FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION OF HIGHER EDUCATION
ITMO UNIVERSITY

Report on learning practice #3
"Sampling of multivariate random variables"

Performed by
Tyulkov Nikita

St. Petersburg
2021

1. Substantiation of chosen sampling.

On the first step we have to choose target and predictor variables. From avocado dataset the next variables was chosen:

large_bags	target
xlarge_bags	target
4770	target
4046	predictor
4225	predictor
total_volume	predictor
small_bags	predictor
average_price	predictor
total_bags	predictor

2. Sampling of chosen target variables using univariate parametric distributions (from practice 2) with 2 different sampling methods.

For the sampling two algorithms were chosen:
On the Figure 1 we can see the results of random sampling algorithm.

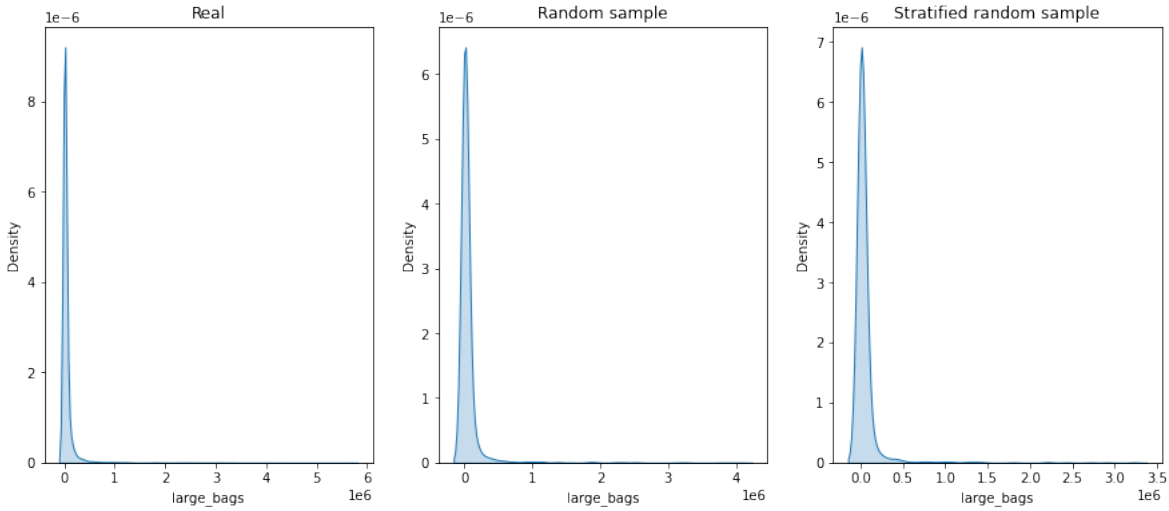

Figure 1: Random sampling algorithm

On the Figure 2 we can see the results of stratified random sampling algorithm.
On the Figure 3 we can see the real data plots.

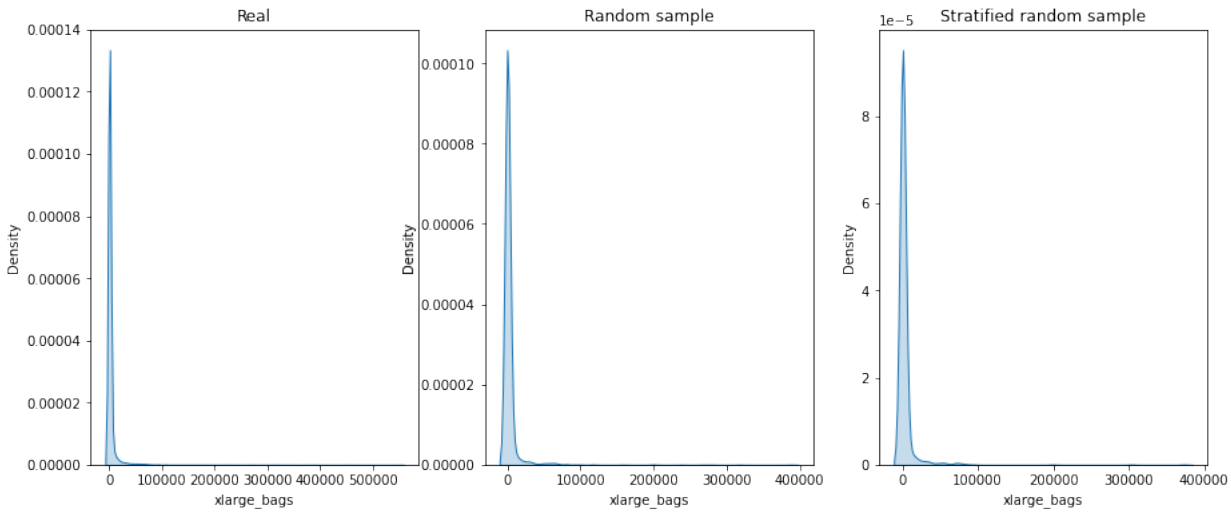

Figure 2: Stratified random sampling algorithm

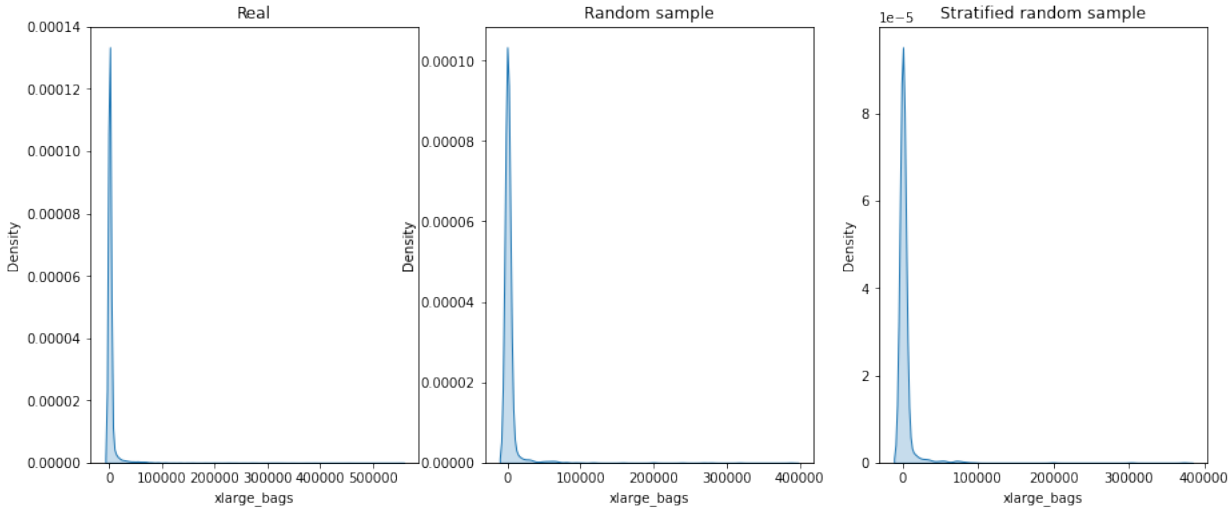

Figure 3: Real data

As we can see, the distribution of the algorithms are similar. That is lead to conclusion that our sampling works well.

3. Estimation of relations between predictors and chosen target variables.

On the third step we calculate the correlation between our variables for next choosing of them to build the Bayesian network. The results shown on Figures 4, 5, 6.

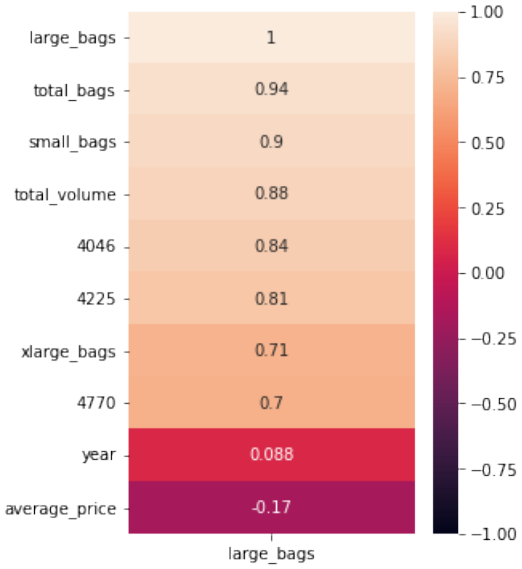

Figure 4: "large_bags" correlations

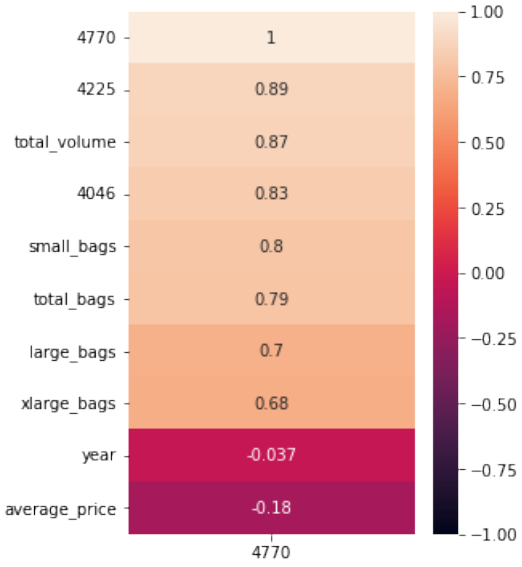

Figure 5: "4770" correlations

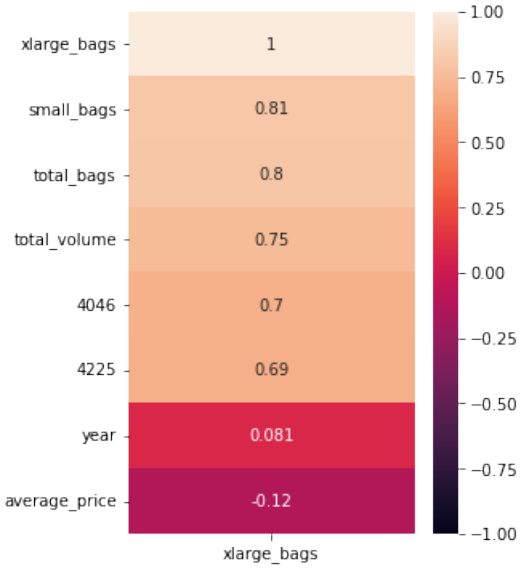

Figure 6: "xlarge_bags" correlations

4. Bayesian network.

In that section we going to build, firstly, manual Bayesian network and than 2 algorithms for structural learning.

For manual Bayesian network the variables were chosen based on correlations that were calculated in Section 3. For target variable "large_bags" predictors like "4046" and "4225" were chosen. For target variable "4770" predictors like "total_bags", "small_bags", "4046" were chosen. For target variable "xlarge_bags" predictors like "4225","4046","total_volume" were chosen. The approach for choosing was simple: if variables have correlation between 0.6 and 0.8 than we choose that variable.

After that the manual Bayesian network was built. On Figure 7 we can see the result.

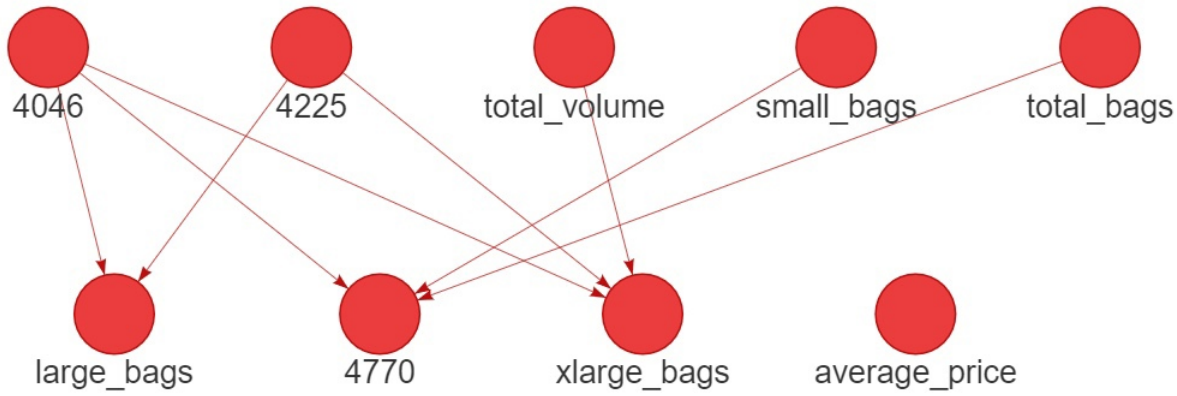

Figure 7: Manual Bayesian network

Than Bayesian network for the same set of variables based on algorithms for structural learning was built. On Figure 8 Bayesian network based on Hill-Climbing with Mutual Information function is shown:

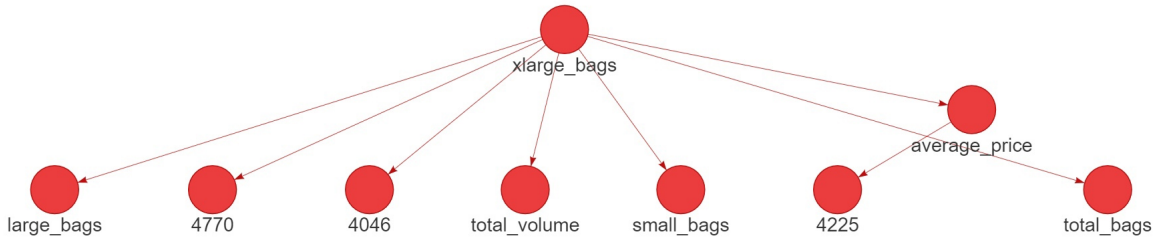

Figure 8: Bayesian network based on Hill-Climbing with Mutual Information function

The last one is Evolutionary algorithm with Mutual Information function. On Figure 9 the result is presented:

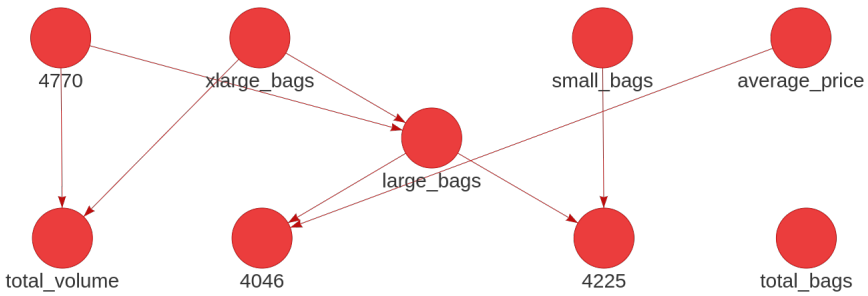

Figure 9: Evolutionary algorithm with Mutual Information function

5. Quality analysis.

In last section we want analyze the quality of our networks. Via methods from BAMT synthetic data was generated and drawn on real data. On Figures 10, 11, 12 results are presented.

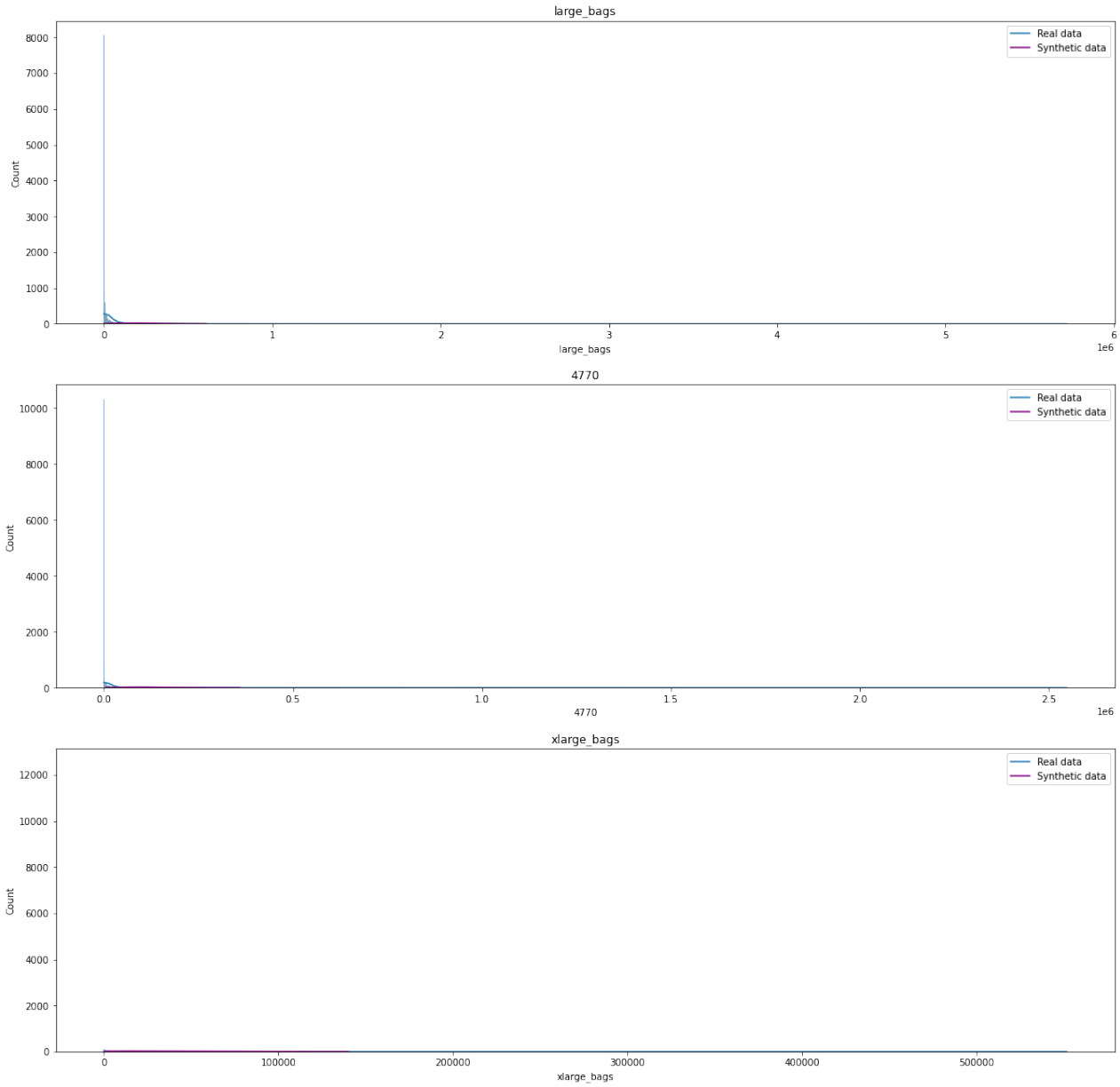

Figure 10: Mutual

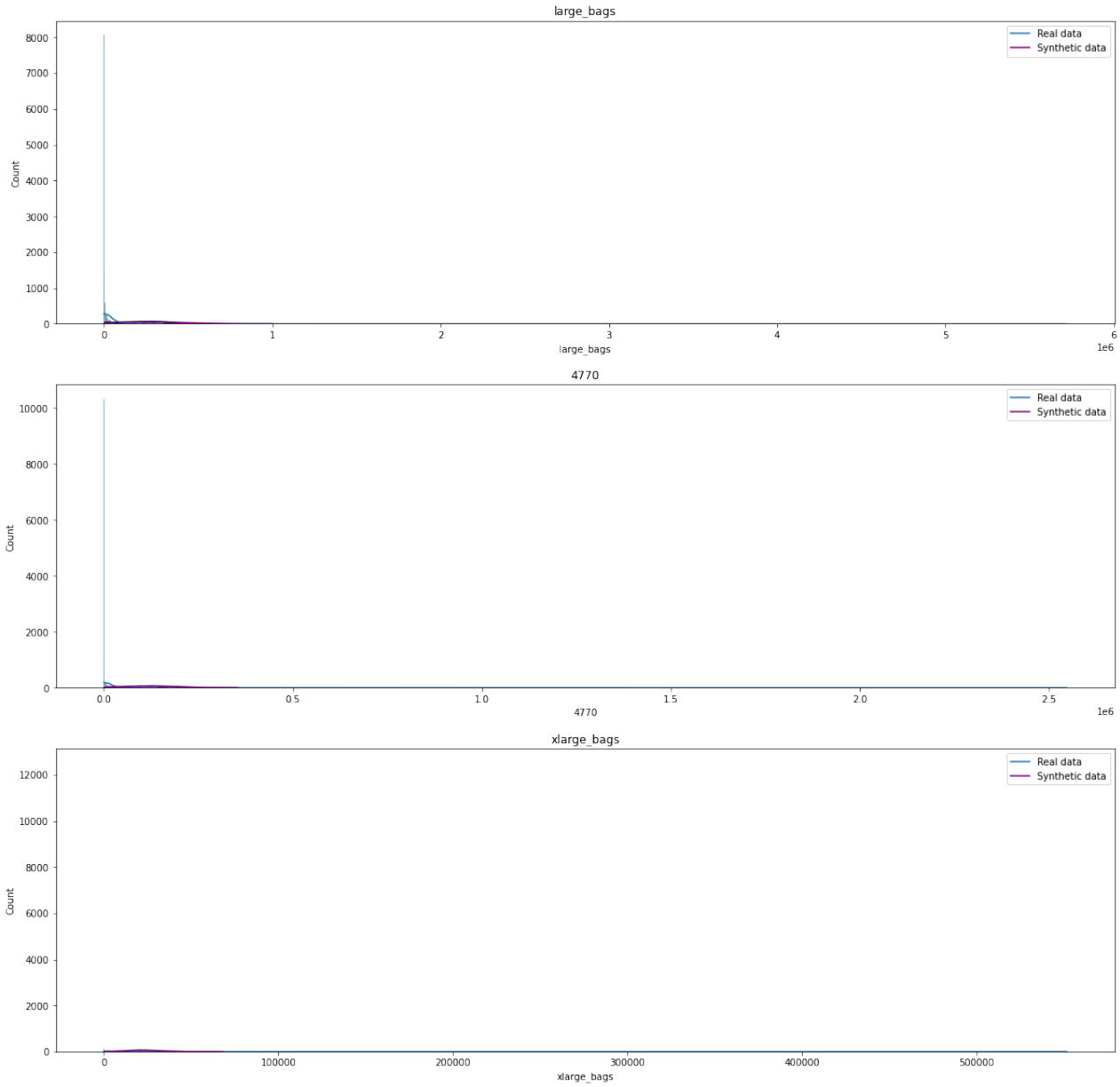

Figure 11: Hill-Climbing with Mutual Information function

The results are not perfect. Generated data does not match with real. Also we can check RMSE values for our networks. As we can see, the best result was provided by manual Bayesian network, since RMSE values for variables are the smallest.

	Manual	Hill-Climbing	Evolutionary
large_bags	163526	230658	182010
xlarge_bags	11256	19133	21936
4770	73650	108952	122924

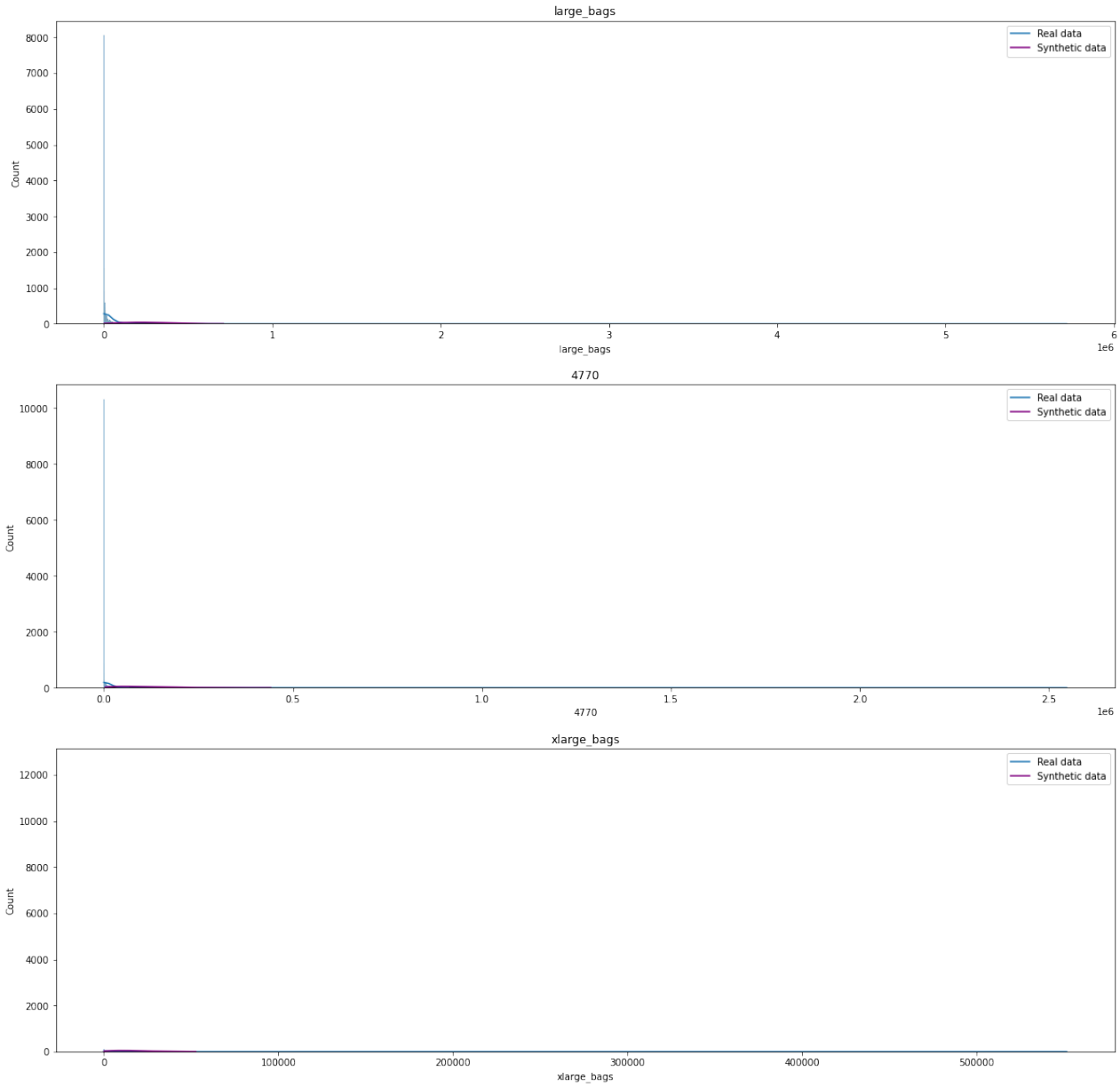

Figure 12: Evolutionary algorithm

Conclusions

In this laboratory work we used several algorithms for sampling data. Then built Bayesian networks using different algorithms and provide quality analysis for these algorithms.

Source code

Source code is located on GitHub: https://github.com/DmitryPogrebnoy/multivariate-data-analysis