NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "Data Science and Business Analytics"

Project N2

Made by:

Khaibrakhmanov Timur,

Tsarev Timur,

Malutin Alexander

Moscow 2024

Contents

1 Introduction

Choosing or selling a car is an important part of a person's life. It is often difficult to assess whether the price of a car you are buying is fair or how much to sell your car for. In our paper we will try to solve these problems. For this purpose, we have collected data from the website for selling and buying cars auto.ru, through which more than one and a half million used cars are sold annually. For our analysis we decided to use the most popular BMW brand and X5 model. This model is from the segment where the probability of incorrect data is significantly reduced, but at the same time the number of offers to analyse is quite significant.

2 Data

The data was sparse from the auto.ru website and collected into a dataset. There are 11 columns in the dataframe, which we will describe in more detail below.

- **Price** - our target variable, which is the price of the car being sold.
- **Used** - variable means whether the machine has been used before. 1 marks machines with a mileage of more than 100 km, 0 marks machines with a mileage of less than 100 km.
- **Millage (It should be mileage, but the mistake occurred and "e" got lost in translation:))** - means the mileage of the car being sold.
- **Transmission** - refers to the type of transmission of the vehicle. Automatic or manual.
- **Colour** - means the colour of the car.
- **Engine_capability** - refers to the capabilities of the vehicle's engine.
- **Horsepower** - refers to the amount of horsepower of the vehicle.
- **Fuel_type** - means the type of fuel consumed by the vehicle engine.
- **Tax** - car tax.
- **State** - means whether the vehicle requires repairs for normal use.
- **Year** - the year of manufacture of the vehicle.

In total, our dataset has 11 columns with features and 825 rows. This is enough to draw some conclusions. In the future, we plan to check the data for various outliers and anomalies, as well as analyse the features to improve the performance of the future model.

3 Expected Results

In the research process, we will explore the data, variables interrelationships and their characteristics. Thus, we will follow a standardized algorithm when examining data.
In the research process, we want to analyze competent pricing, understanding all the nuances and identifying the most significant variables that affect the target variable.
As a result, we will try to predict the price of a car under customized conditions. The forecast will be based on a linear regression model. In order to make it reliable, we will experiment with the parameters and other models, compare the results.

4 Preprocessing

Utilizing Python libraries like Pandas, NumPy, Matplotlib, and Seaborn, we began by mounting the parsed dataset, ensuring seamless integration and analysis. Initial data inspection revealed no missing values, and types of data ranged from integers and floats to categorical variables. Furthermore, as the columns 'transmission' and 'state' have contained almost all identical values, it was decided to leave those variables behind, as they did not possess any relevant information for the model anyways.
Then we have removed the duplicates, if any, as well as fuel types with low occurrences. After that we checked for any anomalies such as inadequate year, or negative prices and the refined dataset comprised 820 observations across 11 variables. Further analysis required the implementation of the correlation matrix.

Figure 4.1: Correlation matrix

From this correlation matrix we managed to derive some insights:

- *price* and *used*: negative correlation of **-0.63**. This indicates that newer models, or those manufactured in more recent years, tend to command higher prices in the market.
- *price* and *milage*: negative correlation of **-0.8**, suggesting that cars with higher mileage are usually priced lower, likely due to wear and tear and perceived reduced reliability.
- *price* and *horsepower*, *price* and *tax* have a **weak positive correlation**, indicating that the more horsepower a vehicle has, the higher price on it will be. Furthermore, the possibility of multicollinearity, particularly between 'horsepower' and 'tax', was noted as a potential concern for regression models, as it could distort the results or the interpretation of these 'variables' effects. High correlation coefficient (**0.93**) could be explained by tax on a vehicle being calculated with respect to its horsepower.
- Another concern is the very strong negative correlation between the *milage* and *year* - (**-0.85**), which is reasonable due to fact that the older car is the more it was used by its previous owner(-s).

After deriving some insights from the correlation matrix we continue further with the price distribution:

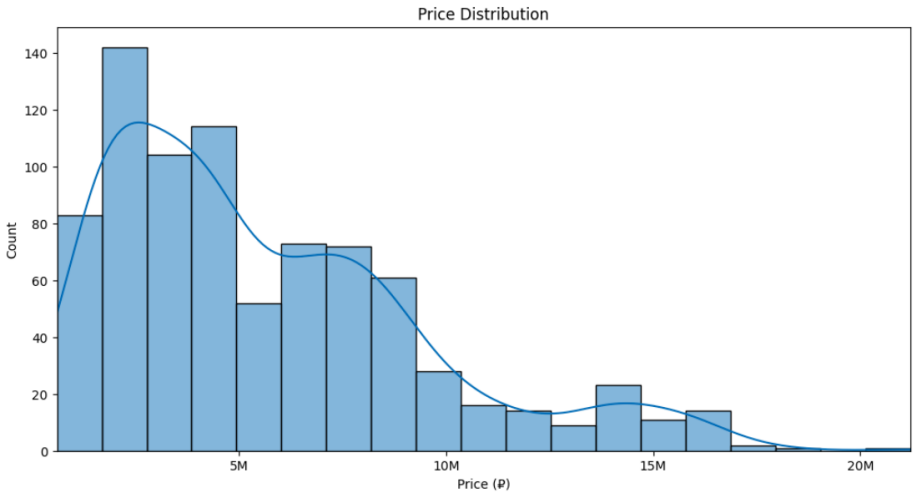

Figure 4.2: Price distribution

The histogram shows the distribution of both used and unused BMW X5 car prices, indicating a right-skewed pattern where most cars are priced lower, with a peak in count between 2M₽ and 3M₽. The long tail towards the higher prices suggests fewer cars are listed at these higher price points. This visualization underscores the prevalence of more affordable options in the market. After we have looked at the price distribution, let us continue with the color vs price and the fuel_type vs price dependency:

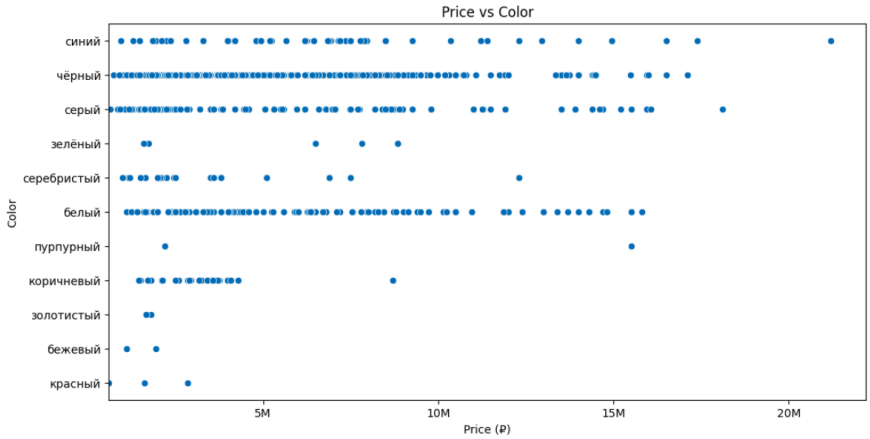

Figure 4.3: Price vs Colour

On the Figure 4.3 we can see the confirmation of the results, acquired in the previous analysis on the price distribution: most of the cars are rather more affordable, which can be accosiated with the specific of the data, most of the vehicles being previously in use (approximately 93.78%). Furthermore, as it can clearly be seen, the prevailing amount of cars are of the either black, or grey, or white color.

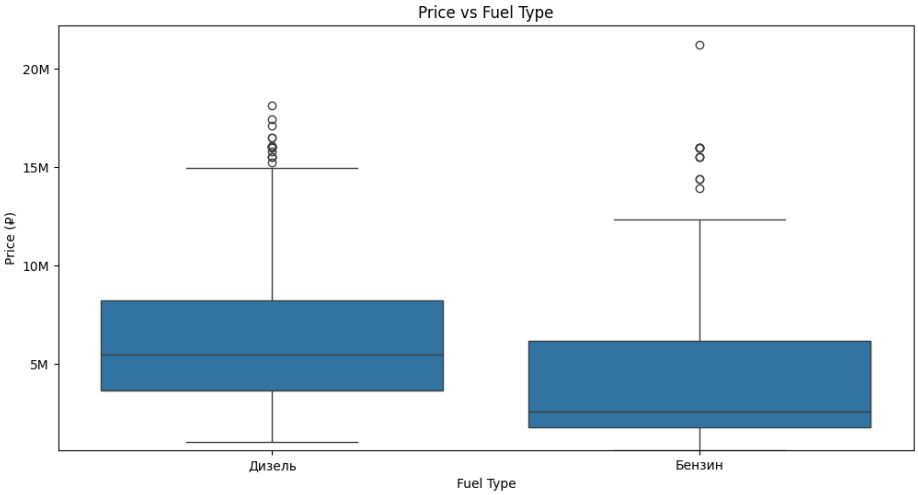

Figure 4.4: Fuel type vs Price

The image is a box plot on the Figure 4.4 that compares the prices of vehicles based on their fuel type: Diesel (Дизель) and Petrol (Бензин). Here are some insights we can obtain from this plot: Diesel vehicles have a higher median price around 5M₽, with an interquartile range extending from approximately 4M₽to 9M₽. This suggests that diesel vehicles tend to be more expensive than petrol vehicles, which have a median price near 2M and an interquartile range less than the diesel cars. What is interesting is that there are several outliers in both categories, which are probably the new, not used cars.

5 Multicollinearity

We should conduct deeper variable analysis. It is extremely important to understand relationships between independent variables and their influence on a dependent variable.

One of the problems is multicollinearity. It is a phenomenon in which there is a strong correlation between traits, thus it is difficult to interpret the constructed model, as some attributes may behave unpredictably and interfere with the adequate perception of the results.

To solve this problem, it was decided to use a heat map based feature exclusion method for correlation coefficients based on threshold. We chose **0.85** as the threshold, which means that the features that correlate more strongly with each other will be considered multicollinear. Earlier we have already constructed a heat map with correlation coefficients (Figure 4.1), so we can use it. The correlation between a car's year of manufacture and its mileage (**-0.85**) is immediately apparent. This is quite obvious, because the older the car is, the more time it has been driven, and therefore the mileage will be higher.

Also, the value of the correlation between horsepower and car tax, namely **0.93**, exceeds the acceptable threshold, and thus is another case of multicollinearity. This correlation also does not raise questions, since the tax is formed based on the horsepower.

So, we have identified the features that cause multicollinearity and now we have to solve this problem by removing one of the two highly correlated features. In the pair 'year' and 'mileage', it was decided to remove the 'year' feature, and in the pair 'horsepower' and 'tax' to remove the 'tax' feature. The decision to remove one or the other attribute was made on the basis of reflection on how important this or that attribute would be in the interpretation of the model and in its further use.

	VIF	variable
0	73.273170	Intercept
1	1.162101	used
2	1.389821	milage
3	1.486004	engine_capacity
4	1.626769	horsepower

Figure 5.1: VIF

To further analyse our traits for multicollinearity, it was decided to perform a VIF test to identify the remaining inappropriate variables on Figure 5.1. The VIF value starts at 1 and has no

upper limit. As we can see, all traits received a VIF value between 1 and 2, indicating a moderate correlation. This means that these attributes will not negatively affect the results of the model. After a quick work on dummifying the qualitative variables we are ready to go to the model itself.

6 Econometric Model Specification

The model specification for predicting the prices of used Ford cars integrates a combination of domain knowledge and insights drawn from exploratory data analysis. The dependent variable in our regression model is the price of the car, which serves as the outcome we seek to predict using various independent variables.

In our regression model aimed at predicting car *prices*, the *price* serves as the dependent variable, influenced by several key factors: the *mileage*, which corresponds to the amount of miles the car has already driven, therefore the higher this number, the lower the price tends to be; the *used* variable, where the value 'True' would typically indicate lower price and vice versa; the *engine size*, where a larger capacity usually suggests a higher price due to increased power, as well as the *horsepower*; the vehicle's *fuel type*, which, as it has showed the data analysis, tends to affect the vehicles price in such a way, that diesel cars would usually be costlier; the *color* variable, representing the color of a car, all of which are quantitatively analyzed using an Ordinary Least Squares regression approach.

The model takes the form: $\mathbf{Price} = \beta_0 + \beta_1 * \mathbf{used} + \beta_2 * \mathbf{milage} + \beta_3 * \mathbf{engine\_capacity} + \beta_4 * \mathbf{horsepower} + \beta_5 * \mathbf{fuel\_type\_Дизель} + \sum_i^n \beta_i * \mathbf{colour}_i$

for i in {белый, черный, синий, серый, зеленый, золотистый, коричневый, красный, пурпурный, серебристый}

The Ordinary Least Squares (OLS) method is utilized to estimate the model parameters. The presence of a constant term (β_0) accommodates the baseline price level. Each coefficient (β_j) represents the expected change in the price given a one-unit change in the predictor, holding all other variables constant.

The use of statsmodels library in Python facilitates the model fitting, with sm.OLS() being employed to perform the regression analysis. The model's adequacy is evaluated based on statistical significance of the coefficients (p-values), the overall model fit (R-squared), and other diagnostic tests. Model summary is above:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.830
Model:	OLS	Adj. R-squared:	0.826
Method:	Least Squares	F-statistic:	208.5
Date:	Fri, 17 May 2024	Prob (F-statistic):	1.50e-234
Time:	17:55:21	Log-Likelihood:	-10296.
No. Observations:	656	AIC:	2.062e+04
Df Residuals:	640	BIC:	2.070e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.733e+07	1.72e+06	10.062	0.000	1.39e+07	2.07e+07
used	-6.019e+06	2.78e+05	-21.656	0.000	-6.56e+06	-5.47e+06
milage	-21.0403	0.780	-26.963	0.000	-22.573	-19.508
engine_capacity	-8.038e+05	1.76e+05	-4.557	0.000	-1.15e+06	-4.57e+05
horsepower	1.334e+04	1362.631	9.789	0.000	1.07e+04	1.6e+04
fuel_type_Дизель	1.525e+06	1.56e+05	9.758	0.000	1.22e+06	1.83e+06
colour_белый	-5.925e+06	1.64e+06	-3.619	0.000	-9.14e+06	-2.71e+06
colour_зелёный	-5.527e+06	1.88e+06	-2.943	0.003	-9.21e+06	-1.84e+06
colour_золотистый	-5.287e+06	2.28e+06	-2.323	0.021	-9.76e+06	-8.17e+05
colour_коричневый	-6.433e+06	1.65e+06	-3.893	0.000	-9.68e+06	-3.19e+06
colour_красный	-5.05e+06	1.81e+06	-2.790	0.005	-8.6e+06	-1.5e+06
colour_пурпурный	-3.643e+06	1.98e+06	-1.841	0.066	-7.53e+06	2.42e+05
colour_серебристый	-5.624e+06	1.66e+06	-3.395	0.001	-8.88e+06	-2.37e+06
colour_серый	-5.297e+06	1.64e+06	-3.238	0.001	-8.51e+06	-2.09e+06
colour_синий	-5.043e+06	1.65e+06	-3.065	0.002	-8.27e+06	-1.81e+06
colour_чёрный	-5.611e+06	1.63e+06	-3.437	0.001	-8.82e+06	-2.4e+06

Omnibus:	104.915	Durbin-Watson:	1.932
Prob(Omnibus):	0.000	Jarque-Bera (JB):	213.139
Skew:	0.906	Prob(JB):	5.22e-47
Kurtosis:	5.124	Cond. No.	1.51e+07

Now it is essential to first of all check whether there are probable non-linear dependencies. The results of the Ramsey RESET test with a high F-statistic and a p-value of almost zero indicate

```
import statsmodels.stats.diagnostic as dg

reset = dg.linear_reset(results, power=3, test_type= 'fitted', use_f=True)

print ('== Correct Functional Form Ramsey-RESET Test ==')
print ('')
print ('Ramsey-RESET Test F-Statistic:', np.round (reset.fvalue, 6))
print ('Ramsey-RESET Test P-Value:', np.round (reset.pvalue, 6))
```

```
== Correct Functional Form Ramsey-RESET Test ==

Ramsey-RESET Test F-Statistic: 6975.554903
Ramsey-RESET Test P-Value: 0.0
```

that our linear model for predicting car prices is misspecified. This suggests that the model may be missing important non-linear relationships or key variables, indicating the need to develop a more complex model or include additional variables.

7 Heteroscedasticity

We have conducted a heteroskedasticity test on our model's residuals to check the assumption of homoskedasticity. Homoskedasticity is an important assumption in ordinary least squares (OLS) regression, which states that the variance of the errors is constant across all levels of the independent variables. Initially, we used the Breusch-Pagan test. It uses the squared residuals

from the regression and regresses them on a set of explanatory variables, in our case, the same variables used in the original regression.

The results of the Breusch-Pagan test are as follows:

- **LM Statistic:** 21.35
- **LM-Test p-value**: 0.13
- **F-Statistic**: 1.44
- **F-Test p-value**: 0.12

Both the LM-Test p-value and the F-Test p-value are greater than 0.05, suggesting that we fail to reject the null hypothesis of homoskedasticity at 5% confidence level. This means that there is not enough evidence to suggest that heteroskedasticity is present in our model's residuals, which is a good sign for the validity of our OLS regression model. However, the values are quite close to the threshold value, thus it was decided to conduct the Glejser test to confirm the results. The results of the test are as follows:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.053
Model:	OLS	Adj. R-squared:	0.031
Method:	Least Squares	F-statistic:	2.393
Date:	Fri, 17 May 2024	Prob (F-statistic):	0.00225
Time:	19:33:44	Log-Likelihood:	-9984.8
No. Observations:	656	AIC:	2.000e+04
Df Residuals:	640	BIC:	2.007e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9573.3180	1.07e+06	0.009	0.993	-2.09e+06	2.11e+06
x1	3.93e+05	1.73e+05	2.273	0.023	5.35e+04	7.32e+05
x2	-1.4499	0.485	-2.987	0.003	-2.403	-0.497
x3	7.065e+04	1.1e+05	0.644	0.520	-1.45e+05	2.86e+05
x4	584.9139	847.644	0.690	0.490	-1079.585	2249.413
x5	3.985e+04	9.72e+04	0.410	0.682	-1.51e+05	2.31e+05
x6	6.228e+05	1.02e+06	0.611	0.541	-1.38e+06	2.62e+06
x7	1.861e+05	1.17e+06	0.159	0.873	-2.11e+06	2.48e+06
x8	-3.875e+05	1.42e+06	-0.274	0.784	-3.17e+06	2.39e+06
x9	1.901e+04	1.03e+06	0.018	0.985	-2e+06	2.04e+06
x10	1.631e+06	1.13e+06	1.449	0.148	-5.8e+05	3.84e+06
x11	9.473e+05	1.23e+06	0.770	0.442	-1.47e+06	3.36e+06
x12	9.408e+05	1.03e+06	0.913	0.362	-1.08e+06	2.96e+06
x13	6.175e+05	1.02e+06	0.607	0.544	-1.38e+06	2.62e+06
x14	7.613e+05	1.02e+06	0.744	0.457	-1.25e+06	2.77e+06
x15	6.389e+05	1.02e+06	0.629	0.529	-1.36e+06	2.63e+06

Omnibus:	317.177	Durbin-Watson:	1.863
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2340.019
Skew:	2.028	Prob(JB):	0.00
Kurtosis:	11.316	Cond. No.	1.51e+07

Figure 7.1: Results of the Glejster test

The small p-values for x1 and x2 (<0.05), which suggests that these variables might be contributing to heteroskedasticity in your model. Thus, we can reject the null hypothesis of homoscedasticity. This means there is evidence of heteroskedasticity in our model. This is an important finding as

it suggests that we may need to adjust our model or use techniques robust to heteroskedasticity for reliable inference.

Considering that according to the Glejster test there is enough evidence of heteroskedasticity, we need to adjust our approach to ensure that our model provides reliable results. For this purpose we introduce the robust standard errors, which adjust the standard errors of the coefficients to account for the heteroskedasticity.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.830
Model:	OLS	Adj. R-squared:	0.826
Method:	Least Squares	F-statistic:	46.82
Date:	Fri, 17 May 2024	Prob (F-statistic):	2.71e-88
Time:	19:31:10	Log-Likelihood:	-10296.
No. Observations:	656	AIC:	2.062e+04
Df Residuals:	640	BIC:	2.070e+04
Df Model:	15		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.733e+07	1.1e+09	0.016	0.987	-2.14e+09	2.17e+09
used	-6.019e+06	3.08e+05	-19.515	0.000	-6.62e+06	-5.41e+06
milage	-21.0403	1.204	-17.482	0.000	-23.399	-18.681
engine_capacity	-8.038e+05	1.66e+05	-4.840	0.000	-1.13e+06	-4.78e+05
horsepower	1.334e+04	1377.960	9.680	0.000	1.06e+04	1.6e+04
fuel_type_Дизель	1.525e+06	1.59e+05	9.617	0.000	1.21e+06	1.84e+06
colour_белый	-5.925e+06	1.1e+09	-0.005	0.996	-2.16e+09	2.15e+09
colour_зелёный	-5.527e+06	1.1e+09	-0.005	0.996	-2.16e+09	2.15e+09
colour_золотистый	-5.287e+06	1.43e+09	-0.004	0.997	-2.8e+09	2.79e+09
colour_коричневый	-6.433e+06	1.1e+09	-0.006	0.995	-2.16e+09	2.15e+09
colour_красный	-5.05e+06	1.1e+09	-0.005	0.996	-2.16e+09	2.15e+09
colour_пурпурный	-3.643e+06	1.1e+09	-0.003	0.997	-2.16e+09	2.15e+09
colour_серебристый	-5.624e+06	1.1e+09	-0.005	0.996	-2.16e+09	2.15e+09
colour_серый	-5.297e+06	1.1e+09	-0.005	0.996	-2.16e+09	2.15e+09
colour_синий	-5.043e+06	1.1e+09	-0.005	0.996	-2.16e+09	2.15e+09
colour_чёрный	-5.611e+06	1.1e+09	-0.005	0.996	-2.16e+09	2.15e+09

Omnibus:	104.915	Durbin-Watson:	1.932
Prob(Omnibus):	0.000	Jarque-Bera (JB):	213.139
Skew:	0.906	Prob(JB):	5.22e-47
Kurtosis:	5.124	Cond. No.	1.51e+07

Figure 7.2: OLS with robust standard errors

8 Potential Issues

Endogeneity and **exogeneity** are key concepts in econometrics that relate to the nature of variables in a regression model. Any variable that has a correlation with the model error is said to be endogenous. Measurement mistakes, reverse causation, or missing variables can all contribute to this association and produce results that are skewed and inconsistent. **The mileage** of an automobile in a model used to estimate its value is an example of a potential endogenous variable because it might depend on a variety of variables not included in the model, such as service quality or operating circumstances.

Conversely, an exogenous variable does not lead to bias in the estimations because it is not connected with model error. It only has an impact on the dependent variable through the independent variables that are part of the model. Engine displacement, which influences a car's value only through mileage and not directly through unexplained factors, is an example of an exogenous

variable.

9 Results

Our econometric study consisted of several important steps in studies of this nature: data preprocessing, multicollinearity testing, econometric model implementation, heteroskedasticity testing, and quality of model's prediction power.

During preprocessing, we examined our independent variables, checked the realism of the variables, and looked for missing values and duplicates. Next, we examined the dependencies of the variables by constructing a heat map with correlations between our variables. After this step, we arrived at multicollinearity. We set the threshhold at 0.85, which means that at or above this correlation, we will remove one of the highly correlated variables.

A critical element of such studies is to verify the existence of heteroskedasticity. We used several approaches for a deeper investigation: Breusch-Pagan test and Glejser test, which has shown the presence of heteroskedasticity.

We chose OLS as our model - as an efficient and simple model. The result of the model was a value of $R^2 = 0.83$. This means that 83% of the variation in car price is explained by the variables in the model. In other words, it means that our model has a high explanatory power.

References

[1] Matt Bogard. "Econometric Research Methods for Agricultural Economics". In: 2005. URL: https://api.semanticscholar.org/CorpusID:113406493.

[2] Christopher Dougherty. *Introduction to econometrics*. English. Fifth. Oxford University Press, 2021. ISBN: 0199676828;9780199676828;

[3] *Heteroskedasticity: Definition, Overview Example*. URL: https://www.freshbooks.com/glossary/financial/heteroskedasticity.

[4] Sabyasachee Mishra. "Statistical and Econometric Methods for Transportation Data Analysis". In: 2014. URL: https://api.semanticscholar.org/CorpusID:123235683.

[5] I. Lesmana Osly Usman. *The Effect of Availability of Teaching Materials, the Use of the Learning Method and the Learning Stimulus of Learning Motivation by Osly Usman, I. Lesmana*. URL: https://www.semanticscholar.org/paper/The-Effect-of-Availability-of-Teaching-Materials%2C-Usman-Lesmana/fde3cfbe81cb0f17490559cc11e0cbda791e6777.

[6] Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[7] Songhao Wu. *Multicollinearity in Regression Why it is a problem? How to check and fix it by Songhao Wu*. URL: https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea.