FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION OF HIGHER EDUCATION
ITMO UNIVERSITY

Report on learning practice № 4
Stationarity of the processes

Performed by:
Denis Zakharov,
Maxim Shitilov,
Sapelnikova Ksenia,
Vdovkina Sofia,
Academic group J4132c, J4133c

Saint-Petersburg
2022

Goals

1. *Predictors and target variables.*
2. *Analysis of stationarity: covariance or correlation function for chosen target variables and mutual correlation functions among predictors and targets.*
3. *Filter high frequencies.*
4. *Built auto-regression model filtered and non-filtered data*
5. *Build a model in a form of linear dynamical system, using chosen predictors.*

Brief theoretical part

In this laboratory work we will work with stationary processes. A stationary process is a stochastic process in which the probability distribution does not change with time displacement.

Obviously, the process can be stationary and non-stationary. To clarify this point, there are several tests, in particular the Dickey-Fuller test, which is based on the fact that when a critical unit is reached, the series $x_t = \rho x_{t-1} + e_t$ ceases to return to its average value. If we subtract x_{t-1} from the left and right parts, we get $x_t - x_{t-1} = (\rho -)x_{t-1} + e_t$, where the expression on the left is the first differences. If $\rho = 1$, then the first differences will give a stationary white noise e_t.

Also in this laboratory, the topic of noise filtering will be touched upon. High-quality filtering of noisy data allows you to reduce the error and increase the quality of models based on noisy data.

Results

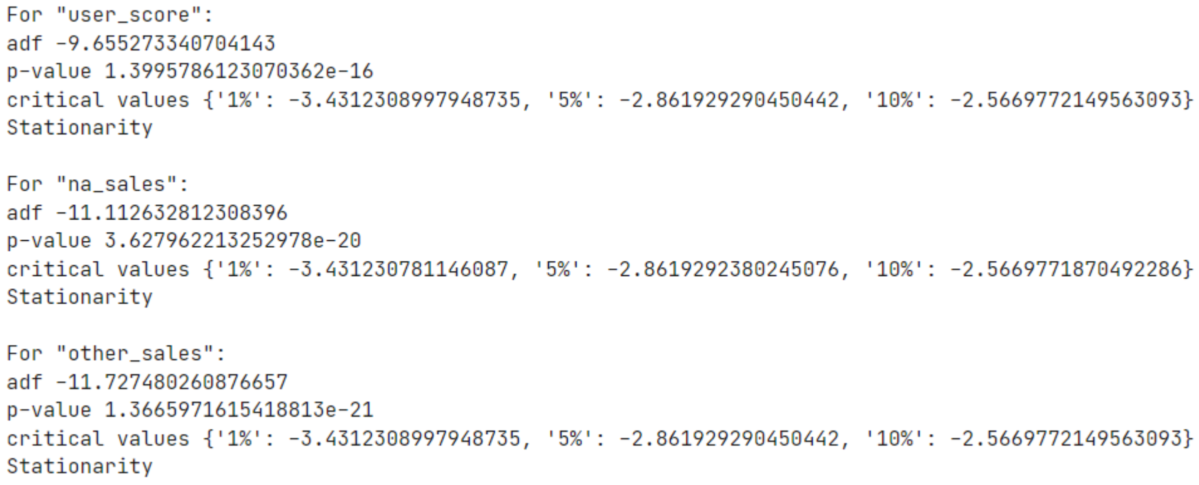
1. ***Choose about 3-6 variables from your dataset (2-3 – target variables, the rest - predictors).***

At this stage, it was necessary to select 2-3 target variables and several predictors. As can be seen from the following code, the target variables are 'user_score', 'na_sales', 'other_sales'.

```
targets = ['user_score', 'na_sales', 'other_sales']

predictors = df[[

    'name', 'platform', 'year_of_release', 'eu_sales',

    'jp_sales', 'critic_score', 'world_sales',

    'genre', 'rating'

]]
```

2. ***Analyze stationarity of a process (for mathematical expectation and variance) for all chosen variables. Make them more stationary if needed.***

At this stage of the laboratory work, a Dickey-Fuller test was performed to determine the stationarity of the process. This test is implemented in the statsmodels module in python. To use it, you just need to select the necessary modules (statsmodels and pandas), load the data, and output the result.



```
For "user_score":
adf -9.655273340704143
p-value 1.3995786123070362e-16
critical values {'1%': -3.4312308997948735, '5%': -2.861929290450442, '10%': -2.5669772149563093}
Stationarity

For "na_sales":
adf -11.112632812308396
p-value 3.627962213252978e-20
critical values {'1%': -3.431230781146087, '5%': -2.8619292380245076, '10%': -2.5669771870492286}
Stationarity

For "other_sales":
adf -11.727480260876657
p-value 1.3665971615418813e-21
critical values {'1%': -3.4312308997948735, '5%': -2.861929290450442, '10%': -2.5669772149563093}
Stationarity
```

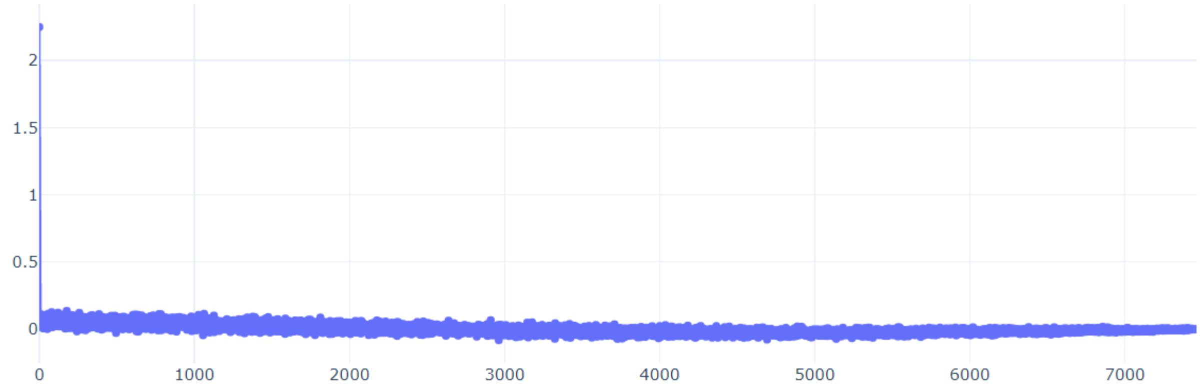
Pic. 1 — The result of the Dickey-Fuller test for target variables.

According to the results of the Dickey-Fuller test, all the target variables we selected are stationary.

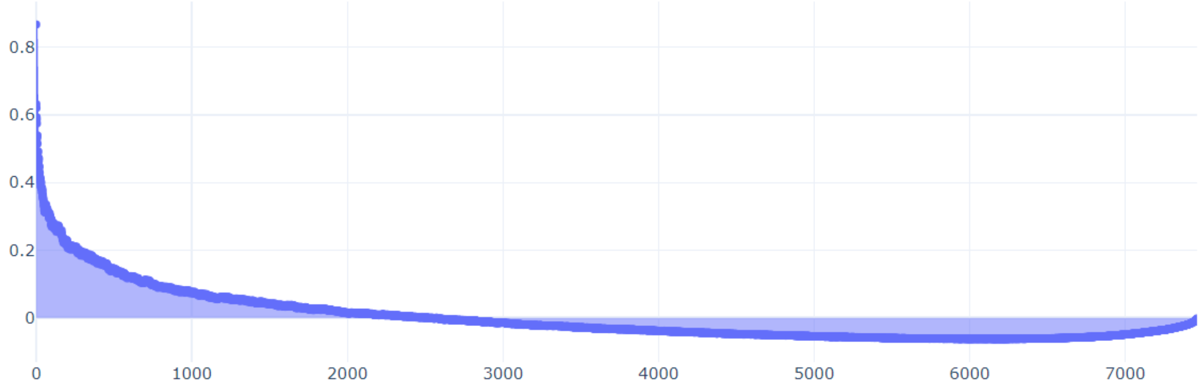
3. *Analyze covariance or correlation function for chosen target variables and mutual correlation functions among predictors and targets.*

In probability theory and statistics, given a random process, autocovariance is a function that gives the covariance of a process with itself in pairs of time points.

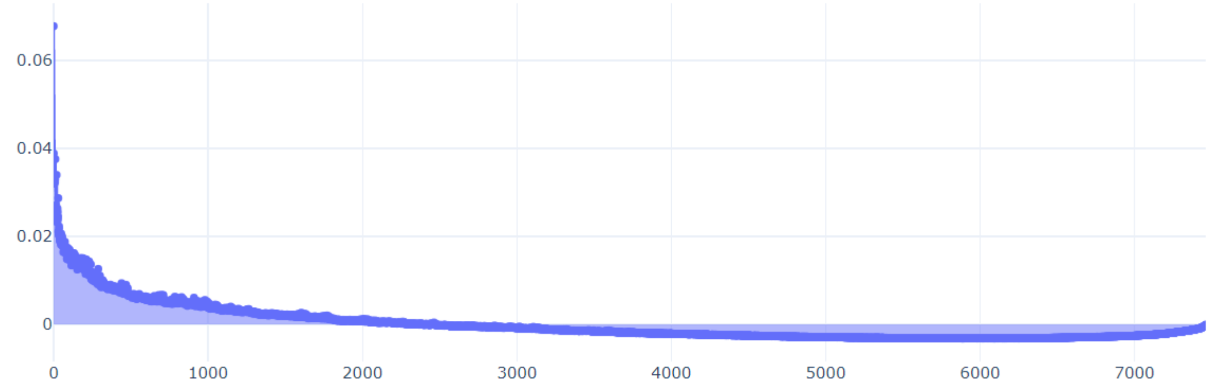
Below you can see graphs of autocovariance functions for the target variables we have selected.



Pic. 2 — The autocovariances of the 'user_score' variable.



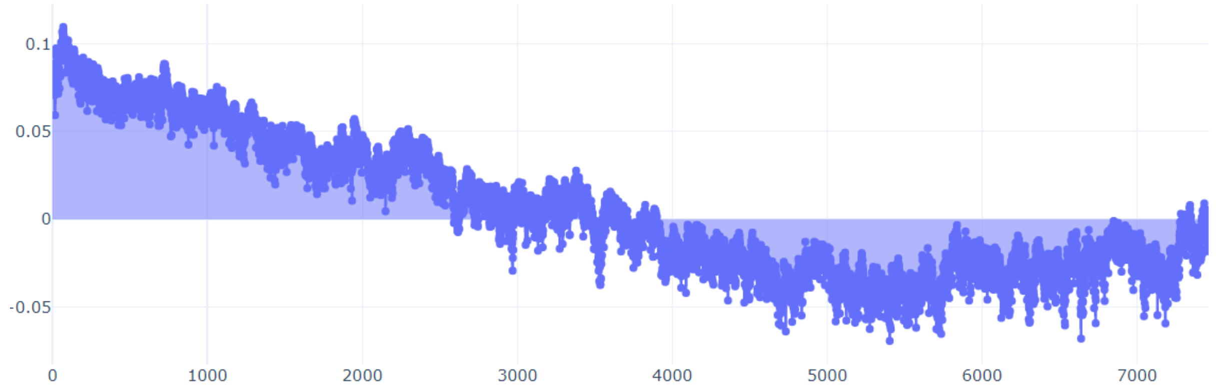
Pic. 3 — The autocovariances of the 'na_sales' variable.



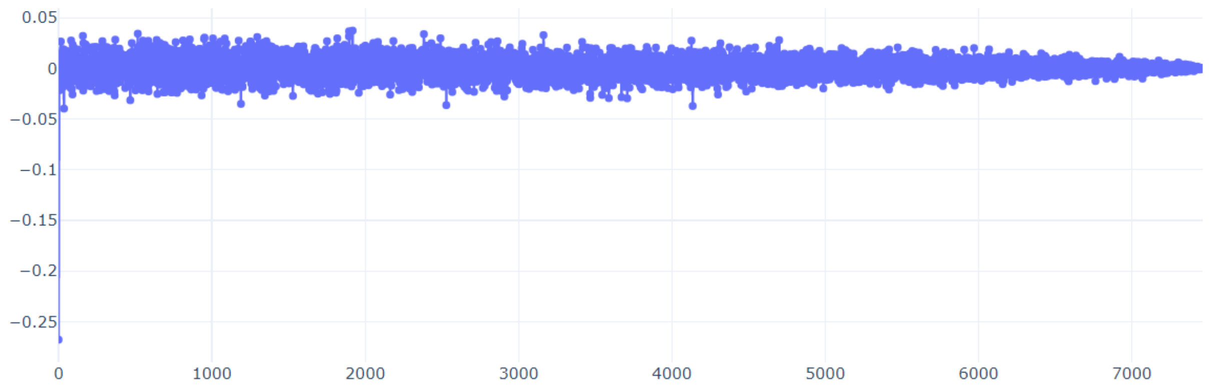
Pic. 4 — The autocovariances of the 'other_sales' variable.

The mutual correlation function of two random functions X(t) and Y(t) is called a non-random function Rxy(tv t2) of two independent arguments R, and t.r whose value for each pair of fixed argument values is equal to the correlation moment of the sections of both functions corresponding to the same fixed argument values.

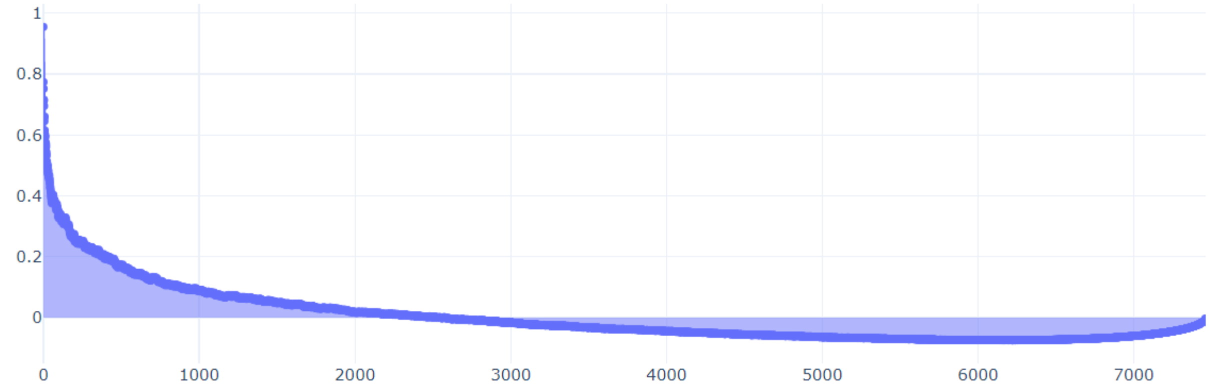
In the graphs below, you can observe graphs of mutual correlation functions for pairs composed of target variables and predictors 'world_sales' and 'year_of_release'.



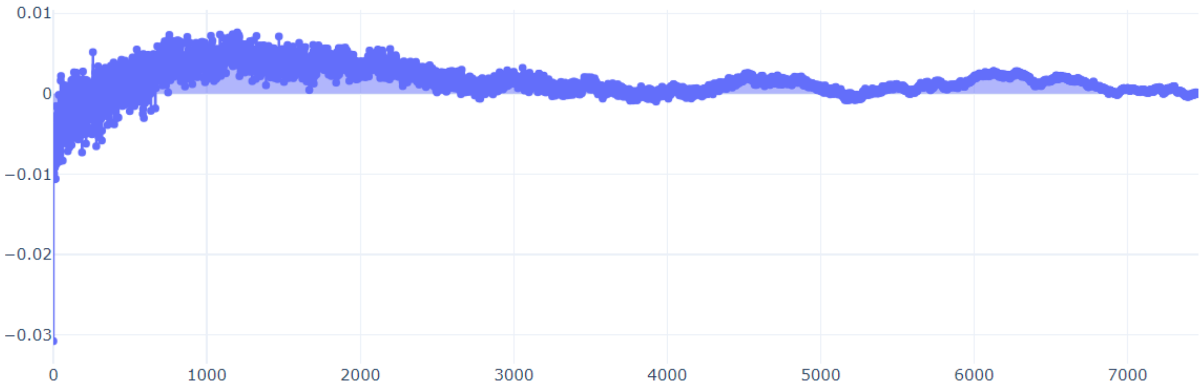
Pic. 5 — The function of mutual correlation of variables 'user_score' and 'world_sales'



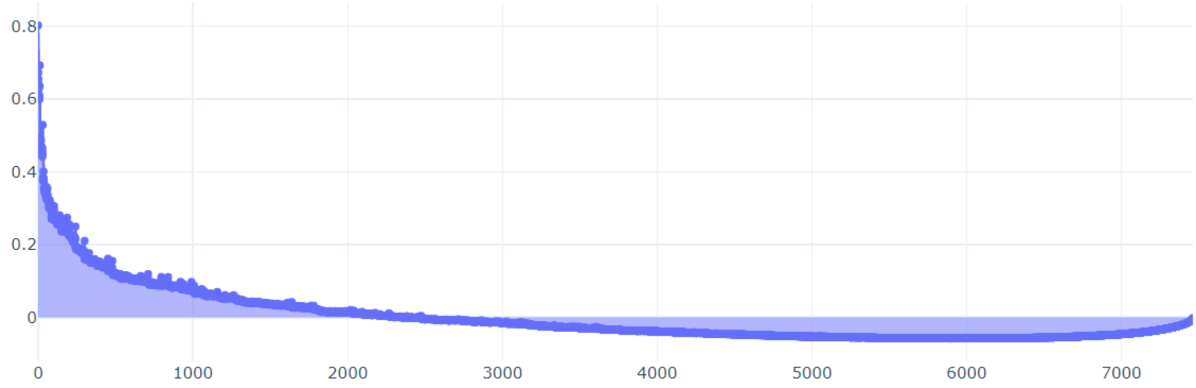
Pic. 6 — The function of mutual correlation of variables ‘user_score’ and ‘year_of_release’



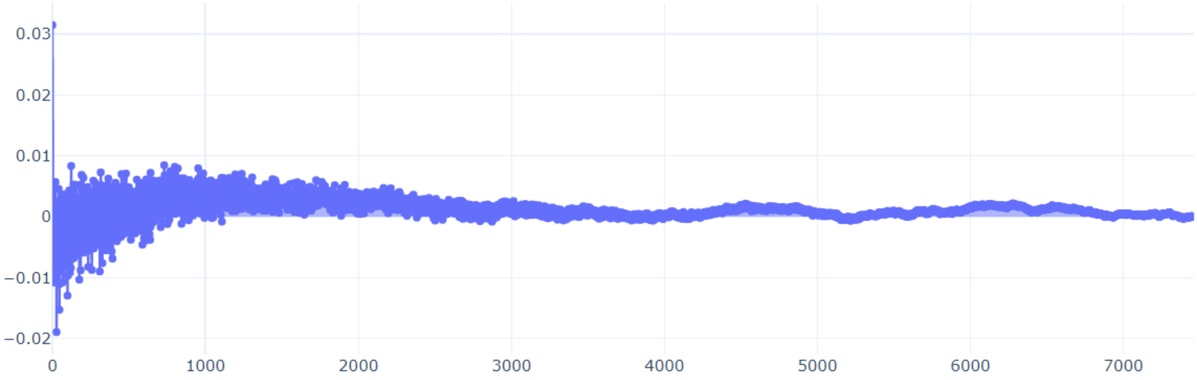
Pic. 7 — The function of mutual correlation of variables ‘na_sales’ and ‘world_sales’



Pic. 8 — The function of mutual correlation of variables ‘na_sales’ and ‘year_of_release’



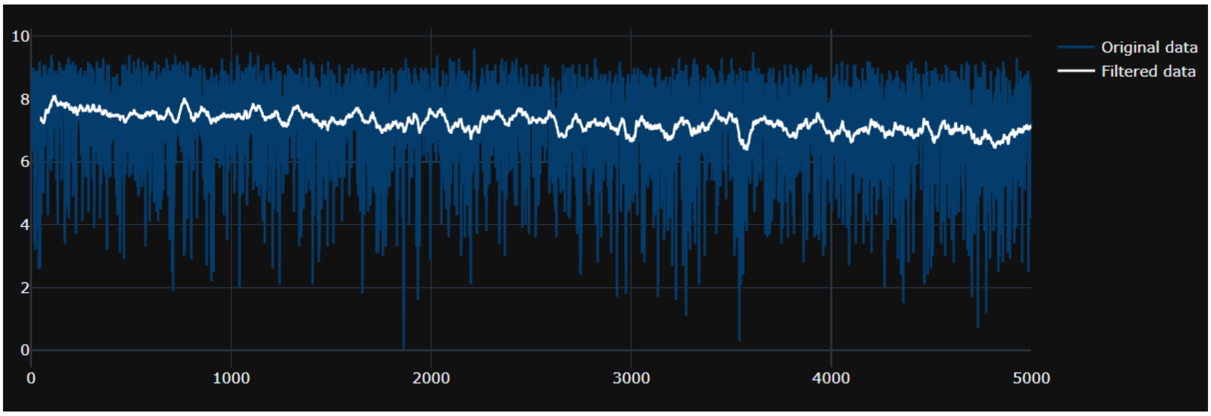
Pic. 9 — The function of mutual correlation of variables ‘other_sales’ - ‘world_sales’



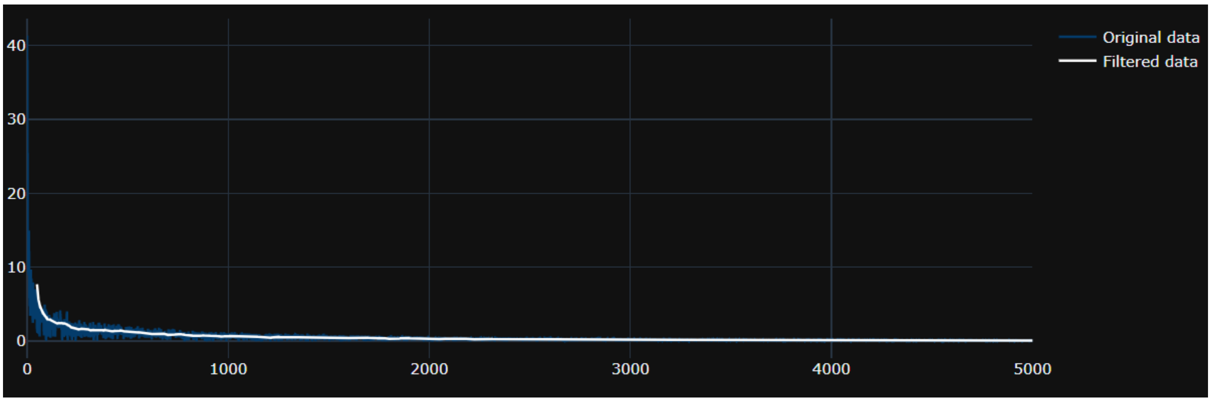
Pic. 10 — The function of mutual correlation of variables ‘other_sales’ - ‘year_of_release’

4. *Filter high frequencies (noise) with chosen 2 filters for target variables.*

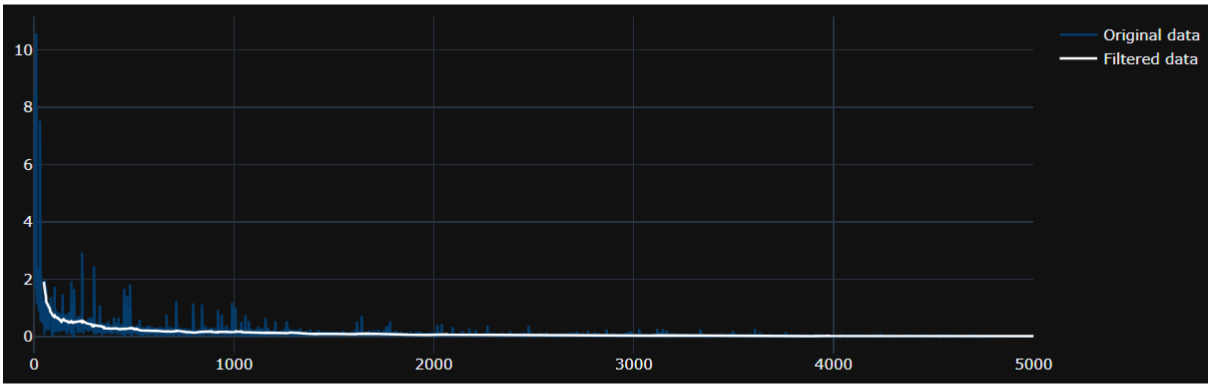
Rolling mean is a common name for a family of functions whose values at each point of definition are equal to some average value of the original function for the previous period.



Pic. 11 — The graph of the variable '' filtered using the method ‘rolling mean’



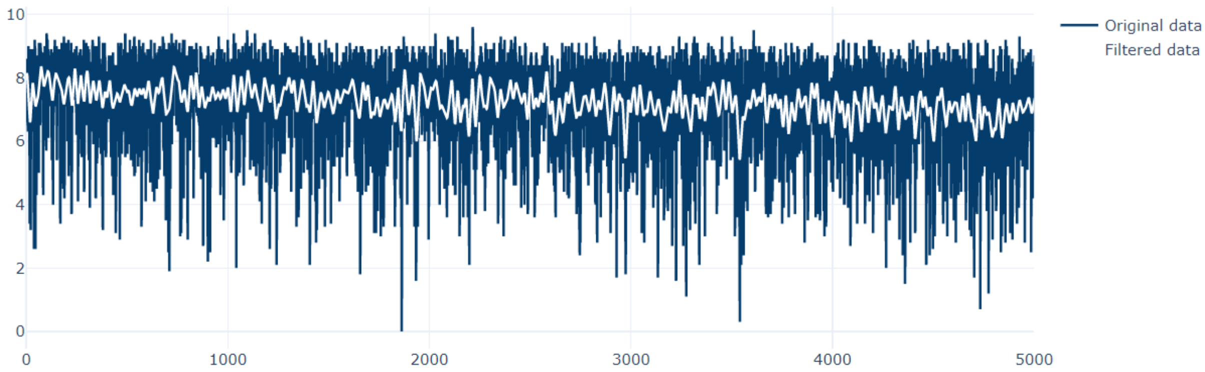
Pic. 12 — The graph of the variable '' filtered using the method 'rolling mean'



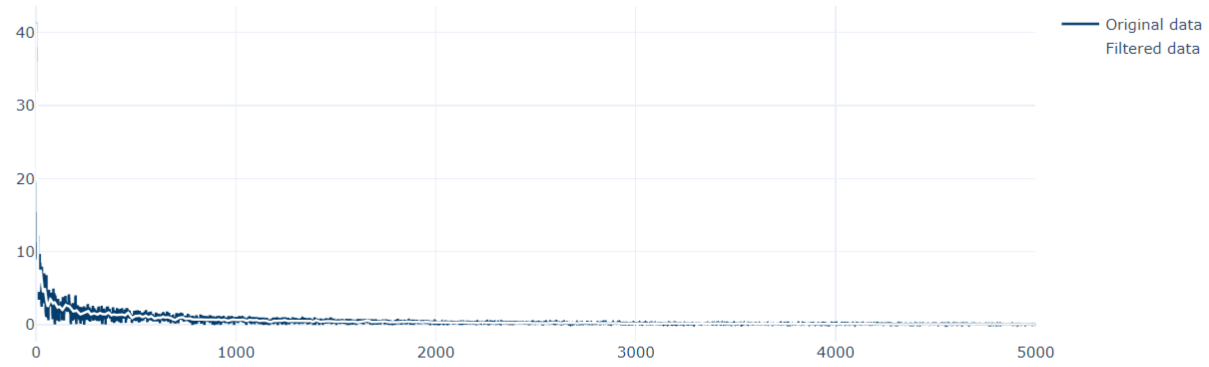
Pic. 13 — The graph of the variable '' filtered using the method 'rolling mean'

The 'filtfilt' method applies a linear digital filter twice, once forward and once backward. The combined filter has a zero phase and a filtration order twice as high as the original one.

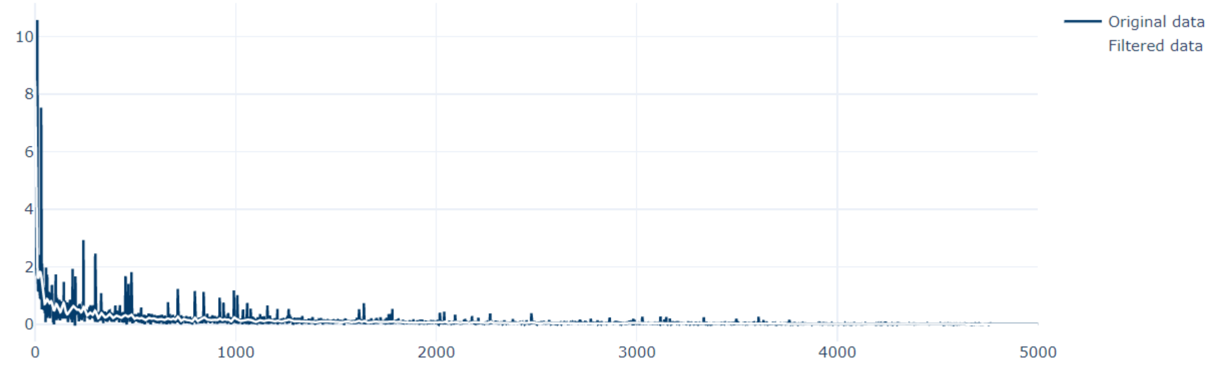
The function provides options for processing the edges of the signal.



Pic. 14 — The graph of the variable '' filtered using the method 'filtfilt'



Pic. 15 — The graph of the variable '' filtered using the method 'filtfilt'



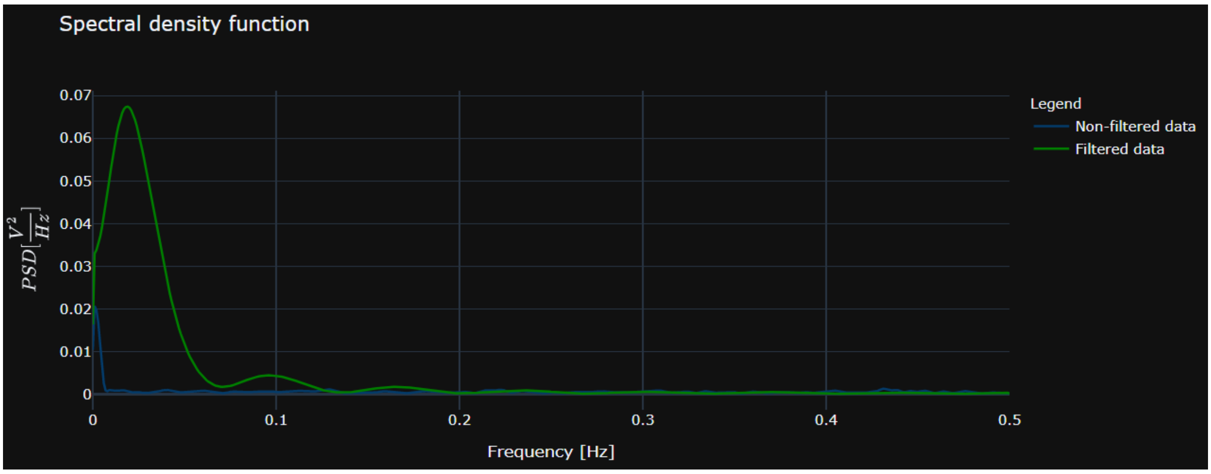
Pic. 16 — The graph of the variable '' filtered using the method 'filtfilt'

Based on the graphs presented above, we can say that the moving average method copes a little better than the 'filtfilt' method. In the case of the moving average method, the filtered data function is smoother.

5. *Estimate spectral density function for with and without filtering.*

Spectral density function; along with probability density, correlation functions, mathematical expectation and variance, the spectral density function refers to the characteristics by which the basic properties of stationary random processes are analyzed. Spectral density is used to analyze systems exposed to random signals, to determine the properties of systems by input and output processes, to identify energy sources and noise; spectral density functions to assess the relationship between the periodic and noise components of a random process.

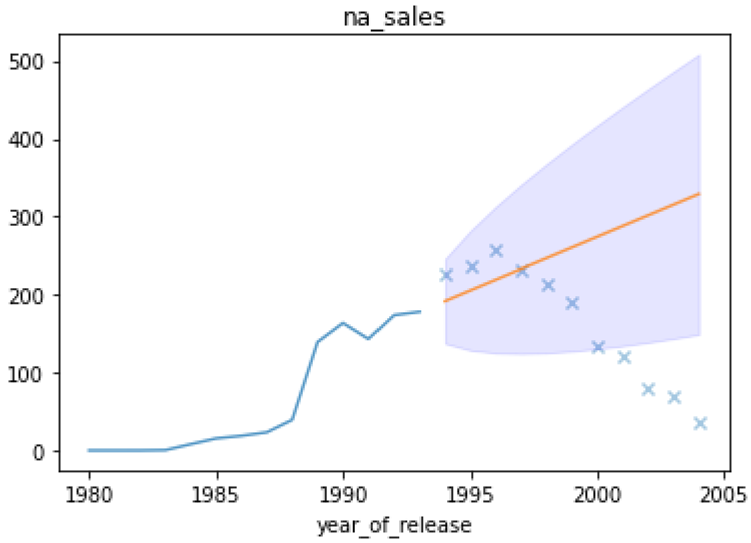
Since the spectral density characterizes the distribution of average energy over frequencies, it can be determined through the procedure of narrow-band filtering and averaging the energy of the filtered signal.



Pic. 16 — Spectral density function

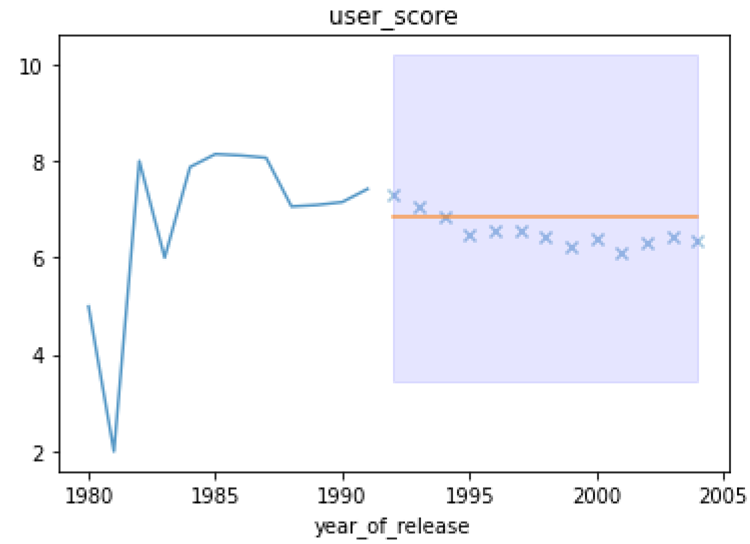
In the graph above, you can see graphs of the spectral density function based on the original and previously filtered dates.

6. ***Built auto-regression model filtered and non-filtered data. To analyze residual error and to define appropriate order of model.***



Pic. 17 — Prediction of na_sales variable

Test RMSE: 151.818



Pic. 18 — Prediction of user_score variable

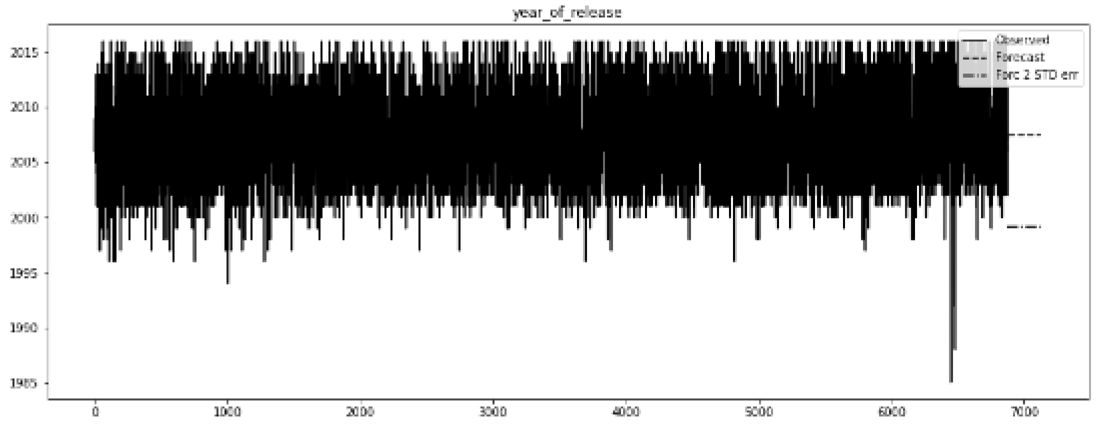
Test RMSE: 0.433

The graphs show the results of an autoregressive model trained on filtered data (na_sales and user_score). The blue line is the initial and training data.

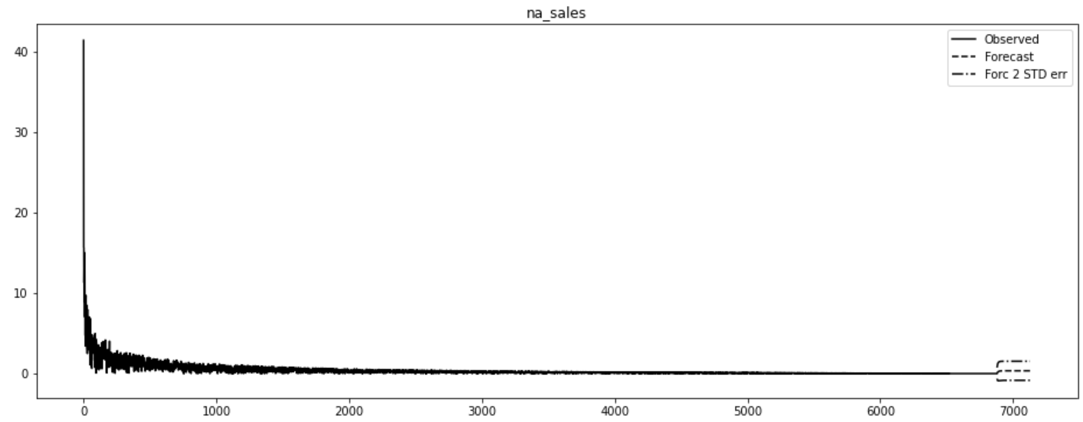
Blue crosses are the source data. The orange line is the predicted data.

Judging by the graphs, this model is a first-order autoregression model and is generally described by an equation of the form $y_{t+1} = \rho y_t + u_{t+1}$.

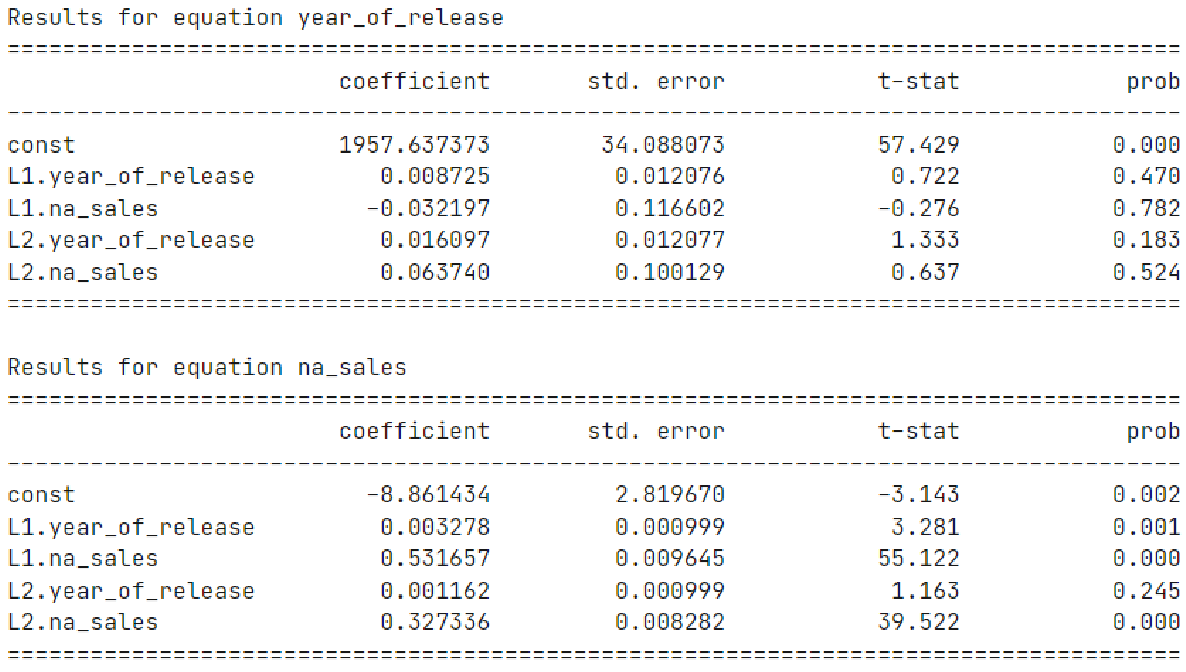
7. ***Build model in a form of linear dynamical system, using chosen predictors. To analyze residual error and to define appropriate order of model.***



Pic. 18 — Linear dynamical system 'year_of_release'



Pic. 19 — Linear dynamical system 'na_sales'



```
Results for equation year_of_release
==========================================================================================
                        coefficient       std. error           t-stat            prob
------------------------------------------------------------------------------------------
const                   1957.637373        34.088073           57.429           0.000
L1.year_of_release         0.008725         0.012076            0.722           0.470
L1.na_sales               -0.032197         0.116602           -0.276           0.782
L2.year_of_release         0.016097         0.012077            1.333           0.183
L2.na_sales                0.063740         0.100129            0.637           0.524
==========================================================================================

Results for equation na_sales
==========================================================================================
                        coefficient       std. error           t-stat            prob
------------------------------------------------------------------------------------------
const                     -8.861434         2.819670           -3.143           0.002
L1.year_of_release         0.003278         0.000999            3.281           0.001
L1.na_sales                0.531657         0.009645           55.122           0.000
L2.year_of_release         0.001162         0.000999            1.163           0.245
L2.na_sales                0.327336         0.008282           39.522           0.000
==========================================================================================
```

Pic. 20 — Evaluation of the results of the linear dynamic system

On the presented graphs you can see the construction of the model in the form of a dynamic system.

A solid black line is the source data. On the right side of the graph, according to the legend, you can see the forecast, as well as the confidence interval, which is called "Fork 2 STD err" on the legend.

Conclusions

As a result of the work, we have worked with stationary analysis of covariance or correlation function. We also have to deal with noise filtration by *filtering it using rolling mean and signal.filtfilt approaches, estimate spectral density function and get our hands on building an auto-regression model and model data in a form of linear dynamical system.*

Appendix

DataLore: site. – URL: https://datalore.jetbrains.com/notebook/RemqSkuJwmr1PM4Gc3cBqB/SDel6aHaXiAVtQ4KpsPD5r/ (circulation date: 03.12.2022)