ReCell

Supervised Learning Foundations

Julie Kistler - November 14, 2023

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

ReCell would like to capture a significant market share of the online marketplace of used devices. According to the IDC (International Data Corporation) it is forecasted that the used phone market will be worth 52.7 billion by 2023. The growth is due to an uptick in demand for used phones and tablets offering consumers a significant saving over new purchases. The first step to remain competitive in this “under-the-radar” marketplace is to create a Machine Learning based solution that will develop a dynamic pricing strategy for used/refurbished devices. We developed a linear regression model to help predict the price of the used/refurbished devices and identify supporting factors that impact/influence pricing.

The model developed proved to have the ability to explain 84% of the variation in the data within 4.54% of the normalized new price which is good for prediction and inference purposes. Devices with a newer release data tend to command higher prices. Devices with main cameras offering 16MP or greater tend to have higher demand than other devices. It appears the Android OS dominates the market with ~ 93% of the market share. Not one brand is dominate; but, the model does show a positive correlation to Nokia and Xiaomi.

Based on the data analysis and model performance, it is recommended to focus on building a product inventory of devices that are newer models, have main cameras with 16MP or more, have an Android OS, and have 5g network capability to generate the maximum revenue and increase market share. In addition, it is recommended to build a supporting linear regression model to further analyze additional services (warranties, insurance, etc) and how they may drive pricing and demand. Furthermore, adding a layer of demographic data to the models may be insightful to understand the consumer market to refine the target market and marketing message.

Business Problem Overview and Solution Approach

Business problem: Build a Machine Learning based solution that will develop a dynamic pricing strategy for used/refurbished devices.

Solution Approach/Methodology: Build a linear regression model to help predict the price of the used/refurbished devices and identify supporting factors that impact/influence pricing.

Data Overview

Data Dictionary	
brand_name	Name of manufacturing brand
os	OS on which the device runs
screen_size	Size of the screen in cm
4g	Whether 4G is available or not
5g	Whether 5G is available or not
main_camera_mp	Resolution of the rear camera in megapixels
selfie_camera_mp	Resolution of the front camera in megapixels
int_memory	Amount of internal memory (ROM) in GB
ram	Amount of RAM in GB
battery	Energy capacity of the device battery in mAh
weight	Weight of the device in grams
release_year	Year when the device model was released
days_used	Number of days the used/refurbished device has been used
normalized_new_price	Normalized price of a new device of the same model in euros
normalized_used_price	Normalized price of the used/refurbished device in euros

Data Overview, Cont…

Column	Dtype
brand_name	object
os	object
screen_size	float64
4g	object
5g	object
main_camera_mp	float64
selfie_camera_mp	float64
int_memory	float64
ram	float64
battery	float64
weight	float64
release_year	int64
days_used	int64
normalized_new_price	float64
normalized_used_price	float64

Rows	Columns
3454	15

- 4 object data types (brand_name, os, 4g, 5g)
- 2 integer data types (release_year, days_used)
- 9 float data types (screen_size, main_camera_mp, selfie_camera_mp, int_memory, ram, battery, weight, normalized_new_price, normalized_used_price

EDA Results _ Univariate Analysis

- There are no duplicate values
- There are some missing values for the following columns

main_camera_mp	selfie_camera_mp	int_memory
ram	battery	weight

Univeriate Analysis/Statistical Summary Insights

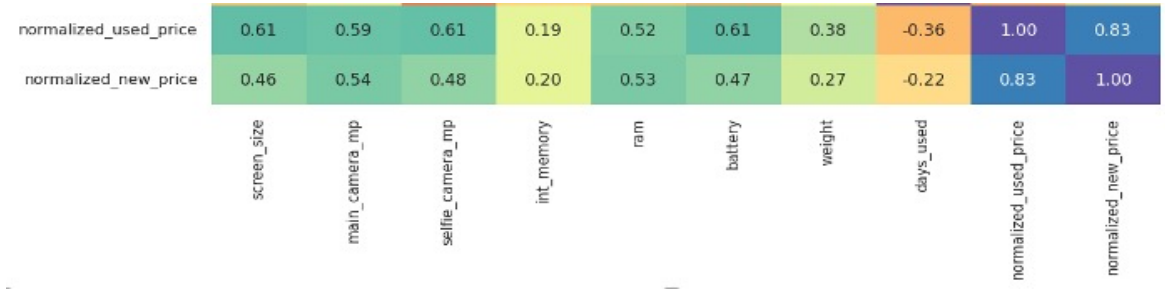
- The normalized used price ranges between 1.54 to 6.62 Euros with an average of 4.36 Euros.
- The normalized new price ranges between 2.90 to 7.85 Euros with an average of 5.23 Euros.
- The average screen size is 13.71 cm with a min of 5.08 cm to a max of 30.71 cm.
- Devices may have 4g or 5g network capability.
- The release year: oldest to be 2013 and newest to be 2020.
- The average amount of RAM is 4.04 GB.
- The average battery capacity is 3133.40 mAh.
- The average main camera resolution is 9.46 MP.
- The average selfie camera resolution is 6.55 MP.
- Average internal memory is ~ 54.57 grams with a min of .01 grams and max of 1024 grams.
- The average weight is ~ 182.75 grams with a min of .69 grams and max of 855 grams.
- There are 4 types of operating systems (OS) – Android is the most dominate.
- There are 34 unique brand names – not one traditional brand name is dominate.

Link to Appendix statistical summary and supporting univariate analysis slides

EDA Results

Bivariate Analysis

- There appears to be a positive correlation with the normalized new price, screen size, selfie camera MP, battery, main camera MP, and RAM. A negative correlation occurs with number of days used.



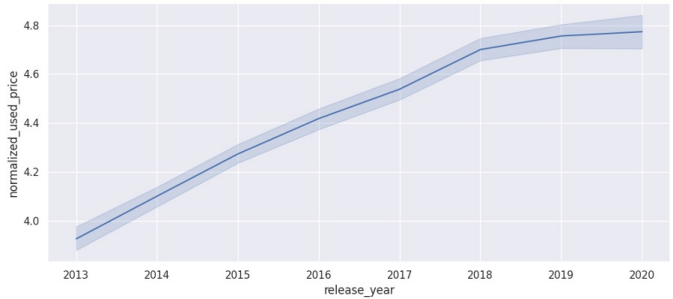
- The brand with the largest screen size is Huawei (13.6% market share) followed by Samsung (10.8% market share).
- The brand with the most devices that have selfie cameras greater than 8 MP is Huawei (13.3%) followed by Vivo (11.9%) and Oppo (11.5%) – these three brands appear to capture over of third of the market share.
- The average battery capacity across all devices is 31.33 mAh.
- The brand with the most devices that have main cameras greater than 16 MP is Sony (39.4%).
- The brand with the most RAM appears to be OnePlus and Celkor appears to have the least.

Link to Appendix slides – Supporting Bivariate Analysis

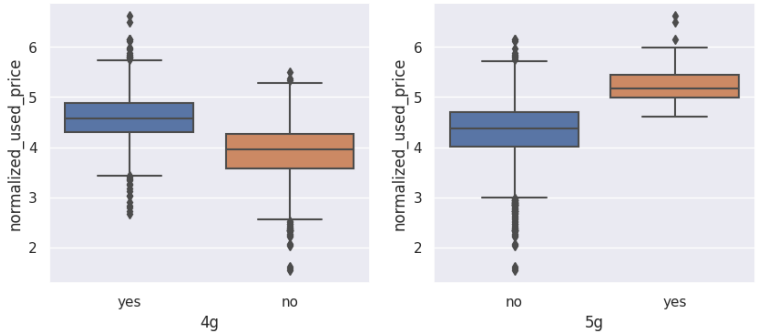
EDA Results

Bivariate Analysis cont,,,

- As the release year of the device rises the prices rises as well demonstrating a positive correlation.



- Devices that offer 5g network availability appear to command higher prices over devices offering 4g network availability.



Link to Appendix slides – Bivariate Analysis

Data Preprocessing

- There are no duplicate values.
- There are some missing values:

main_camera_mp	selfie_camera_mp	int_memory
ram	battery	weight

*Treated the missing values by:

1. Imputed the missing values by column medians grouped by release year and brand name.
2. Imputed the remaining missing values by the column medians grouped by brand name.
3. One remaining value needed to be treated – main_camera_mp we filled with the column median.
4. All missing values were treated.

Link to Appendix slide – Missing Value Treatment

Data Preprocessing Cont...

Feature Engineering: Created a new column labeled “years_since_release” and dropped the ”release_year” column.

* Average years since release: 5 years
* Min years since release: 1 year
* Max years since release: 8 years

- There are outliers in the data.
  * These outliers will not be treated as they are proper variables.

Data Preparation for Modeling

- We want to predict the normalized price of used devices.
- Added intercept
- Created dummy variables: (Object, Category)
- Split the data 70:30 ratio for a train to test data
  *Train Data: 2417
  *Test Data: 1037

Link to Appendix slides – Outliers and Feature Engineering

Data Preprocessing Cont...

- Model Building _ Linear Regression (Train Data (70%) – Test for Performance)

  OLS Regression Results: (Train Data: 2417 rows)
  * R-squared: 0.850
  * Adjusted. R-squared: 0.845. (GOOD)
  * *const* coefficient: 1.6815
  * Coefficient of normalized_new_price: 0.4146

Link to Appendix slide to view OLS Model

Model Performance Summary

- Overview of ML Model and its parameters
  - We wanted to build a ML model that will predict the normalized price of used devices
    - The ML Model split the data into a 70:30 ratio
      *Train Data: 2417 rows
      *Test Data: 1037 rows

- Summary of most important factors used by the ML model for prediction
  - **No Multicollinearity:** Check and remove multicollinearity for variables with VIF of 5 or greater – remove variables with a p-value greater than .05 if needed.
  - **Linearity of Variables:** Predictor variables must have linear relation with the dependent variables.
  - **Independence or error terms:** If they don't follow any pattern, then we say the model is linear and independent.
  - **Normality of error terms:** The shape of the histogram of residuals can give an initial idea about the normality.
  - **No Heteroscedasticity:** If the p-value greater than 0.05, the residuals are homoscedastic. Otherwise, they are heteroscedastic.

Link to Appendix slide on data model tests used for prediction

Model Performance Summary cont...

- Summary of key performance metrics for training and test data
  - **Linear Progression Performance Check**

	index	RMSE	MAE	R-squared	Adj. R-squared	MAPE
Training	0	0.226107	0.17736	0.849942	0.844202	4.247918
Test	0	0.239404	0.185272	0.841094	0.82616	4.495925

  - **Remove Multicollinearity – Performance Check**

	index	RMSE	MAE	R-squared	Adj. R-squared	MAPE
Training	0	0.231053	0.180675	0.843304	0.841798	4.32716
Test	0	0.241528	0.187329	0.838262	0.834589	4.541924

Link to Appendix slide on data model tests used for prediction

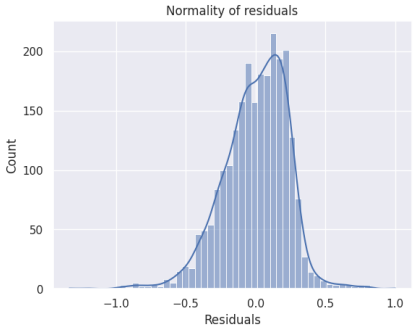
Model Performance Summary cont...

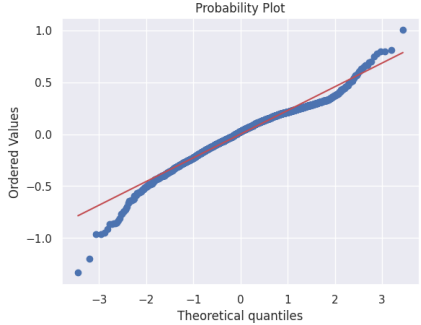
- Summary of key performance assumptions for training and test data
  - **Test for Linearity and Independence**

Assumption is satisfied

  - **Test for Normality**

Assume for normality





  - **Test for Homoscedasticity**

* p-value: 0.15606511727489084

Residuals are homoscedastic

Link to Appendix slide on data model tests used for prediction

Model Performance Summary cont...

- Summary of key performance metrics for training and test data
  - **Linear Progression Performance Check**

	index	RMSE	MAE	R-squared	Adj. R-squared	MAPE
Training	0	0.226107	0.17736	0.849942	0.844202	4.247918
Test	0	0.239404	0.185272	0.841094	0.82616	4.495925

  - **Remove Multicollinearity – Performance Check**

	index	RMSE	MAE	R-squared	Adj. R-squared	MAPE
Training	0	0.231053	0.180675	0.843304	0.841798	4.32716
Test	0	0.241528	0.187329	0.838262	0.834589	4.541924

  - **Final Model Summary**

	index	RMSE	MAE	R-squared	Adj. R-squared	MAPE
Training	0	0.231053	0.180675	0.843304	0.841798	4.32716
Test	0	0.241528	0.187329	0.838262	0.834589	4.541924

- Model Assumptions:
  - The model is able to explain ~ 84% of the variation in the data.
  - The train and test RMSE and MAE are low and comparable. The model is not overfitting.
  - The MPAE on the test set suggest we can predict within 4.54% of the normalized new price.
  - The Final model is good for prediction as well as inference purposes.

Link to Appendix slide on OLS model used for prediction

Actionable Insights and Recommendations

- **INSIGHTS**

1. The model is able to explain ~ 84% of the variation in the data within 4.54% of the normalized new price which is good for prediction and inference purposes.
2. The normalized new price demonstrated a strong positive coefficient with the normalized used price.
3. Phones with a newer release year are able to command a higher price.
4. Positive coefficients are realized for selfie camera, RAM, weight, normalized new price, 4G, and main cameras with 16MP or greater.
5. Positive coefficients are realized for Nokia and Xiaomi name brands.
6. The Samsung brand and devices with main camera configuration of less than 16.0 MP appear to have negative coefficients.
7. Very few of the devices in the data collection have 5g network capability (less than 5%).
8. Android is the dominated OS with ~ 93% of the market share.
9. Approximately a third of the devices do not have 4g network capability.
10. Number of days used appears to have a negative correlation to the price.

Actionable Insights and Recommendations cont…

- **RECOMMENDATIONS**

1. Use this model to make predictions for the price of used devices.
2. Offer new devices as they tend to resale for higher prices.
3. Offer devices with high quality cameras of 16MP or higher. (Sony appears to dominate this area with approximately 39.4% of the market share).
4. Look to elevate the following brands on the online marketplace: Nokia and Xiaomi.
5. The android operating system is dominating the marketplace, focus on offering android devices to the marketplace.
6. Very few devices in the data set have 5g network capability it is recommended to build another model with data collection that include more 5g devices to determine if 5g network capability has a significant impact on demand and resale price.
7. Future data collection and models may also want to expand on additional service offerings (insurance, warranties, repair, etc…) to determine if these offering may drive demand and prices higher.
8. Future data collection and models may also want to expand on income and other demographics to determine a refined target audience to maximize resale efforts.
9. Build a marketing demand around the brand, cost savings, and the positive environmental impact ReCell is providing by recycling and reducing waste.

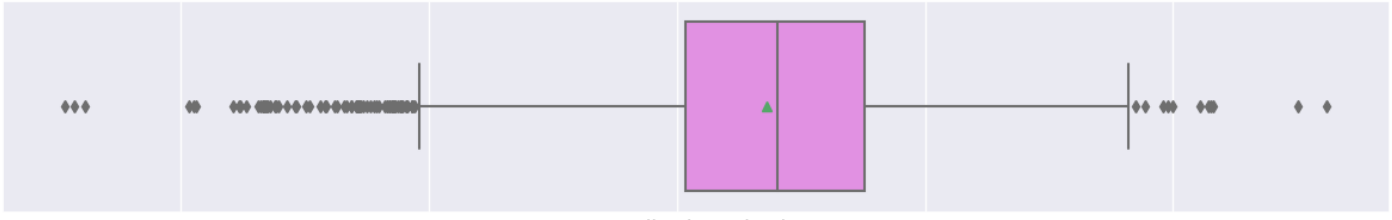
APPENDIX

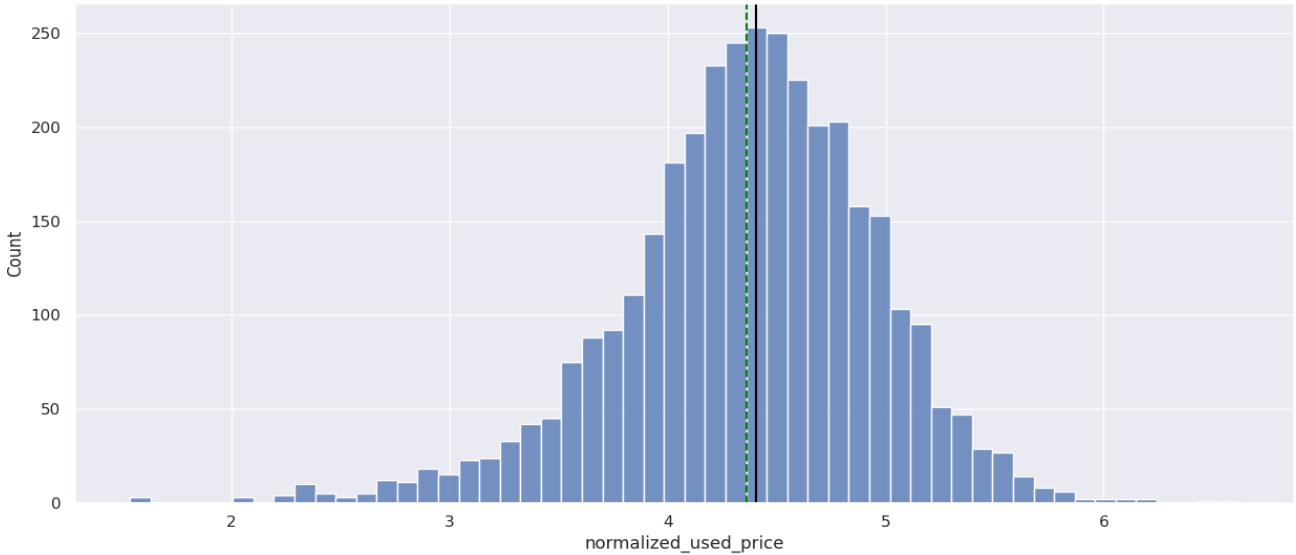
Statistical summary of the dataset

index	count	unique	top	freq	mean	std	min	25%	50%	75%	max
brand_name	3454	34	Others	502	NaN	NaN	NaN	NaN	NaN	NaN	NaN
os	3454	4	Android	3214	NaN	NaN	NaN	NaN	NaN	NaN	NaN
screen_size	3454.0	NaN	NaN	NaN	13.713115228720325	3.805279597858718	5.08	12.7	12.83	15.34	30.71
4g	3454	2	yes	2335	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5g	3454	2	no	3302	NaN	NaN	NaN	NaN	NaN	NaN	NaN
main_camera_mp	3275.0	NaN	NaN	NaN	9.460207633587787	4.8154612416365135	0.08	5.0	8.0	13.0	48.0
selfie_camera_mp	3452.0	NaN	NaN	NaN	6.554229432213209	6.970371918748835	0.0	2.0	5.0	8.0	32.0
int_memory	3450.0	NaN	NaN	NaN	54.57309855072464	84.97237057162442	0.01	16.0	32.0	64.0	1024.0
ram	3450.0	NaN	NaN	NaN	4.036121739130435	1.3651046591081606	0.02	4.0	4.0	4.0	12.0
battery	3448.0	NaN	NaN	NaN	3133.402697215777	1299.6828438751706	500.0	2100.0	3000.0	4000.0	9720.0
weight	3447.0	NaN	NaN	NaN	182.75187119234116	88.41322781358869	69.0	142.0	160.0	185.0	855.0
release_year	3454.0	NaN	NaN	NaN	2015.965257672264	2.298454568407353	2013.0	2014.0	2015.5	2018.0	2020.0
days_used	3454.0	NaN	NaN	NaN	674.8697162709901	248.5801658671337	91.0	533.5	690.5	868.75	1094.0
normalized_used_price	3454.0	NaN	NaN	NaN	4.364712079133227	0.5889136178484645	1.536867219599265	4.033930852723032	4.405132623388357	4.755700007809356	6.619433001642933
normalized_new_price	3454.0	NaN	NaN	NaN	5.233107171901414	0.6836368428653764	2.9014215940827497	4.790341843401424	5.2458918493708575	5.673718250179289	7.847840659422

Univariate Analysis

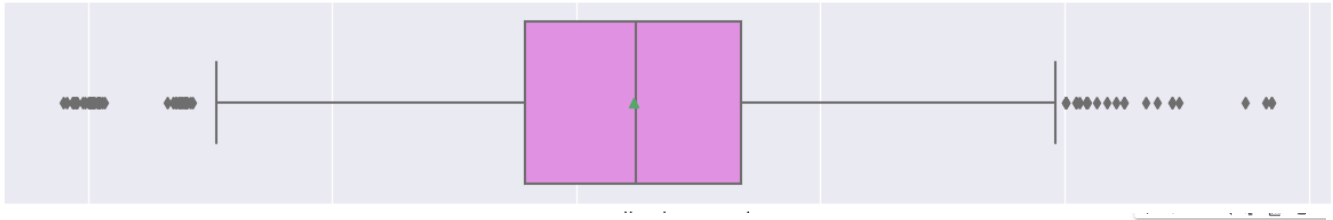
- **normalized_used_price:**
  - The normalized used price looks to have a normal distribution with some outliers at both higher and lower ends.
  - The average normalized new price is $4.36.

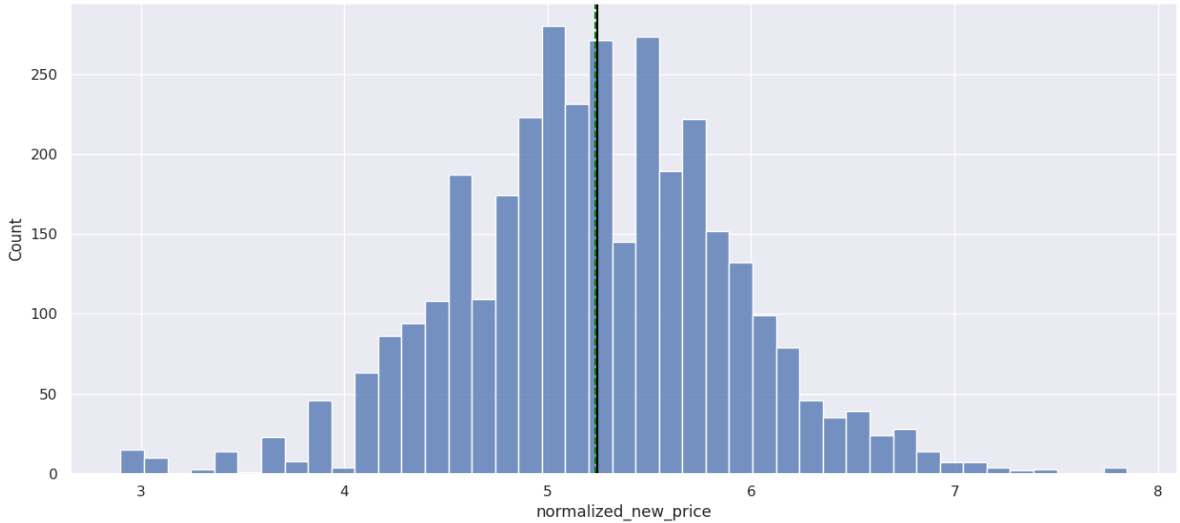




Univariate Analysis Cont...

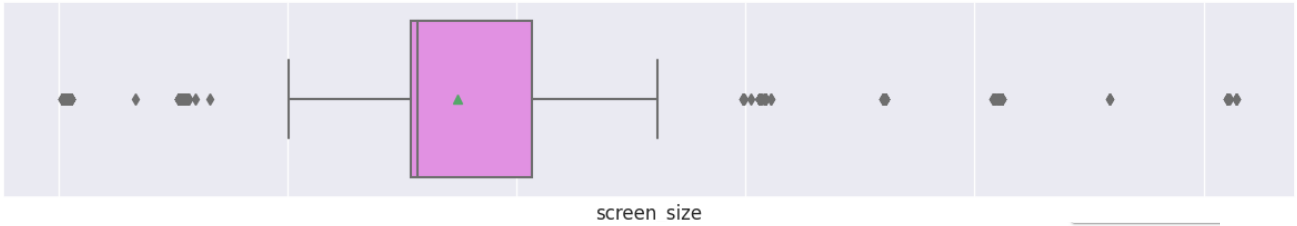
- **normalized_new_price:**
  - The normalized used price looks to have a normal distribution with some outliers at both higher and lower ends.
  - The average normalized new price is $5.23.

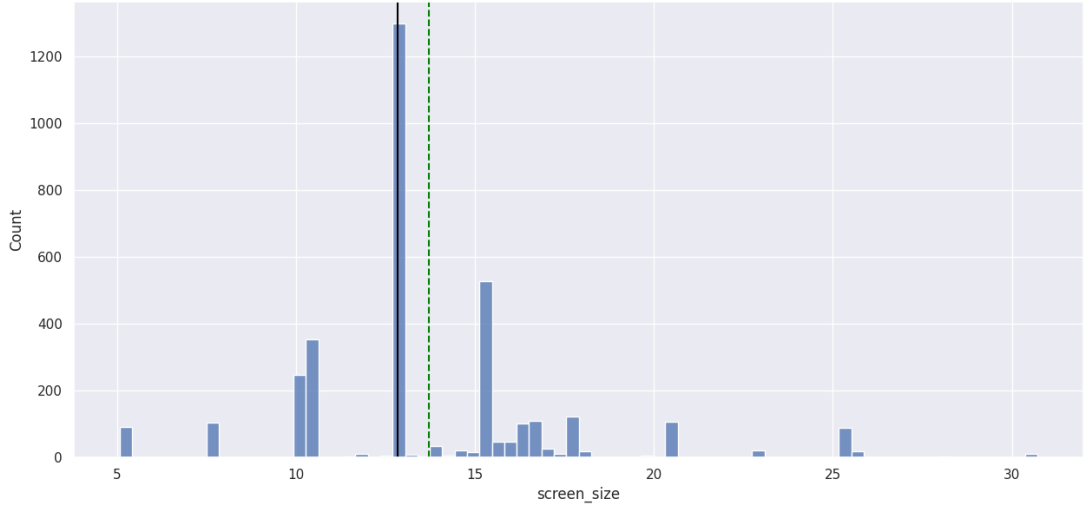




Univariate Analysis Cont...

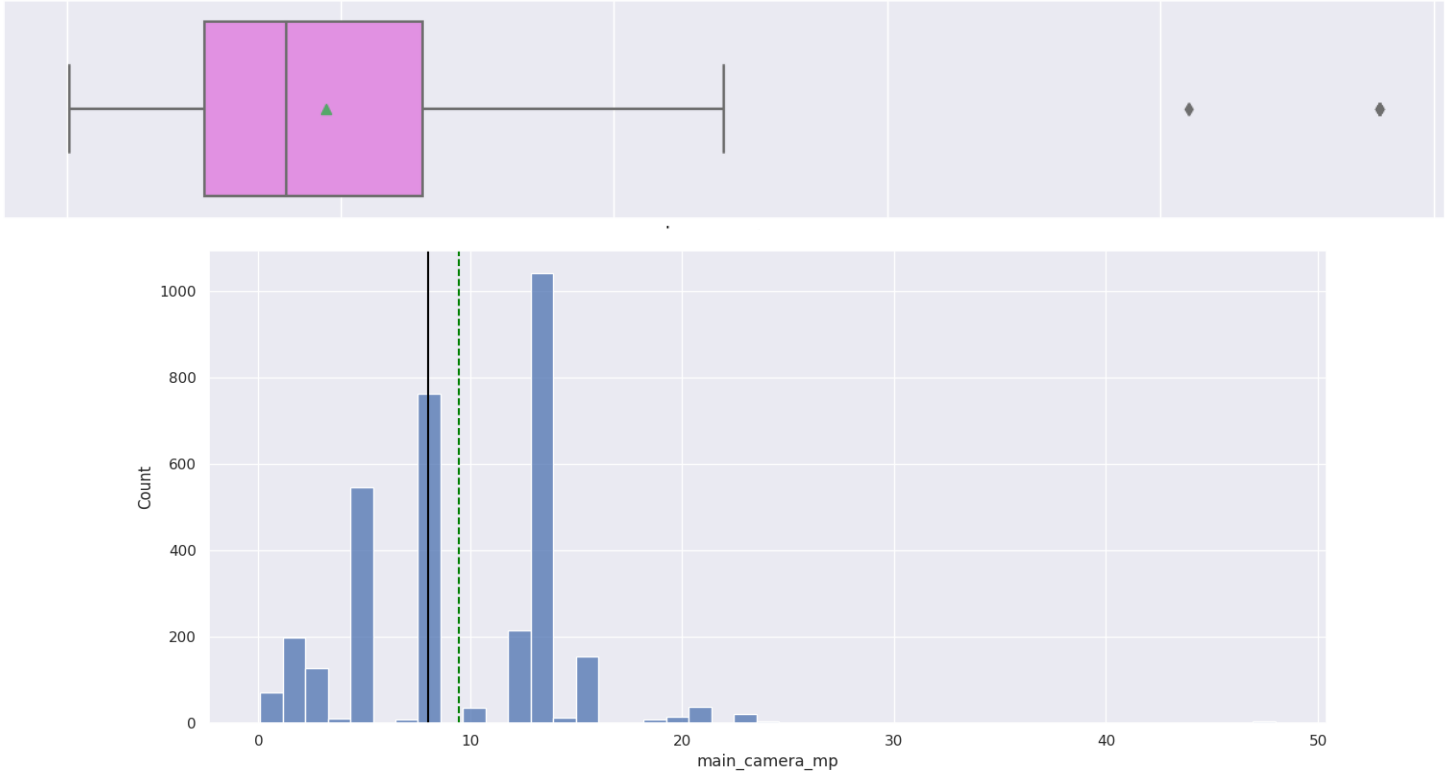
- **screen_size:**
  - The screen sized does not have a clearly defined pattern. It does appear to have normal distribution skewed slightly right. There are outliers on both higher and lower ends.
  - The average screen size is 13.71 cm.





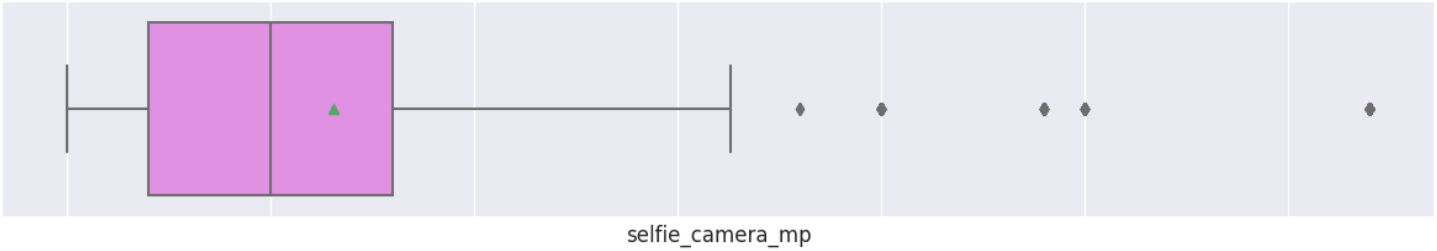
Univariate Analysis Cont...

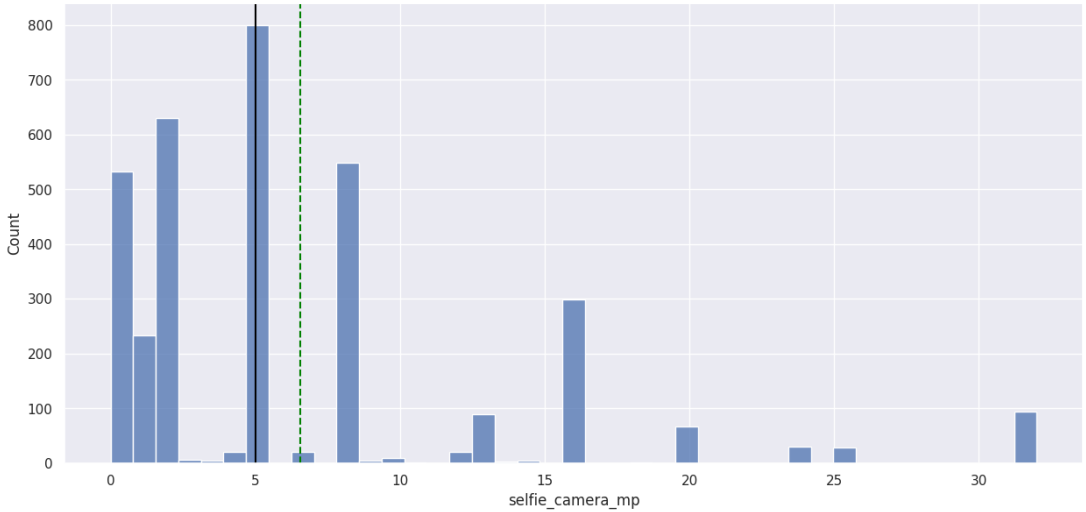
- **main_camera_mp:**
  - The distribution for the resolution of the main camera appears to be skewed slightly left with outliers at the upper end.
  - The average resolution is 9.46 MP.



Univariate Analysis Cont...

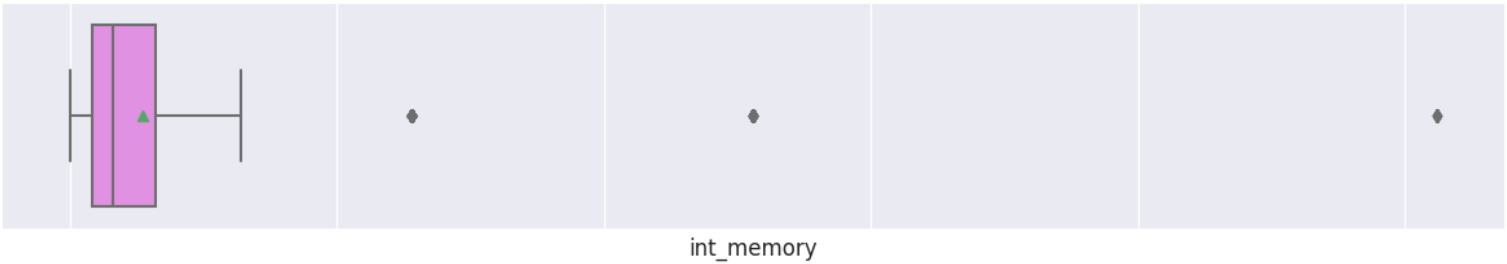
- **selfie_camera_mp:**
  - The distribution for the resolution of the selfie camera appears to be skewed slightly right with outliers at the upper end.
  - The average resolution is 6.55 MP.

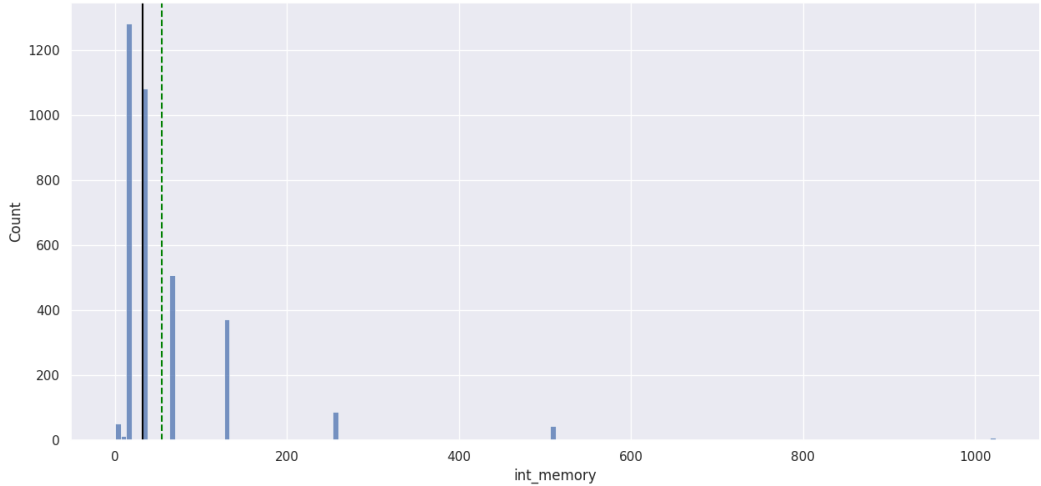




Univariate Analysis Cont...

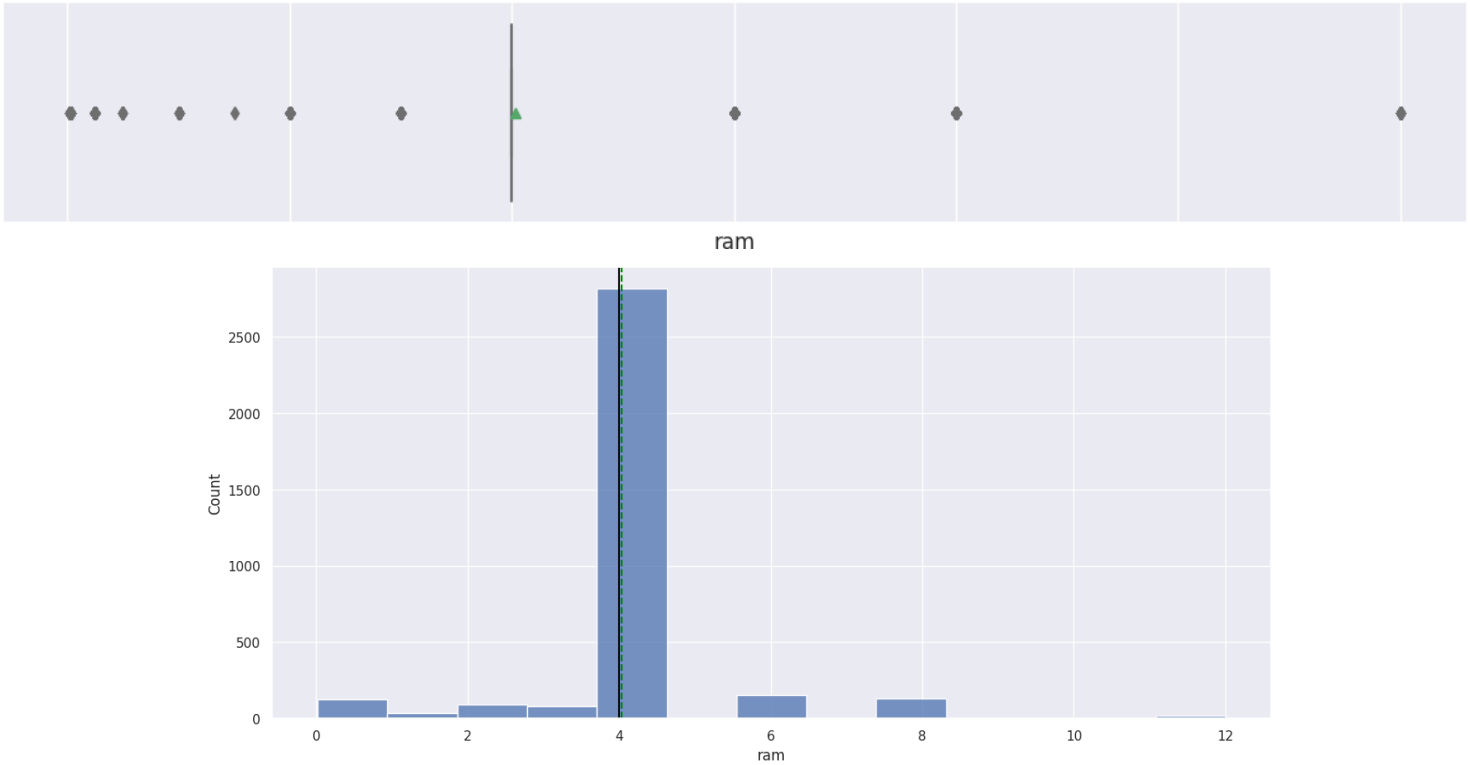
- **int_memory:**
  - The distribution for amount of internal memory appears to be skewed right with outliers at the upper end.
  - The average amount of internal memory is 54.57 GB.





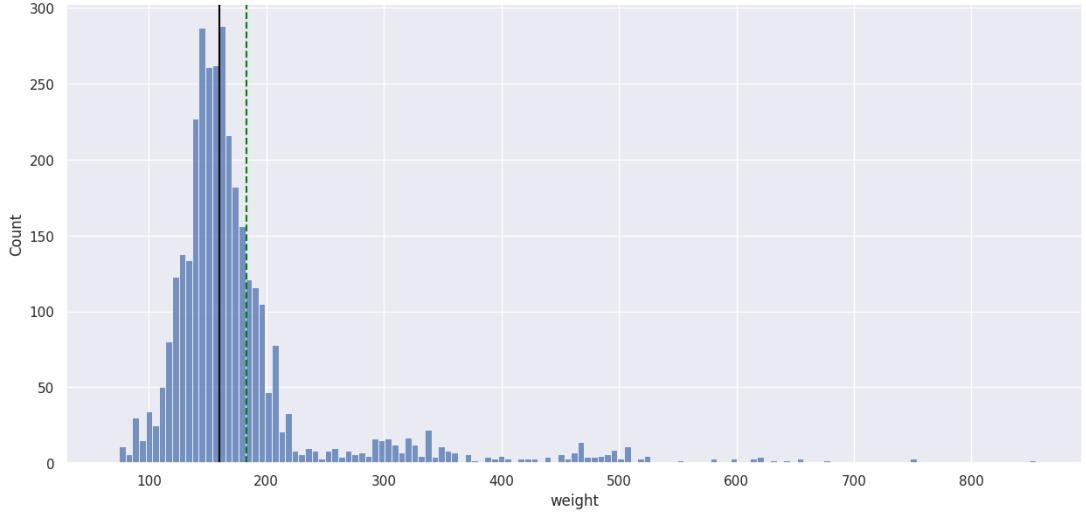
Univariate Analysis Cont...

- **RAM:**
  - The amount of RAM on the devices may have be skewed slightly right; however, the distribution does not appear to be concerning . There are outliers on both the higher and lower ends.
  - The average amount of RAM is 4.04 GB.



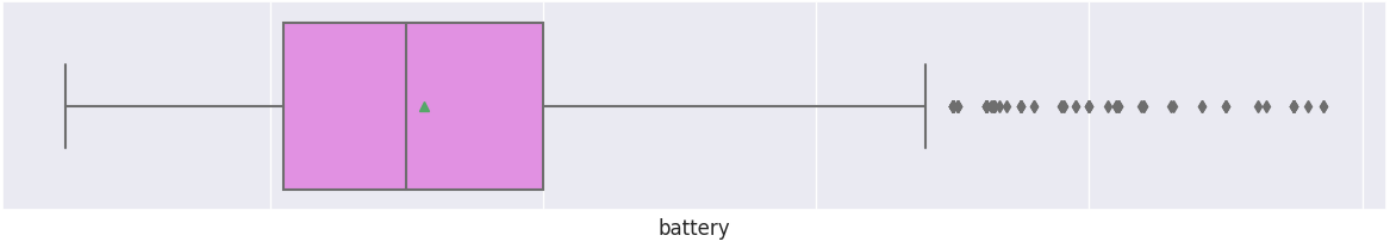
Univariate Analysis Cont...

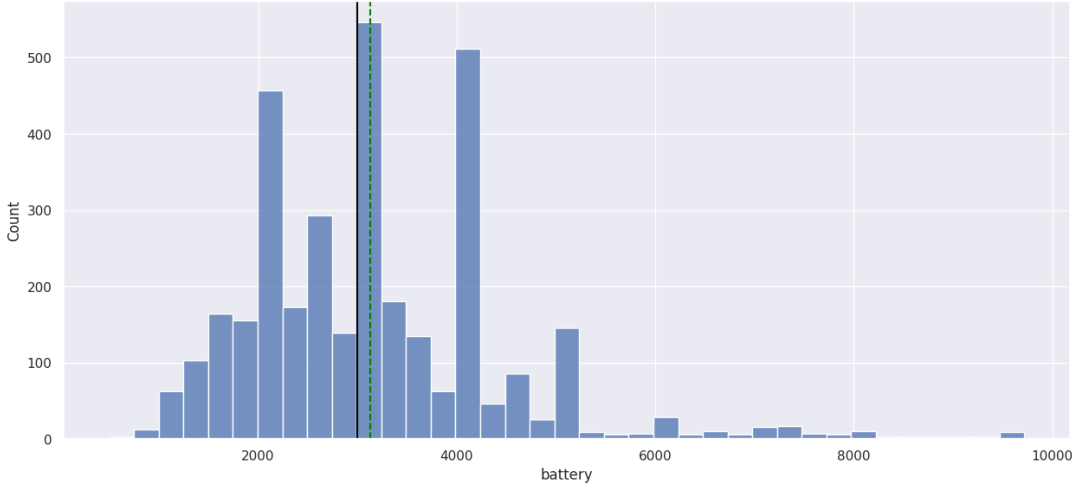
- **Weight**
  - The distribution weight does appear to be skewed right. There appears to be a lot pf outliers on the upper end and fewer on the lower end.
  - The average weight is 182.7g.



Univariate Analysis Cont...

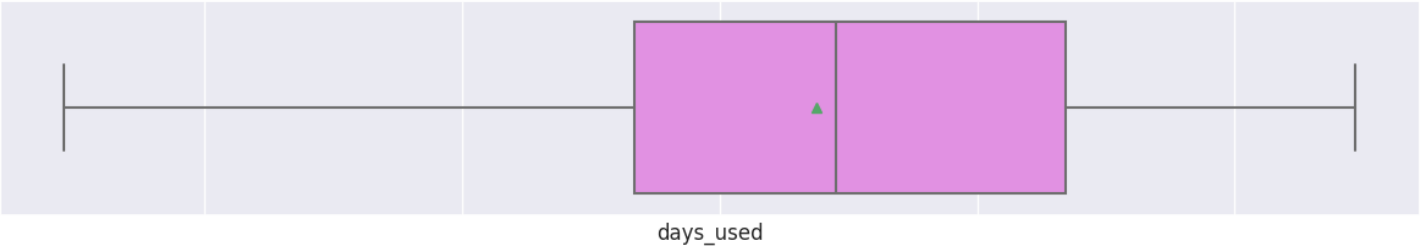
- **Battery:**
  - The distribution of the battery capacity for the devices appears to be a type of Bernoulli distribution with outliers on the upper end.
  - The average battery capacity is 3133.40 mAh.

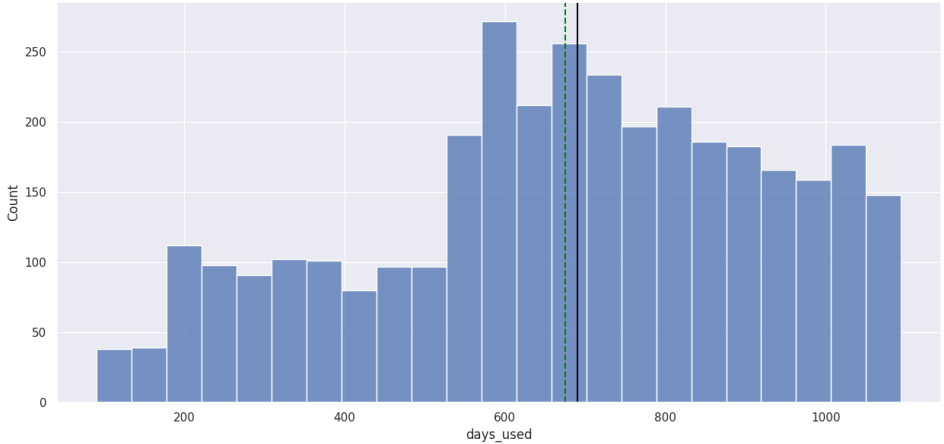




Univariate Analysis Cont...

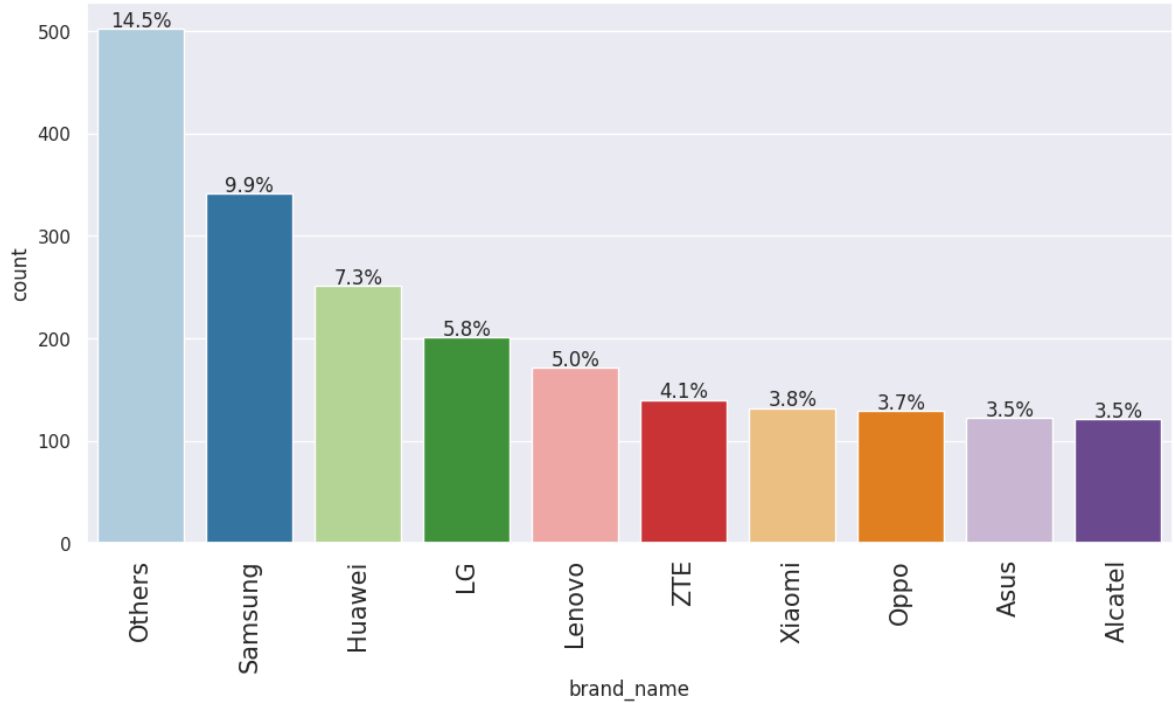
- **days_used**
  - The distribution of the number of days appears to be skewed slightly left with no apparent outliers.
  - The average number of days used is approximately 675 days.





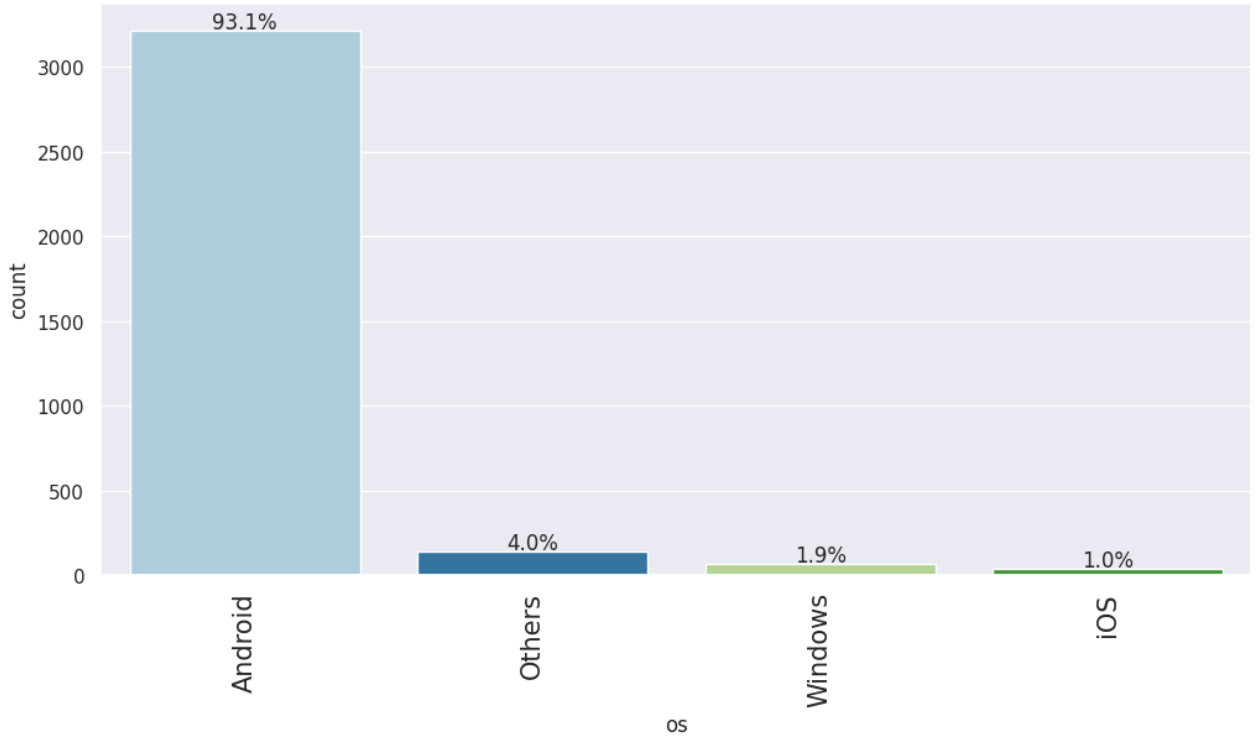
Univariate Analysis Cont...

- **brand_name**
  - It appears not one brand name dominates the market. In fact, brand names that fall under ”others” captures 14.5% of the market. Followed by Samsung at 9.9% and Huawei with 7.3%.



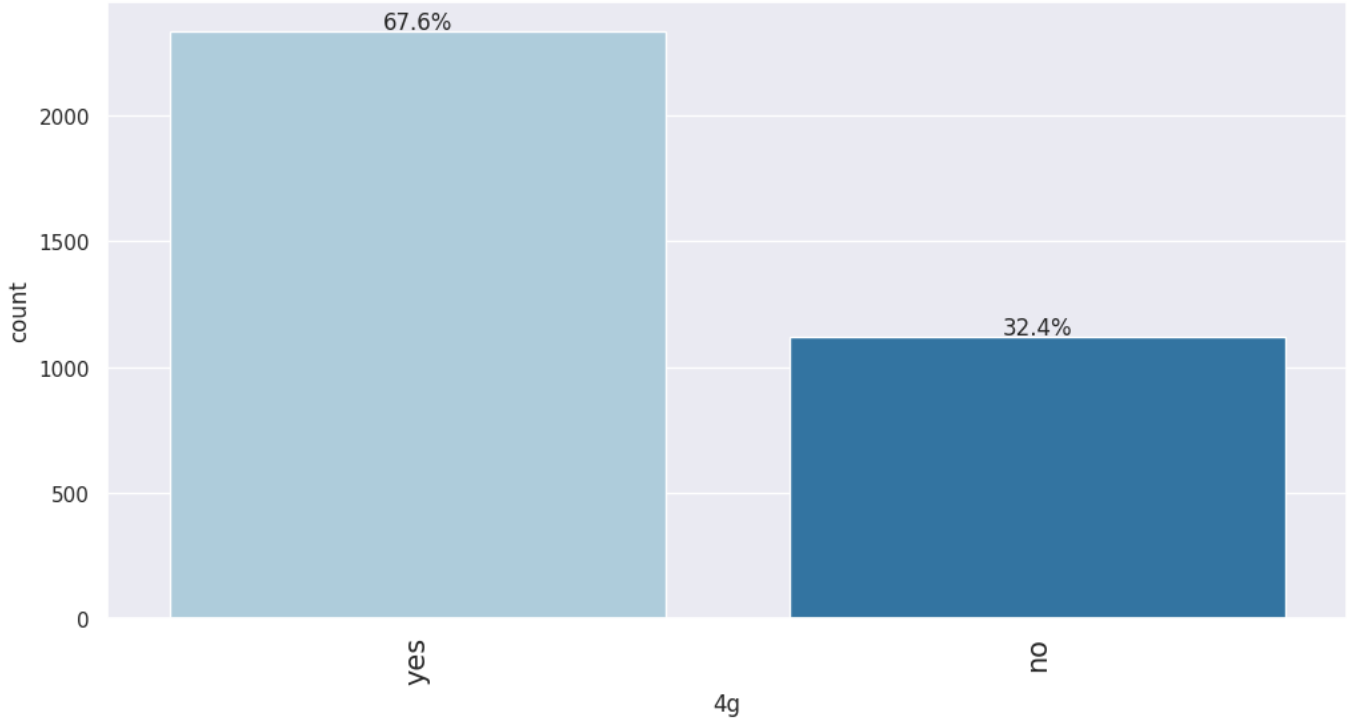
Univariate Analysis Cont...

- OS
  - Android appears to dominate the market at 93.1% of the market share where iOS holds the smallest market share at 1%



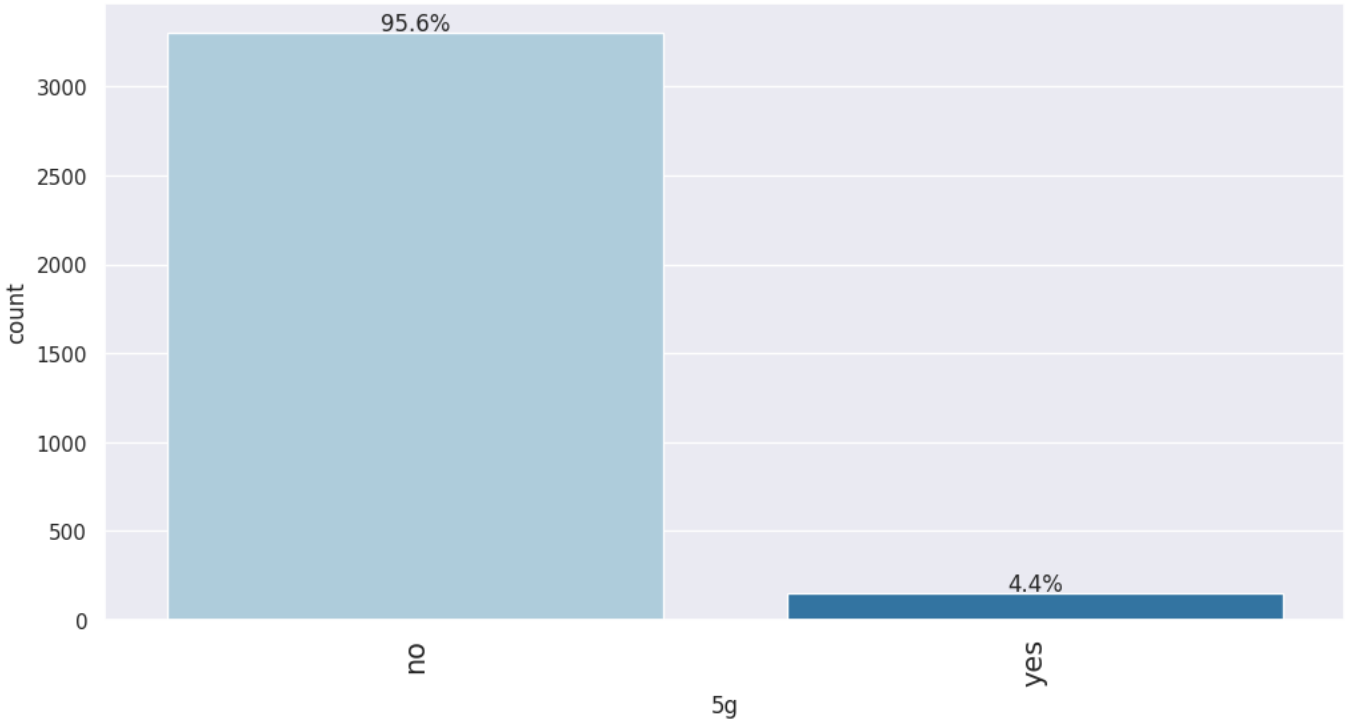
Univariate Analysis Cont...

- 4g
  - Approximately a third of the devices are available with 4g at 67.6%



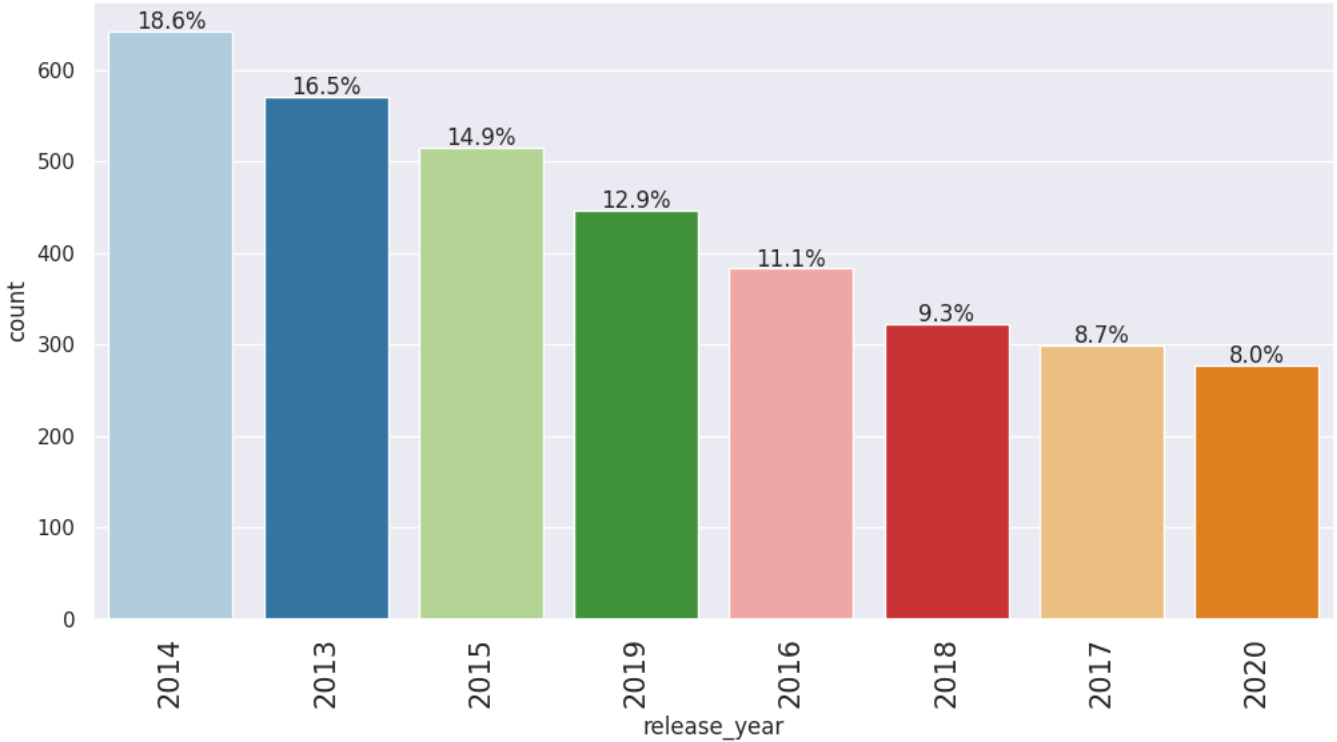
Univariate Analysis Cont...

- 5g
  - Almost all devices are not available with 5g - 95.6%



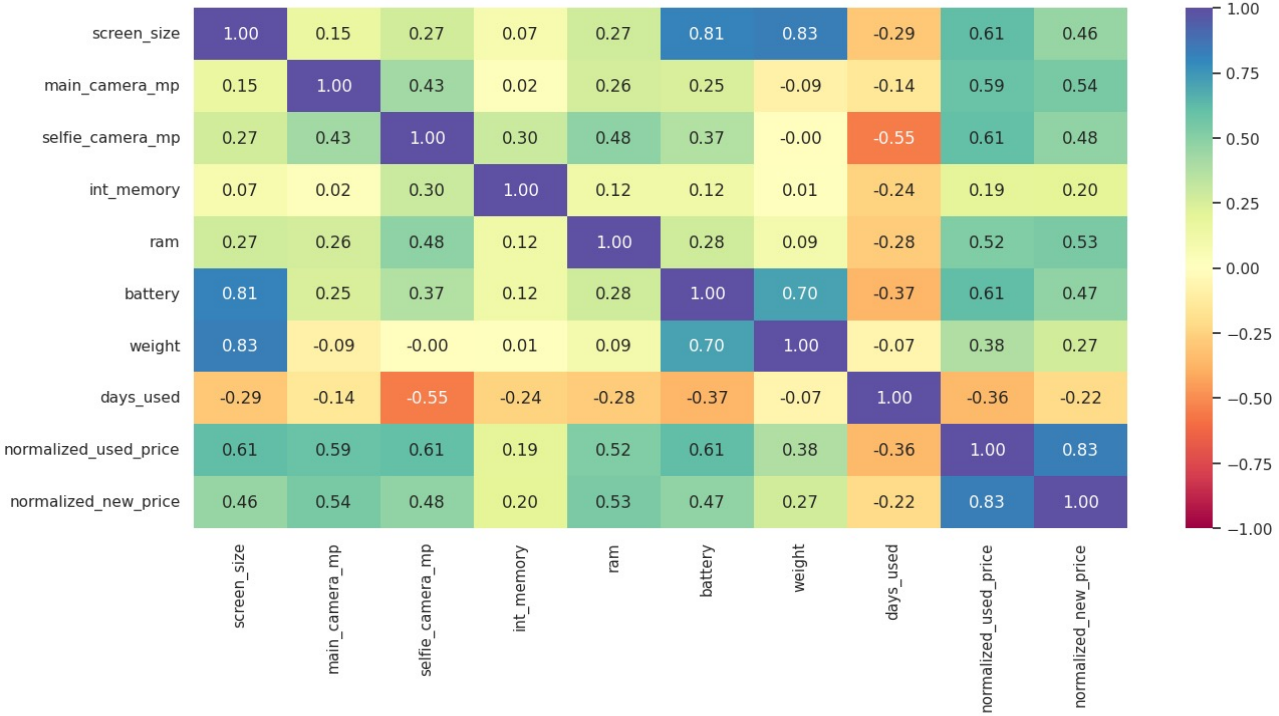
Univariate Analysis Cont...

- release_year
  - The most prominent release year is 2014 at 18.6% followed by 2013(16.5%) and 2015 (14.9%)



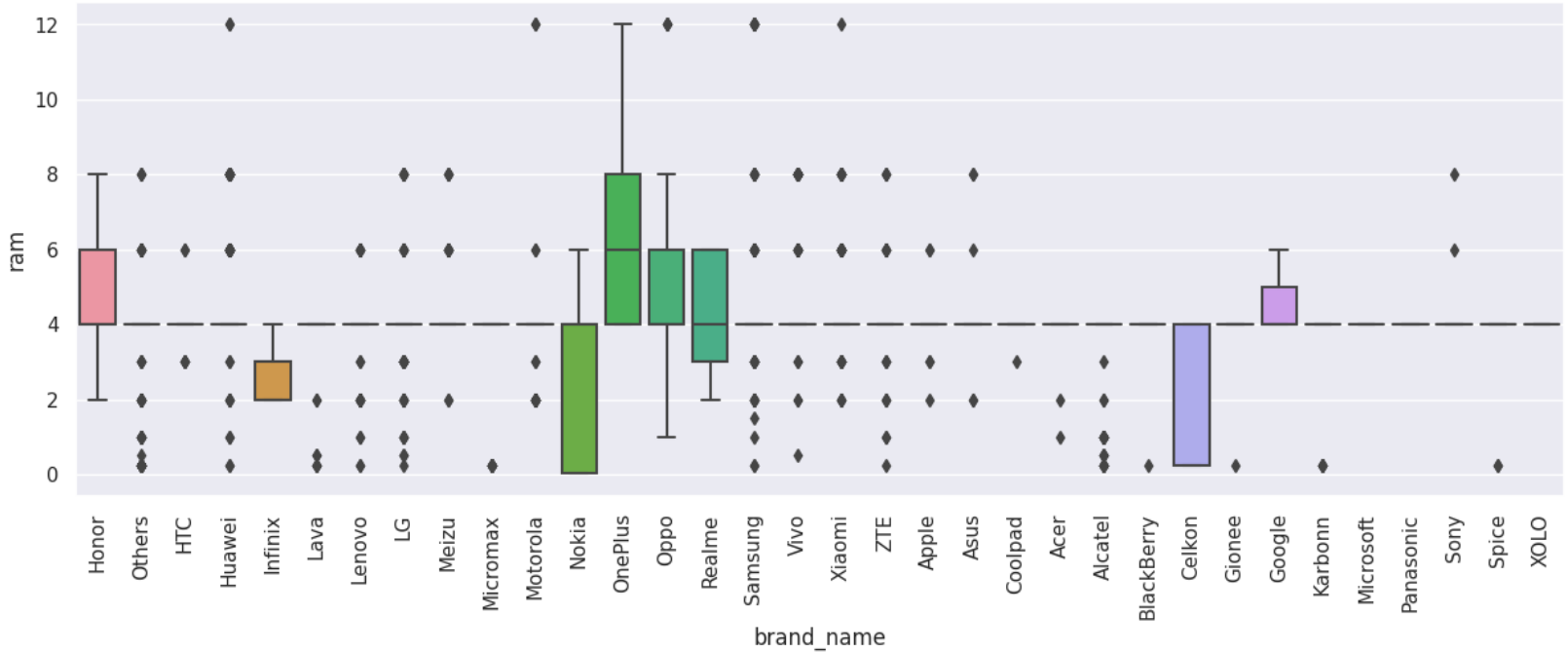
Bivariate Analysis

- **Correlation Check:**
  - The normalized used price demonstrates a high positive correlation with the normalized new price, screen size, selfie camera MP, battery, main camera MP, and RAM.
  - The normalized used price demonstrates a negative correlation with days used.



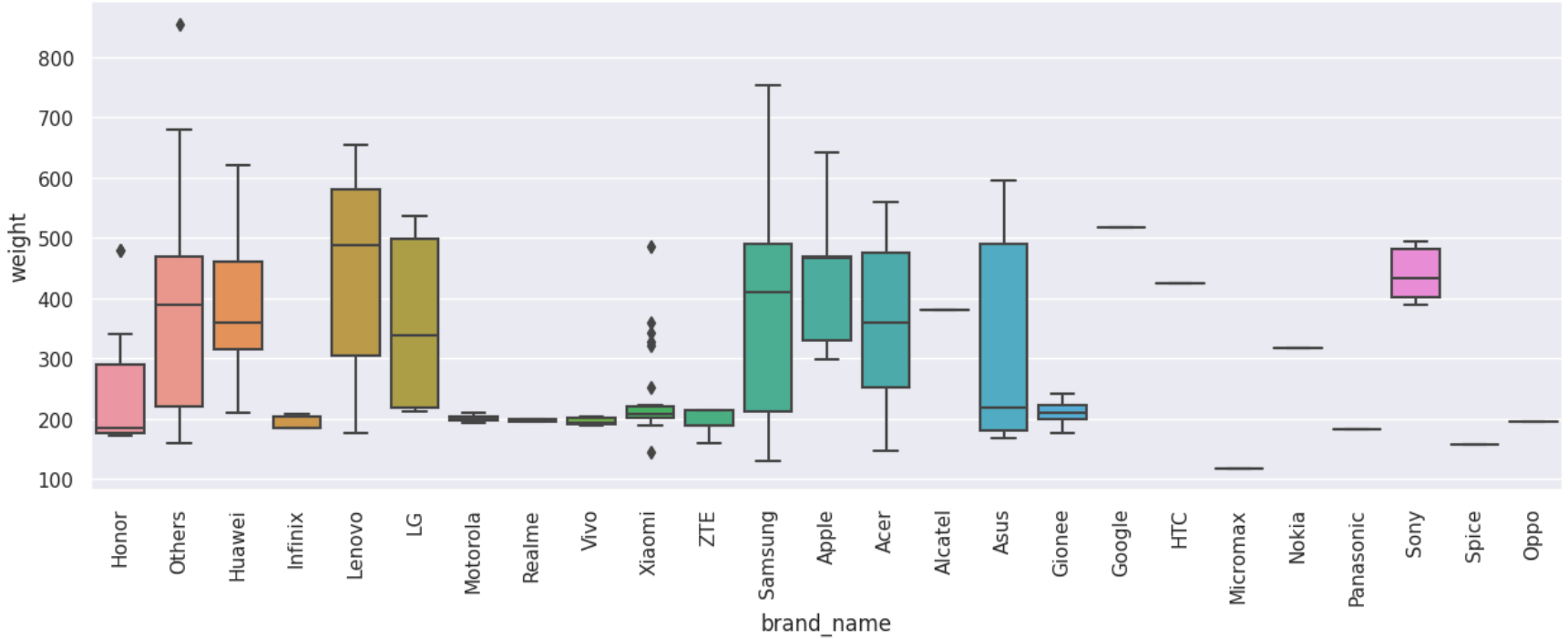
Bivariate Analysis Cont...

- **brand_name / ram:**
  - The brand that has the most RAM is OnePlus and the one that has the least amount of RAM appears to be Celkor.
  - There are many outliers and there does not appear a normal distribution among the data.
  - The distribution of RAM among the brands appear to have highly prevalent differences.



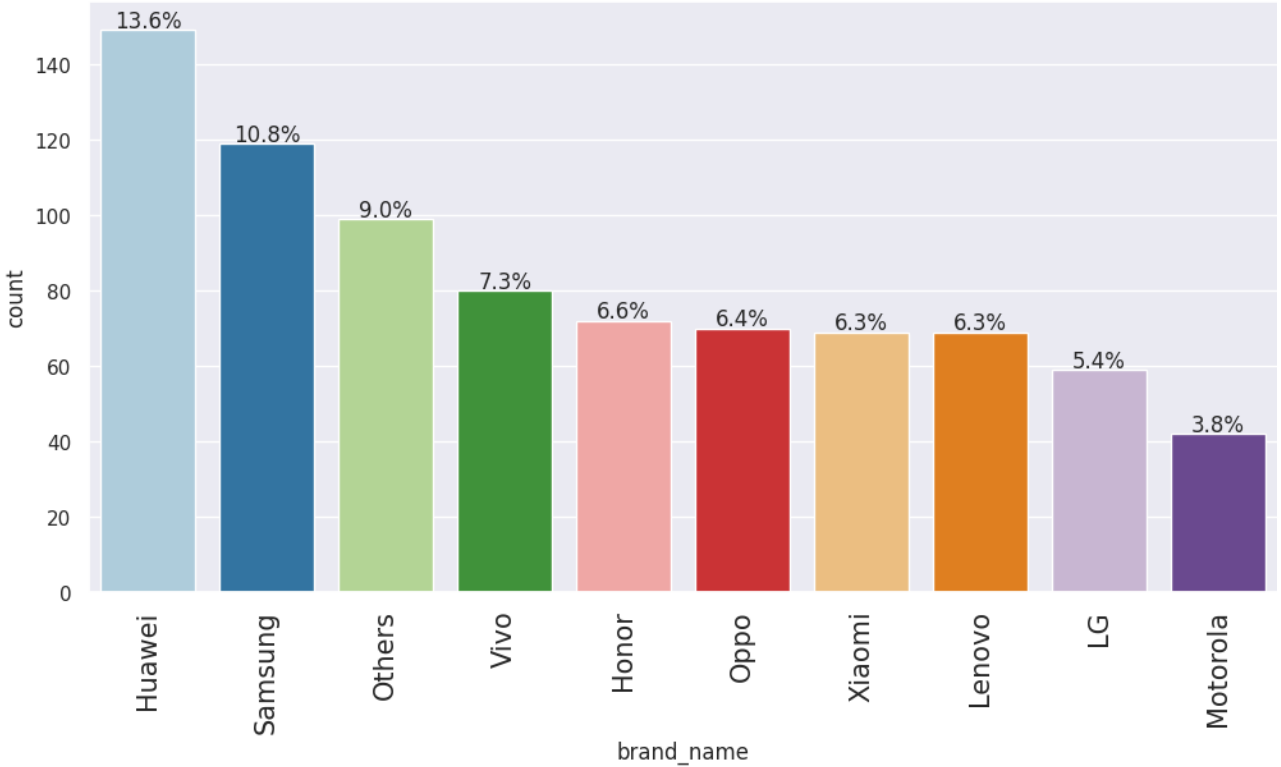
Bivariate Analysis Cont...

- **brand_name / weight**
  - The brand that is the heaviest is Google and the one that weights the least appears to be Micromax.
  - There are many outliers and there does not appear a normal distribution among the data.
  - The distribution of weight among the brands appear to have highly prevalent differences.



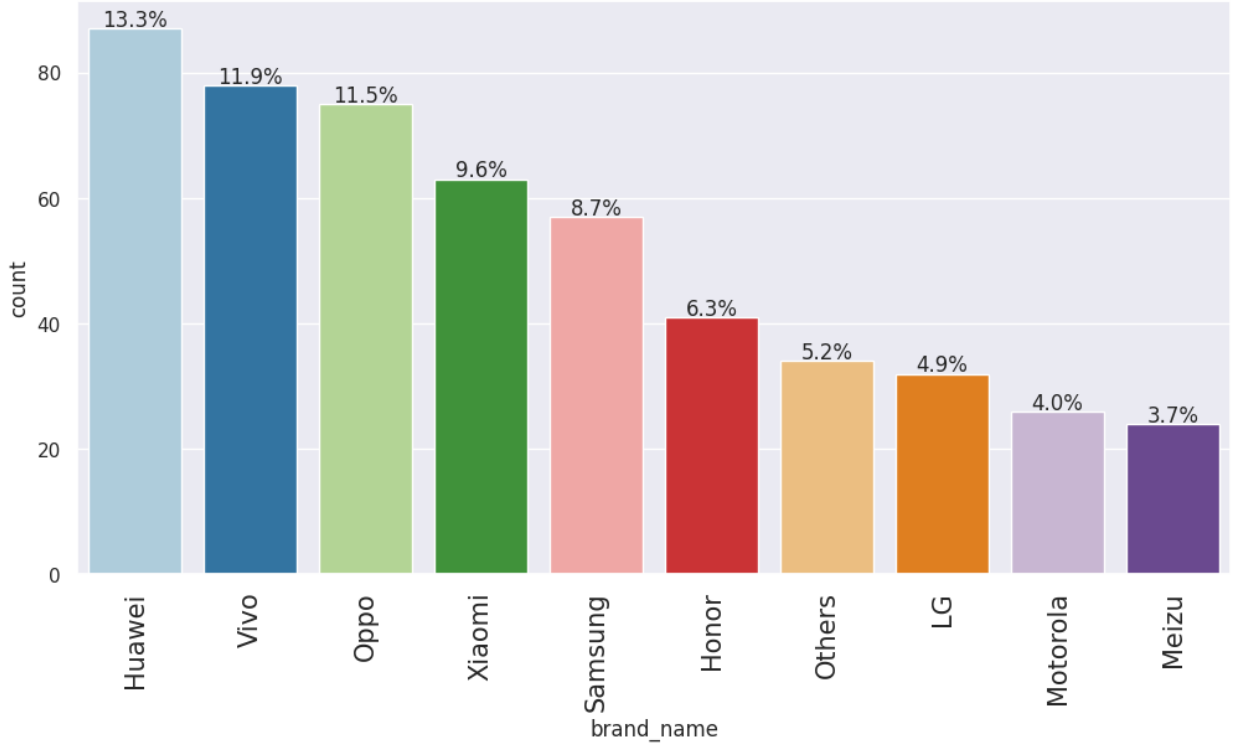
Bivariate Analysis Cont...

- **Large screen / brand_name:**
  - The brand that has largest amount of devices with large screens is Huawei (13.6%) followed by Samsung (10.8%)
    - *Note: Large Screen = Greater than 6 x 2.54 cm*



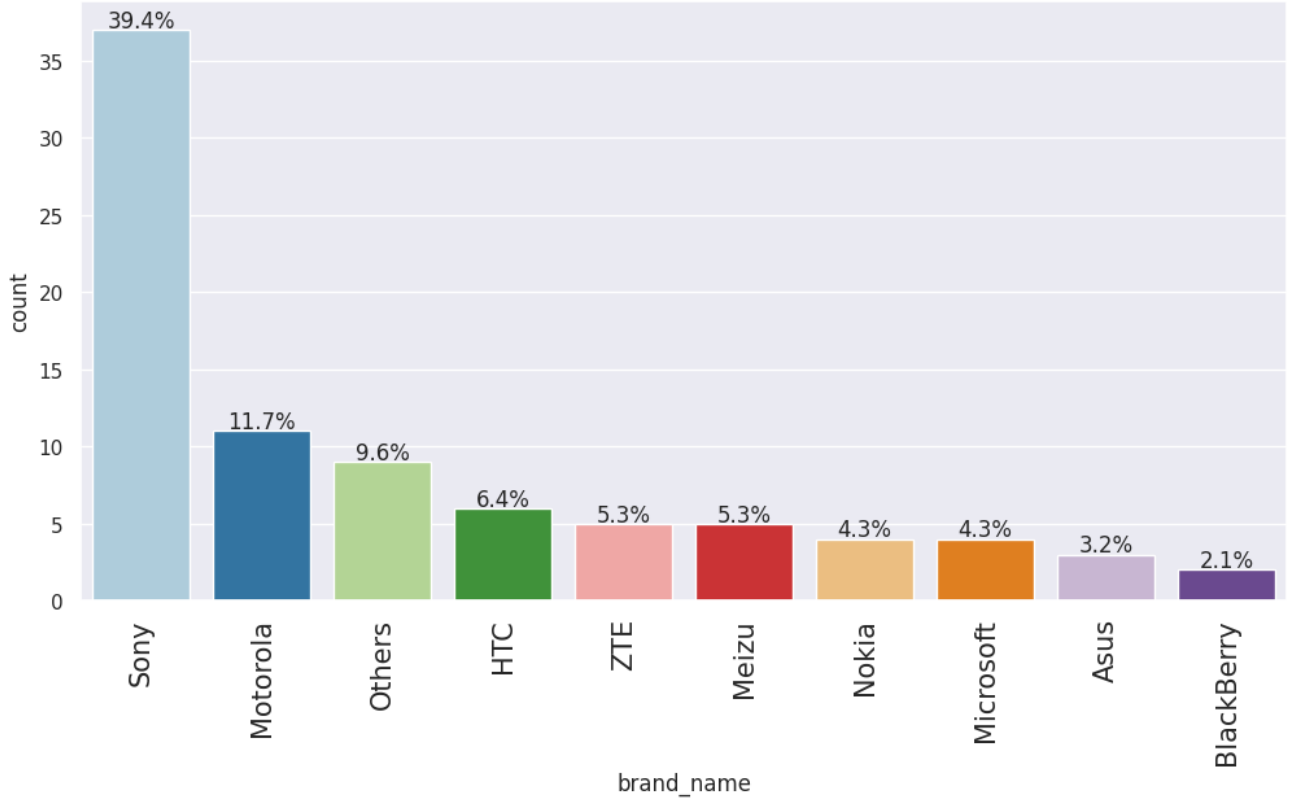
Bivariate Analysis Cont...

- **Selfie Camera / brand_name:**
  - The brand that has the most devices with cameras more than 8 MP is Huawei (13.3%) followed by Vivo (11.9%) and Oppo (11.5%). These three brands appear to capture over third of the market share.



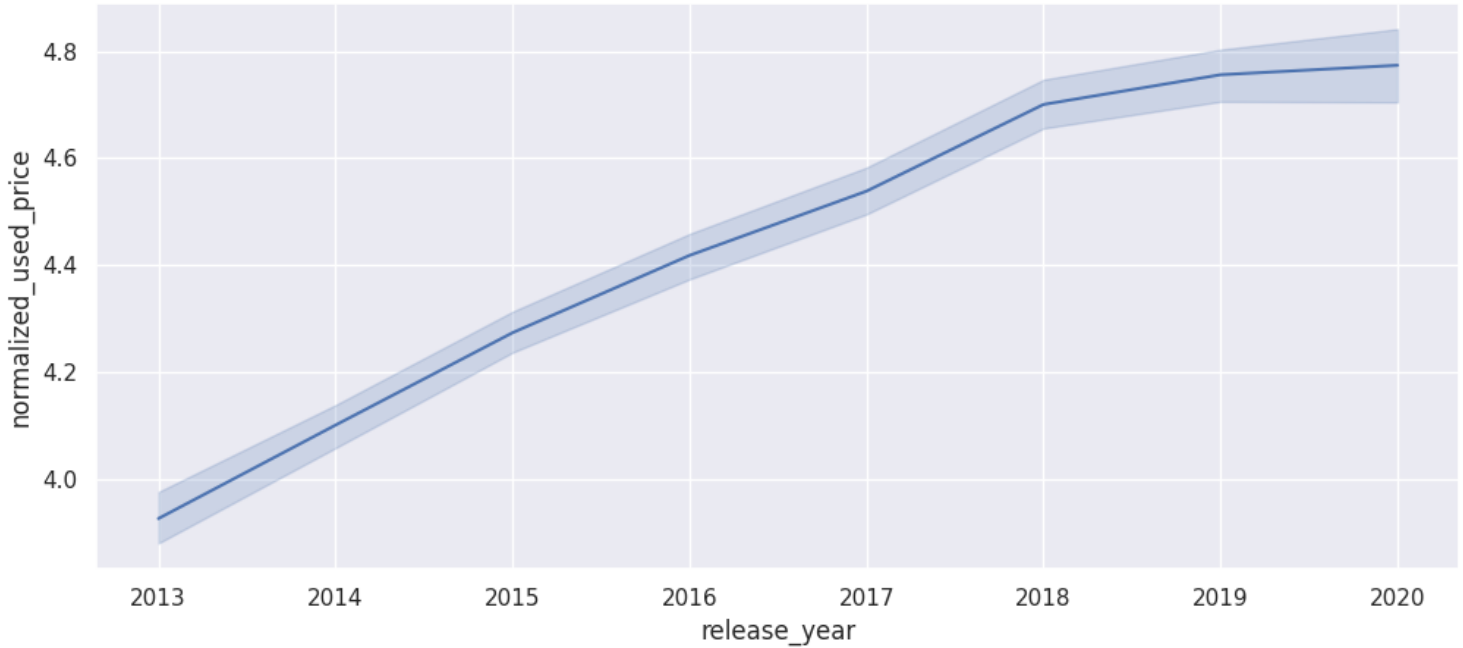
Bivariate Analysis Cont...

- **Main (Rear) Camera / brand_name:**
  - This is one brand that captures a significant portion of the market that has a main camera of more than 16 MP – Sony at 39.4%.



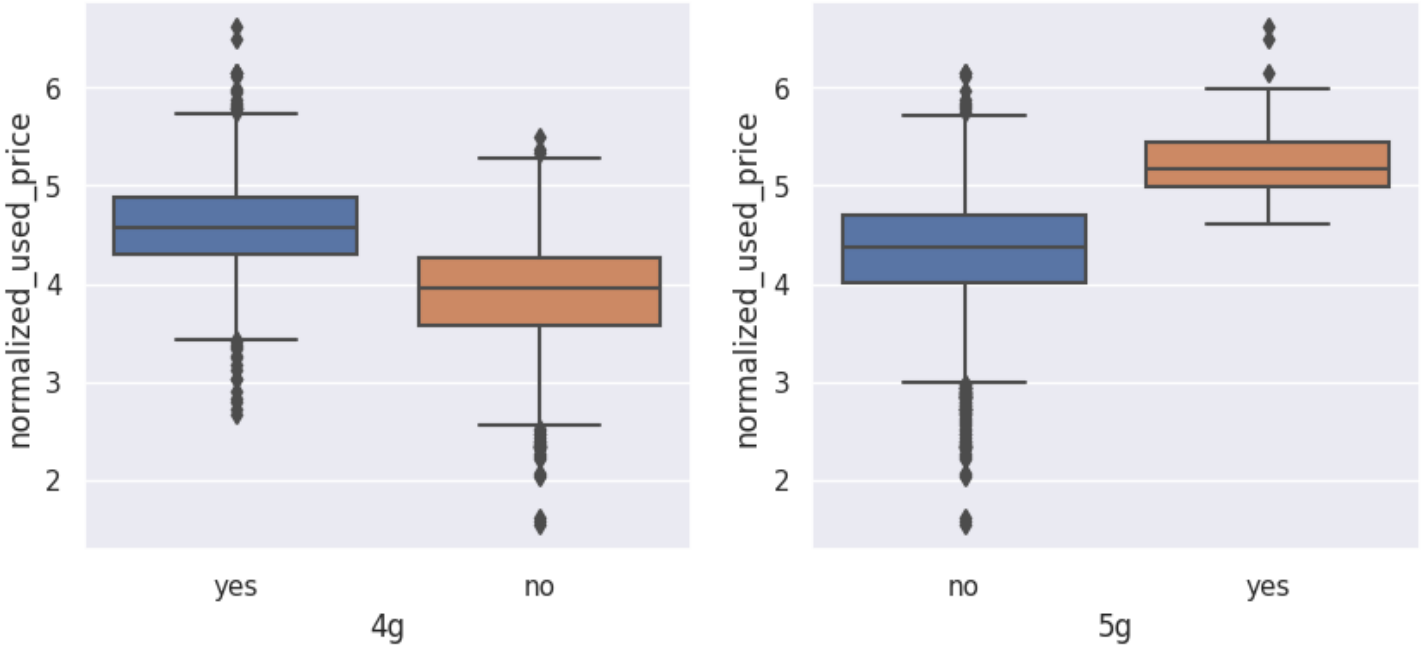
Bivariate Analysis Cont...

- **release_year / normalized_used_price:**
  - The most current release year appears to command the highest Normalized used price. As the release year of the device rises the prices rises as well demonstrating a positive correlation.



Bivariate Analysis Cont...

- normalized_used_price and 4g vs. 5g network available
  - It appears that devices that offer 4g network availability have a higher normalized used price than others.
  - It appears that devices that offer 5g network availability have a higher normalized used price than others.
  - Devices that offer 5g network availability appear to command higher prices over devices offering 4g network availability.



Missing Value Treatment

Initial Missing Value Set

```
brand_name                0
os                        0
screen_size               0
4g                        0
5g                        0
main_camera_mp          179
selfie_camera_mp          2
int_memory                4
ram                       4
battery                   6
weight                    7
release_year              0
days_used                 0
normalized_used_price     0
normalized_new_price      0
dtype: int64
```

Missing Value Set – After Treatment 1

```
brand_name                0
os                        0
screen_size               0
4g                        0
5g                        0
main_camera_mp          179
selfie_camera_mp          2
int_memory                0
ram                       0
battery                   6
weight                    7
release_year              0
days_used                 0
normalized_used_price     0
normalized_new_price      0
dtype: int64
```

Missing Value Set – After Treatment 2

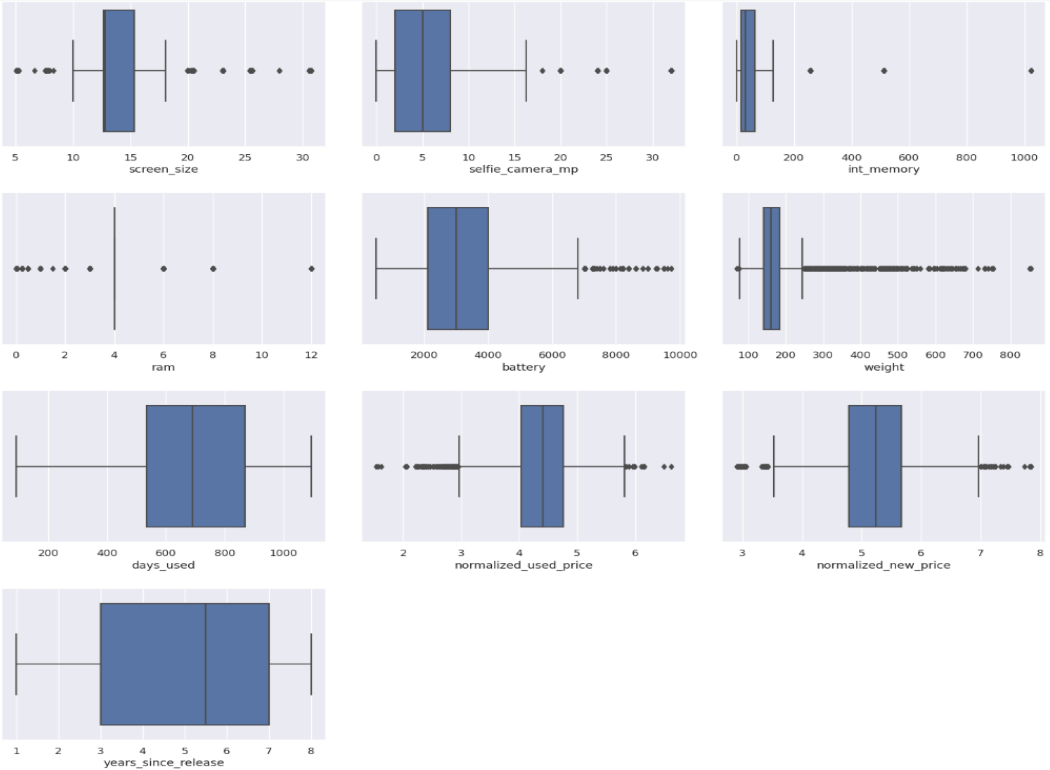
```
brand_name                0
os                        0
screen_size               0
4g                        0
5g                        0
main_camera_mp           10
selfie_camera_mp          0
int_memory                0
ram                       0
battery                   0
weight                    0
release_year              0
days_used                 0
normalized_used_price     0
normalized_new_price      0
dtype: int64
```

Missing Value Set – After Treatment 3

```
brand_name                0
os                        0
screen_size               0
4g                        0
5g                        0
main_camera_mp            0
selfie_camera_mp          0
int_memory                0
ram                       0
battery                   0
weight                    0
release_year              0
days_used                 0
normalized_used_price     0
normalized_new_price      0
dtype: int64
```

Outlier Check

- There are outliers in the data.
- These outliers will not be treated as they are proper variables.



Feature Engineering

Years Since Release

count	3454.000000
mean	5.034742
std	2.298455
min	1.000000
25%	3.000000
50%	5.500000
75%	7.000000
max	8.000000

Model Performance Summary:

OLS Linear Regression Results:

(I could not get this compress to the model looking OLS Results- had to resort to a series of screen shots)

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.850
Model:	OLS	Adj. R-squared:	0.845
Method:	Least Squares	F-statistic:	159.2
Date:	Sun, 12 Nov 2023	Prob (F-statistic):	0.00
Time:	21:46:43	Log-Likelihood:	163.90
No. Observations:	2417	AIC:	-159.8
Df Residuals:	2333	BIC:	326.6
Df Model:	83		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6815	0.080	20.959	0.000	1.524	1.839
screen_size	0.0238	0.003	6.865	0.000	0.017	0.031
selfie_camera_mp	0.0137	0.001	12.146	0.000	0.012	0.016
int_memory	0.0002	7.09e-05	2.259	0.024	2.11e-05	0.000
ram	0.0223	0.005	4.274	0.000	0.012	0.032
battery	-1.753e-05	7.39e-06	-2.373	0.018	-3.2e-05	-3.04e-06
weight	0.0011	0.000	7.982	0.000	0.001	0.001
days_used	2.496e-05	3.11e-05	0.803	0.422	-3.6e-05	8.6e-05
normalized_new_price	0.4146	0.013	32.372	0.000	0.390	0.440

OLS Linear Regression Results Cont…

years_since_release					
-0.0213	0.005	-4.560	0.000	-0.031	-0.012
brand_name_Alcatel					
0.0093	0.047	0.196	0.844	-0.084	0.102
brand_name_Apple					
-0.1129	0.147	-0.766	0.444	-0.402	0.176
brand_name_Asus					
0.0033	0.048	0.070	0.944	-0.090	0.097
brand_name_BlackBerry					
-0.0768	0.072	-1.067	0.286	-0.218	0.064
brand_name_Celkon					
-0.0499	0.067	-0.747	0.455	-0.181	0.081
brand_name_Coolpad					
0.0089	0.073	0.123	0.902	-0.133	0.151
brand_name_Gionee					
0.0425	0.058	0.738	0.461	-0.070	0.155
brand_name_Google					
0.0562	0.157	0.359	0.720	-0.251	0.364
brand_name_HTC					
-0.0112	0.049	-0.230	0.818	-0.107	0.084
brand_name_Honor					
0.0201	0.049	0.407	0.684	-0.076	0.117
brand_name_Huawei					
-0.0225	0.045	-0.502	0.616	-0.110	0.065
brand_name_Infinix					
0.0192	0.046	0.413	0.680	-0.072	0.110
brand_name_Karbonn					
0.0772	0.067	1.155	0.248	-0.054	0.208
brand_name_LG					
-0.0206	0.045	-0.455	0.649	-0.110	0.068
brand_name_Lava					
0.0197	0.062	0.317	0.752	-0.102	0.142
brand_name_Lenovo					
0.0340	0.045	0.751	0.453	-0.055	0.123
brand_name_Meizu					
-0.0198	0.056	-0.353	0.724	-0.130	0.090
brand_name_Micromax					
-0.0254	0.048	-0.533	0.594	-0.119	0.068

OLS Linear Regression Results Cont…

brand_name_Microsoft					
0.0642	0.089	0.720	0.472	-0.111	0.239
brand_name_Motorola					
-0.0037	0.050	-0.074	0.941	-0.102	0.094
brand_name_Nokia					
0.0781	0.052	1.498	0.134	-0.024	0.180
brand_name_OnePlus					
0.0680	0.077	0.881	0.378	-0.083	0.219
brand_name_Oppo					
0.0051	0.048	0.107	0.915	-0.089	0.099
brand_name_Others					
-0.0125	0.042	-0.297	0.767	-0.095	0.070
brand_name_Panasonic					
0.0428	0.056	0.768	0.443	-0.067	0.152
brand_name_Realme					
0.0156	0.062	0.253	0.800	-0.105	0.136
brand_name_Samsung					
-0.0477	0.043	-1.098	0.272	-0.133	0.037
brand_name_Sony					
-0.0476	0.055	-0.862	0.389	-0.156	0.061
brand_name_Spice					
-0.0184	0.063	-0.291	0.771	-0.142	0.106
brand_name_Vivo					
-0.0279	0.049	-0.575	0.565	-0.123	0.067
brand_name_XOLO					
-7.906e-05	0.055	-0.001	0.999	-0.107	0.107
brand_name_Xiaomi					
0.0650	0.048	1.342	0.180	-0.030	0.160
brand_name_ZTE					
-0.0027	0.048	-0.057	0.955	-0.097	0.091
os_Others					
0.0241	0.035	0.683	0.495	-0.045	0.093
os_Windows					
-2.594e-05	0.048	-0.001	1.000	-0.093	0.093
os_iOS					
0.0272	0.146	0.186	0.852	-0.259	0.313
4g_yes					
0.0479	0.016	2.975	0.003	0.016	0.080

OLS Linear Regression Results Cont…

```
5g_yes
-0.0665      0.032     -2.065      0.039      -0.130      -0.003
main_camera_mp_8.0
-0.1088      0.015     -7.331      0.000      -0.138      -0.080
main_camera_mp_5.0
-0.1803      0.019     -9.565      0.000      -0.217      -0.143
main_camera_mp_10.5
0.0275       0.055      0.501      0.616      -0.080       0.135
main_camera_mp_3.15
-0.2440      0.032     -7.688      0.000      -0.306      -0.182
main_camera_mp_<bound method NDFrame._add_numeric_operations.<locals>.median of 0       13.0
1       13.0
2       13.0
3       13.0
4       13.0
        ...
3449    13.0
3450    13.0
3451    13.0
3452    13.0
3453    13.0
Name: main_camera_mp, Length: 3454, dtype: float64>     0.0192      0.046      0.413      0.680
0.072       0.110
main_camera_mp_2.0
-0.2751      0.029     -9.569      0.000      -0.331      -0.219
main_camera_mp_16.0
0.1076       0.025      4.345      0.000       0.059       0.156
main_camera_mp_0.3
-0.4714      0.044    -10.704      0.000      -0.558      -0.385
main_camera_mp_12.0
0.0076       0.024      0.319      0.750      -0.039       0.055
main_camera_mp_14.5
-0.0145      0.078     -0.187      0.852      -0.167       0.138
main_camera_mp_48.0
0.2757       0.118      2.344      0.019       0.045       0.506
main_camera_mp_3.0
-0.1121      0.136     -0.824      0.410      -0.379       0.155
main_camera_mp_21.0
0.0706       0.069      1.028      0.304      -0.064       0.205
```

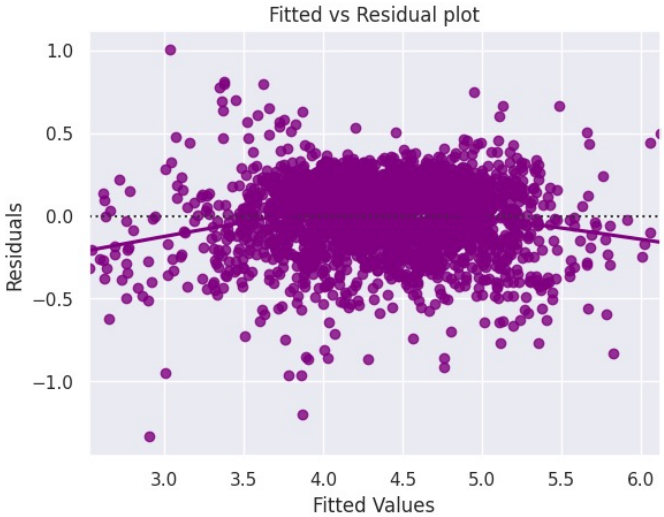
OLS Linear Regression Results Cont…

main_camera_mp_1.3					
-0.4807	0.066	-7.237	0.000	-0.611	-0.350
main_camera_mp_13.1					
0.1410	0.165	0.854	0.393	-0.183	0.465
main_camera_mp_24.0					
0.0398	0.134	0.296	0.767	-0.224	0.303
main_camera_mp_0.08					
-0.4712	0.234	-2.010	0.045	-0.931	-0.011
main_camera_mp_20.7					
0.1023	0.083	1.237	0.216	-0.060	0.264
main_camera_mp_23.0					
0.2705	0.073	3.720	0.000	0.128	0.413
main_camera_mp_1.0					
-4.793e-17	1.28e-16	-0.374	0.708	-2.99e-16	2.03e-16
main_camera_mp_18.0					
0.0682	0.238	0.287	0.774	-0.398	0.535
main_camera_mp_12.2					
-0.1628	0.175	-0.928	0.354	-0.507	0.181
main_camera_mp_12.3					
0.0540	0.108	0.500	0.617	-0.158	0.266
main_camera_mp_20.0					
0.1257	0.096	1.303	0.193	-0.064	0.315
main_camera_mp_20.2					
0.0289	0.232	0.125	0.901	-0.426	0.484
main_camera_mp_4.0					
-0.2934	0.098	-2.981	0.003	-0.486	-0.100
main_camera_mp_12.5					
0.1250	0.117	1.067	0.286	-0.105	0.355
main_camera_mp_10.0					
-0.2221	0.119	-1.868	0.062	-0.455	0.011
main_camera_mp_6.5					
-0.0930	0.164	-0.566	0.572	-0.415	0.229
main_camera_mp_6.7					
-0.2939	0.128	-2.303	0.021	-0.544	-0.044
main_camera_mp_41.0					
1.273e-17	3.36e-17	0.379	0.705	-5.32e-17	7.87e-17
main_camera_mp_20.1					
-7.319e-17	5.28e-17	-1.387	0.165	-1.77e-16	3.03e-17

OLS Linear Regression Results Cont…

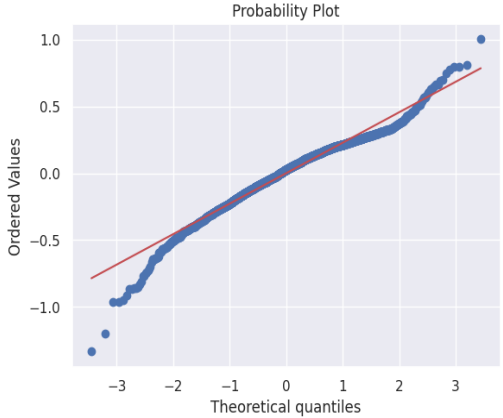
```
main_camera_mp_12.6
-9.248e-17    9.31e-17      -0.993       0.321    -2.75e-16     9.01e-17
main_camera_mp_16.3
0.3742       0.233       1.608       0.108       -0.082        0.830
main_camera_mp_22.6
-0.1479       0.232      -0.639       0.523       -0.602        0.306
main_camera_mp_19.0
-0.0679       0.102      -0.666       0.506       -0.268        0.132
main_camera_mp_21.5
0.1234       0.234       0.527       0.598       -0.335        0.582
main_camera_mp_21.2
0.2065       0.168       1.233       0.218       -0.122        0.535
main_camera_mp_8.1
-0.1930       0.123      -1.572       0.116       -0.434        0.048
main_camera_mp_1.2
-0.1055       0.235      -0.449       0.654       -0.567        0.356
main_camera_mp_22.5
0.1759       0.232       0.759       0.448       -0.279        0.630
==============================================================================
Omnibus:                      205.856   Durbin-Watson:                   1.906
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              389.471
Skew:                          -0.578   Prob(JB):                     2.68e-85
Kurtosis:                       4.591   Cond. No.                     4.21e+19
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.63e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

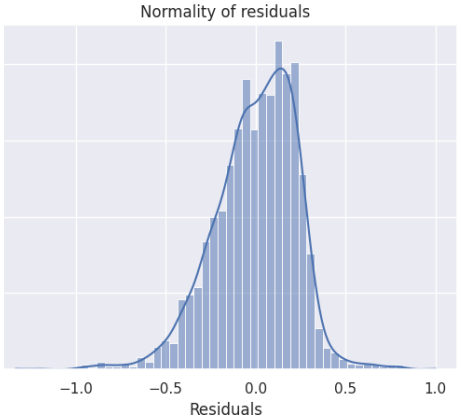
Model Performance Summary



There is no pattern in the plot.

The assumptions of linearity and independence are satisfied.





p-value=4.226222461485406e-21

- The histogram of the residuals does resemble a bell curve.
- The residuals follow a straight line except at the high and low ends.
- p-value is not greater than .05 therefore residuals are not normally distributed per the Shapiro-Wilk test.
- The residuals are not normal (from a literal standpoint)
- However, we can accept this distribution to being “close” to normal.

The assumption is satisfied.

Tests Used

- **Test for Homoscedasticity**

  (goldfeldquandt test – p-value greater than 0.05, we can say that the residuals are homoscedastic. Otherwise, they are heteroscedastic.)

  * p-value: 0.15606511727489084
    Residuals are homoscedastic

- Used the goldfeldquandt test to test for homoscedasticity
- Tested for normality by checking the distribution of residuals by using the Q-Q plot of residuals, and by using the Shapiro-Wilk test.

Final Model OLS Regression Results

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.843
Model:	OLS	Adj. R-squared:	0.842
Method:	Least Squares	F-statistic:	585.6
Date:	Sun, 12 Nov 2023	Prob (F-statistic):	0.00
Time:	21:27:26	Log-Likelihood:	111.59
No. Observations:	2417	AIC:	-177.2
Df Residuals:	2394	BIC:	-44.01
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.9046	0.053	35.745	0.000	1.800	2.009
selfie_camera_mp	0.0148	0.001	13.815	0.000	0.013	0.017
ram	0.0170	0.004	3.790	0.000	0.008	0.026
weight	0.0017	6.02e-05	28.680	0.000	0.002	0.002
normalized_new_price	0.4141	0.011	37.401	0.000	0.392	0.436
years_since_release	-0.0242	0.003	-7.203	0.000	-0.031	-0.018
brand_name_Nokia	0.0710	0.032	2.220	0.027	0.008	0.134
brand_name_Samsung	-0.0350	0.016	-2.136	0.033	-0.067	-0.003
brand_name_Xiaomi	0.0696	0.026	2.725	0.006	0.020	0.120
os_Others	-0.0728	0.031	-2.367	0.018	-0.133	-0.012
4g_yes	0.0394	0.015	2.623	0.009	0.010	0.069
main_camera_mp_8.0	-0.1167	0.014	-8.532	0.000	-0.144	-0.090
main_camera_mp_5.0	-0.1922	0.017	-11.033	0.000	-0.226	-0.158
main_camera_mp_3.15	-0.2671	0.030	-8.905	0.000	-0.326	-0.208
main_camera_mp_2.0	-0.2939	0.027	-10.955	0.000	-0.347	-0.241
main_camera_mp_16.0	0.0995	0.024	4.111	0.000	0.052	0.147
main_camera_mp_0.3	-0.5165	0.042	-12.350	0.000	-0.599	-0.435
main_camera_mp_48.0	0.2916	0.117	2.500	0.012	0.063	0.520
main_camera_mp_1.3	-0.5184	0.063	-8.171	0.000	-0.643	-0.394
main_camera_mp_0.08	-0.4862	0.236	-2.064	0.039	-0.948	-0.024
main_camera_mp_23.0	0.2220	0.068	3.288	0.001	0.090	0.354
main_camera_mp_4.0	-0.3108	0.096	-3.254	0.001	-0.498	-0.123
main_camera_mp_6.7	-0.2925	0.121	-2.408	0.016	-0.531	-0.054

Omnibus:	213.138	Durbin-Watson:	1.913
Prob(Omnibus):	0.000	Jarque-Bera (JB):	413.256
Skew:	-0.587	Prob(JB):	1.83e-90
Kurtosis:	4.650	Cond. No.	1.02e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.02e+04. This might indicate that there are strong multicollinearity or other numerical problems.