FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION OF HIGHER EDUCATION

ITMO UNIVERSITY

Report

on learning practice № 2

Analysis of multivariate random variables

Performed by
Aleksandr Shirokov
Team 26, J4133c

St. Petersburg
2021

Contents

1 Lab 2

1.1 Dataset Info

For labs we have used a dataset which contains information of bitcion price trend indicators. Before doing task, I will write a small tutorial about how to download and use this dataset:

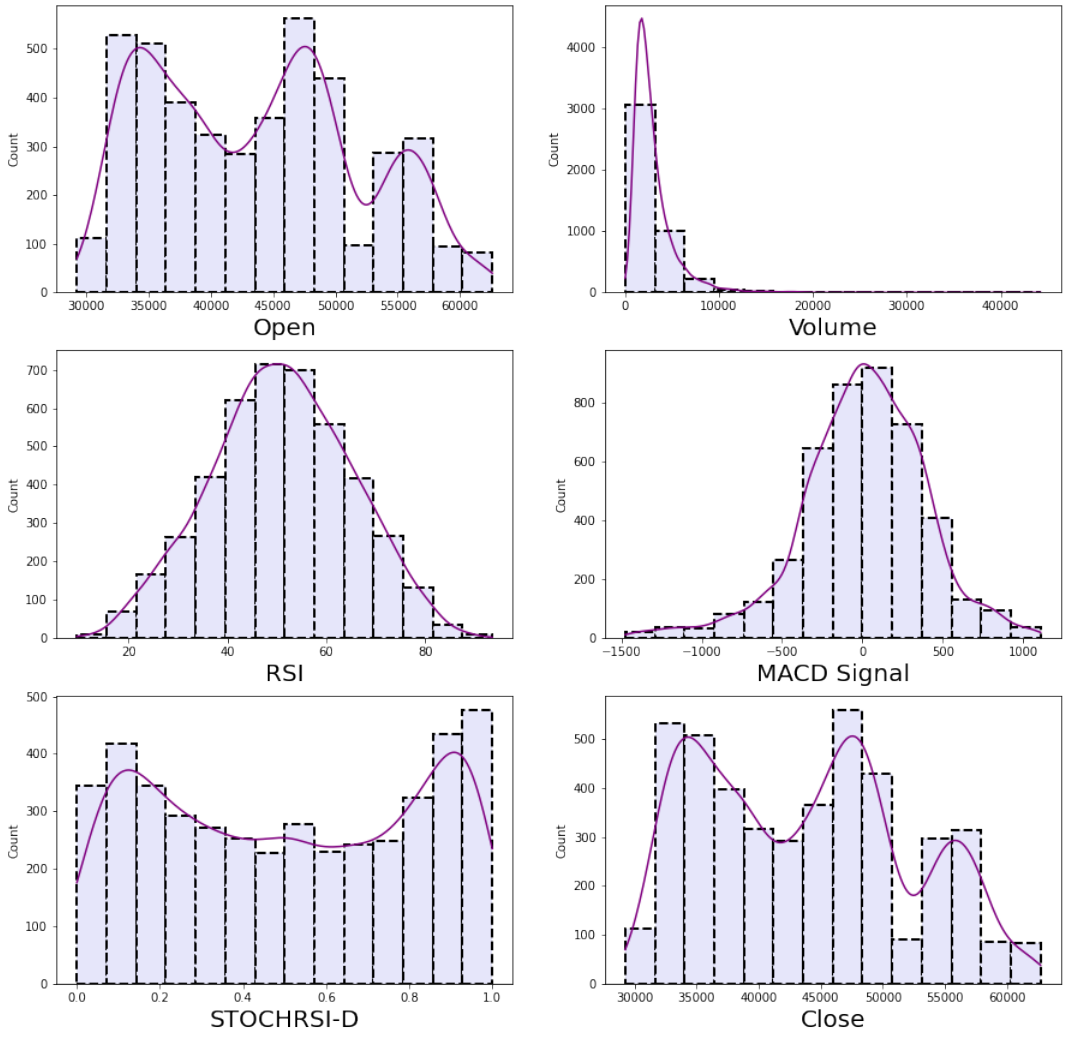
1. Download dataset from this link and save it in ./DATA/ folder with filename DATA.CSV
2. Apply several feature engineering using code in MMA_LAB_2_SHIROKOV.IPYNB notebook

This dataset contains that columns:

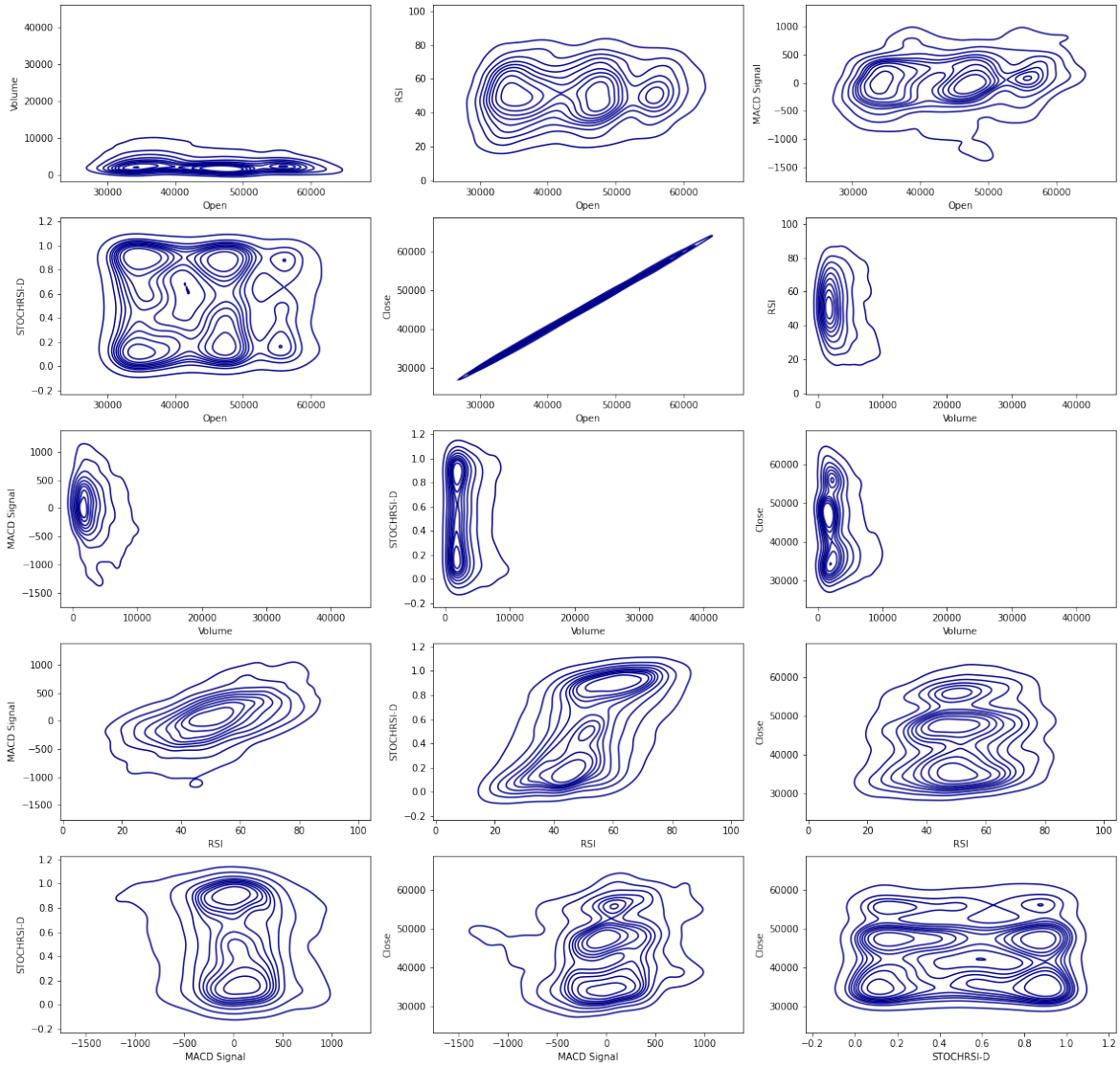
- OPEN - open price
- VOLUME - volume of trading
- RSI - indicator RSI
- MACD SIGNAL - indicator MACD Signal
- MONTH - month number
- STOCHRSI-D - indicator STOCH
- QUANTILE_VOLUME - binned volume
- CLOSE - target variable, close price

1.2 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).

Firstly i have visualised a non-parametric estimation of PDF in form of a histogram and Kerner density function on the same plot for each dataset variable, and after that visualised the KDE distribution between every pair of variables (not using categorical).



Histogram with bins usin Sturges rule and KDE (default epanechnikov kernel)



KDE pairs distributions

Every plot was visualised using python package **Seaborn**.

1.3 Estimation of multivariate mathematical expectation and variance.

In this subsection I have counted the mathematical statistic of expectation and variance using methods of PANDAS.DATAFRAME - MEAN() and VAR().

```
1 data[columns].mean()
```

```
Open                43744.748055
Volume               3018.940768
RSI                    50.929004
MACD Signal             9.319127
month                   7.101575
STOCHRSI-D              0.510336
quantile_Volume         2.499658
Close               43746.064022
dtype: float64
```

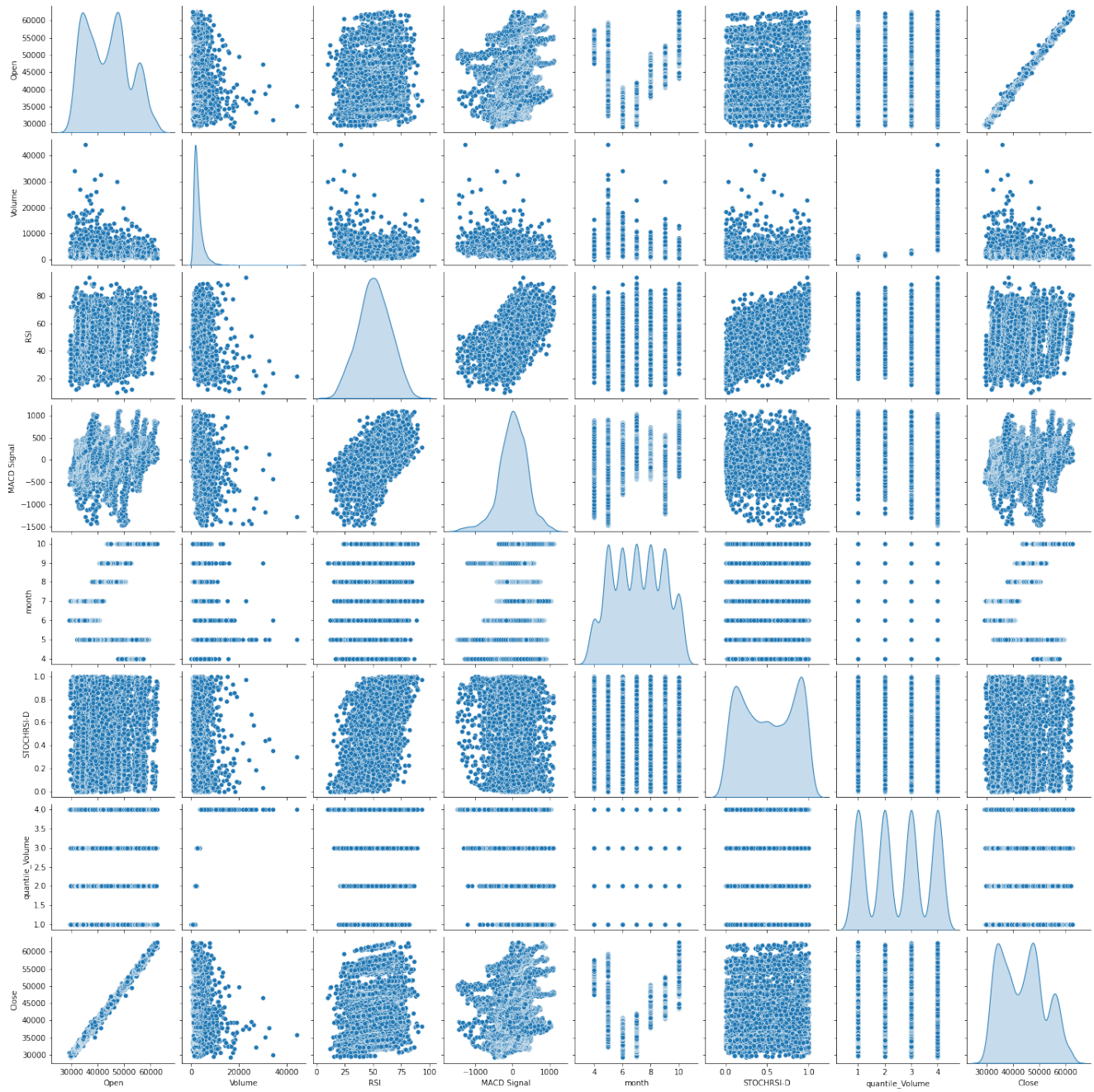
```
1 data[columns].var()
```

```
Open                6.876476e+07
Volume              6.748538e+06
RSI                 2.007612e+02
MACD Signal         1.496410e+05
month               3.127351e+00
STOCHRSI-D          9.952140e-02
quantile_Volume     1.250514e+00
Close               6.880793e+07
dtype: float64
```

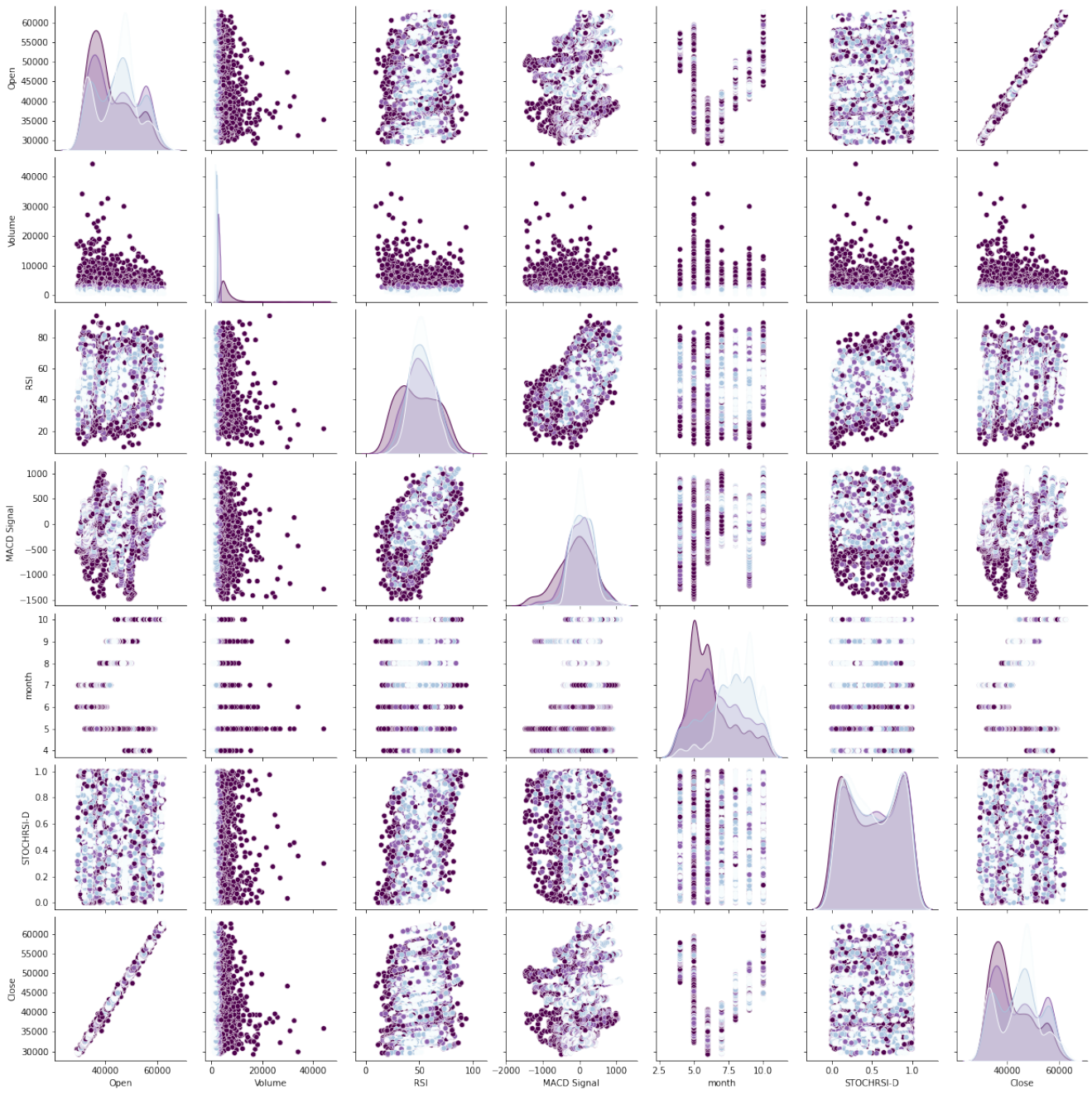
The results of subtask 2

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

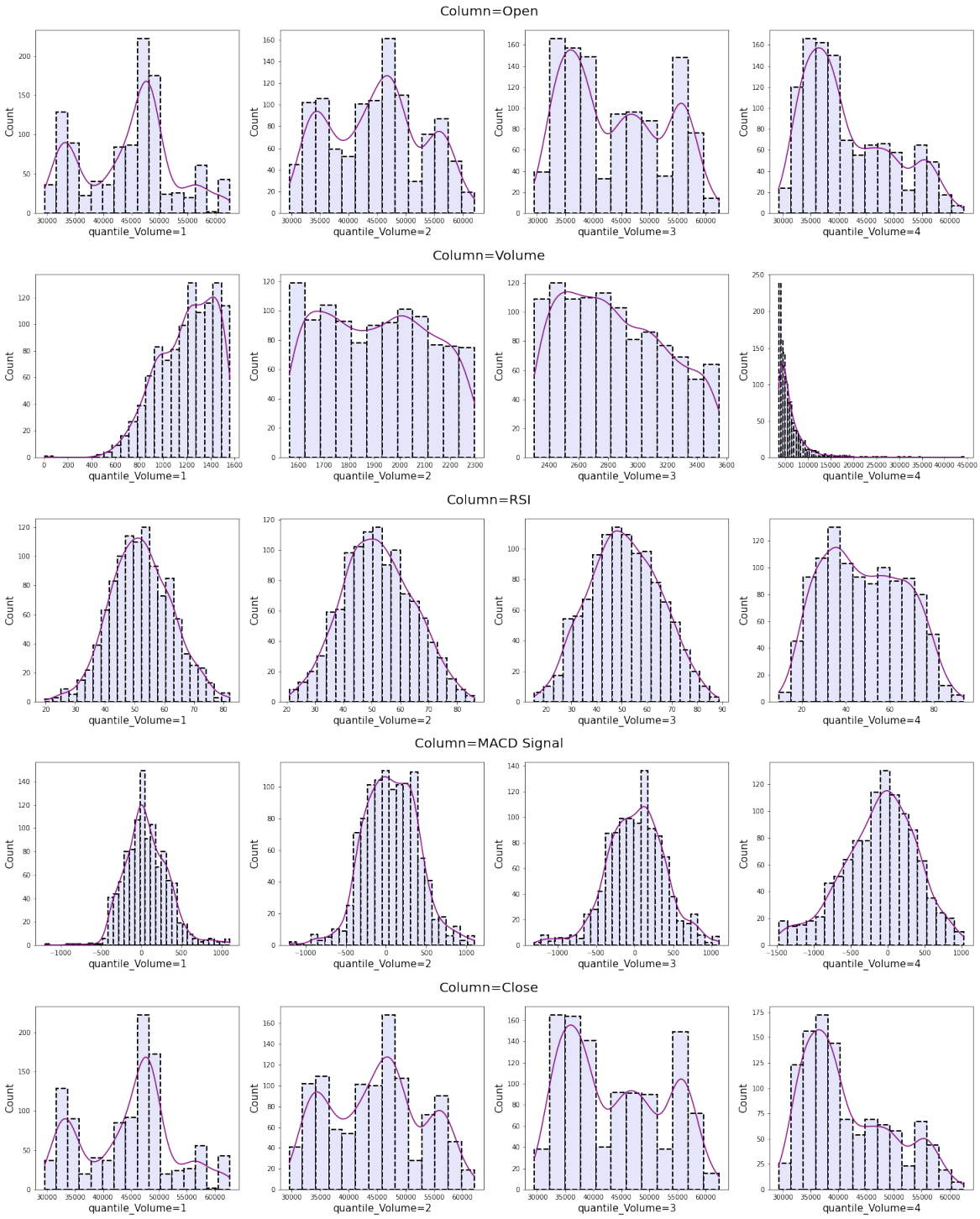
The condition determined the value of the categorical variable, for our dataset this variable is QUANTILE_VOLUME. For each QUANTILE_VOLUME value, step 1 has been reproduced and information for every conditional distribution about mean and var has been counted.



Distribution of each pairs of variables



This is distribution of pairs of variables for each cateofry of QUANTILE_VOLUME using seaborn method SEABORN.PAIRPLOT



Some conditional distributions for each variable, which depends on QUANTILE_VOLUME (more pictures in notebook)

After I have counted conditional mathematil expectation and variance and visualised it as DATAFRAME.

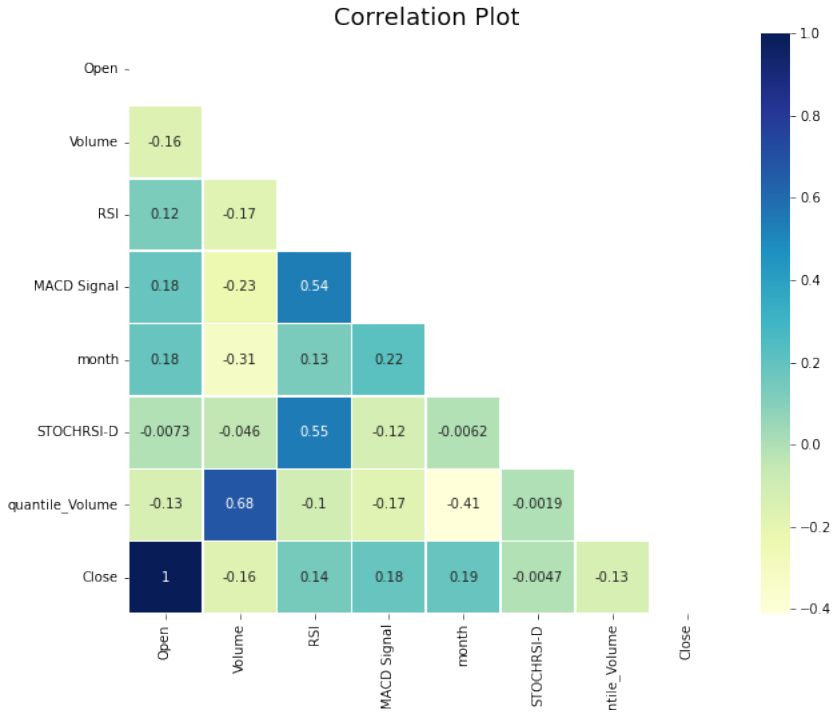
	quantile_Volume=1	quantile_Volume=2	quantile_Volume=3	quantile_Volume=4
mean_Open	6.470311e+07	6.686893e+07	7.559168e+07	6.151362e+07
mean_Volume	5.712776e+04	4.520472e+04	1.198644e+05	1.261243e+07
mean_RSI	1.078101e+02	1.530043e+02	1.963207e+02	3.359093e+02
mean_MACD Signal	7.512455e+04	1.122182e+05	1.432975e+05	2.466097e+05
mean_month	1.830923e+00	2.996524e+00	3.077789e+00	2.480985e+00
mean_STOCHRSI-D	9.287287e-02	9.792741e-02	9.968853e-02	1.077437e-01
mean_Close	6.472727e+07	6.693694e+07	7.560658e+07	6.151082e+07

	quantile_Volume=1	quantile_Volume=2	quantile_Volume=3	quantile_Volume=4
var_Open	44555.475383	44868.400411	43955.143799	41599.232237
var_Volume	1202.160961	1909.361322	2847.455732	6118.444216
var_RSI	52.202464	52.222315	51.065099	48.224974
var_MACD Signal	69.277297	64.243045	17.243959	-113.542551
var_month	8.128650	7.380822	6.715068	6.180822
var_STOCHRSI-D	0.506549	0.513545	0.517946	0.503308
var_Close	44567.078905	44867.676795	43960.847087	41587.903516

Conditional mathematical expectation and variance for each value of categorical variable

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

In this task first I have visualised correlation plot using SNS.HEATMAP and after that for each correlation i have counted confidence intervals. To find the numerical characteristics of the confidence intervals, the stats module of the SciPy package was used.



Correlation plot

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

$$\text{th}\left(\text{arcth}(r) - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right) < \rho < \text{th}\left(\text{arcth}(r) + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right)$$

where r is correlation coefficient.

	Open	Volume	RSI	MACD Signal	month	STOCHRSI-D	quantile_Volume	Close
Open	(1.0, 1.0)	(-0.189, -0.131)	(0.0945, 0.153)	(0.156, 0.213)	(0.155, 0.213)	(-0.0369, 0.0223)	(-0.161, -0.103)	(0.999, 0.999)
Volume	(-0.189, -0.131)	(1.0, 1.0)	(-0.2, -0.143)	(-0.262, -0.207)	(-0.341, -0.287)	(-0.0757, -0.0166)	(0.659, 0.691)	(-0.193, -0.135)
RSI	(0.0945, 0.153)	(-0.2, -0.143)	(1.0, 1.0)	(0.518, 0.56)	(0.101, 0.159)	(0.524, 0.566)	(-0.133, -0.0739)	(0.113, 0.171)
MACD Signal	(0.156, 0.213)	(-0.262, -0.207)	(0.518, 0.56)	(1.0, 1.0)	(0.192, 0.248)	(-0.145, -0.087)	(-0.201, -0.143)	(0.156, 0.213)
month	(0.155, 0.213)	(-0.341, -0.287)	(0.101, 0.159)	(0.192, 0.248)	(1.0, 1.0)	(-0.0358, 0.0234)	(-0.436, -0.387)	(0.157, 0.214)
STOCHRSI-D	(-0.0369, 0.0223)	(-0.0757, -0.0166)	(0.524, 0.566)	(-0.145, -0.087)	(-0.0358, 0.0234)	(1.0, 1.0)	(-0.0315, 0.0277)	(-0.0343, 0.0249)
quantile_Volume	(-0.161, -0.103)	(0.659, 0.691)	(-0.133, -0.0739)	(-0.201, -0.143)	(-0.436, -0.387)	(-0.0315, 0.0277)	(1.0, 1.0)	(-0.162, -0.103)
Close	(0.999, 0.999)	(-0.193, -0.135)	(0.113, 0.171)	(0.156, 0.213)	(0.157, 0.214)	(-0.0343, 0.0249)	(-0.162, -0.103)	(1.0, 1.0)

Confidence interval of each correlation

1.6 Task formulation for regression, multivariate correlation.

Due to results in previous subtask, we will try to predict **target** column TARGET=CLOSE by **predictors**:

$$\text{Open}, \text{Volume}, \text{RSI}, \text{MACD Signal}, \text{month}, \text{quantile\_Volume}$$

without STOCHRSI-D because of very low correlation.

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

In this task I have used python package SCIKIT-LEARN and its implementation of LINEAR-REGRESSION, LASSO REGRESSION and RIDGE Regression. The cross-validation technique on 10 folds with differentseeds was used to count a *Mean Absolute Error* (MAE) and *Root Mean Squared Error* (RMSE) to compare results. Te Ridge and Lasso regression was used for regularization cause *Volume* and *quantile_Volume* has very big multicolleration, so we want to make weights of each predictor be contolled by optimization task.

For Lasso and Ridge Regression the best α (coefficient of regularization) was found from grid $\alpha_i = [0.001, 5]$.

Let's see the results.

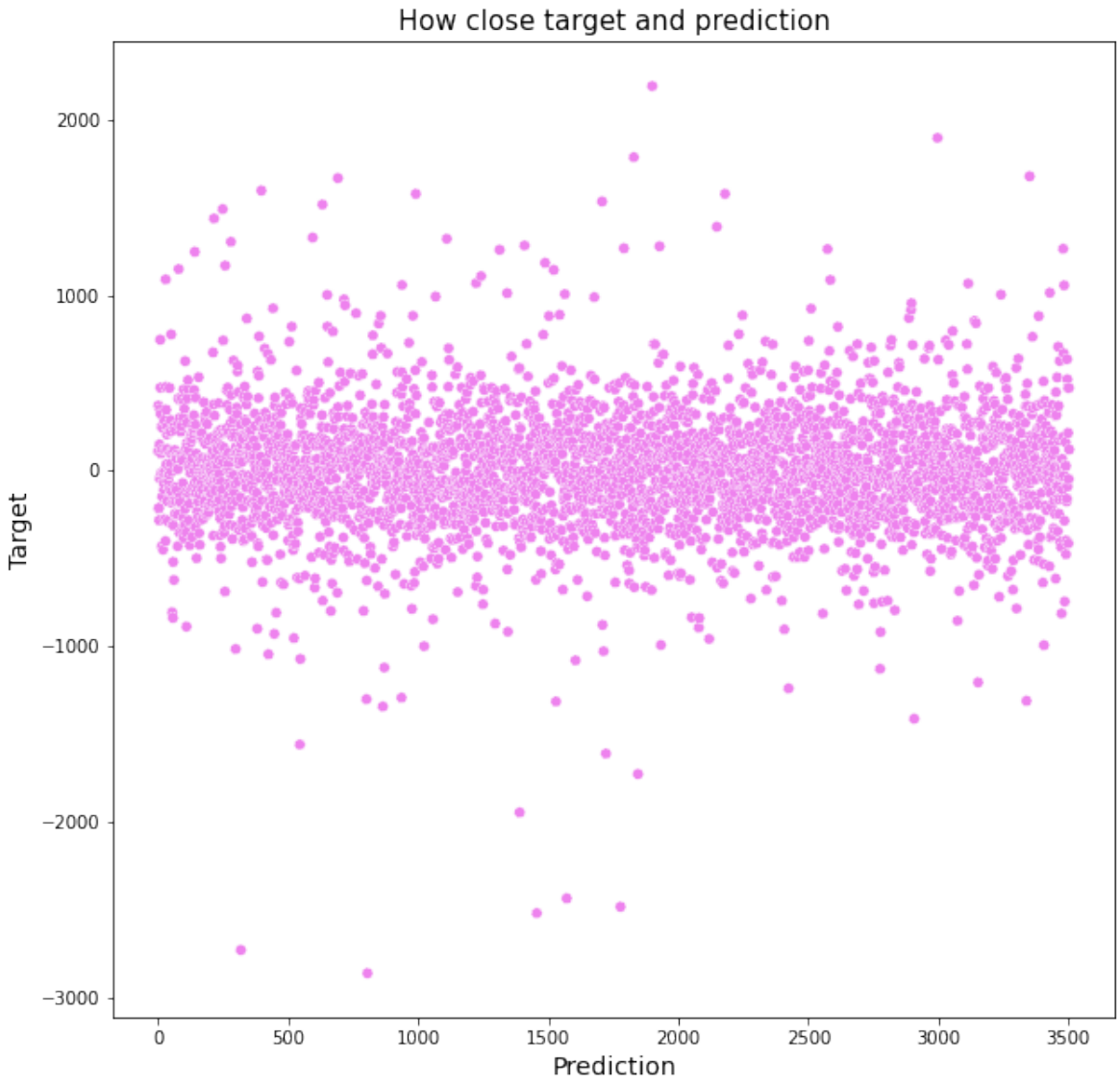
	RMSE	MAE	R2
LinReg	341.005639	245.695283	0.998296
Lasso(α=4.8)	340.924986	245.617767	0.998297
Ridge(α=5.0)	341.003954	245.694665	0.998297

Results of implementation of each method. The Lasso method with counted value of regularization coefficient $\alpha = 4.79$ was the best due to minimizing MAE and RMSE metrics.

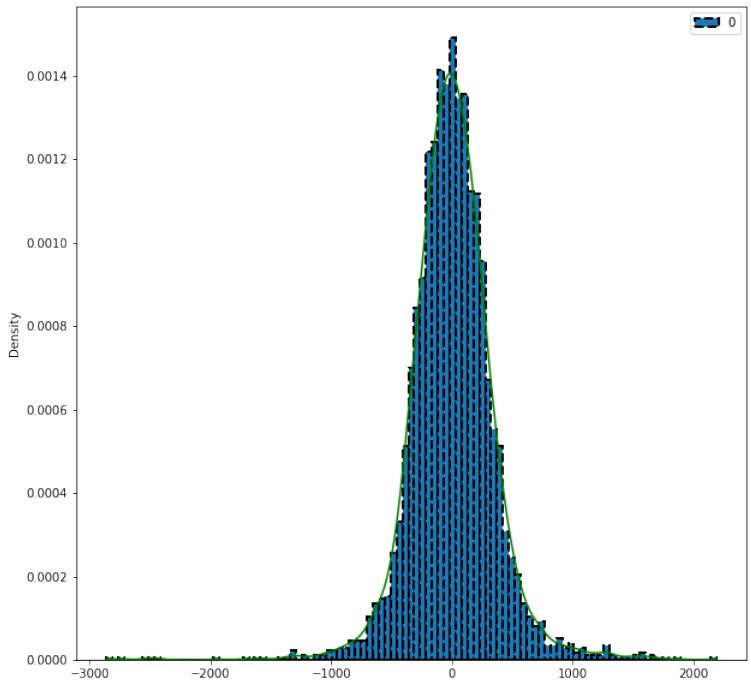
	features	VIF
0	Open	21.201130
1	Volume	4.473409
2	RSI	15.427064
3	MACD Signal	1.308642
4	month	15.230902
5	quantile_Volume	10.358325

Also, using VARIANCE_INFLATION_FACTOR from **StatsModels** package the variance inflation factor has been counted. We can see that with OPEN, RSI, MONTH, and QUANTILE_VOLUME multicollinearity is high.

1.8 Quality analysis.



How close target and prediction



Residuals of model - difference between target and predict.

As you can see in the plot, the model errors seemst to have a good normal distribution with mean around 0, but Shapiro-Wilk's test says that, it is abosutely not normal.

```
from scipy.stats import shapiro, ttest_ind

shapiro(residuals)
```

```
ShapiroResult(statistic=0.9295549988746643, pvalue=8.527424680522784e-3
8)
```

By the way the student mean test says that the hypothesis that mean of residuals is 0 is accepted, so only variance isn't normally satisfied.

```
mu = 0
mean = np.mean(residuals)
var = np.var(residuals)

T = np.abs(np.sqrt(len(residuals)) * (mean - mu) / var)

z = stats.t.ppf(1 - alpha / 2, len(residuals) - 1)
```

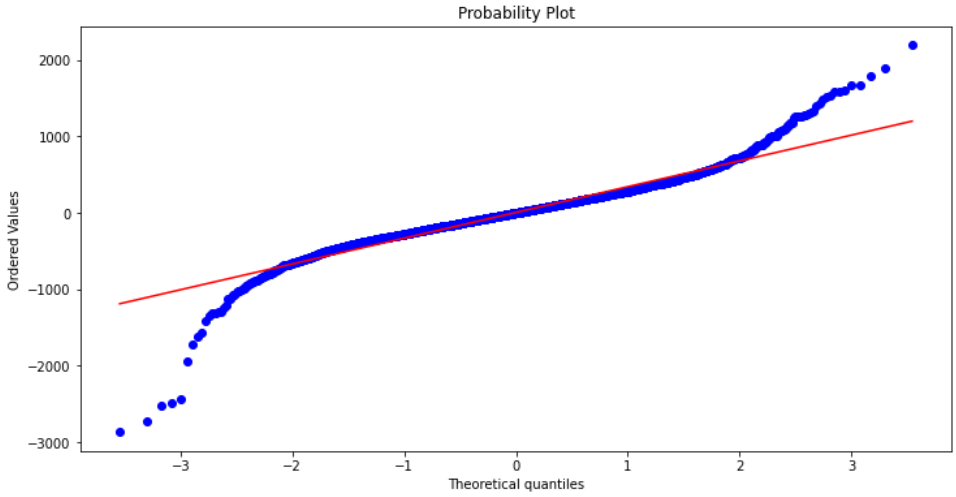
```
T, z
```

```
(0.002287033982773346, 1.9606414259263407)
```

$T < z_{quantile}$, so the hypothesis is accepted

The coefficient of determination fixes the proportion of the explained variance of the effective feature due to the factors under consideration, and since R2 is practically equal to one, there are

no unaccounted factors in the model and all the signs explain the target variable. From the point of view of the prerequisites of the model, the assumption of zero mathematical expectation of residuals is fulfilled and then if you look at the qq-plot



then the assumption of normality of residuals with zero mean and variance is practically fulfilled, with the exception of the tails of the distribution, by the way when increasing the sample size, the assumption of normality for the residuals is not really required, so this assumption is also works for our model.

Therefore, based on the result of the analysis, we can say that the model is suitable for explaining dependencies using linear regression - from the point of view of the explained variance and from the point of view of the prerequisites for building this model.

2 Source code

The link to the sourse code which is placed on my github.

3 Conclusion

I have learned about some methods of multivariate data analysis, used some popular Python packages, have understood the data of the dataset, used linear regression model with regularization and best coeffecient of refulatization and applied some algorithms of mathematical statistic for calculating confidence interval.