Report on learning practice # 2

Analysis of multivariate random variables

Performed by:

Dmitry Pogrebnoy

Group J132c

Saint-Petersburg

2021

**Table of contents:**

[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).](#Step1)

[2. Estimation of multivariate mathematical expectation and variance.](#Step2)

[3. Non-parametric estimation of conditional distributions, mathematical expectations and variances.](#Step3)

[4. Estimation of pair correlation coefficients, confidence intervals for them and significance levels.](#Step4)

[5. Task formulation for regression, multivariate correlation.](#Step5)

[6. Regression model,multicollinearity and regularization (if needed).](#Step6)   
[7. Quality analysis.](#Step7)

**Sourcecode**

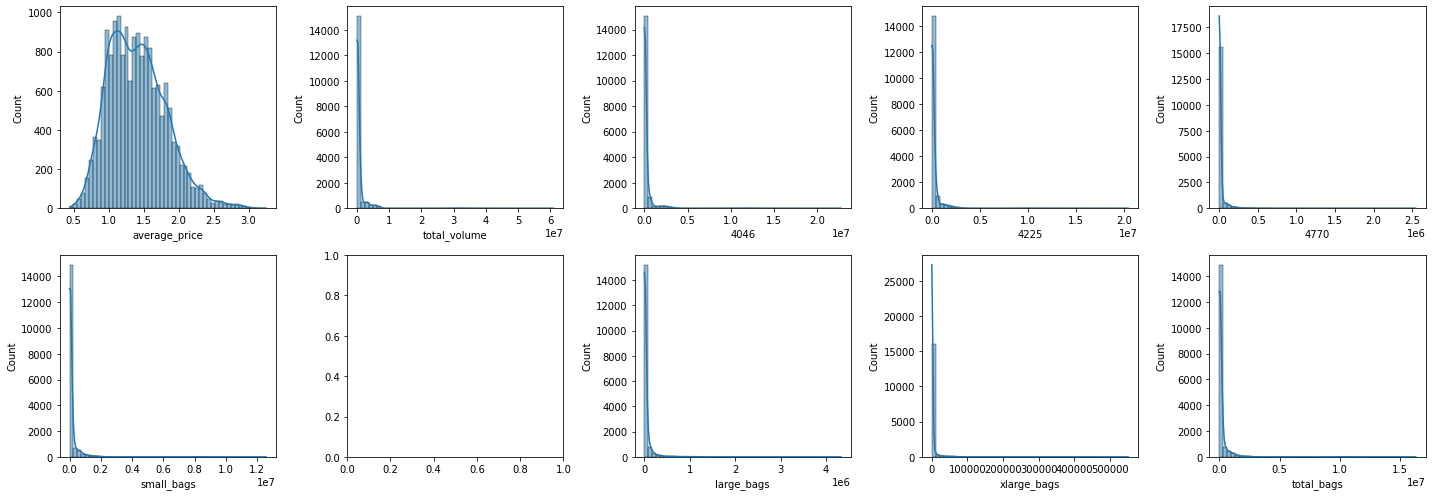
The source code is available at <https://github.com/DmitryPogrebnoy/multivariate-data-analysis/blob/main/task2/task2.ipynb>

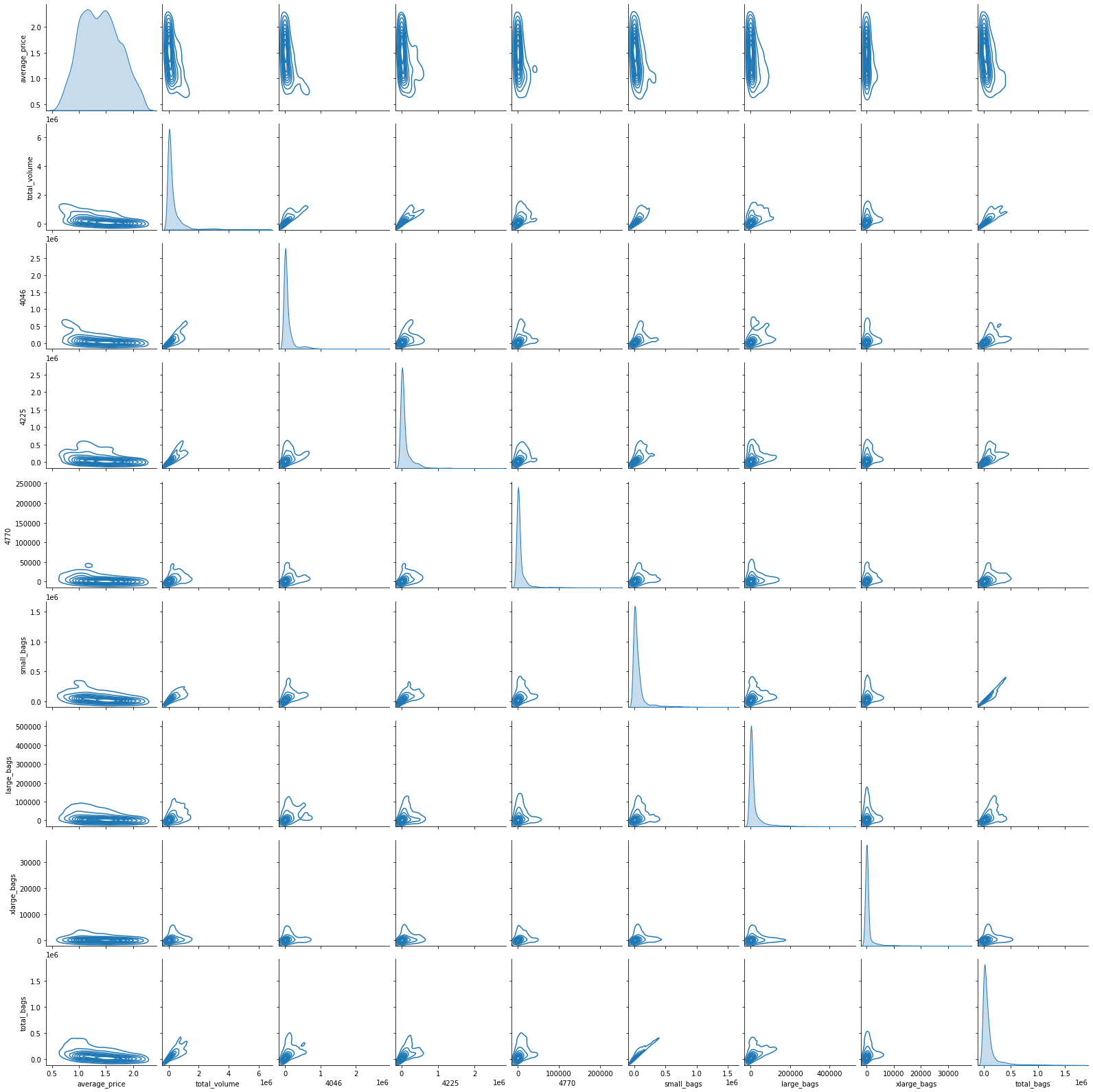
1. Subsample variables

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 |

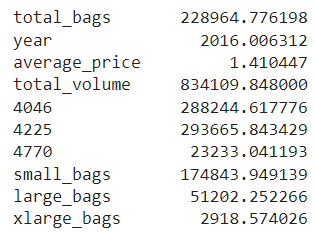
1. Plotting a non-parametric estimation of PDF in form of a histogram and ernel density function for MRV (or probability law in case of discrete MRV).



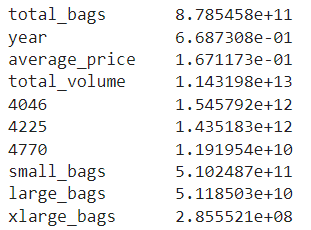


1. Estimation of multivariate mathematical expectation and variance

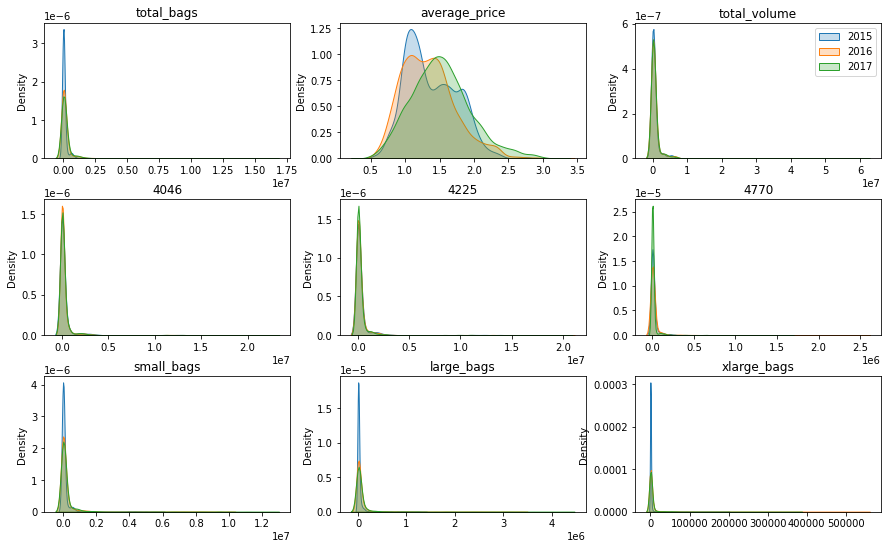
Mathematical expectation:

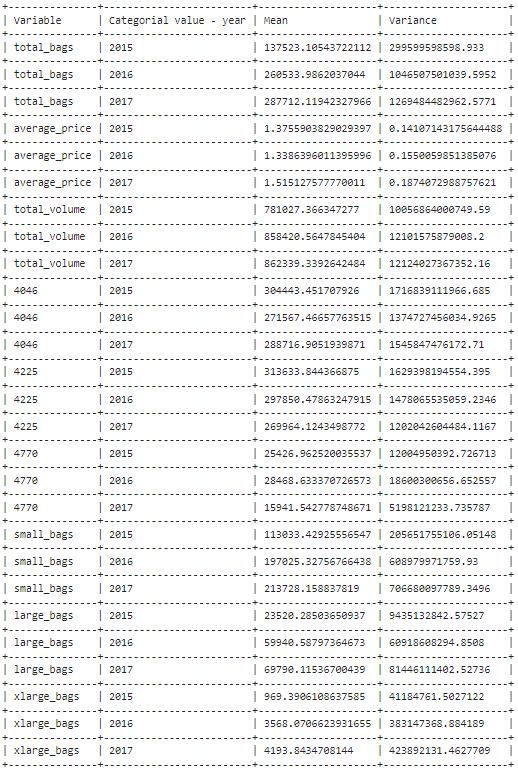


Variance:

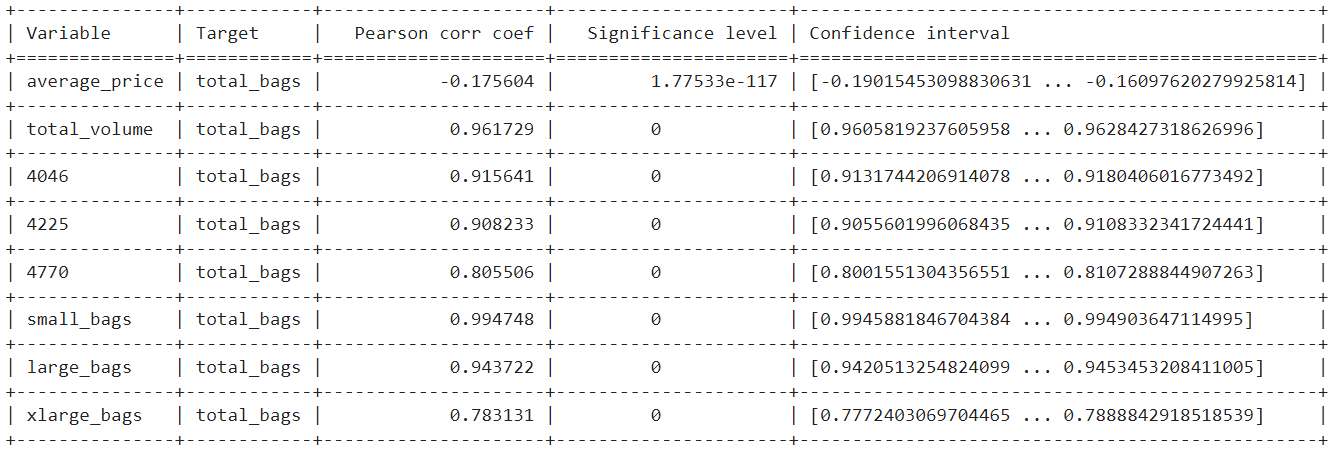


1. Non-parametric estimation of conditional distributions, mathematical expectations and variances.



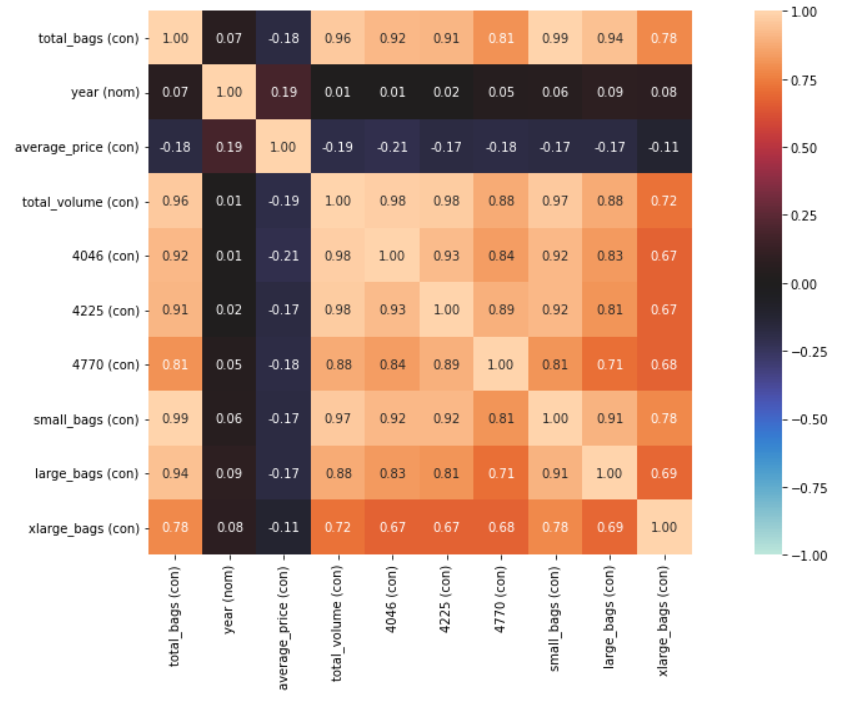


1. Estimation of pair correlation coefficients, confidence intervals for them and significance levels.

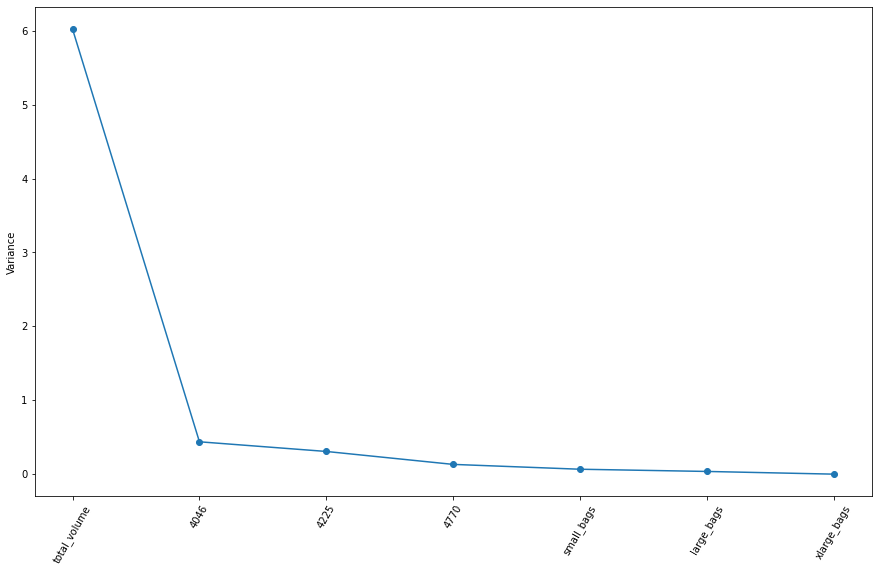


1. Task formulation for regression, multivariate correlation.

Task for regression: predict **total\_bags** based on total\_volume, year, average\_price, 4046, 4225, 4770, small\_bags, large\_bags, xlarge\_bags variables.



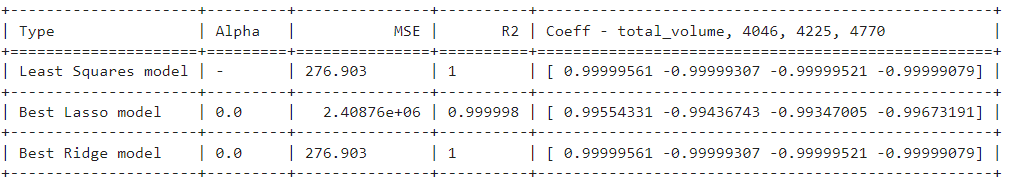
We applied PCA and found out which variables to take as predictors for the regression task. To begin with, uncorrelating variables (yaer, average\_price) and the target variable (total\_bags) were removed from the dataset. Then the values of each variable were standardized and PCA was applied. The result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



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

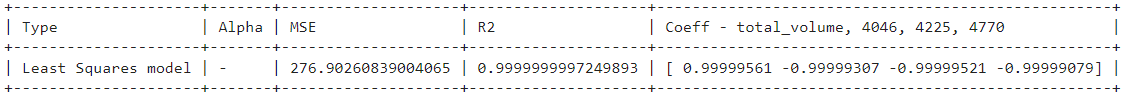
1. Regression model, multicollinearity and regularization (if needed).

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.

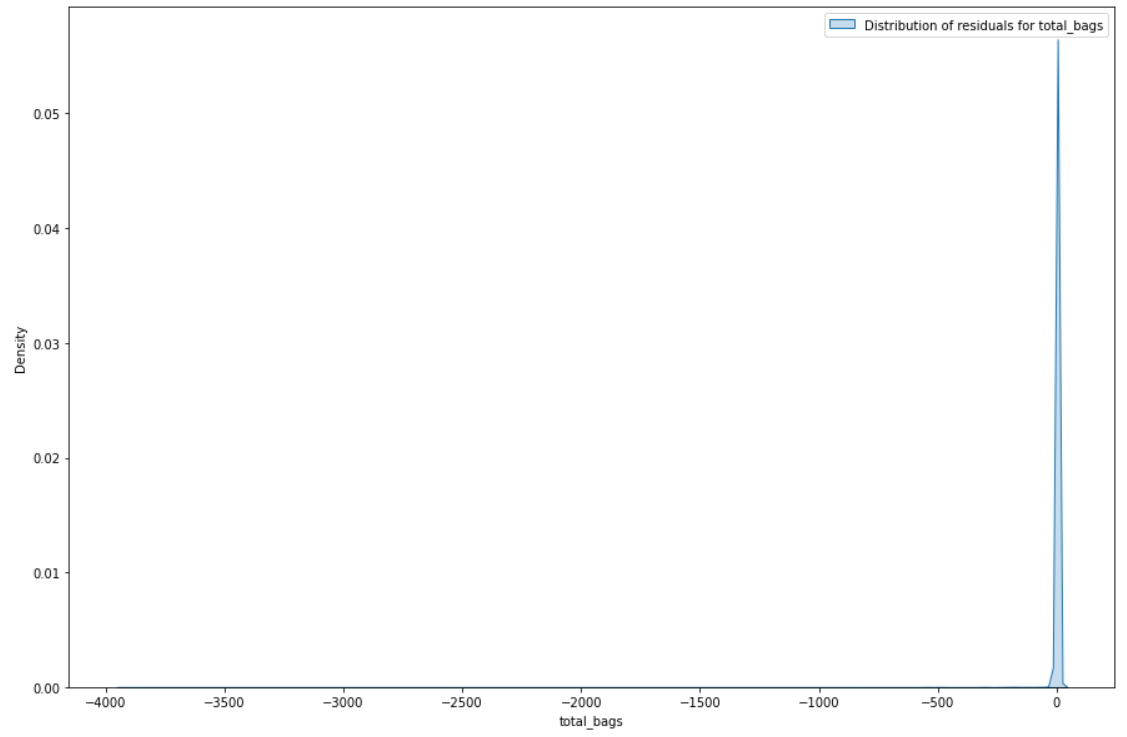


As a result, the linear model showed the best metrics.

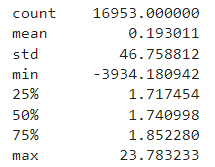
1. Quality analysis

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

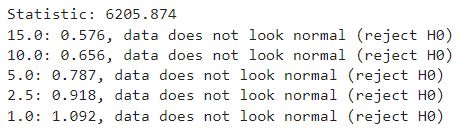
The density of residues is shown in the following graph.



The residuals have the following metrics.



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



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



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.