Report on learning practice # 2

Analysis of multivariate random variables

Performed by:

Dmitry Pogrebnoy

Group J132c

Saint-Petersburg

2021

Table of contents:

Sourcecode

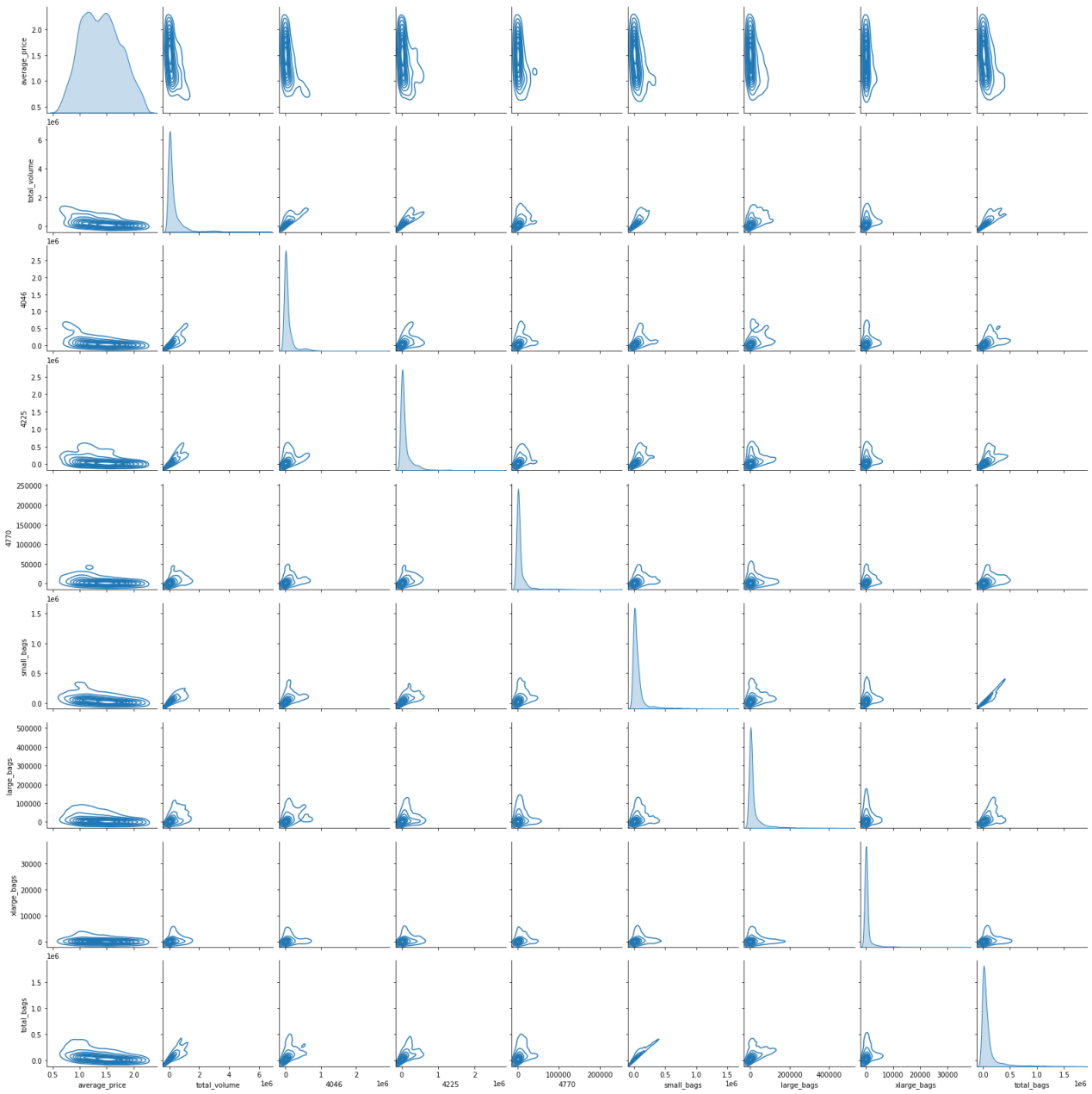
The source code is available at https://github.com/DmitryPogrebnoy/multivariate-data-analysis/blob/main/task2/task2.ipynb

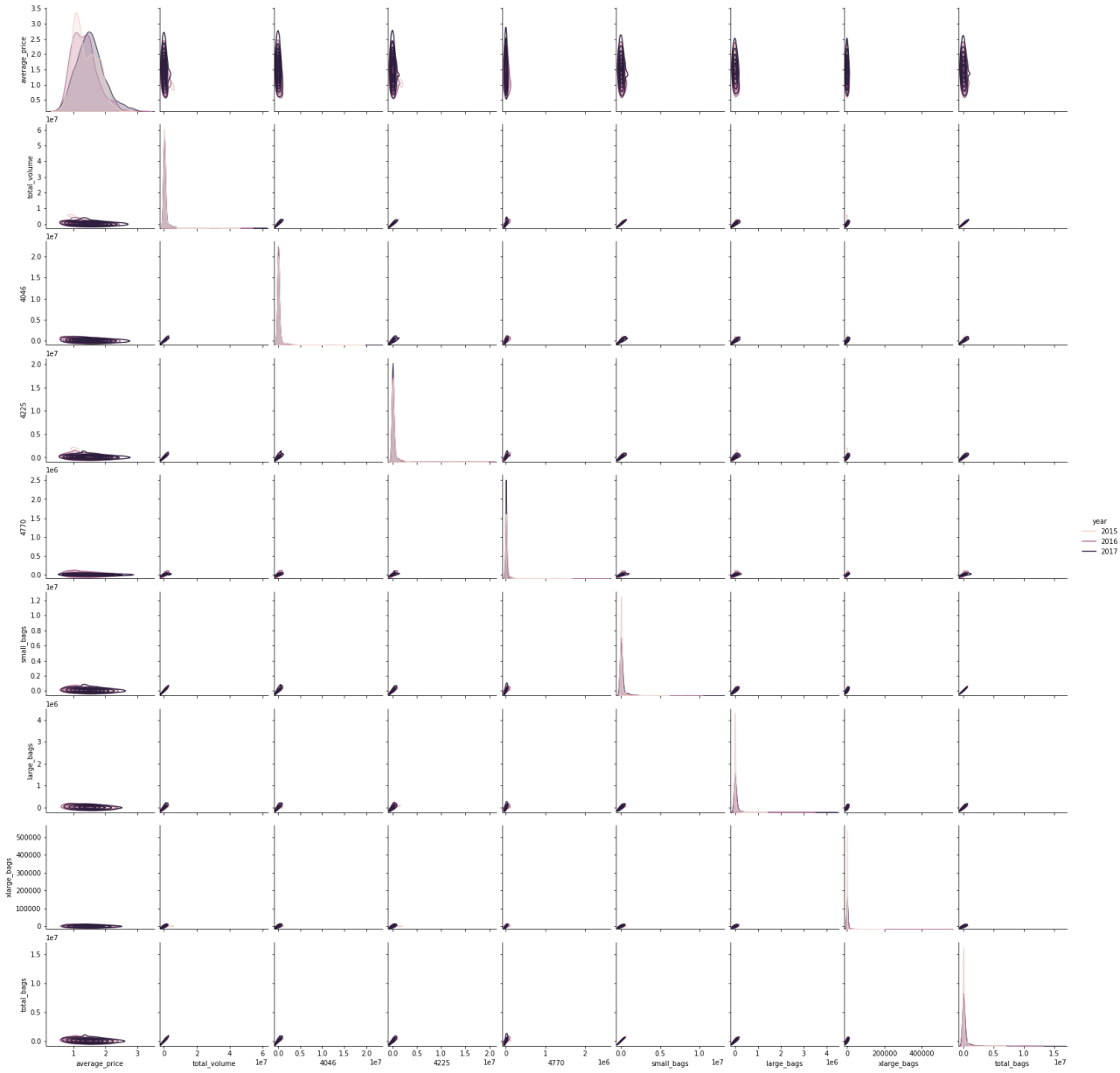
0. Subsample variables

We used a dataset that presents retail scan data and Hass avocado prices from 2015 to 2017. For the second laboratory work, we selected the following variables.

Name	Type	Role	Description
total_bags	continuation	target	number of total bags sold
year	nominal	categorical	year of sale
average_price	continuation	predictor	average selling price
total_volume	continuation	predictor	total sales volume
4046	continuation	predictor	total number of avocados sold with PLU 4046 (PLU - Product Lookup Code)
4225	continuation	predictor	total number of avocados sold with PLU 4225
4770	continuation	predictor	total number of avocados sold with PLU 4770
small_bags	continuation	predictor	number of small bags sold
large_bags	continuation	predictor	number of large bags sold
xlarge_bags	continuation	predictor	number of xlarge bags sold

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


average_price
total_volume
4046
4225
4770
small_bags
large_bags
xlarge_bags
total_bags
1e6



2. Estimation of multivariate mathematical expectation and variance

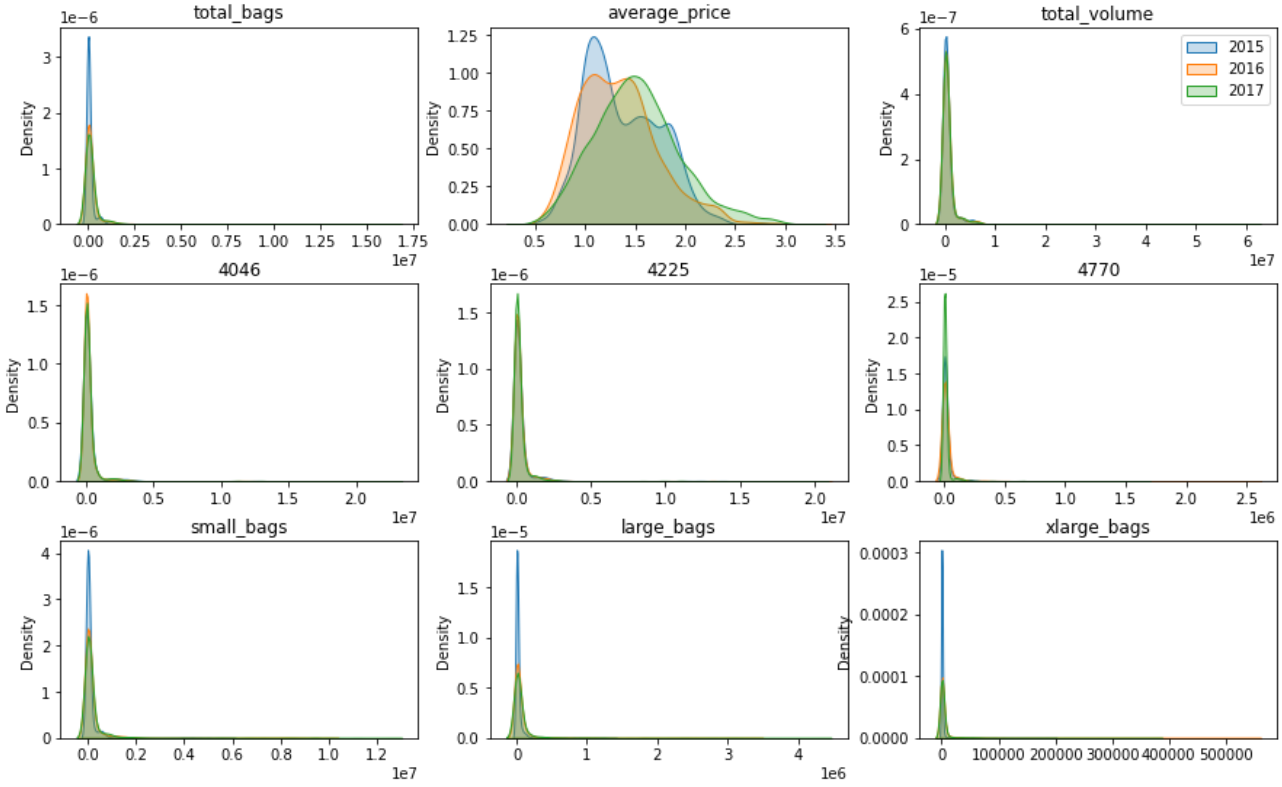
Mathematical expectation:

total_bags	228964.776198
year	2016.006312
average_price	1.410447
total_volume	834109.848000
4046	288244.617776
4225	293665.843429
4770	23233.041193
small_bags	174843.949139
large_bags	51202.252266
xlarge_bags	2918.574026

Variance:

total_bags	8.785458e+11
year	6.687308e-01
average_price	1.671173e-01
total_volume	1.143198e+13
4046	1.545792e+12
4225	1.435183e+12
4770	1.191954e+10
small_bags	5.102487e+11
large_bags	5.118503e+10
xlarge_bags	2.855521e+08

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



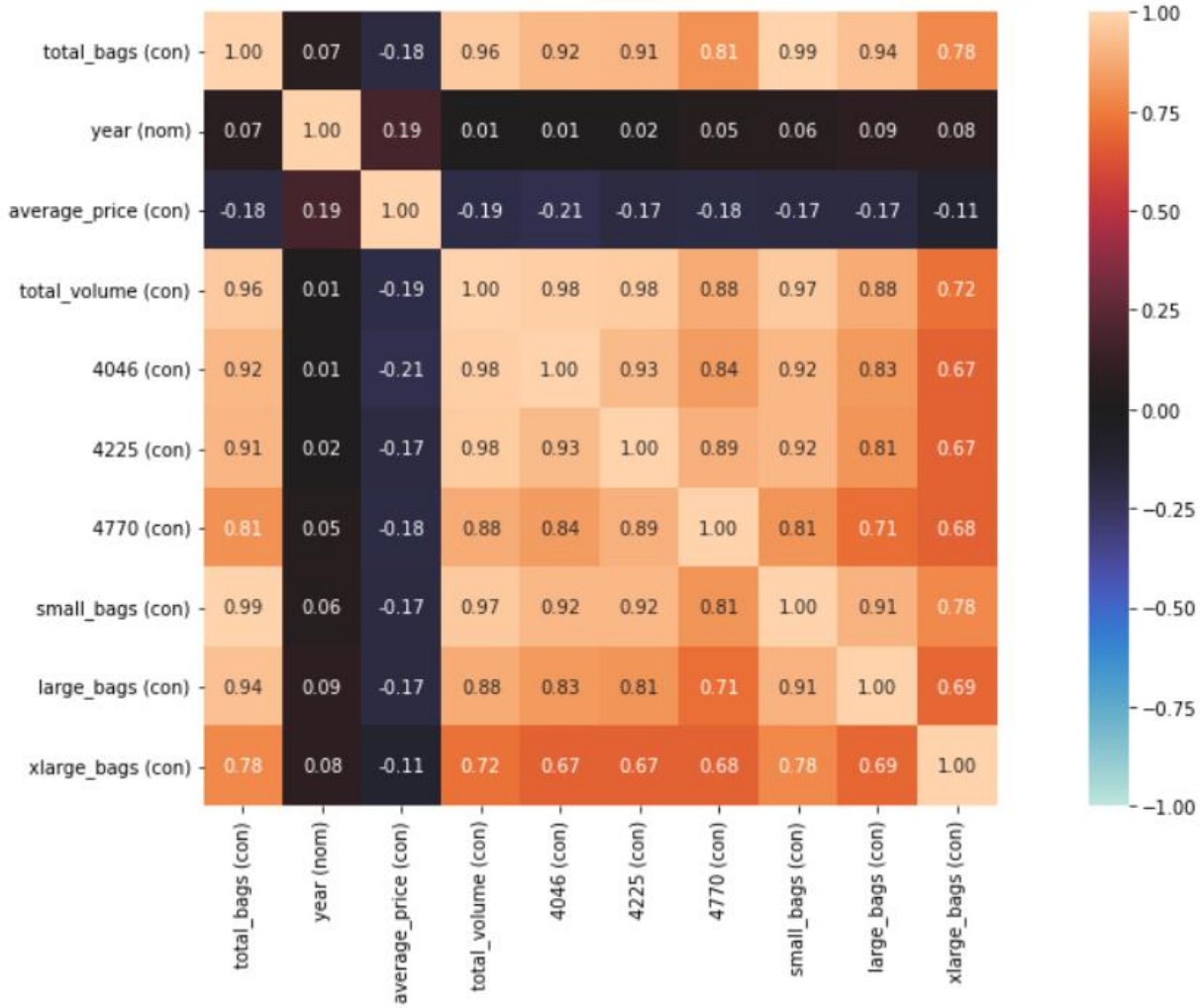
Variable	Categorial value - year	Mean	Variance
total_bags	2015	137523.10543722112	299599598598.933
total_bags	2016	260533.9862037044	1046507501039.5952
total_bags	2017	287712.11942327966	1269484482962.5771
average_price	2015	1.3755903829029397	0.14107143175644488
average_price	2016	1.3386396011395996	0.1550059851385076
average_price	2017	1.515127577770011	0.1874072988757621
total_volume	2015	781027.366347277	10056864000749.59
total_volume	2016	858420.5647845404	12101575879008.2
total_volume	2017	862339.3392642484	12124027367352.16
4046	2015	304443.451707926	1716839111966.685
4046	2016	271567.46657763515	1374727456034.9265
4046	2017	288716.9051939871	1545847476172.71
4225	2015	313633.844366875	1629398194554.395
4225	2016	297850.47863247915	1478065535059.2346
4225	2017	269964.1243498772	1202042604484.1167
4770	2015	25426.962520035537	12004950392.726713
4770	2016	28468.633370726573	18600300656.652557
4770	2017	15941.542778748671	5198121233.735787
small_bags	2015	113033.42925556547	205651755106.05148
small_bags	2016	197025.32756766438	608979971759.93
small_bags	2017	213728.158837819	706680097789.3496
large_bags	2015	23520.28503650937	9435132842.57527
large_bags	2016	59940.58797364673	60918608294.8508
large_bags	2017	69790.11536700439	81446111402.52736
xlarge_bags	2015	969.3906108637585	41184761.5027122
xlarge_bags	2016	3568.0706623931655	383147368.884189
xlarge_bags	2017	4193.8434708144	423892131.4627709

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

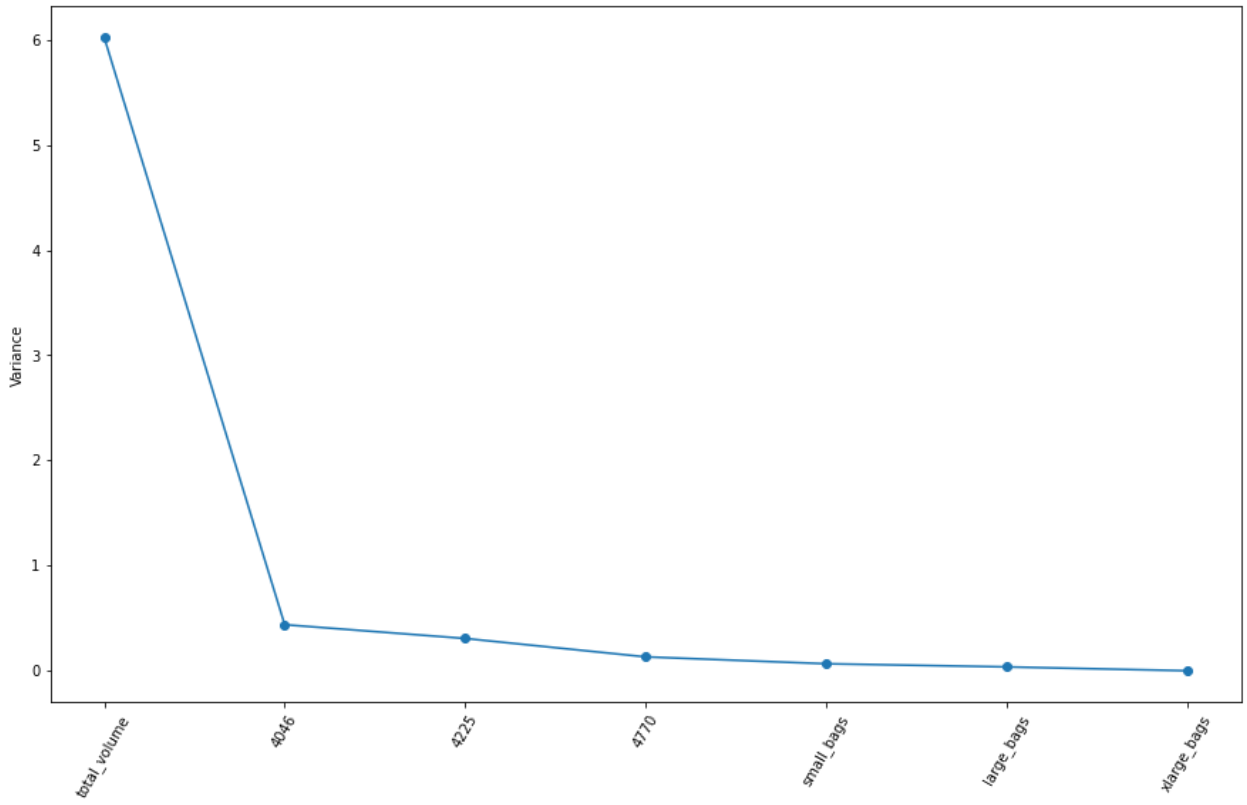
Variable	Target	Pearson corr coef	Significance level	Confidence interval
average_price	total_bags	-0.175604	1.77533e-117	[-0.19015453098830631 ... -0.16097620279925814]
total_volume	total_bags	0.961729	0	[0.9605819237605958 ... 0.9628427318626996]
4046	total_bags	0.915641	0	[0.9131744206914078 ... 0.9180406016773492]
4225	total_bags	0.908233	0	[0.9055601996068435 ... 0.9108332341724441]
4770	total_bags	0.805506	0	[0.8001551304356551 ... 0.8107288844907263]
small_bags	total_bags	0.994748	0	[0.9945881846704384 ... 0.994903647114995]
large_bags	total_bags	0.943722	0	[0.9420513254824099 ... 0.9453453208411005]
xlarge_bags	total_bags	0.783131	0	[0.7772403069704465 ... 0.7888842918518539]

5. Task formulation for regression, multivariate correlation.

Task for regression: predict **total_bags** based on total_volume, year, average_price, 4046, 4225, 4770, small_bags, large_bags, xlarge_bags variables.



We applied PCA and found out which variables to take as predictors for the regression task. To begin with, uncorrelating variables (yaer, average_price) and the target variable (total_bags) were removed from the dataset. Then the values of each variable were standardized and PCA was applied. The result was the following.



As a result, four predictors were taken for the regression task: total_volume, 4046, 4225, 4770.

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

We used a linear regression model, Lasso and Ridge. For Lasso and Ridge, we iterated over the alpha parameter in 0.001 increments and selected the best result for the MSE and R2 metrics. The result was as follows.

Type	Alpha	MSE	R2	Coeff - total_volume, 4046, 4225, 4770
Least Squares model	-	276.903	1	[0.99999561 -0.99999307 -0.99999521 -0.99999079]
Best Lasso model	0.0	2.40876e+06	0.999998	[0.99554331 -0.99436743 -0.99347005 -0.99673191]
Best Ridge model	0.0	276.903	1	[0.99999561 -0.99999307 -0.99999521 -0.99999079]

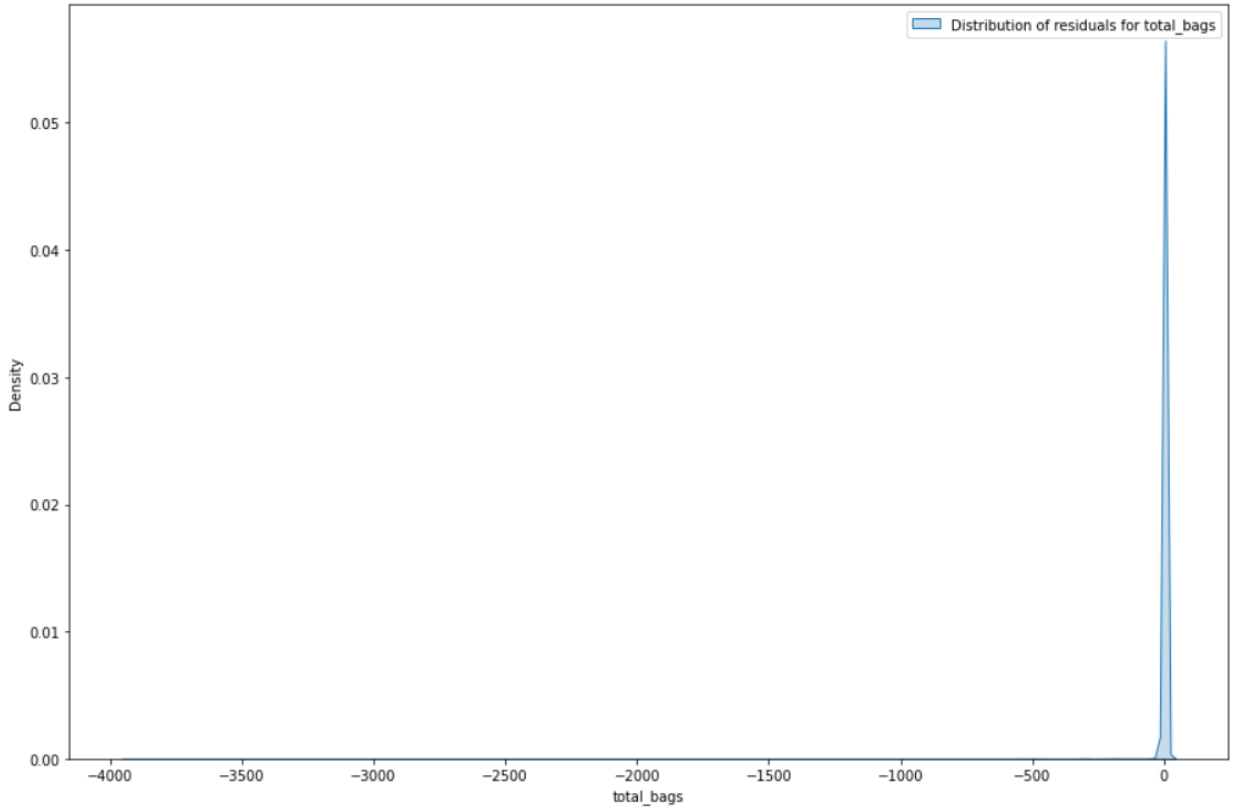
As a result, the linear model showed the best metrics.

7. Quality analysis

Let's analyze the resulting linear regression model, or rather the residuals.

Type	Alpha	MSE	R2	Coeff - total_volume, 4046, 4225, 4770
Least Squares model	-	276.90260839004065	0.9999999997249893	[0.99999561 -0.99999307 -0.99999521 -0.99999079]

The density of residues is shown in the following graph.



The residuals have the following metrics.

```
count    16953.000000
mean         0.193011
std         46.758812
min      -3934.180942
25%          1.717454
50%          1.740998
75%          1.852280
max         23.783233
```

Let's apply the Anderson test to check the distribution of residuals for normality.

```
Statistic: 6205.874
15.0: 0.576, data does not look normal (reject H0)
10.0: 0.656, data does not look normal (reject H0)
5.0: 0.787, data does not look normal (reject H0)
2.5: 0.918, data does not look normal (reject H0)
1.0: 1.092, data does not look normal (reject H0)
```

We also apply K-test to check the distribution of residuals for normality.

```
KstestResult(statistic=0.49636831316265173, pvalue=0.0)
```

According to the test results, it can be said that the residuals are not distributed normally. This means that the linear model does not reflect reality well and the real dependence is most likely nonlinear and requires more complex models.