shown both as point cloud plot (figure 2.8), where the direction and size of cloud characterize the dependence, and heatmap (figure 2.9), where pair correlation coefficients are colored in order to size of coefficient.

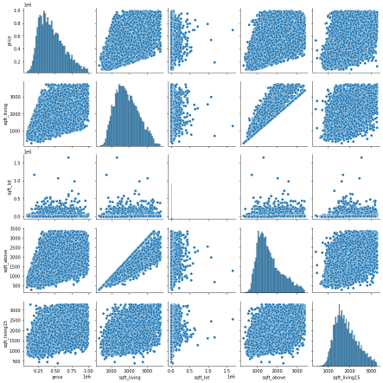


Figure 2.8 – Point clouds

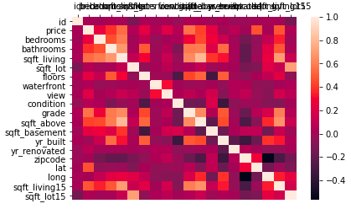


Figure 2.9 – Heatmap

Pair correlation coefficients, confidence intervals for them and significance levels are represented below.



# Task formulation for regression, multivariate correlation.

As we can see on the heatmap on the previous step, there is strong multicollinearity among factors. In the picture 2.10 the most correlated with Price variables and their correlation coefficients are printed.

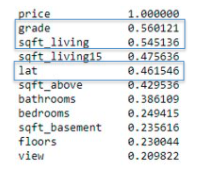


Figure 2.10 – Correlation with target variable Price

It wasn’t easy to choose factors for regression variables correlate with each other more then with price or the correlation with all variables including target one, is very low. Variable sqft\_living correlates with price but also with other variables such as sqft\_living15, sqft\_above, bathrooms, bedrooms, ect. These factors correlated with sqft\_living more than price, so we can`t include it into the model. So, as model factors grade, sqft\_living and latitude were chosen. Multivariate correlation considering these factors is 0.75.

# Regression model, multicollinearity and regularization (if needed).

For regression model Price variable is target is price, it`s logically and obvious. After analyzing point clouds and histograms of the variables we decide to build simple multiple regression.

For building regression model statsmodels python library was used. Multicollinearity was analyzed on previous step, we also don`t need regularization because overfitting is not a problem in our case and the number of factors is small because of multicollinearity so we don`t need any other restrictions.

# 2.7 Quality analysis

After building the model we can calculate R-squared and adjusted R-squared.



Quality of residuals is rather well. Mathematical expectation of residuals is equal 0. Kolmogorov-Smirnov test shows the distribution is normal (p-value = 0,6), so, these features correspond to the classical normal linear model

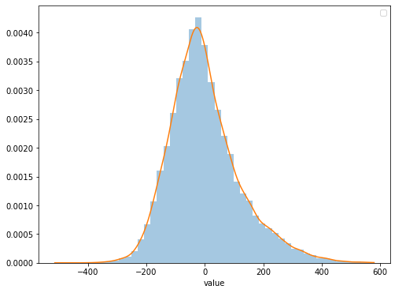


Figure 2.11 – Distribution residuals

# 2.8 Brief conclusions on 2 laboratory work

Quality of regression is not very high according to not very big amount of R-squared. Point cloud shows the data spread is very large, so probably, it`s reasonable to try build not leaner regression but another statistical and AI models.

# 3. SAMPLING OF MULTIVARIATE RANDOM VARIABLES

# 3.1 Choosing variables for sampling

The following columns were used:

“price” - Price of each home sold.

“bathrooms” - Number of bathrooms, where .5 accounts for a room with a toilet but no shower.

“sqft\_living” - Square footage of the apartments interior living space.

“sqft\_lot” - Square footage of the land space.

“grade” - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have high quality level of construction and design.

“sqft\_above” - The square footage of the interior housing space that is above ground level.

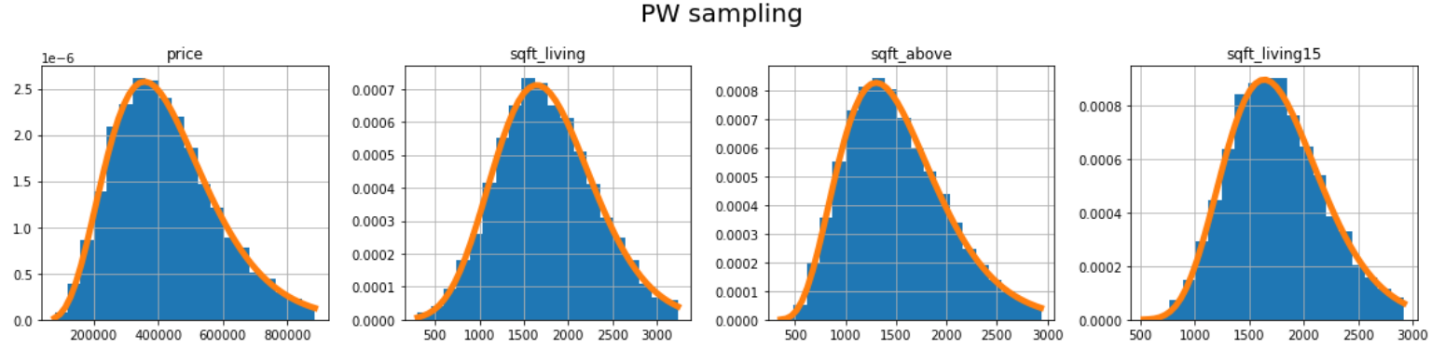
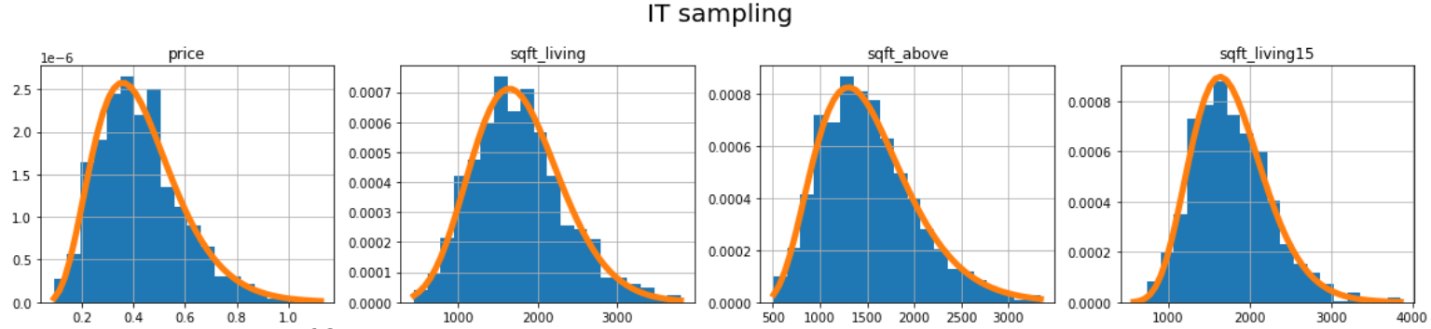
“sqft\_living15” - The square footage of interior housing living space for the nearest 15 neighbors.

“sqft\_lot15” - The square footage of the land lots of the nearest 15 neighbors.

“sqft\_living” and “sqft\_lot” – targets, other variables – predictors.

# 3.2 Sampling of variables from task 1

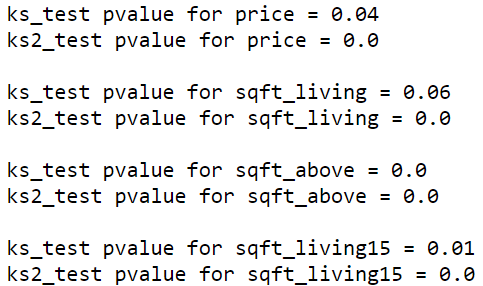
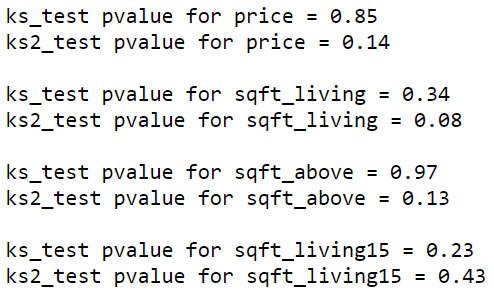
Variables “price”, “sqft\_living”, “sqft\_above” and “sqft\_living15” were fitted by gamma distribution. The sampling was performed from evaluated probability distributions. Graphs are represented below:



Picture 3.1 - Distributions of sampled data that were built via PW and IT sampling algorithms.

Two methods of sampling were used: Piece-wise discretization (PW) and Inverse transform sampling (IT). For each evaluated sample two Kolmogorov-Smirnov tests were performed: for comparison with fitted distribution (ks\_test) and with real data distribution(ks2\_test). Here you can see the results:

*IT sampling: PW sampling:*

**

IT sampling has quite high p-values, PW sampling – not. PW sampling is based on the technique, where the continuous probability distribution is treated as discrete, that is why zero p-values can occur. Despite on this, it fits data very well as well as IT sampling.

# 3.3 Estimating of correlations

For choosing target and predictor variables the correlation analysis was performed. Only highly correlated variables were taken. The correlation table for chosen variables is represented below (values less than 0.5 are indicated as ‘NAN’):

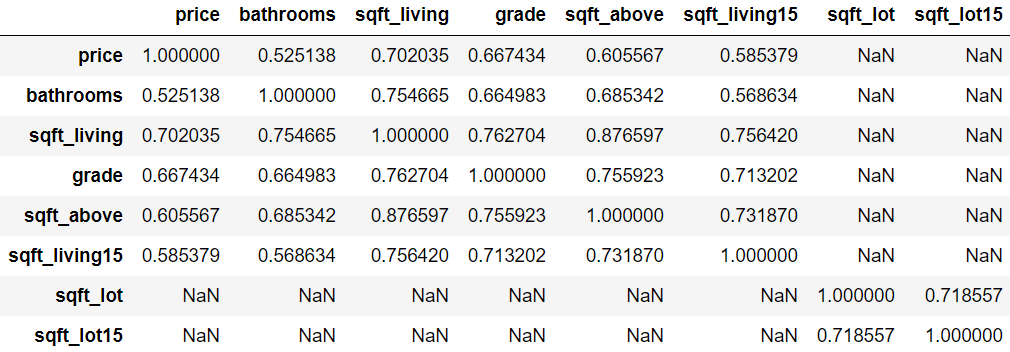


Table 3.1 - Correlation table of data before

For creating a Bayesian network via Python library, it is obligatory to have discrete variables. So, all continuous variables were transformed into discrete ones and then the correlation matrix was built. Correlations almost did not change, thus used transformation did not “spoil” original data.

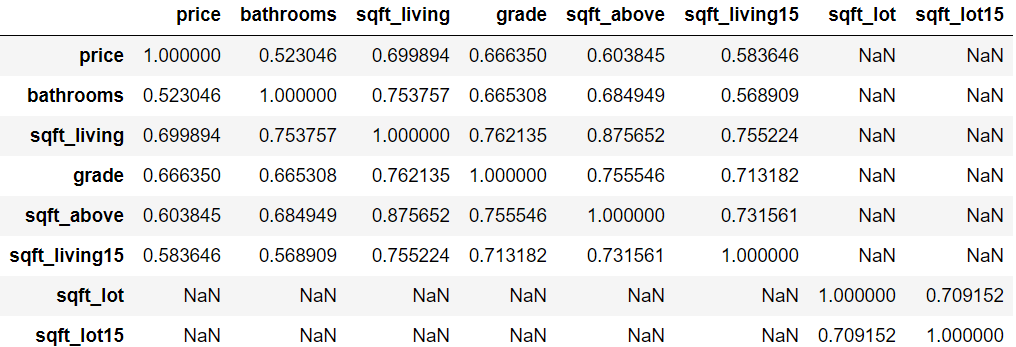
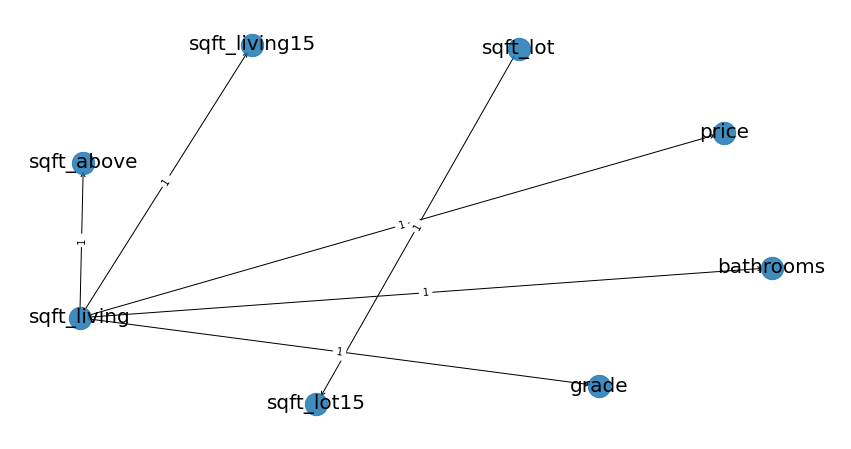


Table 3.2 - Correlation table of data after preprocessing

# 3.4 Constructing of Bayesian network manually

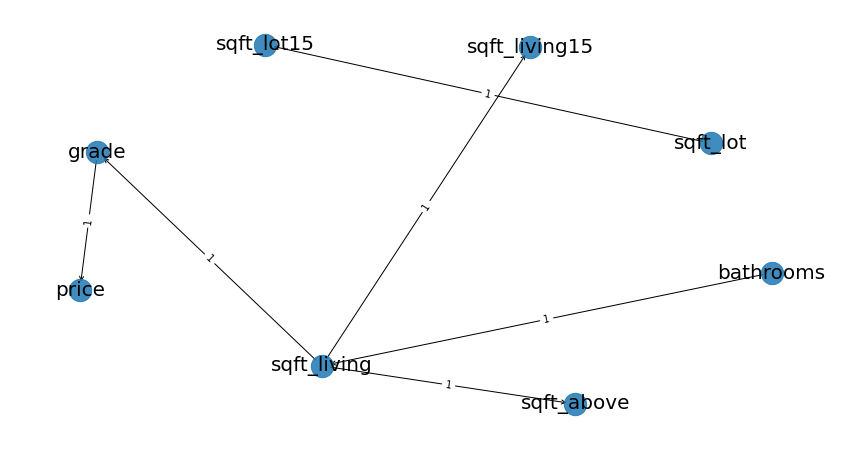
Variable “sqft\_living” has the strongest correlations with different variables. It is logical to suppose that variables “bathrooms”, “sqft\_living 15”, “sqft\_above” “grade” and “price” depend on “sqft\_living”. Moreover, we have “sqft\_lot” and “sqft\_lot15” correlated only with each other. Let “sqft\_lot15” be dependent on “sqft\_lot”.



Picture 3.2 - Handmade Bayesian network structure.

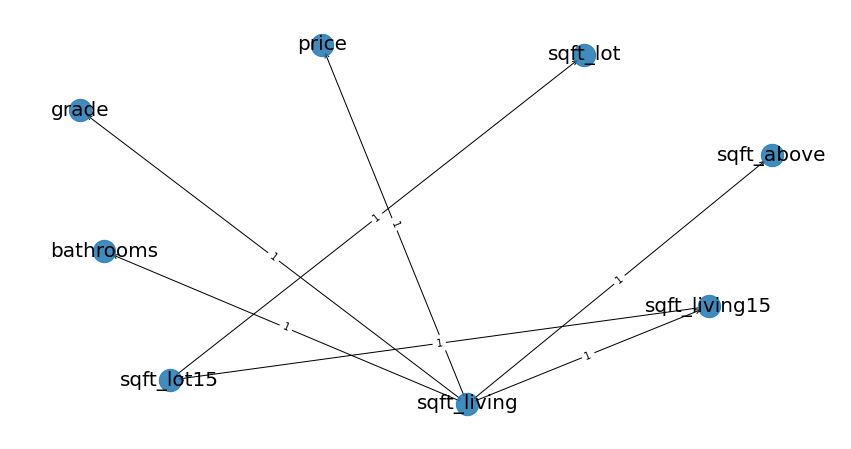
# 3.5 Constructing of Bayesian network algorithmically

Two Bayesian networks were constructed. The first one – by Hill climb search, the second – by Chow-Liu method. K2 score function was used for both algorithms.

Hill climb search:

Picture 3.3 - Hill climb search Bayesian network structure

Chow-Liu method:

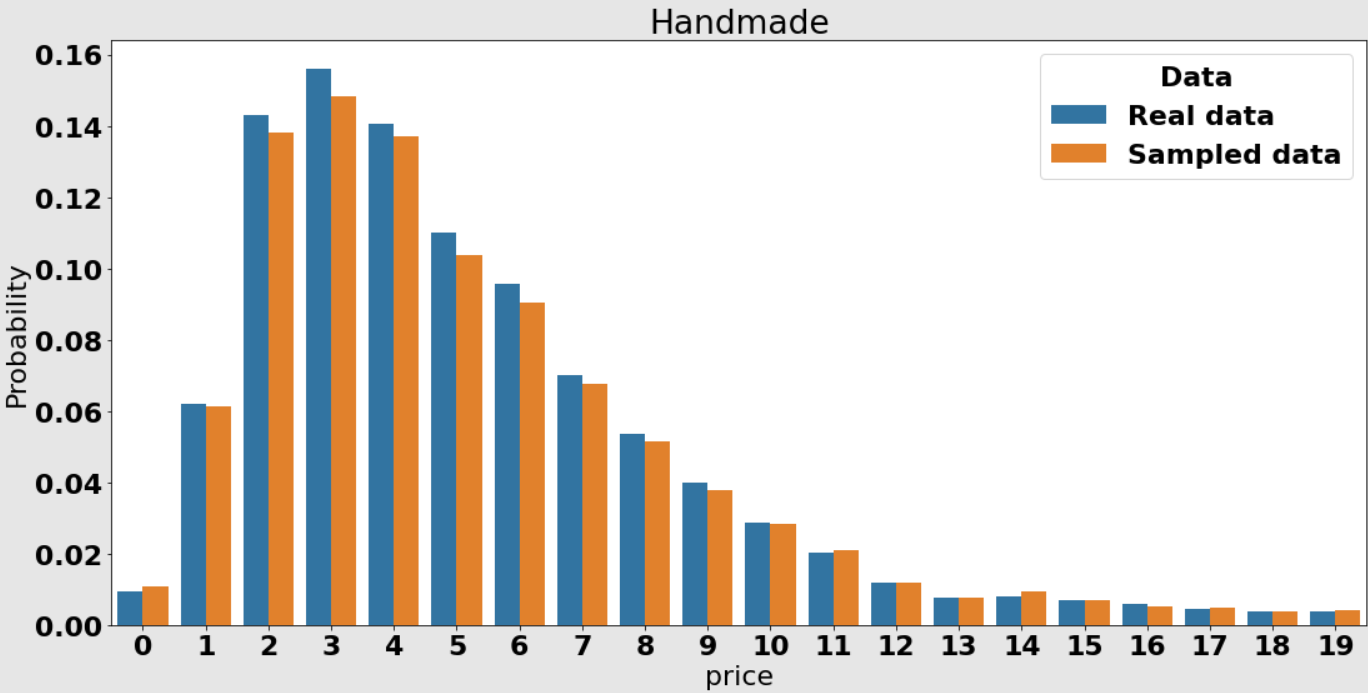


Picture 3.4 - Chow-Liu Bayesian network structure

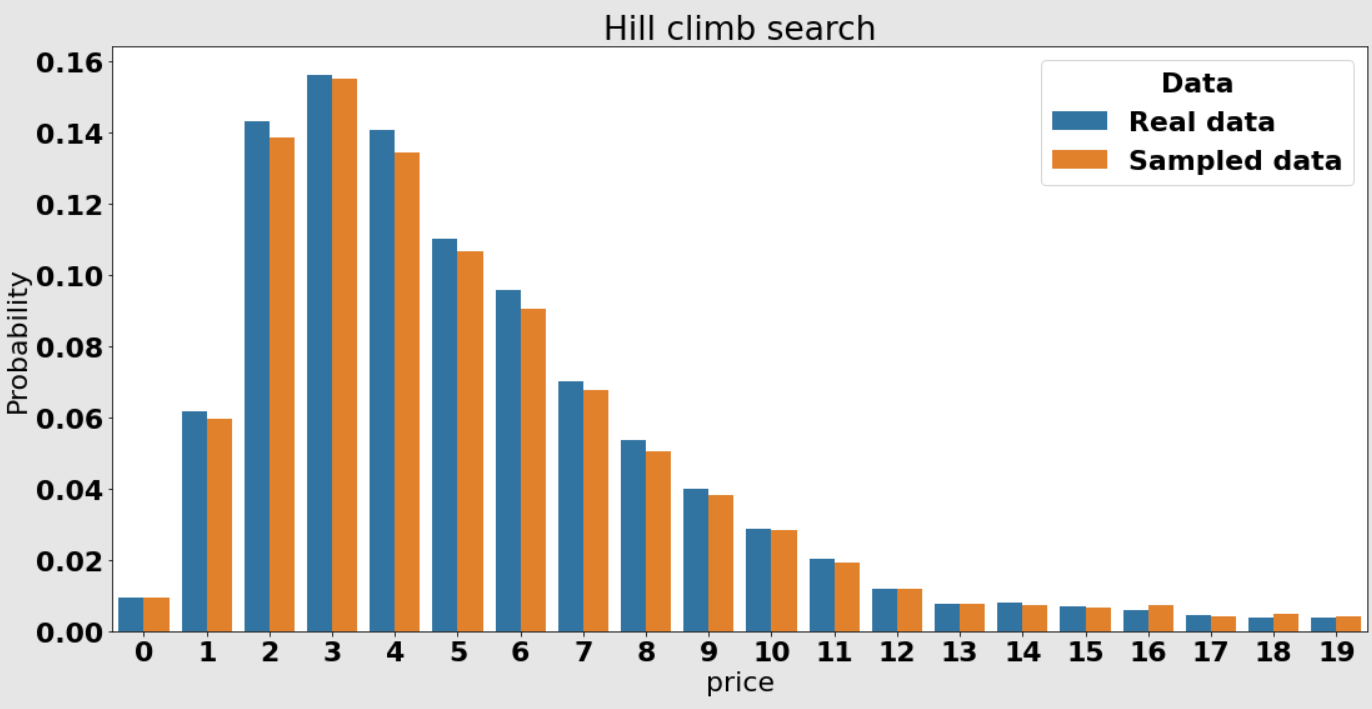
Algorithmically constructed models are very familiar with handmade model.

# 3.6. Analyzing a quality of constructed Bayesian networks

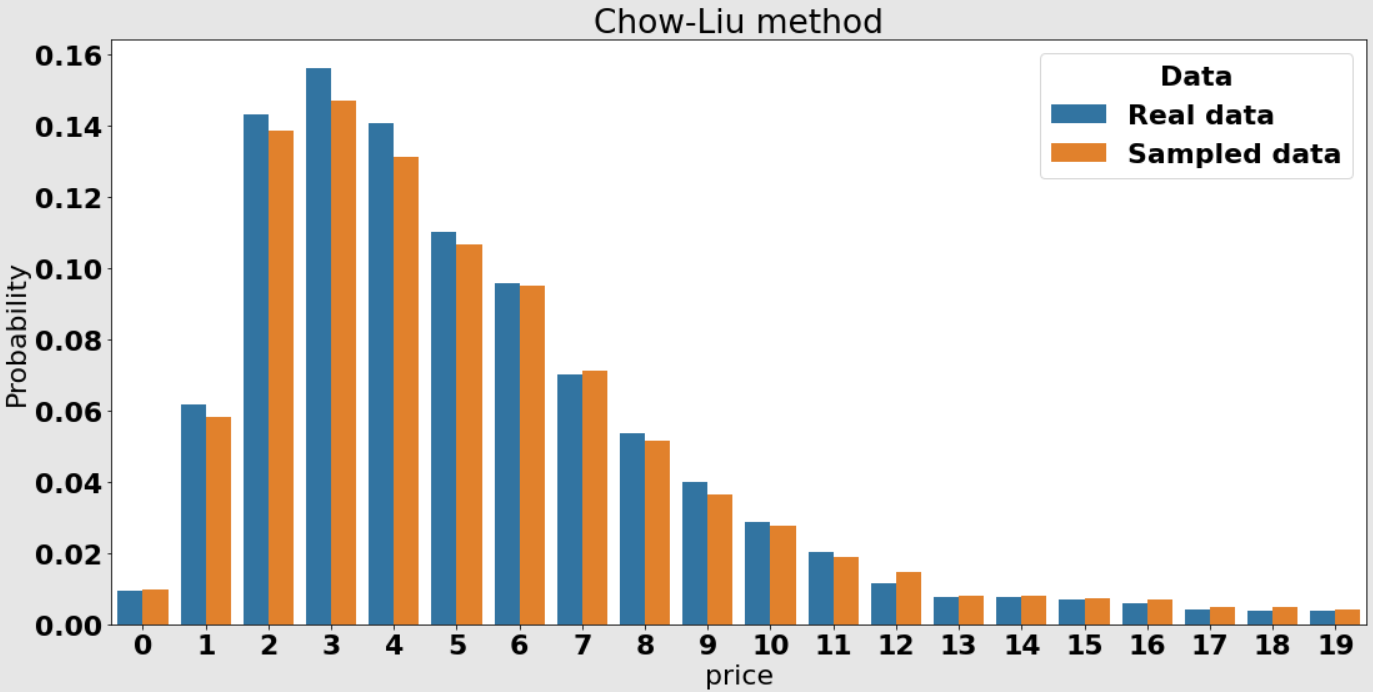
For variable “price” comparison diagrams were constructed. For each created Bayesian network was constructed a plot, where you can see probability density distribution diagrams for real and sampled data.



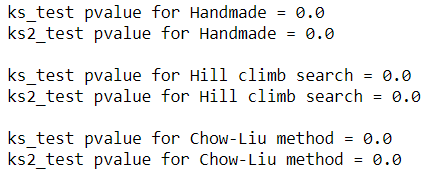
Picture 3.5 - Comparison of real data distribution with sampled via handmade Bayesian network data distribution



Picture 3.6 - Comparison of real data distribution with sampled via hill climb search Bayesian network data distribution.



Picture 3.7 - Comparison of real data distribution with sampled via hill climb search Bayesian network data distribution



P-values are zeros because of data transformation. On the graphs you can see that Bayesian networks sample data quite well. For other variables graphs look almost the same, so they are not included into the report.

# 3.7 Brief conclusions on 3 laboratory work

* The big number of chosen dataset variables are weakly correlated.
* A lot of variables depend on “sqft\_living” as expected.
* Algorithms for Bayesian networks constructing built networks that were very familiar with a handmade one.
* For univariate and multivariate cases sampling models were built. They can sample data very well and can be used for prediction, gap filling and synthetic generation.

# 4. STATIONARITY OF THE PROCESSES

# 4.1 Substantiation of chosen sampling

For this practicing task as target variables have been chosen price of house sold and square footage of living space. Predictors for the first one are grade and square footage above ground level and for the second one are number of bedrooms, number of bathrooms and square footage above ground level as well. This choice could be substantiating by the correlation matrix which is shown strong dependencies between the selected variables rather with the others.

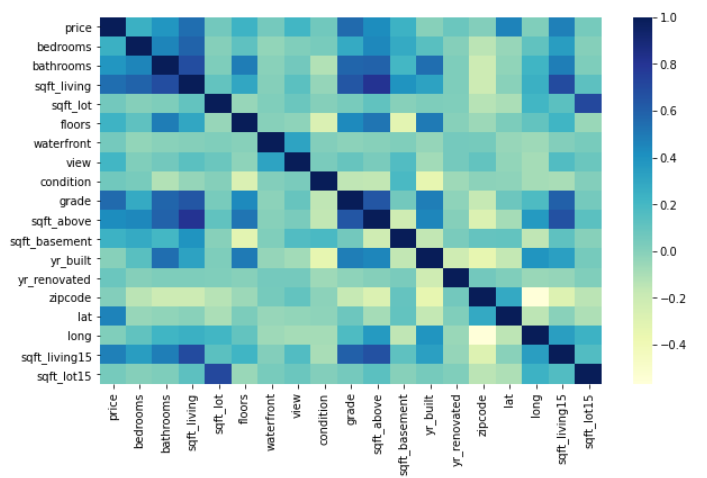


Figure 4.1 – Correlation matrix

# 4.2 Stationary analysis

The first step to process chosen targets as a time series is a stationary analysis which means constant values of such statistical properties as mean and variance according to the time shifting. Stationarity of the processes allow make forecasts for time series much easier and could be estimated not only by visualization of the process but also with the statistical tests. To check whether the timeseries are stationary Dickey-Fuller test has been used. The null hypothesis that declares non-stationarity of the series can be rejected if there is no unit root is presented in an autoregressive model of a given time series.

Firstly, all variables have been grouped by date and aggregated, e.g. summarized, and after that visualized on figure 2 below.

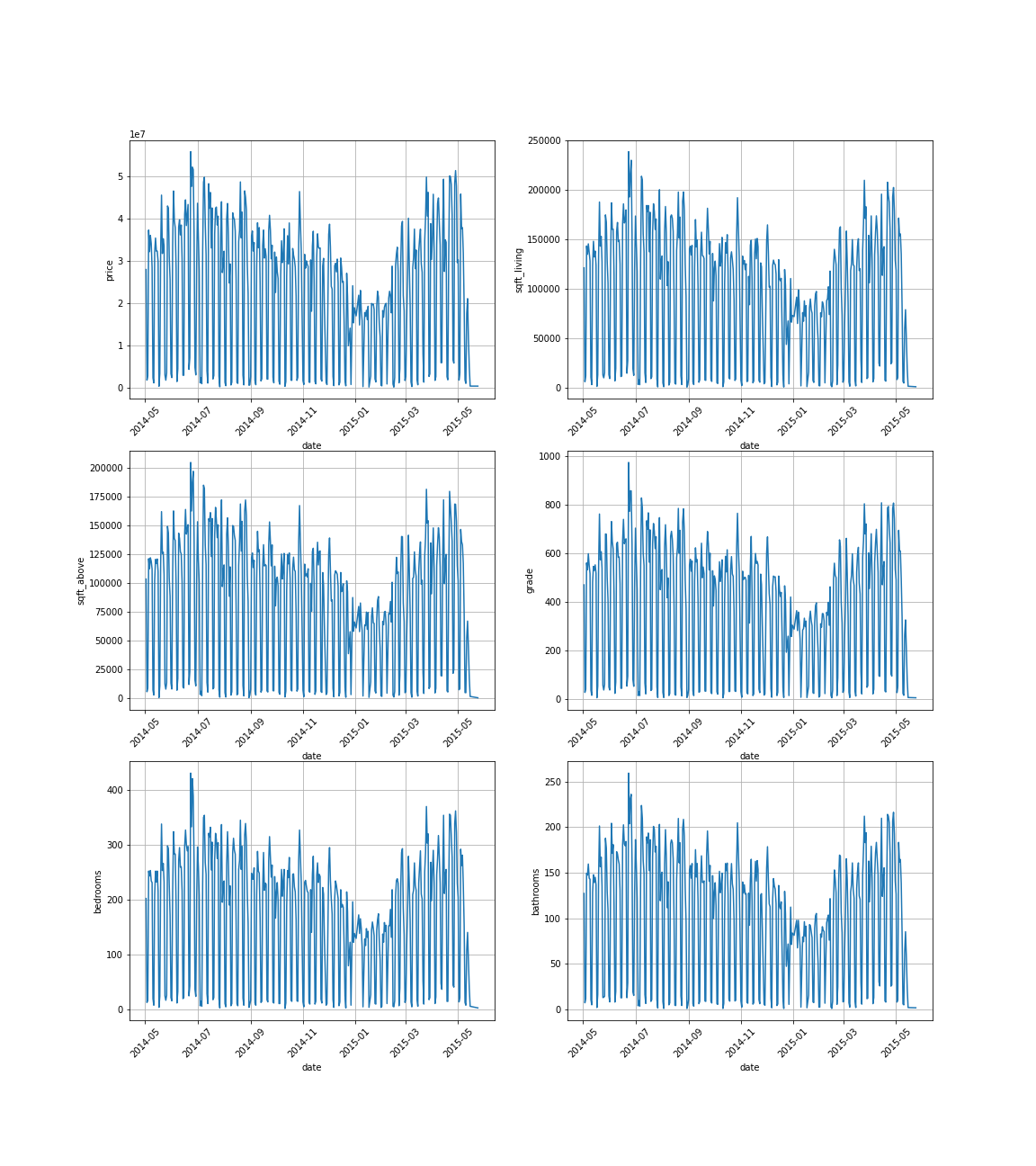


Figure 4.2 – Graphs of the targets and predictors

In our case plotting the data does not give clear understanding of the stationarity of the processes thus Augmented Dickey-Fuller test was provided. The results for each variable can be seen below in the figure 3.

Изображение выглядит как текст

Автоматически созданное описание

Figure 4.3 – Results of ADF tests for subsample

In order to analyze the structure of the data, it was decomposed into the following components: trend, seasonality and differences.

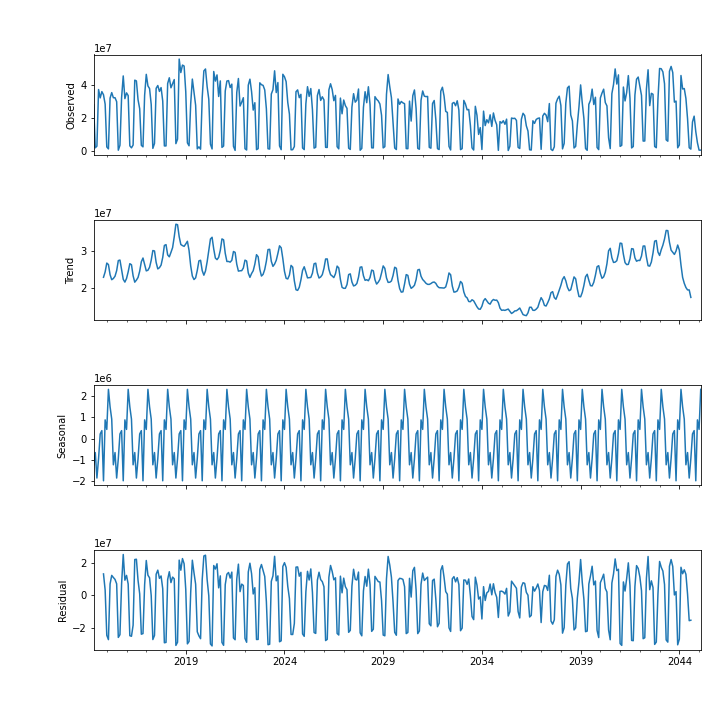


Figure 4.4 – Decomposed “price” timeseries

Such visual decomposition can help in further model selection and its tunning. One of the approaches to reach stationarity of the process is substruction of differences. This step is also meaningful for the setting one of the parameters of the model which is indicating the integration order of the process. After the first substruction the ADF tests give better results thus trend substruction have been applied and provided total stationarity according the test results.

Изображение выглядит как текст

Автоматически созданное описание

Figure 4.5 – Results of ADF tests

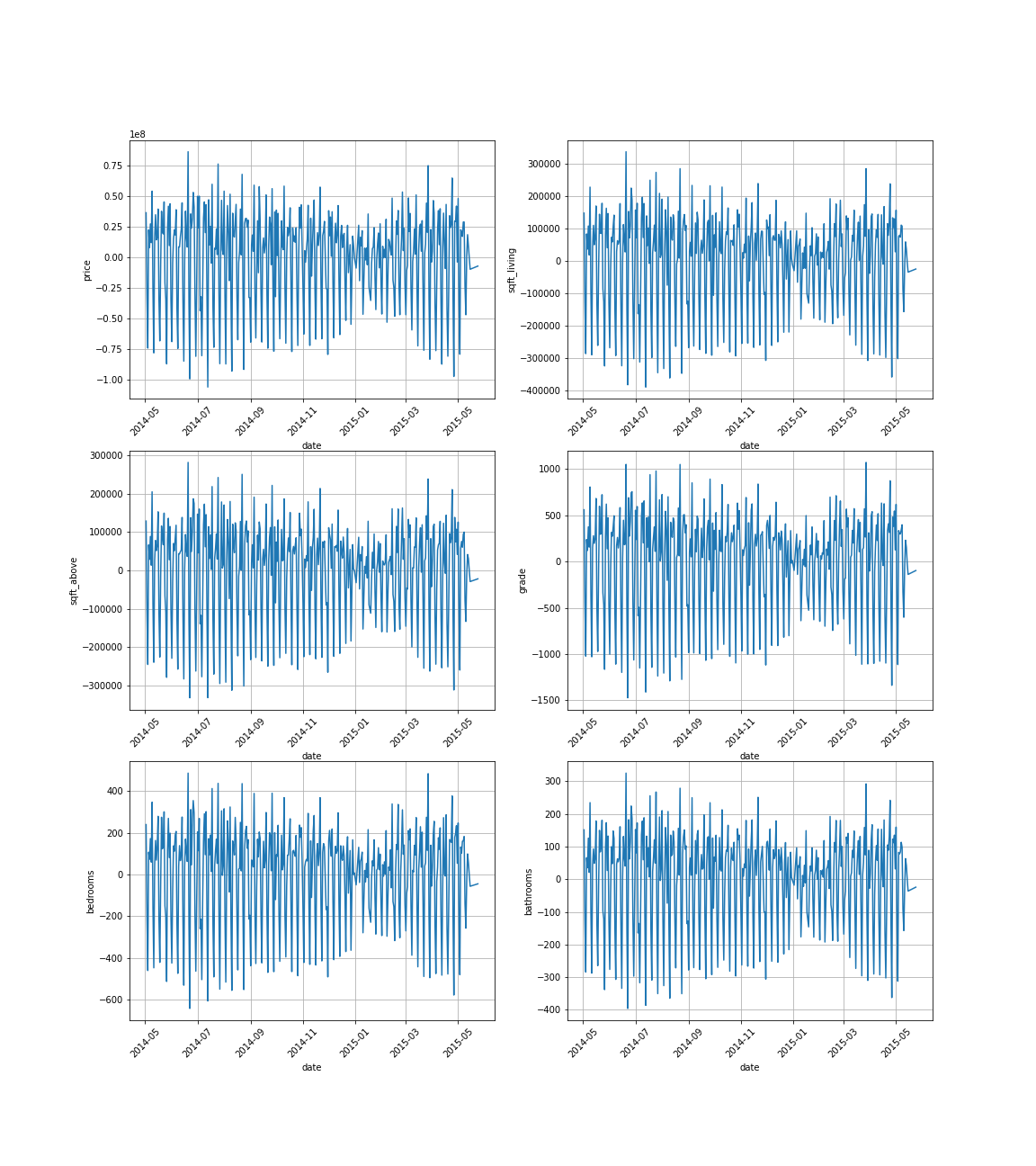


Figure 4.6 – Stationary processes

# 4.3 Covariance or correlation function analysis.

The next step after obtaining the stationary series is the analysis of correlation functions and partial correlation function. Here we can estimate the dependency of the variable with its own values in time shifted interval. This analysis is also connected with the initial assumptions about model parameters. To be more spesific, estimation of the autocorrelation function(AFC) of target variables has been done and for both the most significant value of the ACF and PACF appears on the 7th lag. The ACFs and PACFs are depicted in figure 7.

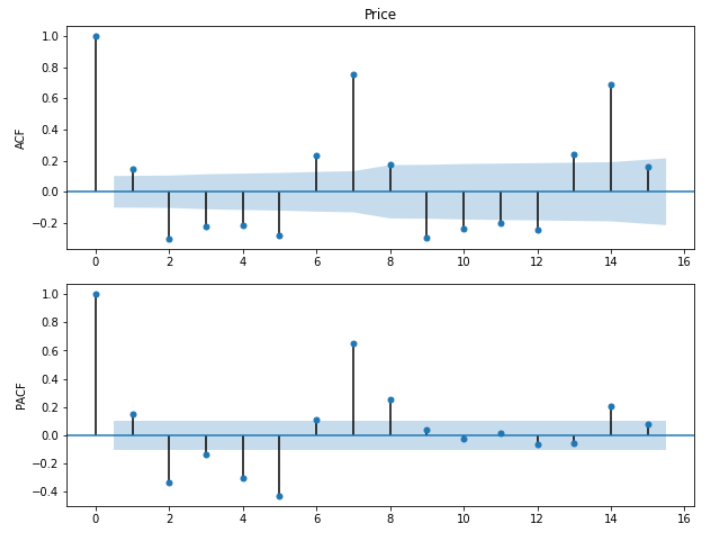
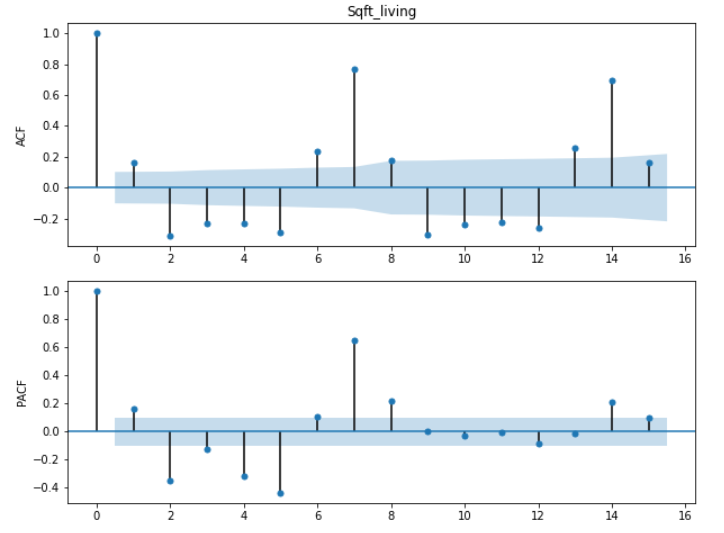
 

Figure 4.7 – ACF and PACF for “price” (left) and “sqft\_living” (right)

Dependencies among targets and chosen predictors for them have been plotted as well and illustrated in figure 8 and 9.

Изображение выглядит как текст, щуп

Автоматически созданное описание

Figure 4.8 – Mutual correlation between target “sqft\_living” and its predictors

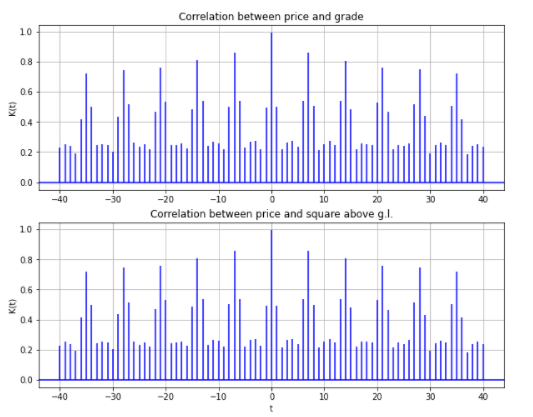


Figure 4.9 – Mutual correlation between target “price” and its predictors

# 4.4 Noise filtration and estimation of spectral density function

As all real data our dataset contains noise that could prevent the model from the well performance. Therefore, the next stage in the constructing pipeline is data filtering. There are a lot of techniques directed on removing of the meaningless data with different appearing frequency. In terms of task applied Gaussian filter and Butterworth low-pass filter to exclude the noise with high frequency. Decision about cutoff frequency should be made after evaluation of periodogram of the data. Periodograms of the targets could be seen in figure 10.

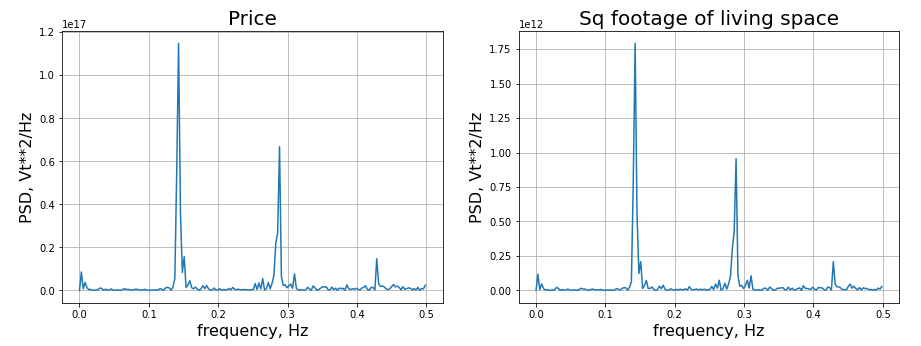


Figure 4.10 – Periodograms of initial data

Observations of the targets in frequency-domain let us set the cutoff frequency about 100 Hz for Butterworth filter. Gaussian filter has been chosen with the window size equals to 10 points as for standard deviation. Initial and filtered data are shown further in figure 11.

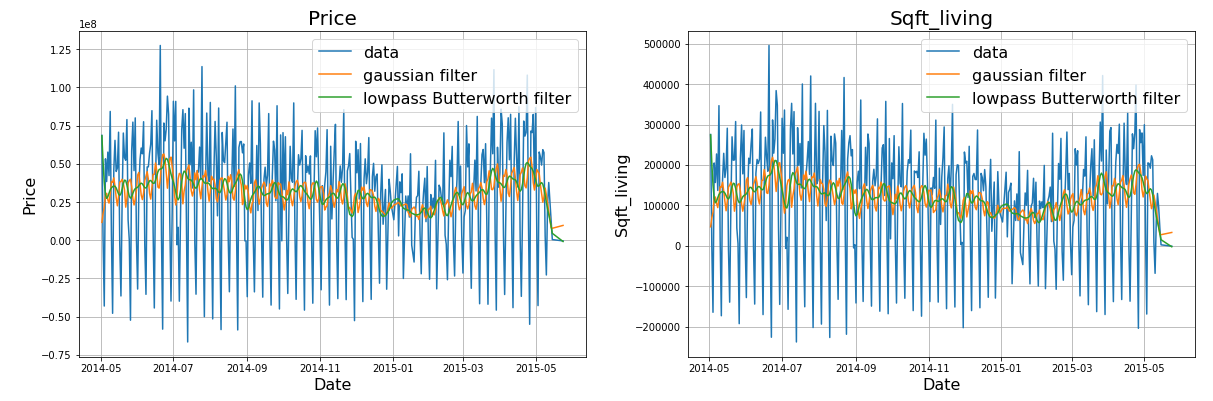
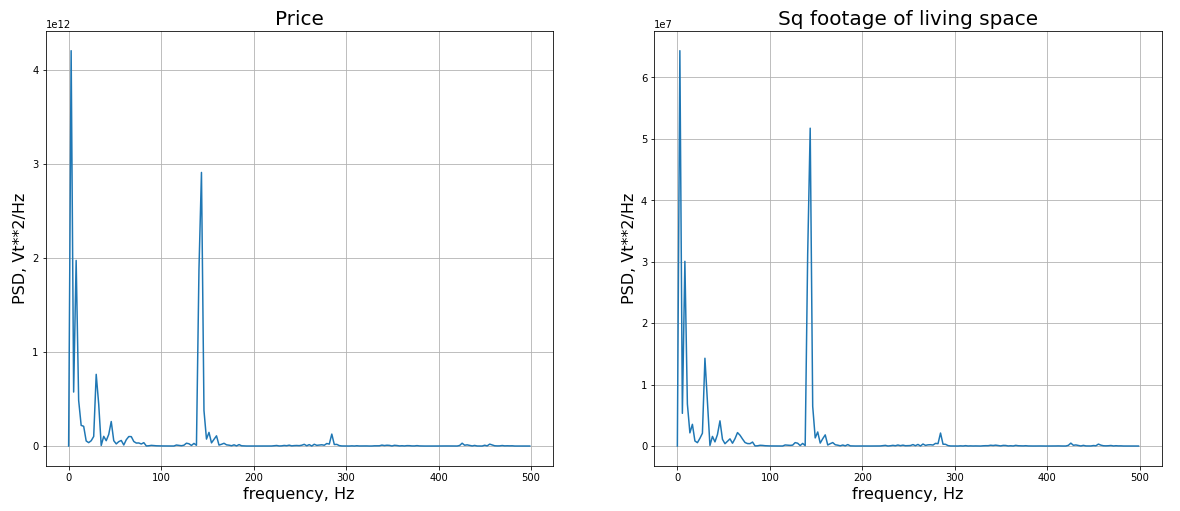


Figure 4.11 – Stationary processes before and after filtering

To be sure about obtained results, periodograms should be observed again. In our case noise was removed well with both filters.

(a)

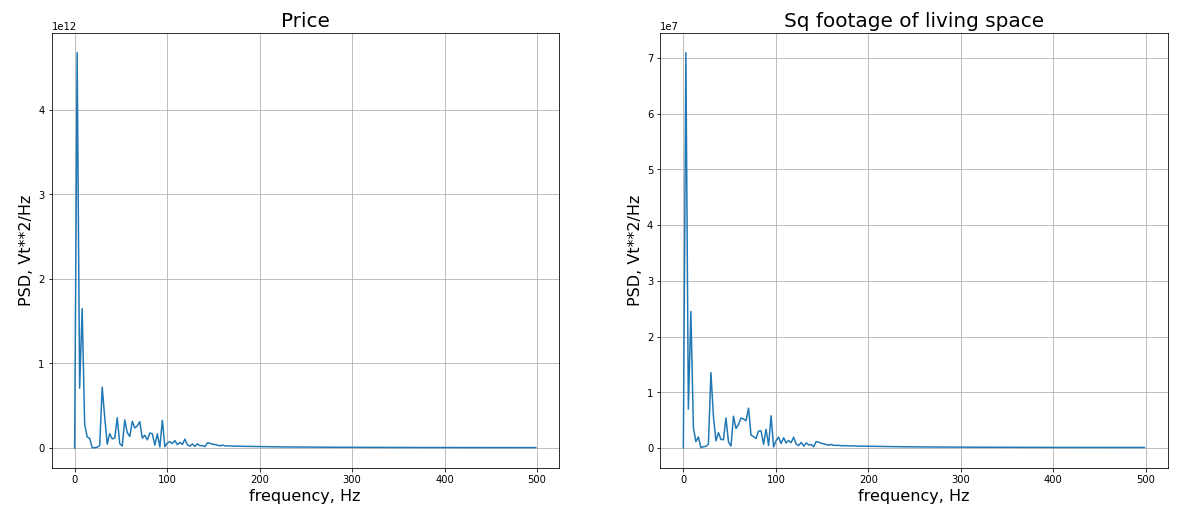
(b)

Figure 4.12 – Periodograms of filtered data with (a) Gaussian filter and (b) Butterworth filter

# 4.5 Auto-regression model

The analysis of timeseries includes forecasts for a future behavior of the observed feature. There is a big group of classical time series prediction methods. For our experiments SARIMA has been chosen to consider seasonality of the data. Parameters for AR which is p and MA which is q and d as integration order of the process were estimated on the previous steps. *p* and *q* have been chosen by plotting PACF and ACF respectively. Another four parameters (*P, D, Q, s*) of the model are the order of the seasonal component for the AR parameters, differences, MA parameters and periodicity. D is still one since we differentiated the data once. Seasonality *s* is given 12 because of the monthly data.

The optimal order of the model was selected by comparison of Akaike information criterion (AIC) through the list of the possible combinations of the parameters. At first, model was fitted with initial data with the following parameters:



In figure 13 - 14 there are plotted initial data of the “price” target in blue and values predicted by the model.

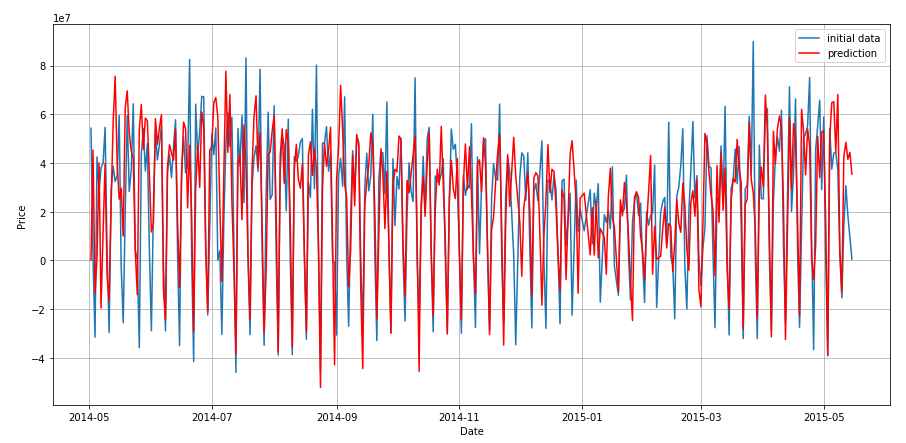


Figure 4.13 – Predicted values of the “price” variable

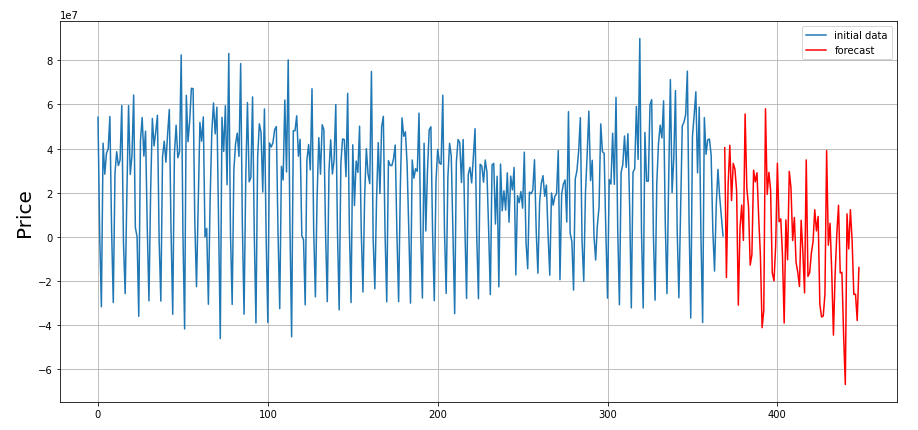


Figure 4.14 – Forecast for the “price” variable

An “adequality” of the model can be checked via residual analysis. The “residuals” in a time series model are what is left over after fitting a model. A good forecasting method will yield residuals with the following properties:

* The residuals are uncorrelated. If there are correlations between residuals, then there is information left in the residuals which should be used in computing forecasts.
* The residuals have zero mean. If the residuals have a mean other than zero, then the forecasts are biased.

According to this properties, estimation of residuals shows their normality by absence of the autocorrelation and normal distribution.

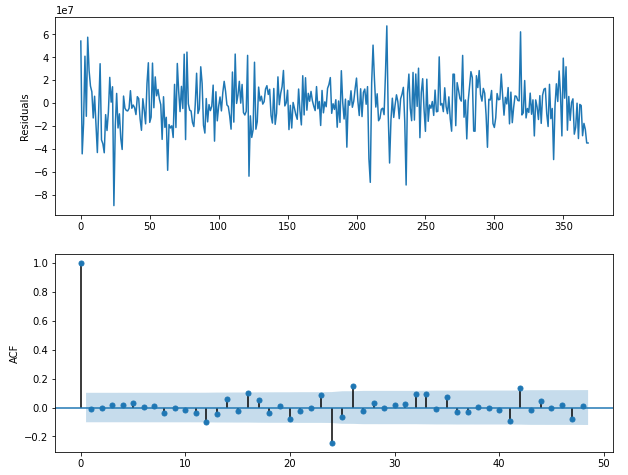


Figure 4.15 – ACF of residuals for “price” prediction

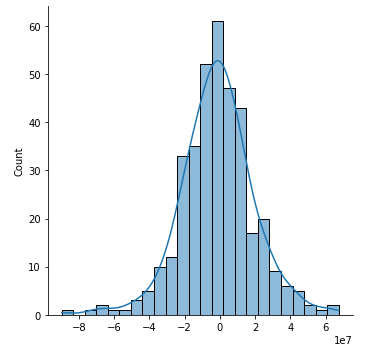
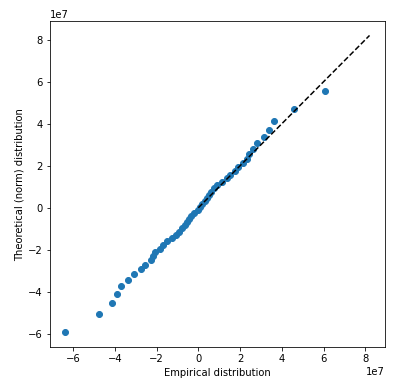
 

Figure 4.16 –Distribution and QQ biplot of residuals for “price” prediction

Metrics of the model are following:

Изображение выглядит как текст

Автоматически созданное описание

Performance of the model is not bad but in order to increase R^2 it has been fitted on filtered data. The result is shown in figure 17.

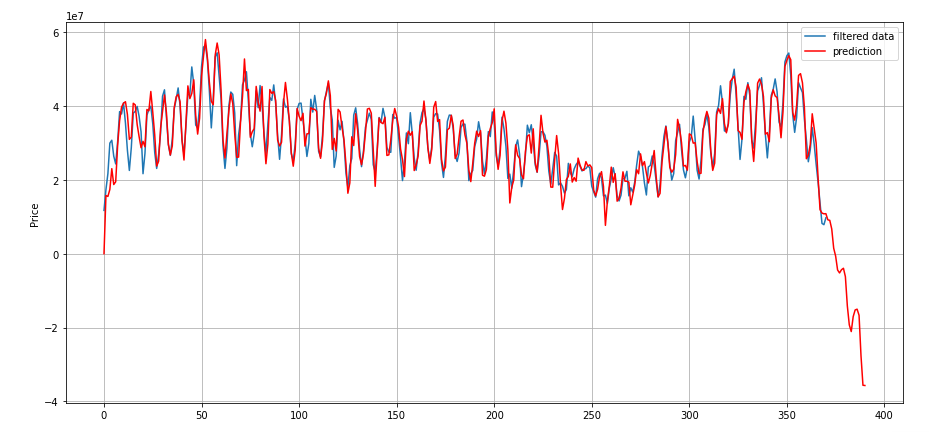


Figure 4.17 – Forecast for the filtered “price” variable

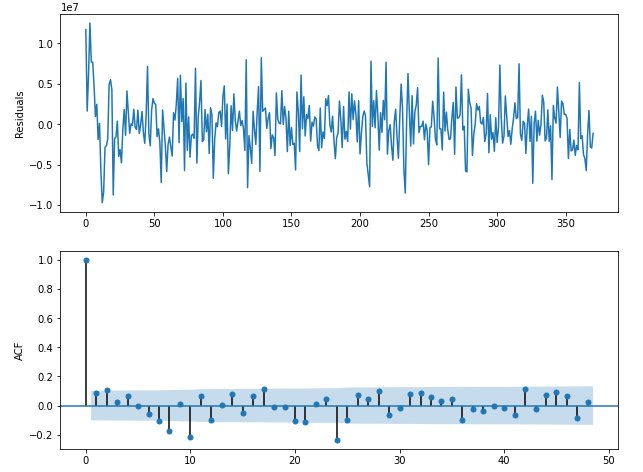


Figure 4.18 – ACF of residuals for “price” prediction

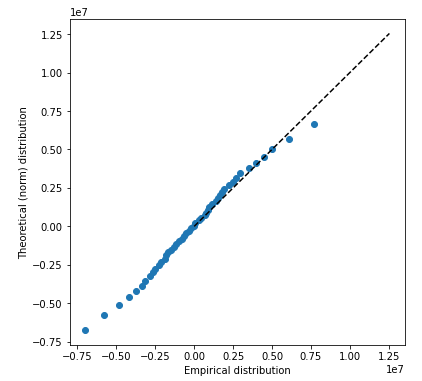
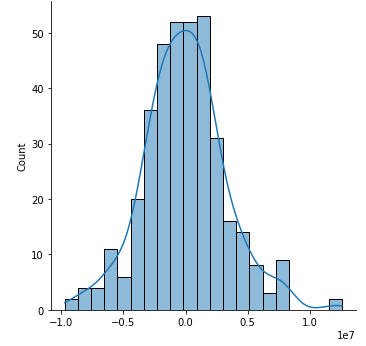


Figure 4.19 –Distribution and QQ biplot of residuals for “price” prediction

R^2 score in this case is much higher and is equal to 0.875.

Further steps are the same for “sqft\_living” target prediction and analysis. All results are depicted in figures 20 -22. The model parameters for this target are chosen:



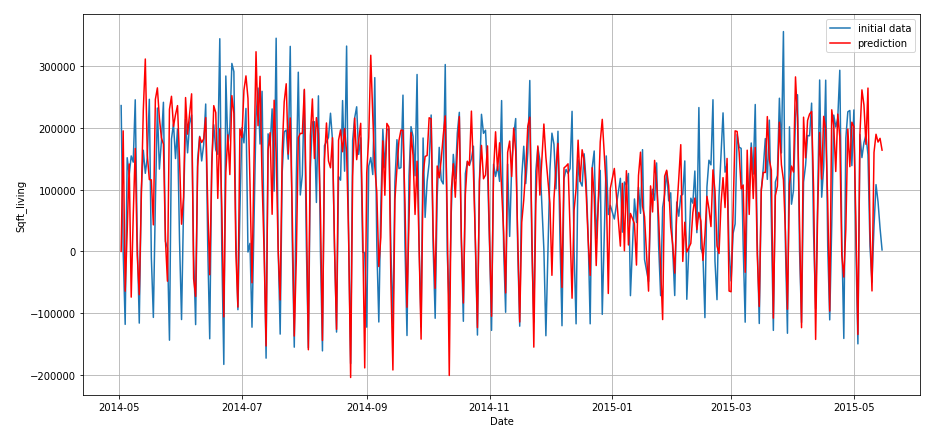




Figure 4.20 – Forecast for the “sqft\_living” variable

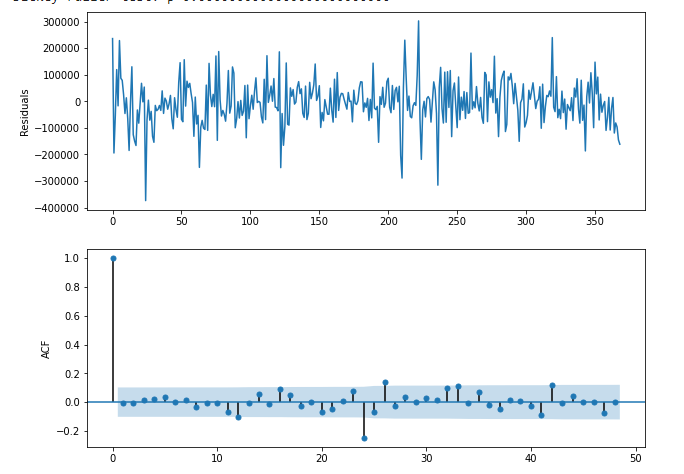


Figure 4.21 – ACF of residuals for “sqft\_living” prediction

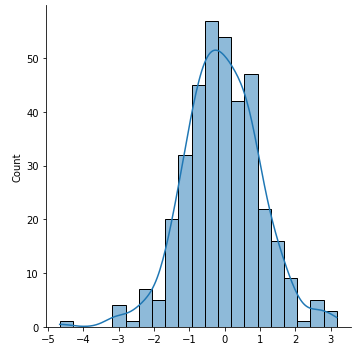
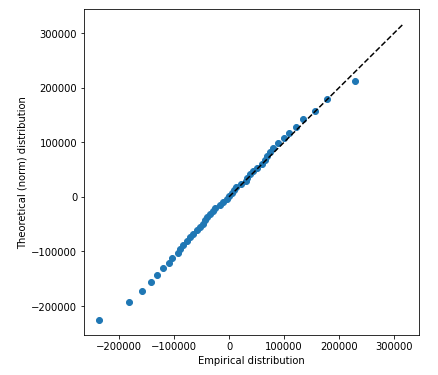
 

Figure 4.22 –Distribution and QQ biplot of residuals for “sqft\_living” prediction

Metrics of the model performance are following:

Изображение выглядит как текст

Автоматически созданное описание

Sqft\_living as output of gaussian filter was fitted to the model as well. The results are described in figures 23 – 25.

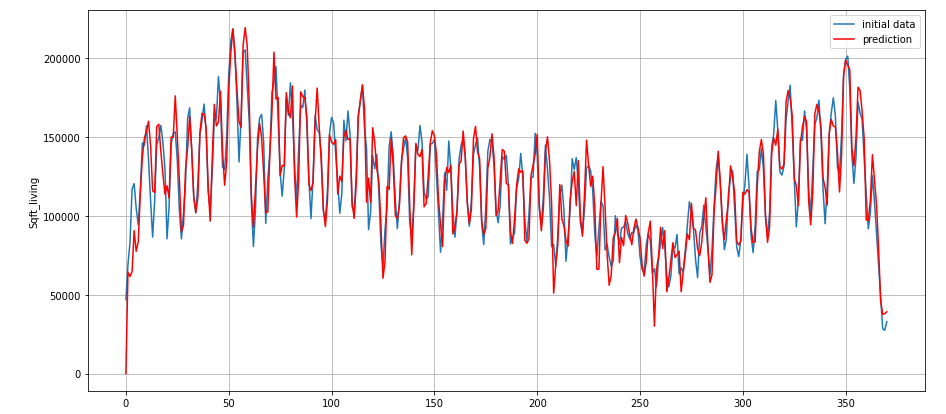


Figure 4.23 – Forecast for the filtered “sqft\_living” variable

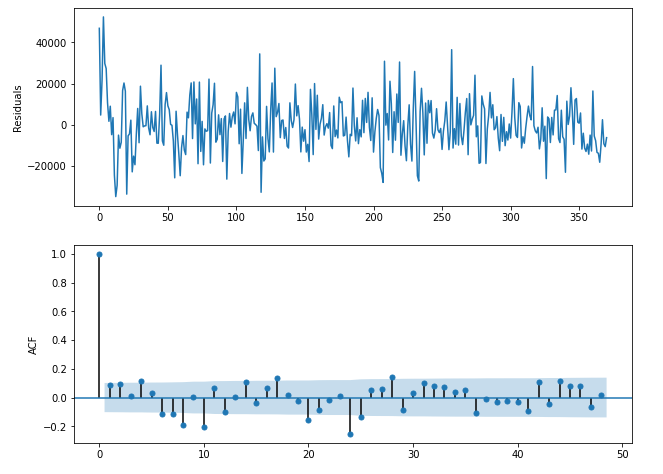


Figure 4.24 – ACF of residuals for filtered “sqft\_living” prediction

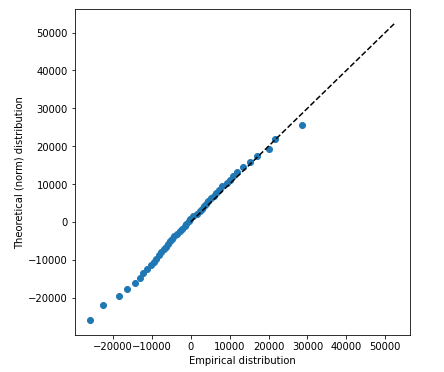
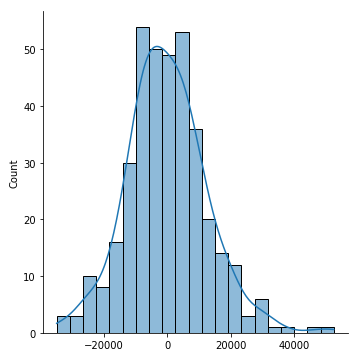


Figure 4.25 –Distribution and QQ biplot of residuals for filtered “sqft\_living” prediction

Metrics are following:

Изображение выглядит как текст

Автоматически созданное описание

# 4.6 Brief conclusions on 4 laboratory work

* Subsample of the dataset have been chosen.
* Stationarity of targets and predictors have been examined and provided for all variables relying on the ADF tests.
* Target variables have been filtered from high frequencies.
* SARIMA model has been constructed and performed on initial and filtered targets.
* Quality and “adequality” of the model have been estimated.

# CONCLUSION

Dataset contains house sale prices for King County. It includes homes sold between May 2014 and May 2015. The shape of data is (17571,21). There are a lot of variables including categorial and continuous variables.

The reasons why this dataset was chosen:

* there are continuous, categorical variables (predictors and factors);
* there is a time series;
* data is suitable for training regression and autorepression models.

For 1 laboratory work the following variables were selected: price (rice of each home sold), sqft\_living (square footage of the apartments interior living space), sqft\_above (the square footage of the interior housing space that is above ground level), sqft\_living15 (the square footage of interior housing living space for the nearest 15 neighbors).

To estimate the PDF of the variables histograms were constructed and kernel density estimating was used. The distribution of the first variable (price) is similar to the gamma distribution, and the normal distribution can also be examined. The second variable (sqft\_living) is characterized normal distribution. The gamma distribution is suitable for variables sqft\_above and sqft\_living15.

Ordinal statistics and plot “box with whiskers” are presented for each variable. This analysis helps better select the appropriate distribution. Gamma distribution describes the variable price, sqft\_above and sqft\_living15. Both distributions can be applied to variable sqft\_living.

We estimated parameters for the distributions of variables. For this we used 2 methods: maximum likelihood technique and LS method. The results of the evaluations are presented.

Finally, we plotted QQ-plots and used statistical tests to determine whether our assumptions about the distribution of the variables were correct.The first variable (price) passed both tests (gamma and normal distribution). However, it should be noted that the p-value for the gamma distribution is larger than for the normal distribution. As already mentioned, the second variable (sqft\_living) can be described by both the gamma distribution and the normal distribution. Passed both tests. The next variable passed both tests, similarly, the p-value for the gamma distribution is greater than for the normal one. We can see the same case with the last variable.

Quality of regression is not very high according to not very big amount of R-squared. Point cloud shows the data spread is very large, so probably, it`s reasonable to try build not leaner regression but another statistical and AI models.

The big number of chosen dataset variables are weakly correlated. A lot of variables depend on “sqft\_living” as expected. Algorithms for Bayesian networks constructing built networks that were very familiar with a handmade one. For univariate and multivariate cases sampling models were built. They can sample data very well and can be used for prediction, gap filling and synthetic generation.

The analysis of stationarity of the processes has been made on the subsample where the targets variables are price and sqft\_living and the predictors are sqft\_above, grade, number of bedrooms and number of bathrooms.

Initial Augmented Dickey-Fuller (ADF) test on stationarity has not rejected the hypothesis about dependency of some variables from time and, therefore, subtraction of differences was provided. After that, the results of repeated ADF showed less non-stationary behavior and analysis of correlation functions was followed.

On this stage we observed autocorrelations of target variables and mutual correlation between them and chosen predictors. Autocorrelation functions (ACF) as well as partial autocorrelation functions (PACF) let us make some assumptions about the coefficients of autoregression (AR), differentiating and moving average (MA) parts of the model.

The following step in our pipeline was to filter high frequencies of the data where Gaussian and Butterworth filters have been used. The estimation of initial processes in frequency domain was done by plotting periodograms for both targets which give us the knowledge about cutoff frequency for low-pass Butterworth filter and evaluate the quality of filtering techniques. The Gaussian filter gives less smoother output comparing to the Butterworth.

The construction of Seasonal Autoregressive Integrated Moving Average (SARIMA) was based on analysis of ACF and PACF for targets. The optimal parameters through the set of probable coefficients for the model have been chosen according to the Akaike’s information criterion, e.g. the set with the least criterion was examined as the best solution. The quality and “adequality” of derived model were estimated by metrics and analysis of residuals, respectively.