A model development Analysis

ReCell and Supervised Learning – Foundations

04/03/2022

Contents / Agenda

Business Problem Overview and Solution Approach

- The used and refurbished device market has grown considerably over the past decade, and a new IDC (International Data Corporation) forecast predicts that the used phone market would be worth $52.7bn by 2023 with a compound annual growth rate (CAGR) of 13.6% from 2018 to 2023. This growth can be attributed to an uptick in demand for used phones and tablets that offer considerable savings compared with new models. The rising potential of this comparatively under-the-radar market fuels the need for an ML-based solution to develop a dynamic pricing strategy for used and refurbished devices.

- Based on the data provided an ML Solution based on a linear regression model was built to predict the price of a used phone/tablet and identify factors that significantly influence it.

Executive Summary

- A linear regression model was developed in order to predict the prices of the used price market of the different cellphones. As a result of the model developed, one of the variables with the most impact is the released price of the cellphone. This is understandable as this gives a perspective of the quality and features of the cellphone.

- The model highlights two brands with a negative impact on the model which are Samsung and Sony. The brands Nokia and Xiaomi are shown on the model as the ones preferred by the consumers of the used market.

- The model currently considers 4g as a variable of interest for the model. This variable is expected to decline in the future as more cellphones will use a 5g network as this type of connection becomes more popular around the world.

- Due to the changes and disruptions that occur on the market of cellphones it is recommended to update the database constantly and run the model at least every semester in order to update the parameters of the model to obtain the maximum benefit of the market.

EDA Results

- The provided data set contains 3454 rows and 15 variables.
- The following table shows a statistical summary of the data:

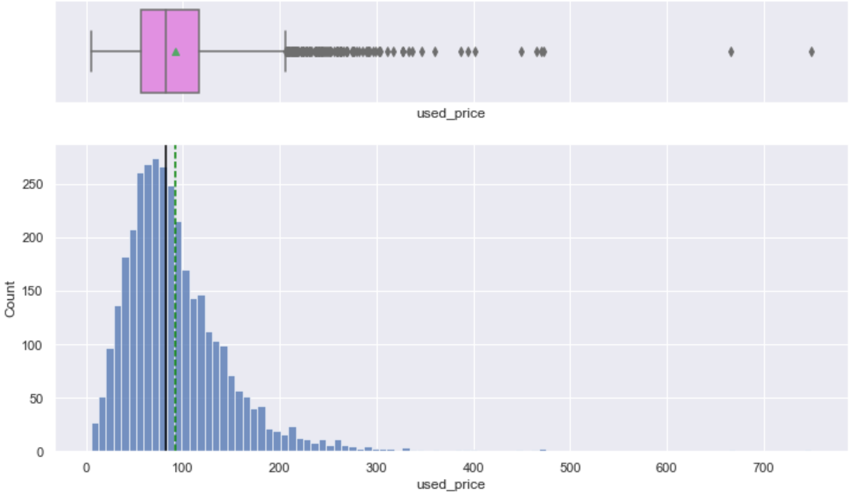
Variables	count	mean	std	min	25%	50%	75%	max
screen_size	3454.0	13.713115	3.805280	5.08	12.7000	12.830	15.340	30.71
main_camera_mp	3275.0	9.460208	4.815461	0.08	5.0000	8.000	13.000	48.00
selfie_camera_mp	3452.0	6.554229	6.970372	0.00	2.0000	5.000	8.000	32.00
int_memory	3450.0	54.573099	84.972371	0.01	16.0000	32.000	64.000	1024.00
ram	3450.0	4.036122	1.365105	0.02	4.0000	4.000	4.000	12.00
battery	3448.0	3133.402697	1299.682844	500.00	2100.0000	3000.000	4000.000	9720.00
weight	3447.0	182.751871	88.413228	69.00	142.0000	160.000	185.000	855.00
release_year	3454.0	2015.965258	2.298455	2013.00	2014.0000	2015.500	2018.000	2020.00
days_used	3454.0	674.869716	248.580166	91.00	533.5000	690.500	868.750	1094.00
new_price	3454.0	237.038848	194.302782	18.20	120.3425	189.785	291.115	2560.20
used_price	3454.0	92.302936	54.701648	4.65	56.4825	81.870	116.245	749.52

Link to Appendix slide on data background check

EDA Results – Univariate Analysis

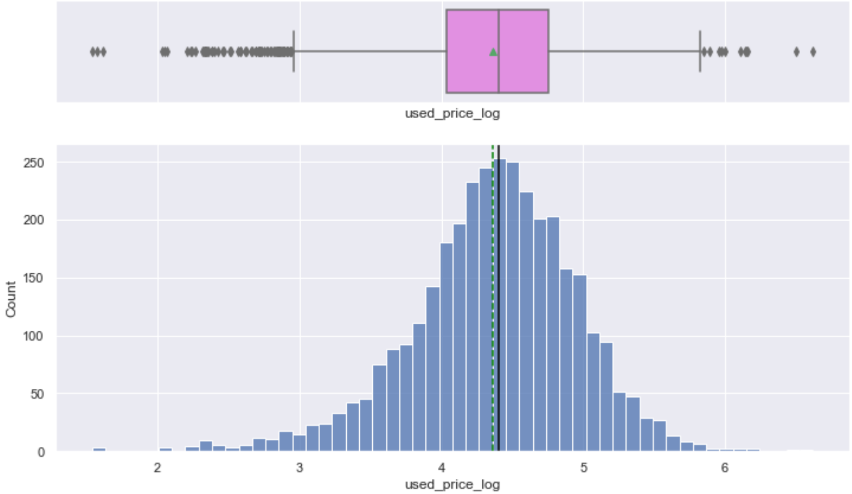
- Used_price

The used price variables show how this variable is skewed to the right. Therefore, there are many outliers for this variable.



- Used_price_log

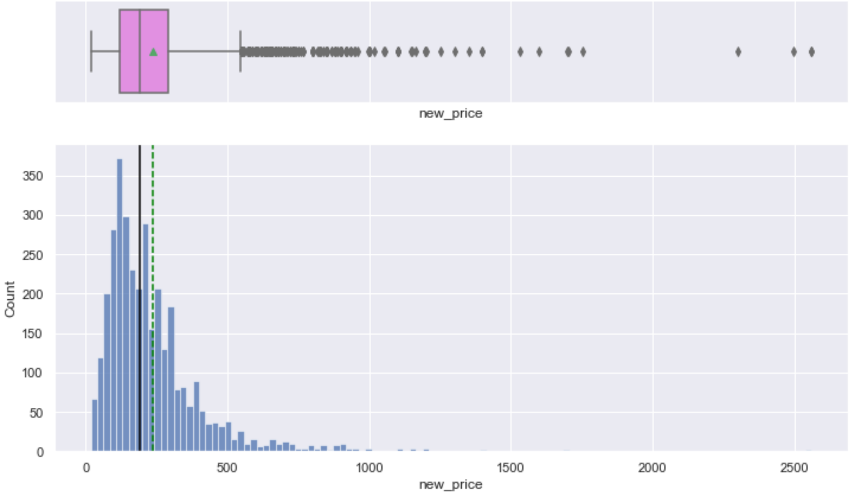
A log transformation was used. This shows a normal distribution for the used price variable, a little bit skewed to the left.



EDA Results – Univariate Analysis

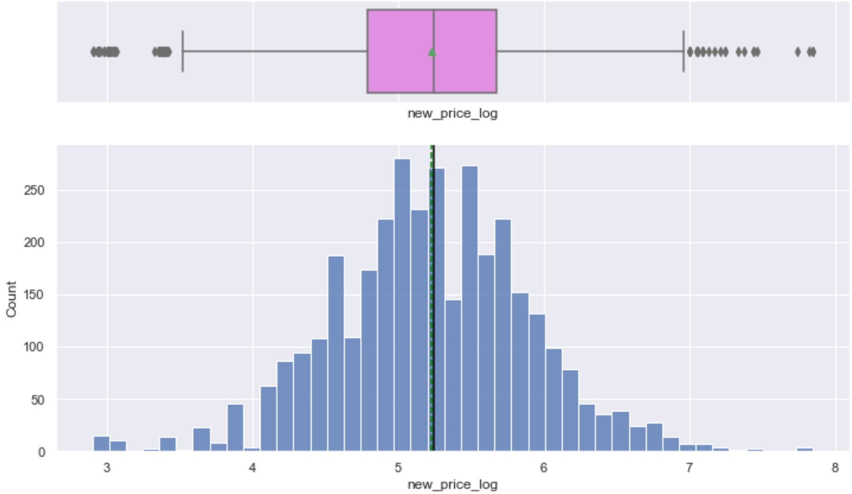
- New_price

The new price variables show how this variable is skewed to the right. Therefore, there are many outliers for this variable.



- New_price_log

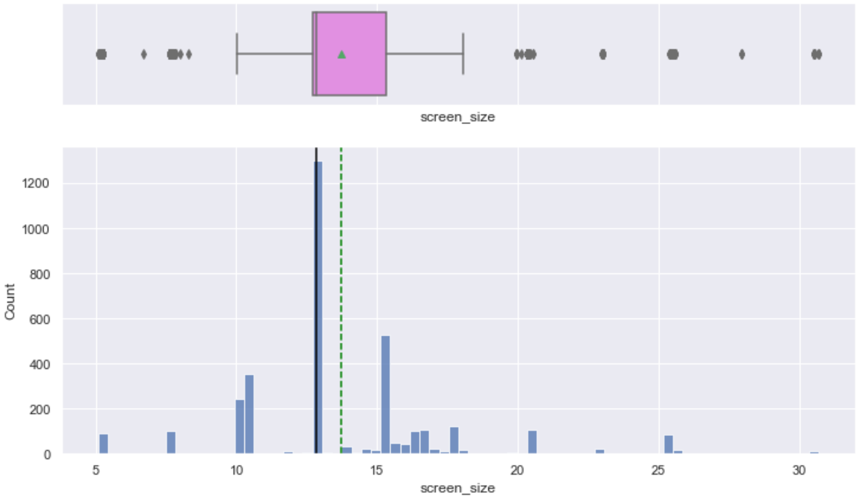
A log transformation was used. This shows a normal distribution for the new price variable, this would be useful for the model.



EDA Results – Univariate Analysis

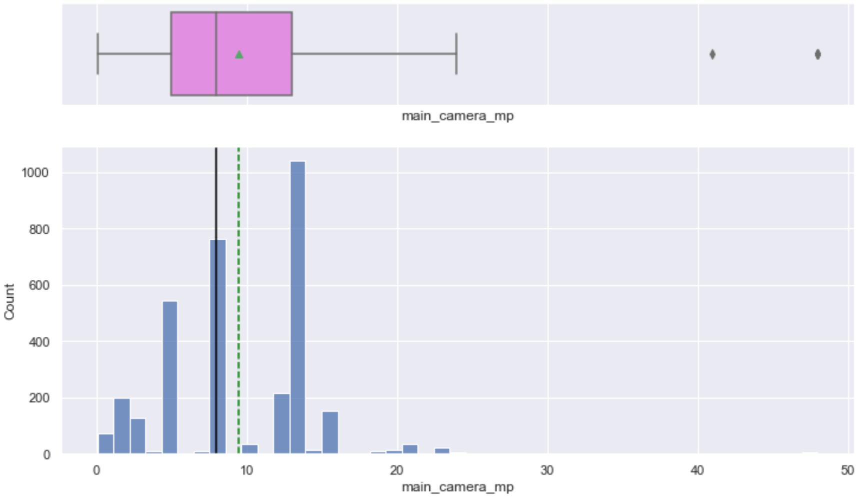
- Screen_size

This variable shows that the median and the mean are relatively close, but the 15 cm phones are the ones moving the mean.



- Main_camera_mp

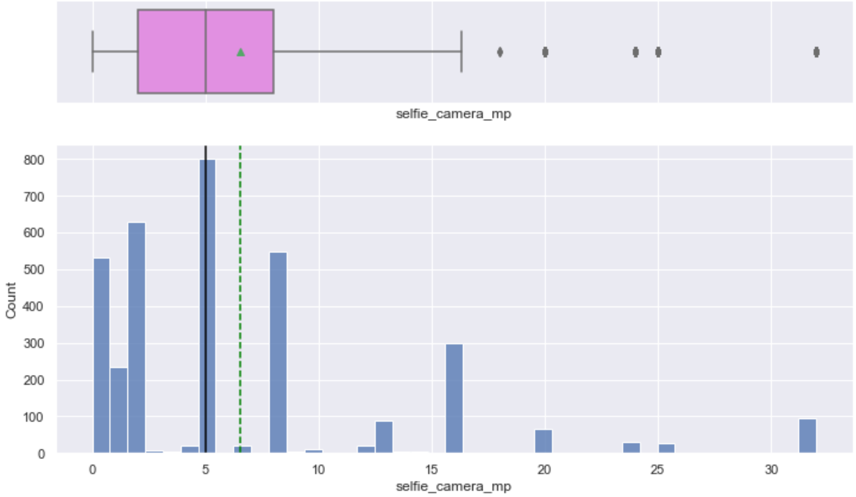
This variable shows how more phones are getting more mp in the main camera. As result, the median is to the left, but the mode is to the right.



EDA Results – Univariate Analysis

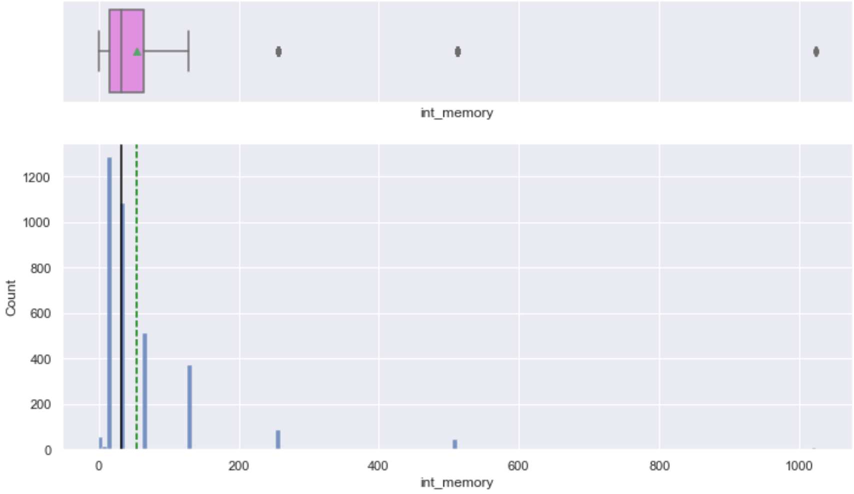
- Selfie_camera_mp

Most of the selfie cameras stay under 17 mp, the 20 mp and over cameras shown to be for more premium phones.



- Int_memory

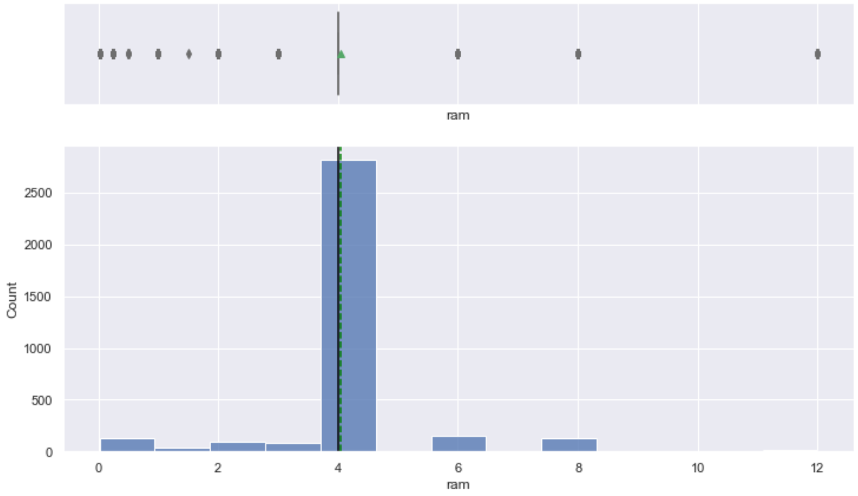
The must common internal memory is under the 150 GB. The premium cellphones must have over 200 GB of memory.



EDA Results – Univariate Analysis

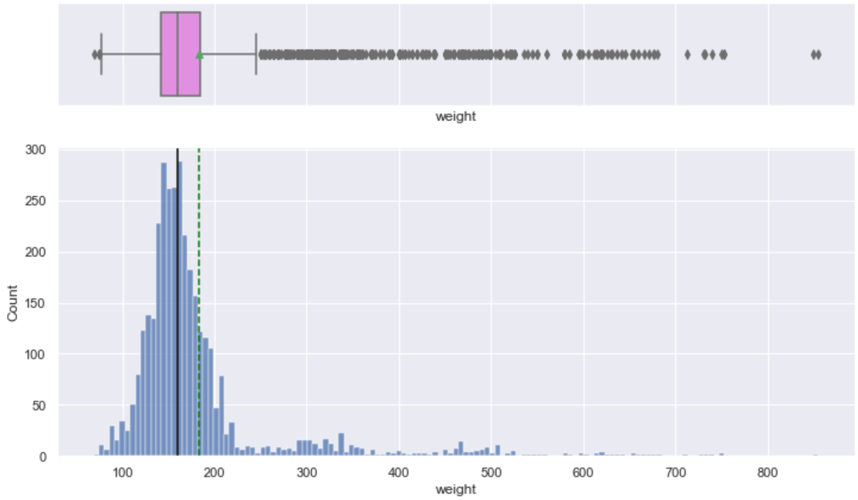
- Ram

The most common size of memory if 4 gb, leaving any other number as an outlier.



- Weight

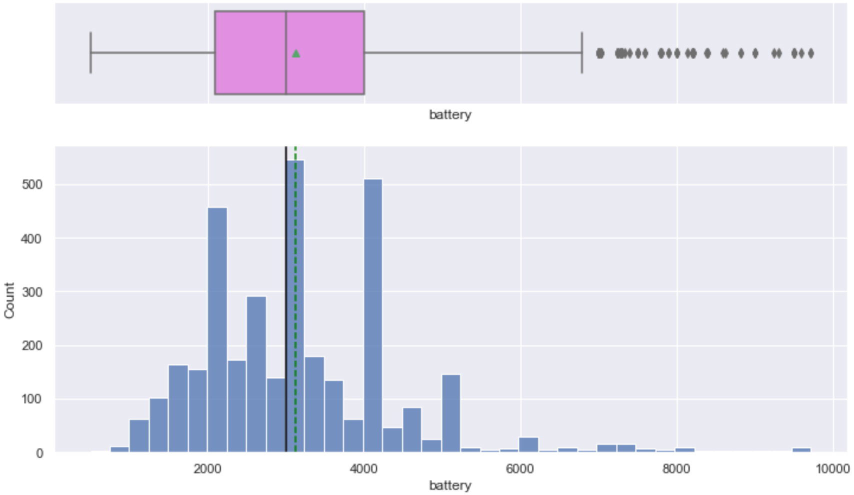
Most of the data concentrates under 250 grams of weight, with this the variables shows it is skewed to the right.



EDA Results – Univariate Analysis

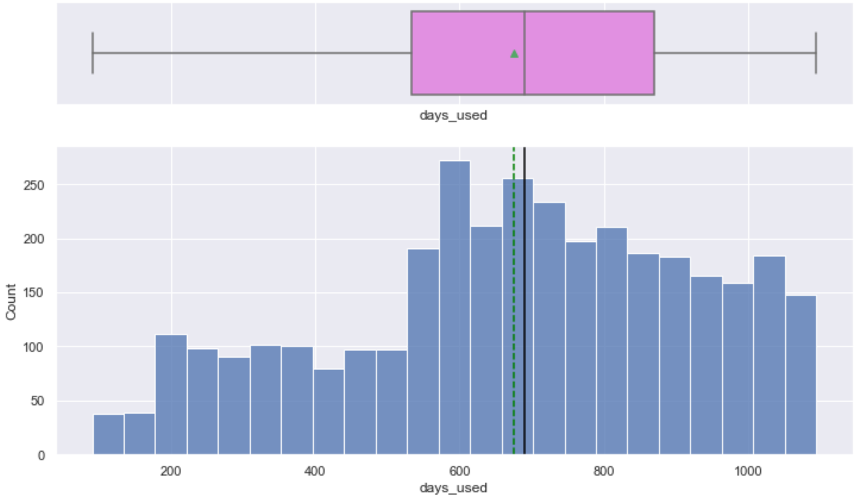
- Battery

The battery capacity of the cellphones shows a tendency of being grow with of the bigger phones on the market.



- Days_used

Most of the users spend almost two years with their current cellphone before looking to sell it.



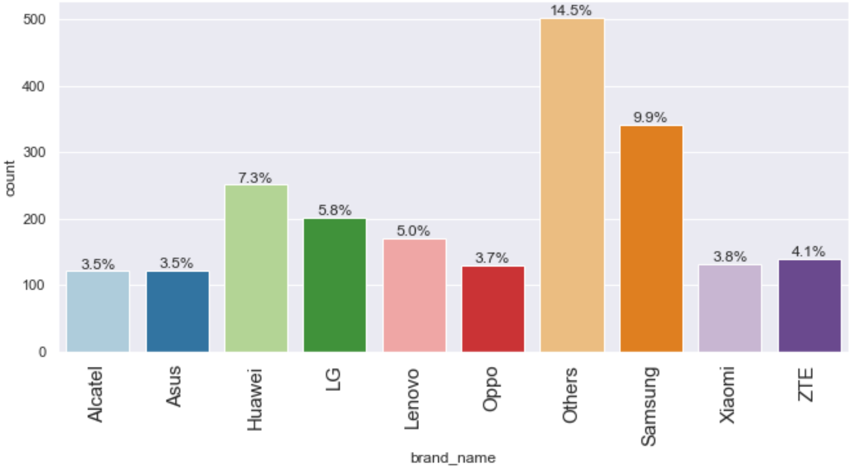
EDA Results – Univariate Analysis

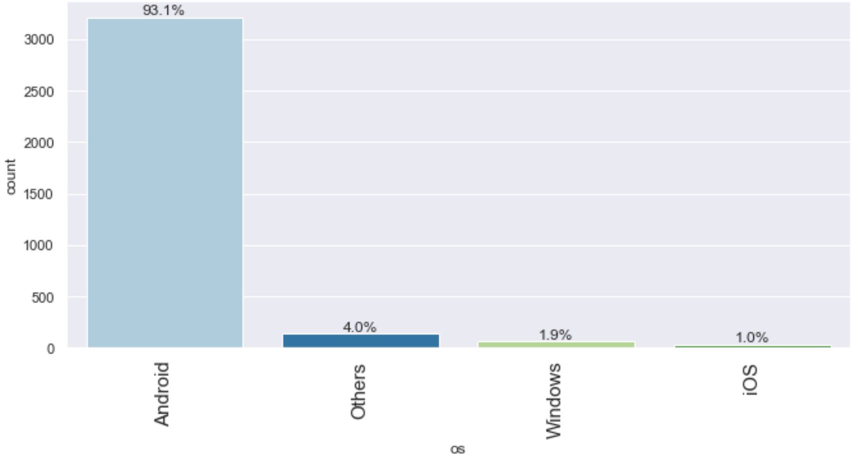
- Brand_name

The prinicpals brands of the used markets are Samsung, Huawei and LG.

- Os

Over 90% of the phones used the OS Android. Even thought iOS, is well regarded by their users it has the smallest share of the market.

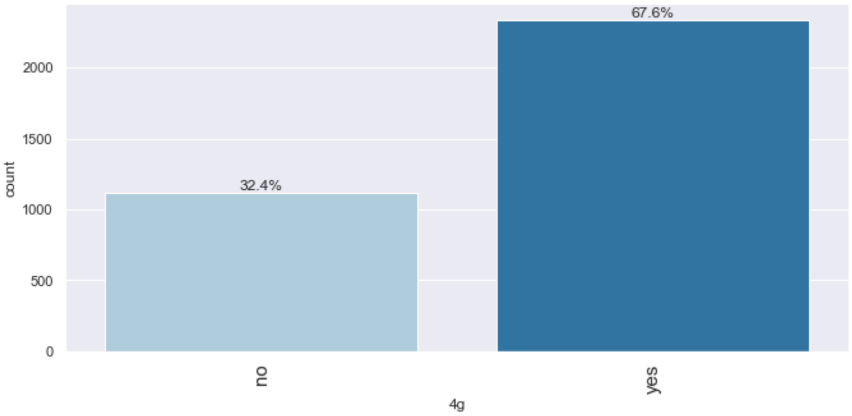




EDA Results – Univariate Analysis

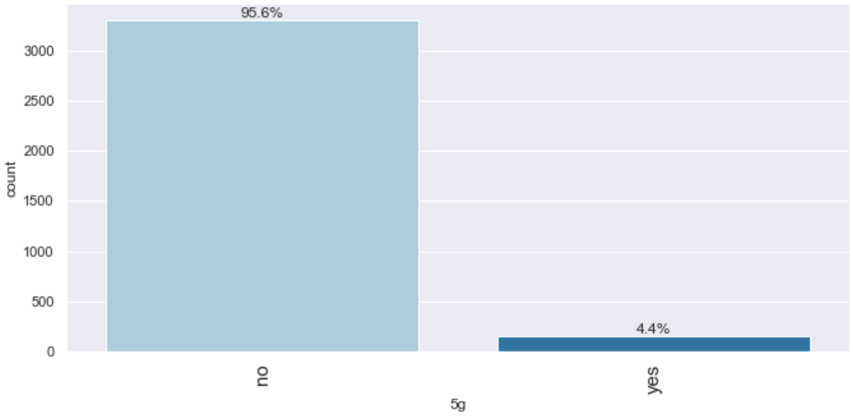
- 4g

The 67.6% of the used phone market have a 4g connection.



- 5g

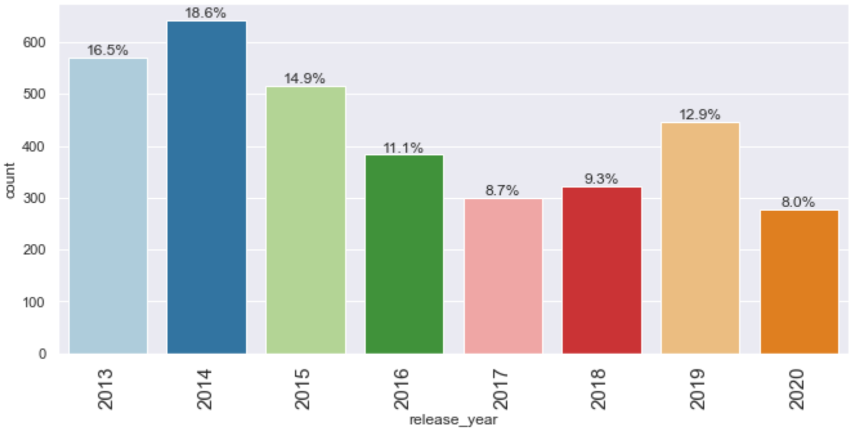
Only a 4.4% of the used phones has a 4g connection.



EDA Results – Univariate Analysis

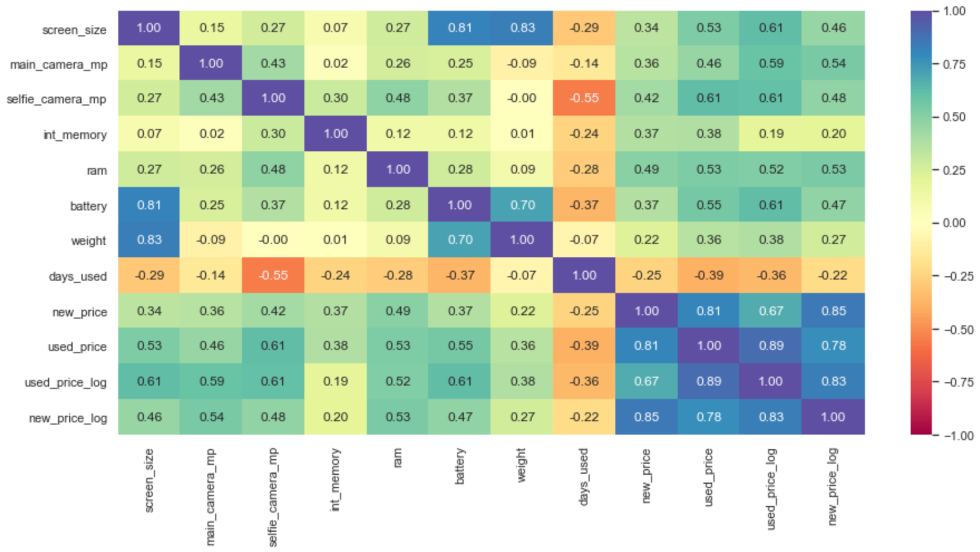
- Release_year

The used market shows that the phones of between 2013 – 2015 have the 50% of the market. But a change on this distributions is expected with the pass of time.



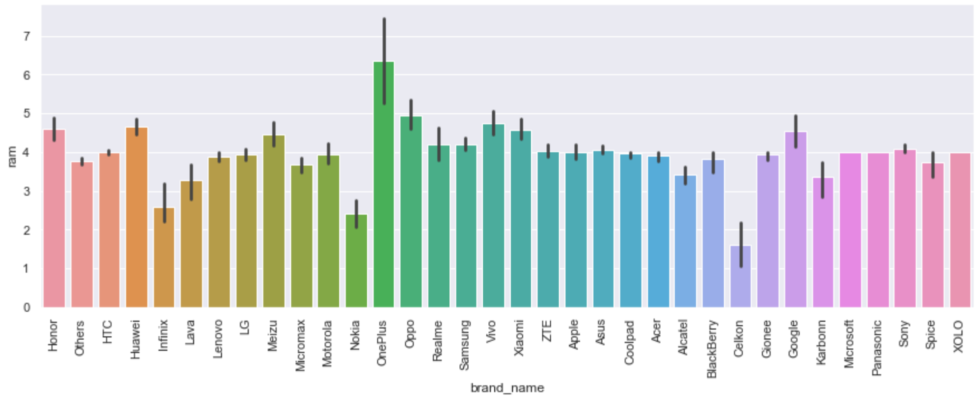
EDA Results - Bivariate Analysis

- The results between correlation between the variables shows some expected relationships such as the screen_size with weight and battery. One relation ship that shows the interest in the used market is the relation ship between used_price with the selfie_camara_mp. Also, the relationship between price and battery shows a that a bigger phones has a bigger price. The must interesting negative relationships are the selfie_camara_mp and days_used and used_price and days_used the last one is the expected with the used market.



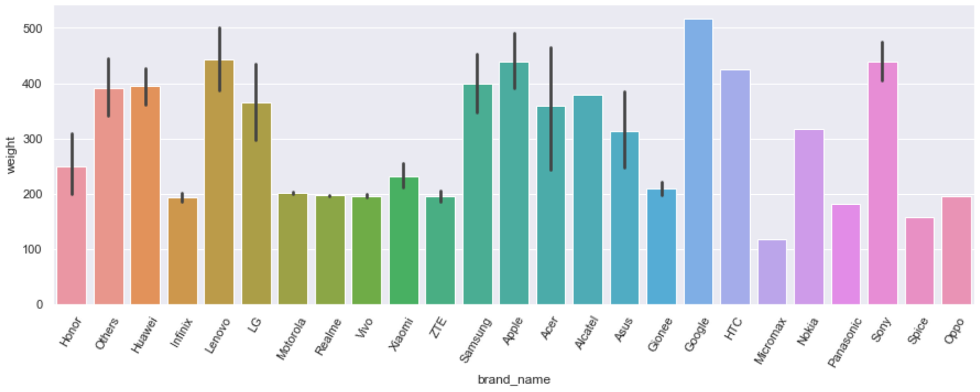
EDA Results - Bivariate Analysis

- The relationship between ram memory and brand_name is to see if a brand has as a differentiate characteristic the amount of memory. In this case OnePlus is the one with the must ram. The rest of the brands are bellow the 6gb and almost all stay near 4 gb. The exceptional cases are Nokia and Celkon.



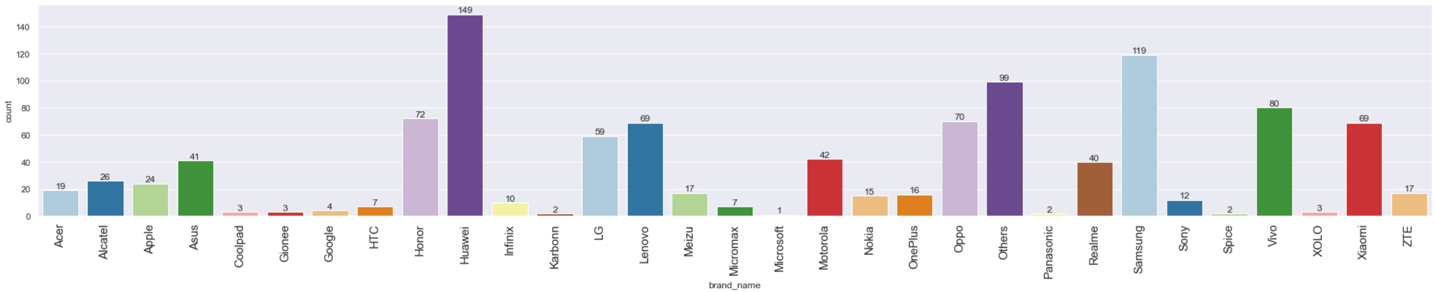
EDA Results - Bivariate Analysis

- The weigh between different brands shows that the brands with known premium phones have the must features and therefore the phones are heavier. For example, Samsung with their top-of-the-line phones that have an included stylus.



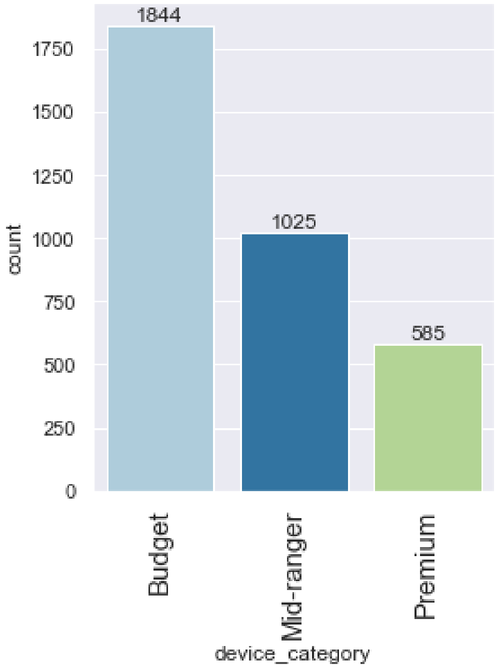
EDA Results - Bivariate Analysis

- The relationship between the used market brands and the number of cellphones on the data base. This show us that Huawei has the higher count for cellphones available in the market with 149. The following brands includes Samsung with 119, Honor with 72and Vivo with 80.



Data Preprocessing – Feature engineering

- For the preprocessing it's necessary to understand for which category the cellphone is designed. Therefore, three categories were designated: Budget, Mid-ranger and Premium. The First category consists of phones under € 200, the Mid-ranger between € 200 and € 350, and premium over € 350.



Data Preprocessing – Missing values

- A duplicate analysis was realized, and it found 0 duplicated values in the data set.
- For the missing value treatment an analysis of the variables was realized, and it was found that the following variables had some missing values:

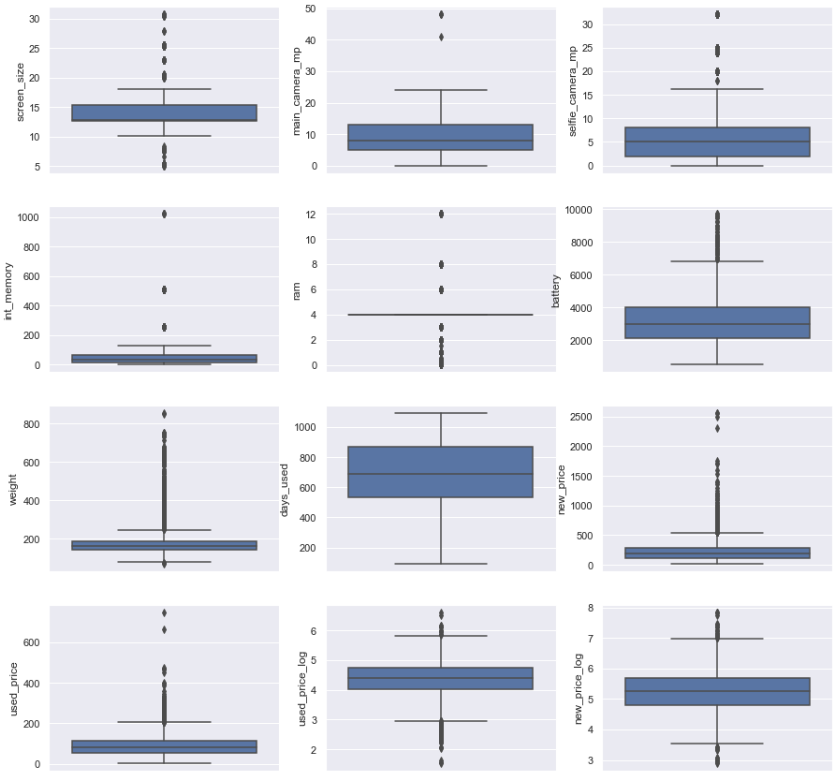
Variables	Total Missing Values
Main_camera_mp	179
Selfie_camera_mp	2
Int_memory	4
Ram	4
Battery	6
Weight	7

Data Preprocessing – Missing values

- For the first treatment of the missing values the cellphones were grouped by their release_year and brand_name to fill the missing values with their respective medians. This approach solve the issue for int_memory and ram.
- Then the same procedure was used for the previous variables but, only grouping them by brand_name. This solved the issue for the following variables: selfie_camera_mp, battery and weight. main_camera_mp had still 10 missing values after this point.
- Finally, the missing values of main_camera_mp were filled with the median of the variable.

Data Preprocessing – Outlier Check

- For the outlier check of all the variables the one that was highlighted the most for our analysis was the used price. As it is our target variable, we need to be normalized in order to achieve a better result for our analysis. The rest of the variables even though, some can some signs of outliers such as weight, or ram are in the normal range for these kinds of products.



Data Preprocessing – Data Preparation for modeling

- For the preparation of the model the independent variable used_price_log is dropped. These variables were also dropped new_price, used_price, and device_category. The variable new_price_log will remain as is a normalized version of the new_price variable.

- The target values will be used_price_log as it's the normalized version of the used_price.

- With the previous treatment dummy variable was added to the model.

- As the last step taken the data was split into a 70:30 ratio for training and testing.

Model Performance Summary

- The result obtained through linear regression is a model with an R-Squared of 0.997 and an Adj. R-Squared of 0.997 which is considered very precise as it explains 99.7% of the variation in the sets.
- The principal variables includes screen_size, main_camera_mp, selfie_camera_mp, ram, weight, release_year, new_price_log, brand_name_Nokia, brand_name_Samsung, brand_name_Sony, brand_name_Xiaomi, and 4G.
- The overall performance of the model brings the following results for the train and test set:

Data	RMSE	MAE	MAPE
Train	25.772441	16.828725	19.15624
Test	24.489763	16.616165	19.43771

- With the results of the RMSE it can be observed that the model fits well as the results are similar.
- Our model can predict used_price with a mean error of € 16.62.
- With the MAPE the model can predict within the 19.44% of the value of the used price of the cellphone.

Link to Appendix slide on model assumptions

Model Performance Summary

- Based on the coefficients the most important variable is new_price_log. The connection to a 4g network as one of the relevant components to be considered in the model. The brands Samsung and Sony have a negative perspective on the expected price of the used price.

```
                            OLS Regression Results
=======================================================================================
Dep. Variable:         used_price_log   R-squared (uncentered):                   0.997
Model:                            OLS   Adj. R-squared (uncentered):              0.997
Method:                 Least Squares   F-statistic:                          7.159e+04
Date:                Wed, 02 Mar 2022   Prob (F-statistic):                        0.00
Time:                        07:53:10   Log-Likelihood:                          93.022
No. Observations:                2417   AIC:                                     -162.0
Df Residuals:                    2405   BIC:                                     -92.56
Df Model:                          12
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
screen_size            0.0310      0.003     10.911      0.000       0.025       0.037
main_camera_mp         0.0216      0.001     15.410      0.000       0.019       0.024
selfie_camera_mp       0.0160      0.001     17.594      0.000       0.014       0.018
ram                    0.0207      0.004      4.723      0.000       0.012       0.029
weight                 0.0007      0.000      5.826      0.000       0.000       0.001
release_year           0.0006   2.14e-05     28.302      0.000       0.001       0.001
new_price_log          0.4102      0.011     38.980      0.000       0.390       0.431
brand_name_Nokia       0.0723      0.029      2.458      0.014       0.015       0.130
brand_name_Samsung    -0.0364      0.016     -2.238      0.025      -0.068      -0.005
brand_name_Sony       -0.0727      0.030     -2.398      0.017      -0.132      -0.013
brand_name_Xiaomi      0.0793      0.026      3.104      0.002       0.029       0.129
4g_yes                 0.0780      0.013      5.812      0.000       0.052       0.104
==============================================================================
Omnibus:                      214.839   Durbin-Watson:                   1.906
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              385.503
Skew:                          -0.616   Prob(JB):                     1.95e-84
Kurtosis:                       4.520   Cond. No.                     1.30e+04
==============================================================================
```

Link to Appendix slide on model assumptions

Conclusions and Recomendations

- The generated model can explain the 99.7% of the variance that occurred in the dataset. With this, we can obtain a price value of the phones in the used market.

- The results of the model show that the current model values the original price greatly to calculate a future used price. Usually, this correlates to a higher-quality cellphone. Right now, 4g is considered the standard and 5g capability is in less than 5% of the cellphones. But this is expected to be changing in the future as this type of network becomes more popular around the world.

- Due to the changes in technology, an update of the model should be done consistently. This update should help to understand what consumers of the used market value the most. For example, in the used market place a Xiaomi cellphone is more valued than a Samsung cellphone. But his can be cahning in the future as the brands develop their name and quality over time.

Link to Appendix slide on model assumptions

APPENDIX

Data Background and Contents - Ductionary

- The variables of the data set are the following:

1. brand_name: Name of manufacturing brand
2. os: OS on which the device runs
3. screen_size: Size of the screen in cm
4. 4g: Whether 4G is available or not
5. 5g: Whether 5G is available or not
6. main_camera_mp: Resolution of the rear camera in megapixels
7. selfie_camera_mp: Resolution of the front camera in megapixels
8. int_memory: Amount of internal memory (ROM) in GB
9. ram: Amount of RAM in GB
10. battery: Energy capacity of the device battery in mAh
11. weight: Weight of the device in grams
12. release_year: Year when the device model was released
13. days_used: Number of days the used/refurbished device has been used
14. new_price: Price of a new device of the same model in euros
15. used_price: Price of the used/refurbished device in euros

Data Background and Contents

- The variables used are dived as follows

Object	Float64	int64
• Brand Name • Os • 4g • 5g	• Scree_size • Main_camera_mp • Selfie_camera_mp • Int_memory • Ram • Battery • Weight • New_price • Used_price	Realease_year Days_used

Model Assumptions

- The variables used_price and new_price were transformed into a log variable in order to seek better treatment through normalization of the values obtained.
- The variable device_category was created using the new_price to determine the type of segment of the market that the cellphone is intended to be placed.
- The independent variables of the dataset are new_price, used_price, used_price_log, and device_category. The new_price is considered independent as it is defined by the manufacturer, and it can’t be explained with the information provided in the data set.
- The dummy variables were created for the object and category type of variables. This resulted in a total of 48 variables for the model.

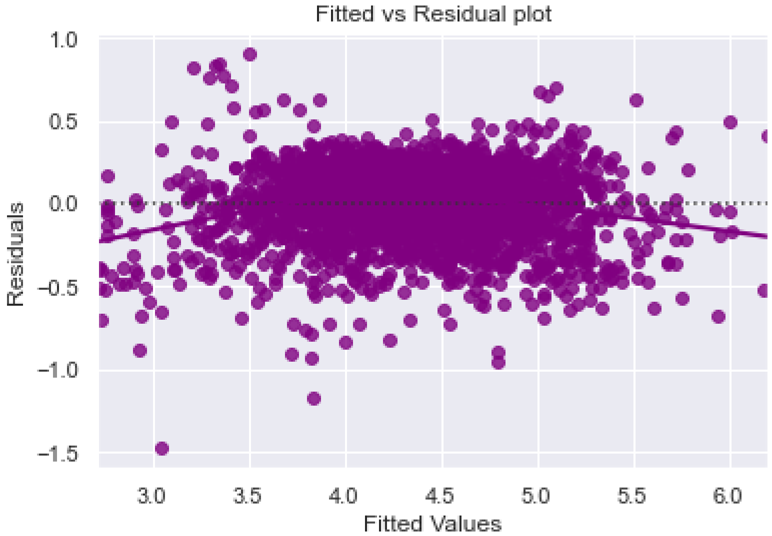
Model Assumptions – No Multicollinearity

- A test for Multicollinearity was released on the data set after the first drop of variables was done. The results of this test show that none of the variables obtained a value higher than 10 in the variance of inflation factor (VIF).
- Following this a test was realized to drop the variables that had a p-value over 0.05. this reduces the 48 variables of the model to 12 variables (final model).
- These changes brought the following results for the model:

Data	RMSE	MAE	MAPE
Train	25.772441	16.828725	19.15624
Test	24.489763	16.616165	19.43771

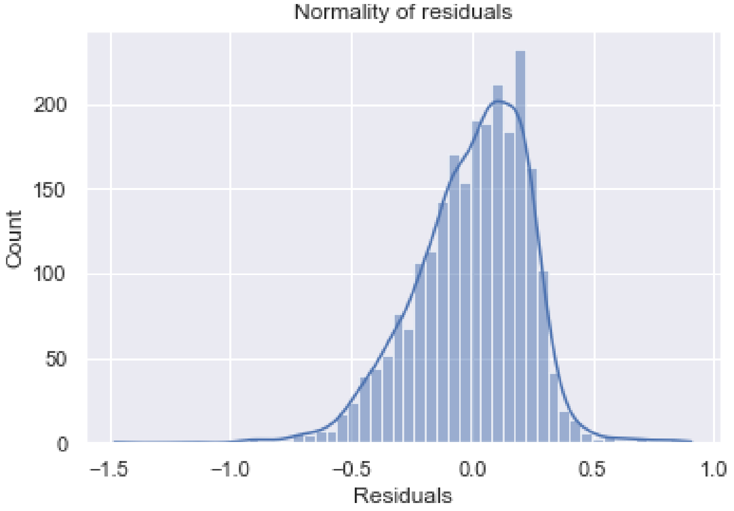
Model Assumptions – Linearity and Independence Test

- For this test the difference between the fitted values and actual values was obtained and the residuals were plotted.
- The values of the residuals don't show any pattern and are distributed between negative and positive values.

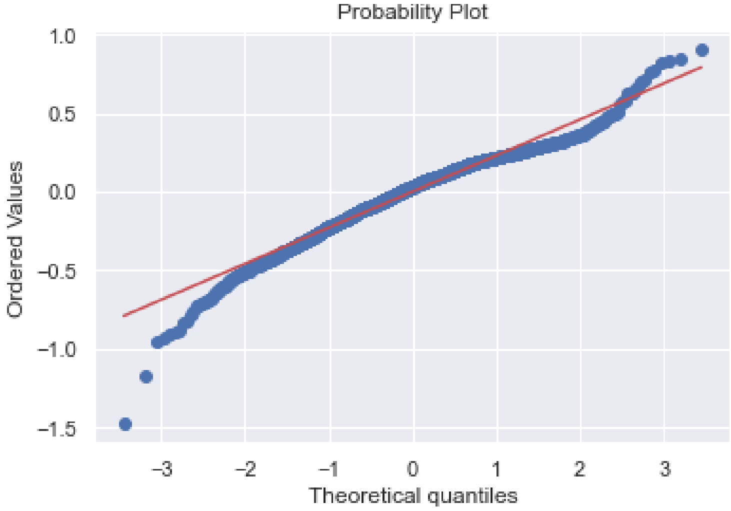


Model Assumptions – Normality of error terms

- The residuals show a normal distribution with a skew to the right.



- The probability plot shows that the values are around the diagonal line. Showing the normality of the error.



Model Assumptions – No Heteroscedasticity

- In order to check the homoscedasticity, the Goldfeld-Quandt test was realized. This test indicated with a p-value of **0.1924** that there is no statistical evidence of heteroscedasticity in the variables for the model.

OLS Model – Original vs Final

Dep. Variable:	used_price_log	R-squared (uncentered):	0.997
Model:	OLS	Adj. R-squared (uncentered):	0.997
Method:	Least Squares	F-statistic:	1.792e+04
Date:	Wed, 02 Mar 2022	Prob (F-statistic):	0.00
Time:	07:53:08	Log-Likelihood:	112.56
No. Observations:	2417	AIC:	-129.1
Df Residuals:	2369	BIC:	148.8
Df Model:	48		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
screen_size	0.0275	0.003	8.161	0.000	0.021	0.034
main_camera_mp	0.0216	0.002	14.288	0.000	0.019	0.025
selfie_camera_mp	0.0154	0.001	14.473	0.000	0.013	0.018
int_memory	0.0002	6.96e-05	2.175	0.030	1.49e-05	0.000
ram	0.0224	0.005	4.323	0.000	0.012	0.033
battery	-1.062e-05	7.19e-06	-1.477	0.140	-2.47e-05	3.48e-06
weight	0.0009	0.000	6.689	0.000	0.001	0.001
release_year	0.0006	3.51e-05	17.886	0.000	0.001	0.001
days_used	-4.078e-05	2.62e-05	-1.554	0.120	-9.22e-05	1.07e-05
new_price_log	0.4132	0.012	34.660	0.000	0.390	0.437
brand_name_Alcatel	0.0166	0.048	0.347	0.729	-0.077	0.110
brand_name_Apple	0.0262	0.148	0.178	0.859	-0.263	0.316
brand_name_Asus	0.0174	0.048	0.361	0.718	-0.077	0.112
brand_name_BlackBerry	-0.0508	0.071	-0.720	0.472	-0.189	0.087
brand_name_Celkon	-0.0702	0.066	-1.057	0.291	-0.200	0.060
brand_name_Coolpad	0.0346	0.073	0.472	0.637	-0.109	0.178
brand_name_Gionee	0.0429	0.058	0.740	0.460	-0.071	0.157
brand_name_Google	-0.0035	0.085	-0.041	0.967	-0.170	0.163
brand_name_HTC	-0.0153	0.048	-0.316	0.752	-0.110	0.080
brand_name_Honor	0.0333	0.049	0.674	0.501	-0.064	0.130
brand_name_Huawei	-0.0007	0.045	-0.015	0.988	-0.088	0.087
brand_name_Infinix	0.1704	0.094	1.821	0.069	-0.013	0.354
brand_name_Karbonn	0.0757	0.067	1.124	0.261	-0.056	0.208
brand_name_LG	-0.0113	0.046	-0.249	0.803	-0.101	0.078
brand_name_Lava	0.0335	0.063	0.535	0.593	-0.089	0.156
brand_name_Lenovo	0.0422	0.045	0.930	0.352	-0.047	0.131
brand_name_Meizu	-0.0098	0.056	-0.174	0.862	-0.120	0.101
brand_name_Micromax	-0.0354	0.048	-0.736	0.462	-0.130	0.059
brand_name_Microsoft	0.1183	0.089	1.336	0.182	-0.055	0.292
brand_name_Motorola	-0.0074	0.050	-0.149	0.882	-0.105	0.090
brand_name_Nokia	0.0945	0.052	1.822	0.069	-0.007	0.196
brand_name_OnePlus	0.0530	0.078	0.681	0.496	-0.100	0.206
brand_name_Oppo	0.0171	0.048	0.356	0.722	-0.077	0.111
brand_name_Others	-0.0048	0.042	-0.115	0.909	-0.088	0.078
brand_name_Panasonic	0.0592	0.056	1.055	0.292	-0.051	0.169
brand_name_Realme	0.0427	0.062	0.690	0.490	-0.079	0.164
brand_name_Samsung	-0.0310	0.043	-0.714	0.476	-0.116	0.054
brand_name_Sony	-0.0704	0.051	-1.388	0.165	-0.170	0.029
brand_name_Spice	-0.0264	0.063	-0.417	0.677	-0.151	0.098
brand_name_Vivo	-0.0155	0.049	-0.319	0.750	-0.111	0.080
brand_name_XOLO	0.0049	0.055	0.088	0.930	-0.103	0.113
brand_name_Xiaomi	0.0882	0.048	1.825	0.068	-0.007	0.183
brand_name_ZTE	-0.0075	0.048	-0.157	0.875	-0.101	0.086
os_Others	-0.0492	0.033	-1.491	0.136	-0.114	0.015
os_Windows	-0.0363	0.045	-0.802	0.423	-0.125	0.052
os_iOS	-0.0771	0.147	-0.524	0.600	-0.366	0.211
4g_yes	0.0822	0.015	5.563	0.000	0.053	0.111
5g_yes	-0.0452	0.032	-1.434	0.152	-0.107	0.017

Omnibus:	215.114	Durbin-Watson:	1.912
Prob(Omnibus):	0.000	Jarque-Bera (JB):	397.240
Skew:	-0.607	Prob(JB):	5.50e-87
Kurtosis:	4.571	Cond. No.	2.01e+05

OLS Regression Results

Dep. Variable:	used_price_log	R-squared (uncentered):	0.997
Model:	OLS	Adj. R-squared (uncentered):	0.997
Method:	Least Squares	F-statistic:	7.159e+04
Date:	Wed, 02 Mar 2022	Prob (F-statistic):	0.00
Time:	07:53:10	Log-Likelihood:	93.022
No. Observations:	2417	AIC:	-162.0
Df Residuals:	2405	BIC:	-92.56
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
screen_size	0.0310	0.003	10.911	0.000	0.025	0.037
main_camera_mp	0.0216	0.001	15.410	0.000	0.019	0.024
selfie_camera_mp	0.0160	0.001	17.594	0.000	0.014	0.018
ram	0.0207	0.004	4.723	0.000	0.012	0.029
weight	0.0007	0.000	5.826	0.000	0.000	0.001
release_year	0.0006	2.14e-05	28.302	0.000	0.001	0.001
new_price_log	0.4102	0.011	38.980	0.000	0.390	0.431
brand_name_Nokia	0.0723	0.029	2.458	0.014	0.015	0.130
brand_name_Samsung	-0.0364	0.016	-2.238	0.025	-0.068	-0.005
brand_name_Sony	-0.0727	0.030	-2.398	0.017	-0.132	-0.013
brand_name_Xiaomi	0.0793	0.026	3.104	0.002	0.029	0.129
4g_yes	0.0780	0.013	5.812	0.000	0.052	0.104

Omnibus:	214.839	Durbin-Watson:	1.906
Prob(Omnibus):	0.000	Jarque-Bera (JB):	385.503
Skew:	-0.616	Prob(JB):	1.95e-84
Kurtosis:	4.520	Cond. No.	1.30e+04

Happy Learning !