Report on learning practice # 2

Analysis of multivariate random variables

Performed by:

Aleksandr Shirokov

J4133c

Saint-Petersburg

2021

TABLE OF CONTENTS

[LAB 2 3](#_Toc88684825)

[1. Plotting a non-parametric estimation of PDF in form of a histogram and kernel density function for MRV (or probability law in case of discrete MRV). 3](#_Toc88684826)

[2. Estimation of multivariate mathematical expectation and variance. 3](#_Toc88684827)

[3. Non-parametric estimation of conditional distributions, mathematical expectations, and variances. 3](#_Toc88684828)

[4. Estimation of pair correlation coefficients, confidence intervals for them and significance levels. 4](#_Toc88684829)

[5. Task formulation for regression. 4](#_Toc88684830)

[6. Regression model, multicollinearity, and regularization (if needed). 5](#_Toc88684831)

[7. Quality analysis. 5](#_Toc88684832)

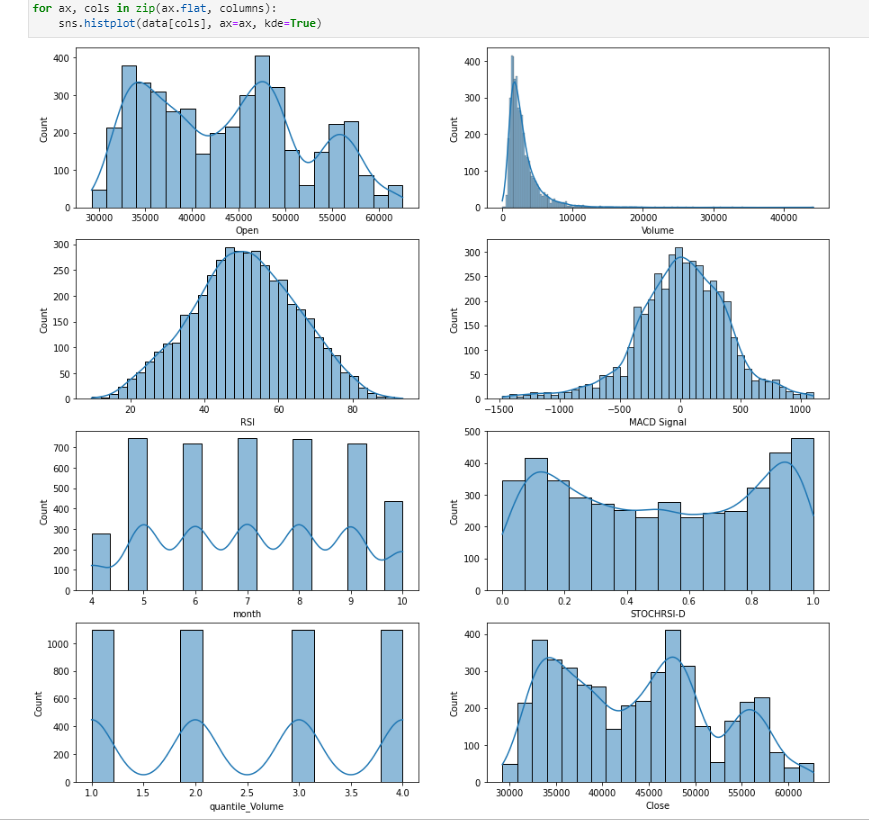
[SOURCE CODE 5](#_Toc88684833)

**LAB 2**

**1. Plotting a non-parametric estimation of PDF in form of a histogram and k****ernel density function for MRV (or probability law in case of discrete MRV).**

Visualization of *non-parametric estimation* of *PDF* can be very useful for data analysis. *Kernel density* function is a data visualization technique in which the histogram becomes smoother.

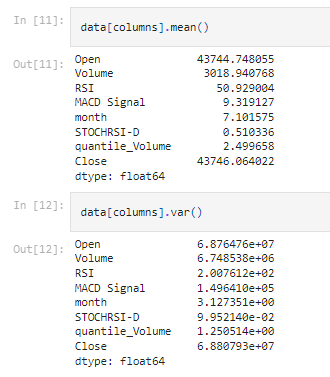
**Seaborn** library was used for visualization in this and subsequent steps.



Picture 1 – *Non-parametric estimation of PDF and Kernel Density*

**2. Estimation of** **multivariate mathematical expectation and variance.**

This sub-task was accomplished with the **pandas** DataFrame functions: and .

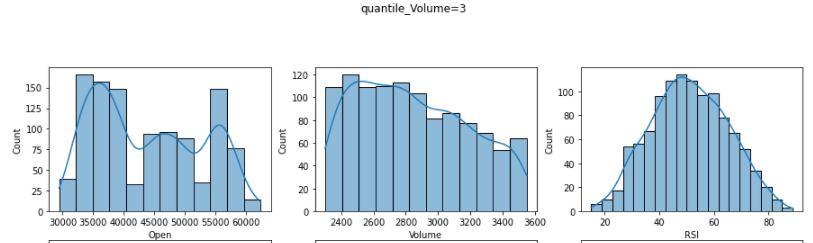


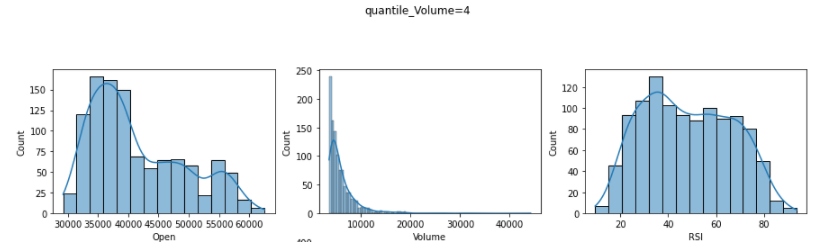
Picture 2 – *Second task*

**3. Non-parametric estimation of conditional distributions, mathematical expectations, and variances.**

The condition determined the value of the categorical variable

For each value, step has been reproduced and information for every conditional distribution about *mean* and *var* has been counted.

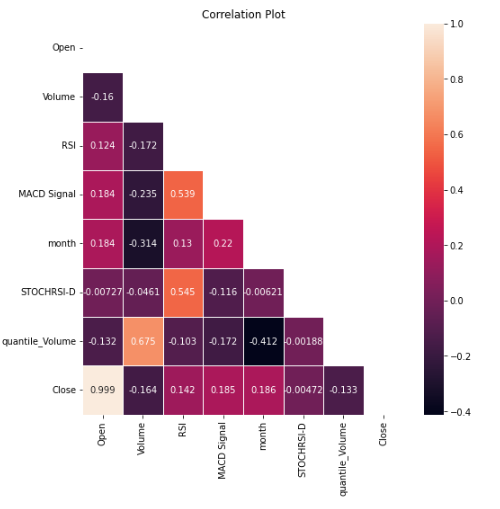




Picture 3 – Distributions for each parameter of (more in notebook)

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

To find the numerical characteristics of the confidence intervals, the stats module of the **SciPy** package was used.



Picture 4 – *Correlation Plot*

As we can see on the *correlation Plot*, column has a very high correlation with target value. Other columns have low correlations between each other. For each pair of feature correlation interval has been counted.

Изображение выглядит как текст

Автоматически созданное описание

Picture 5 – *First 3 Correlation coefficients and confidence intervals*

**5. Task formulation for regression.**

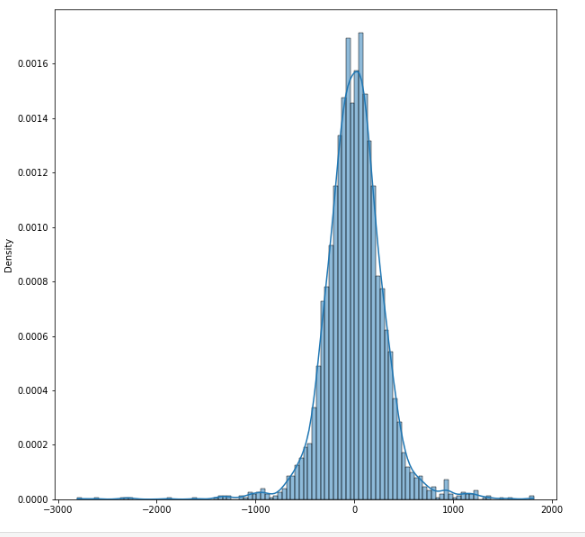
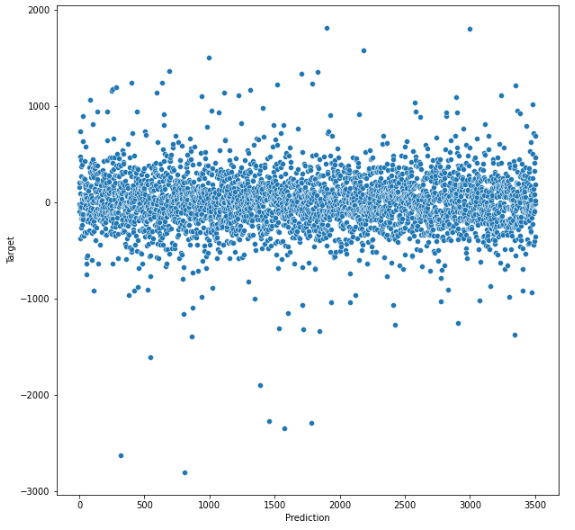
Predict target column by columns .

**6. Regression model,** **multicollinearity, and regularization (if needed).**

Regression models evaluated with -fold validation with python package **Scikit-Learn**. , was calculated for each validation fold.

Next, Lasso regularization was applied, with alpha equals to (-search, and *Ridge* regularization was also done).

**7. Quality analysis.**



Picture 6 – Residuals of prediction

As you can see in the graph, the model errors have a good normal distribution around , with little variance.

In the future, to improve the quality of the model, one can analyze the instances on which the model makes the most mistakes, and, for example, remove it from the training set.

Изображение выглядит как текст

Автоматически созданное описание

Picture 7 – Variance Inflation Factor

Also, using *variance\_inflation\_factor* from **StatsModels** package the variance inflation factor has been counted as we can see on *picture 7.* We can see that with *Open*, *RSI*, *month,* and *quantile\_Volume* multicollinearity is high.

**SOURCE CODE**

[aptmess/MMA: Methods and models for multivariate data analysis. Second lab (github.com)](https://github.com/aptmess/MMA)