ReCell Project

PGP - Data Science & Business Analytics

October 10, 2024

Leslieane Beltran

Contents / Agenda

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

Executive Summary

- Using a linear regression model, univariate analysis, and bivariate analysis we gained insight on the factors that influence the price of used devices.
- The linear regression model concluded that we can explain 84% of the variance in the resale price of a device.
- Features such as camera resolution, screen size, RAM, weight, release year, battery, and brand are significant factors that can be used in the future to predict the used price of a device.
- The linear regression model and the exploratory data analysis led to meaningful results which will aid in predicting used price as well as important device features ReCell should focus on.

Business Problem Overview and Solution Approach

- ReCell is a startup in the growing used and refurbished device market. ReCell faces the challenge of accurately pricing its products. While setting prices too high may lead to losing customers, pricing too low cuts into profits.
- There are various device features and conditions influencing value, therefore, ReCell needs a data-driven solution to predict market prices and stay competitive.
- The solution involves analyzing the key attributes of used devices, such as brand, memory, screen size, and usage, to understand how they impact pricing.
- Next a predictive pricing model will be created using linear regression analysis.
- This model will be used moving forward to analyze new data and keep ReCell’s business strategy up to date.

Data Overview

The data contains the different attributes of used/refurbished phones and tablets. The data was collected in the year 2021. Data dictionary:

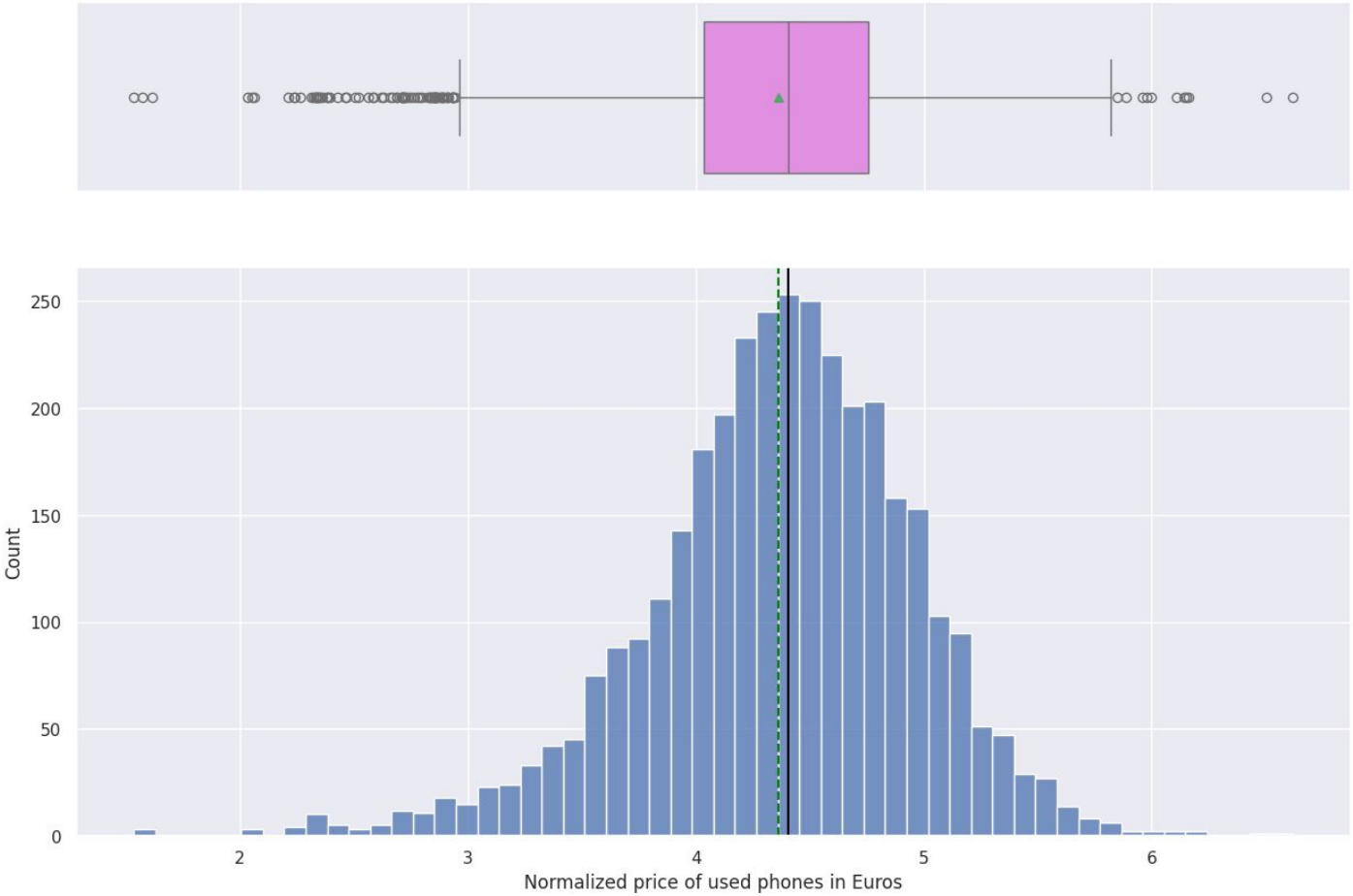
- brand_name: Name of manufacturing brand
- os: OS on which the device runs
- screen_size: Size of the screen in cm
- 4g: Whether 4G is available or not
- 5g: Whether 5G is available or not
- main_camera_mp: Resolution of the rear camera in megapixels
- selfie_camera_mp: Resolution of the front camera 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

Data Dictionary Cont.

- 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

EDA Results - Univariate Analysis

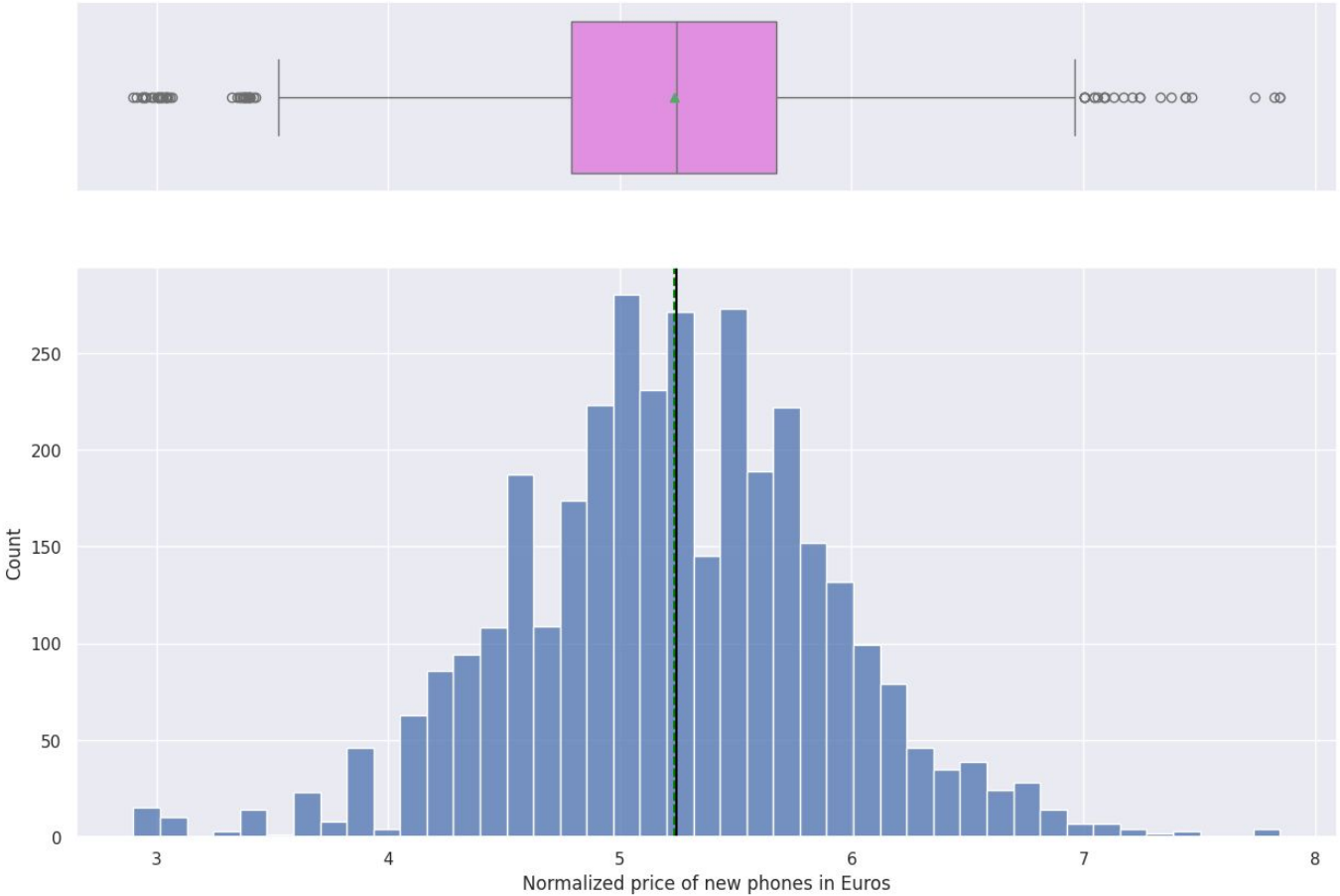
NORMALIZED USED PRICE



Observations:

- The data is normally distributed
- Based on the boxplot, there does not seem to be any extreme outliers

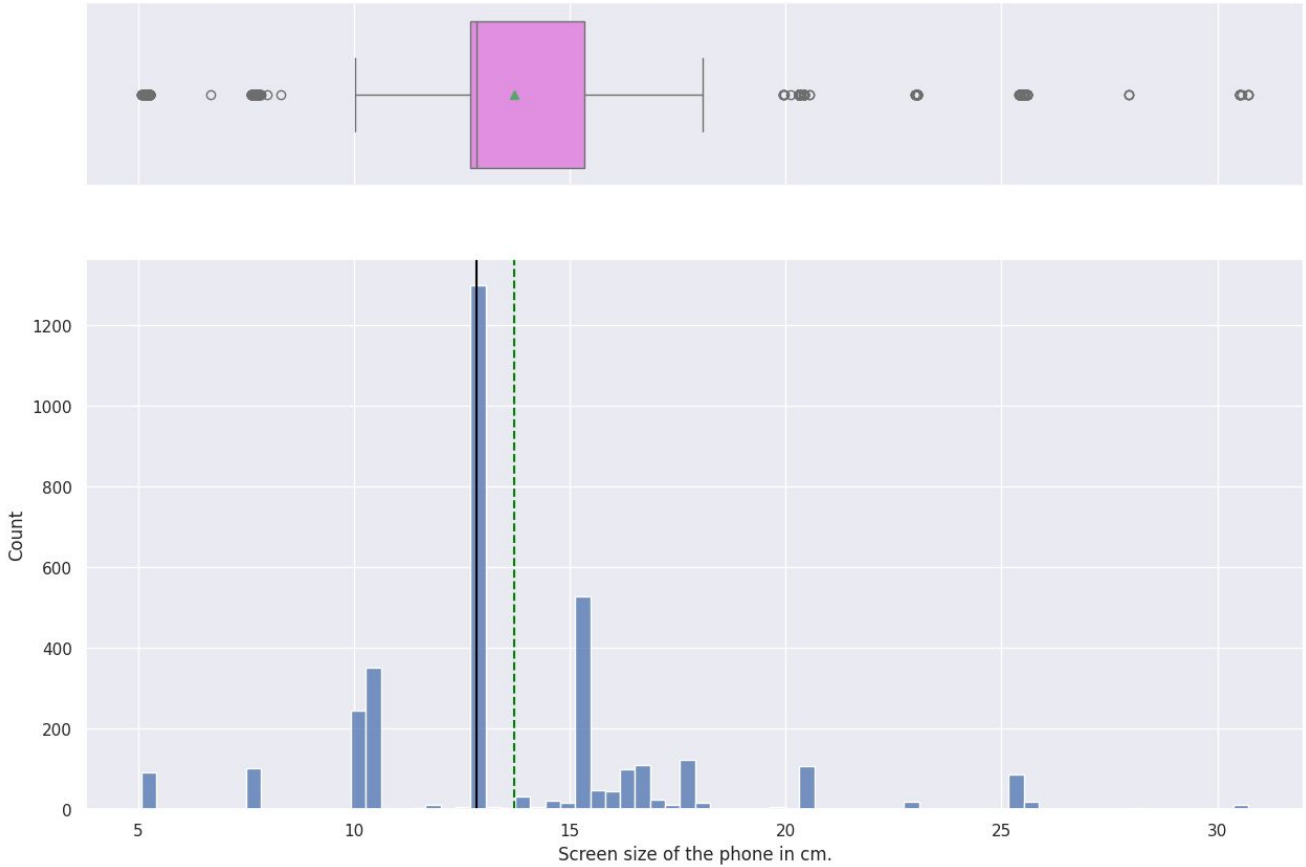
Normalized New Price



Observation

- The data is normally distributed

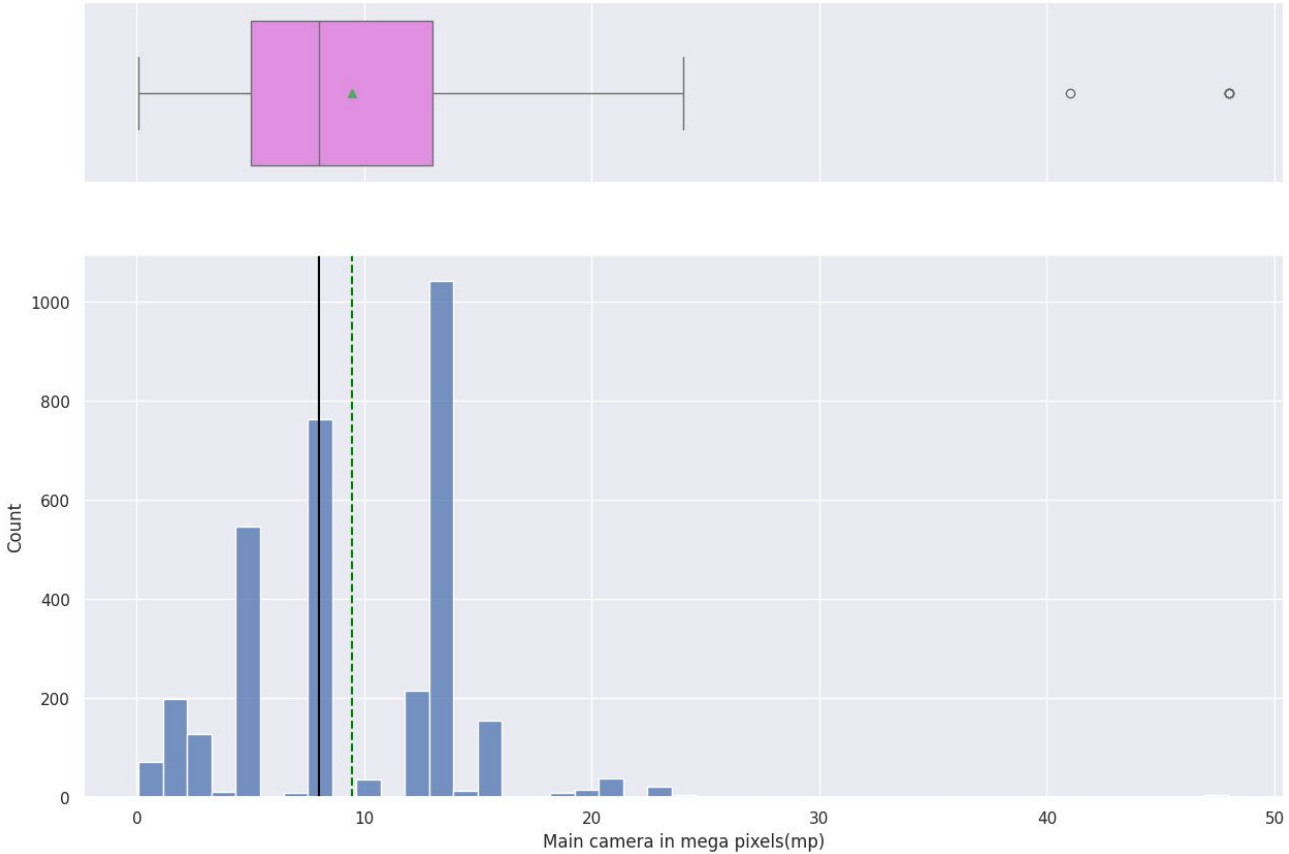
Screen Size



Observations:

- The data is close to normally distributed.
- Most devices seem to have a screen size of about 12 cm (in between 10 cm and 15 cm.

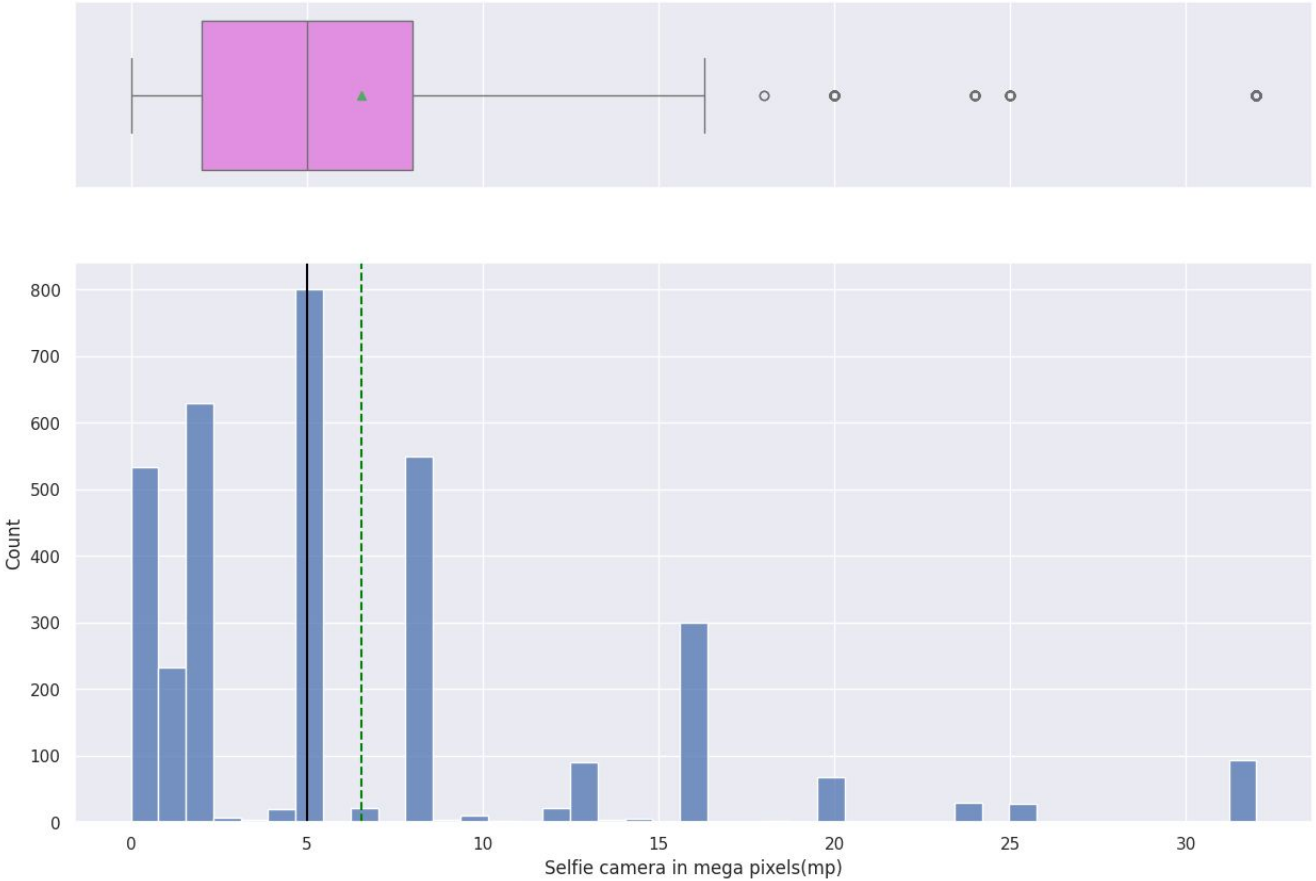
Main Camera



Observation:

- Based on the graphs, most devices have main cameras between 10 and 20 megapixels (mp).

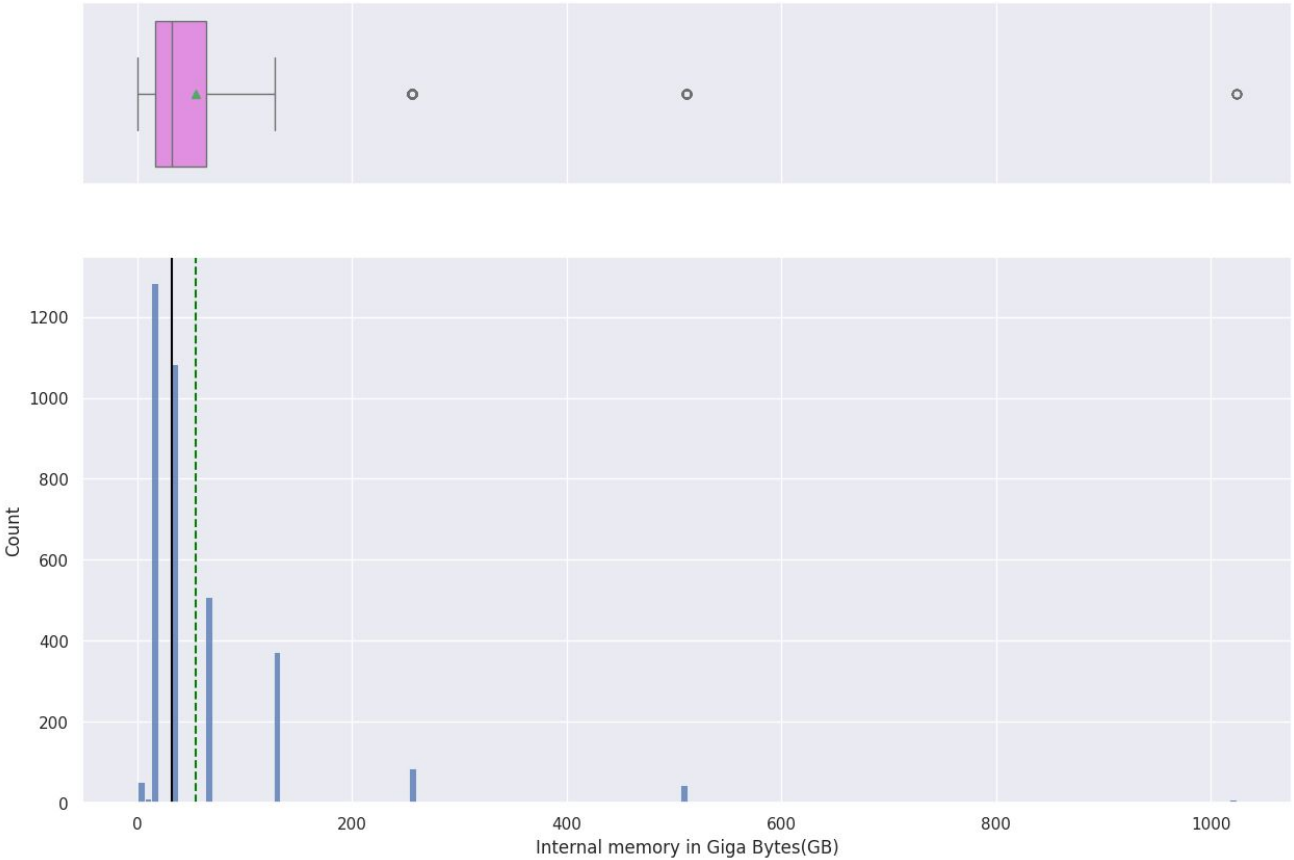
Selfie Cameras



Observation:

- Based on the graphs, most devices have a self camera that is 5 megapixels.

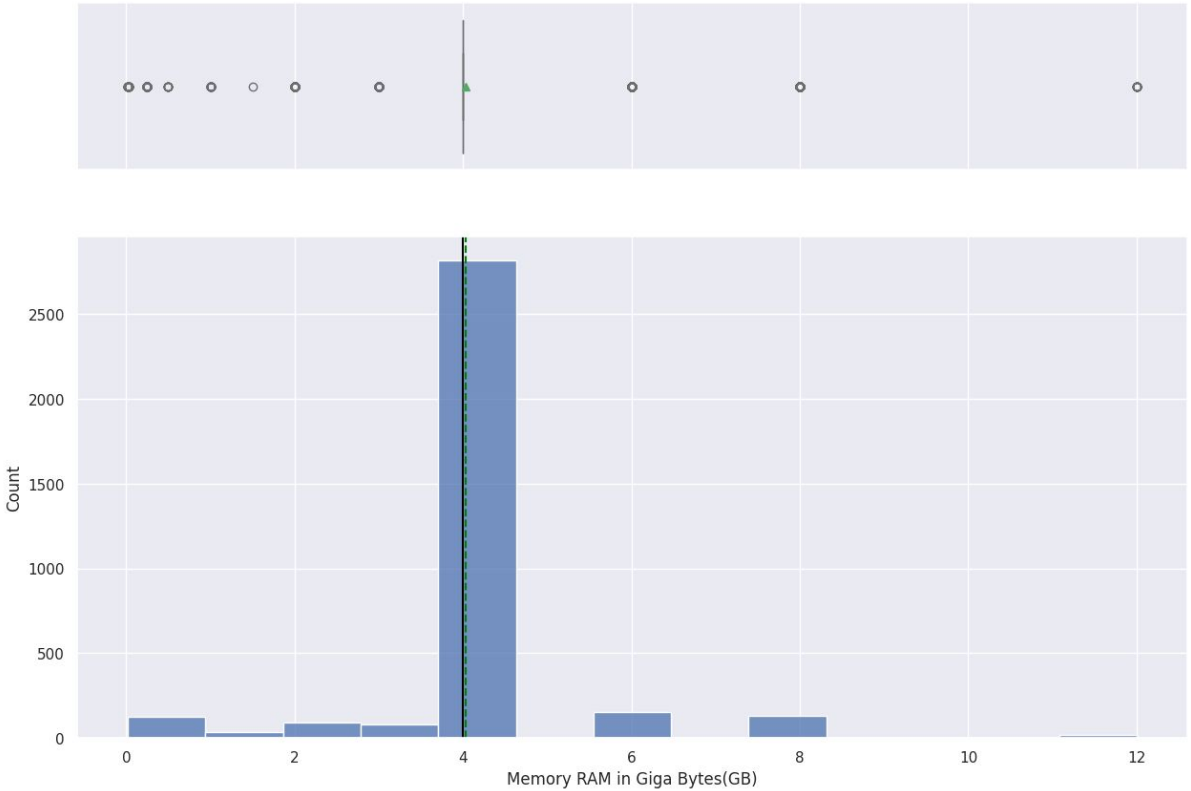
Internal Memory



Observation:

- The devices seem to have an internal memory of about 25 GB to 150 GB.

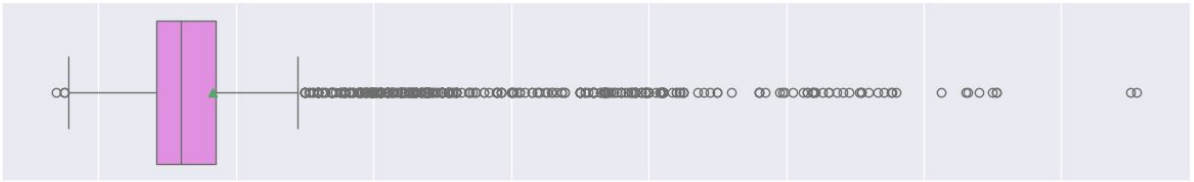
Ram

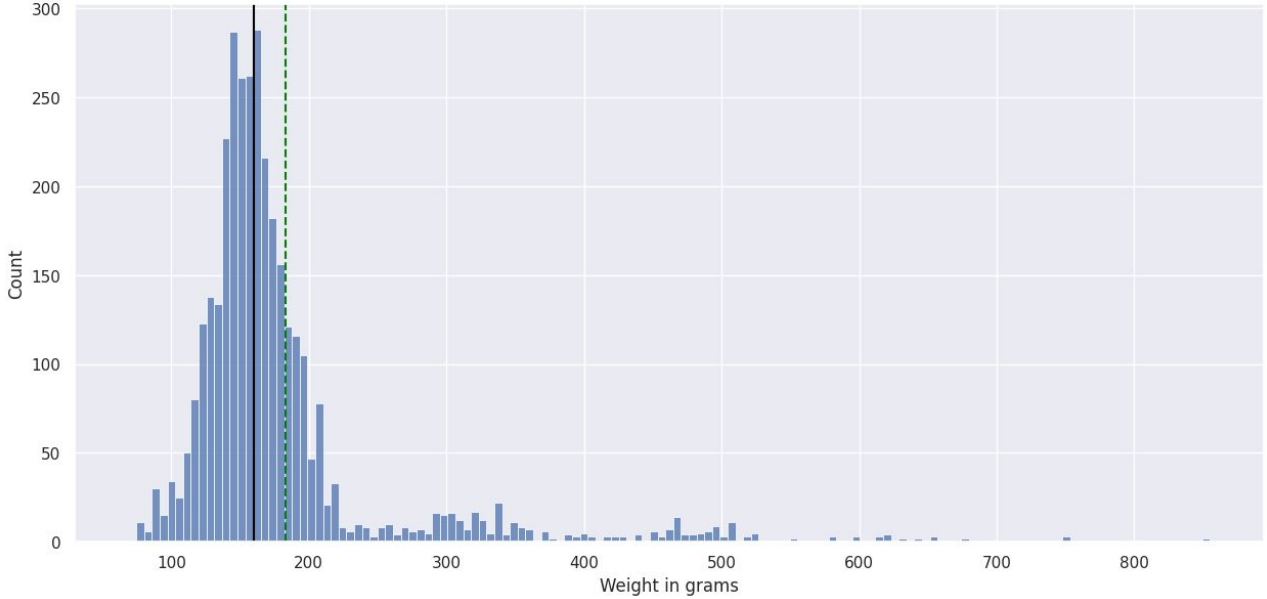


Observation:

- It can be determined that most devices have 4GB of ram memory.

Weight

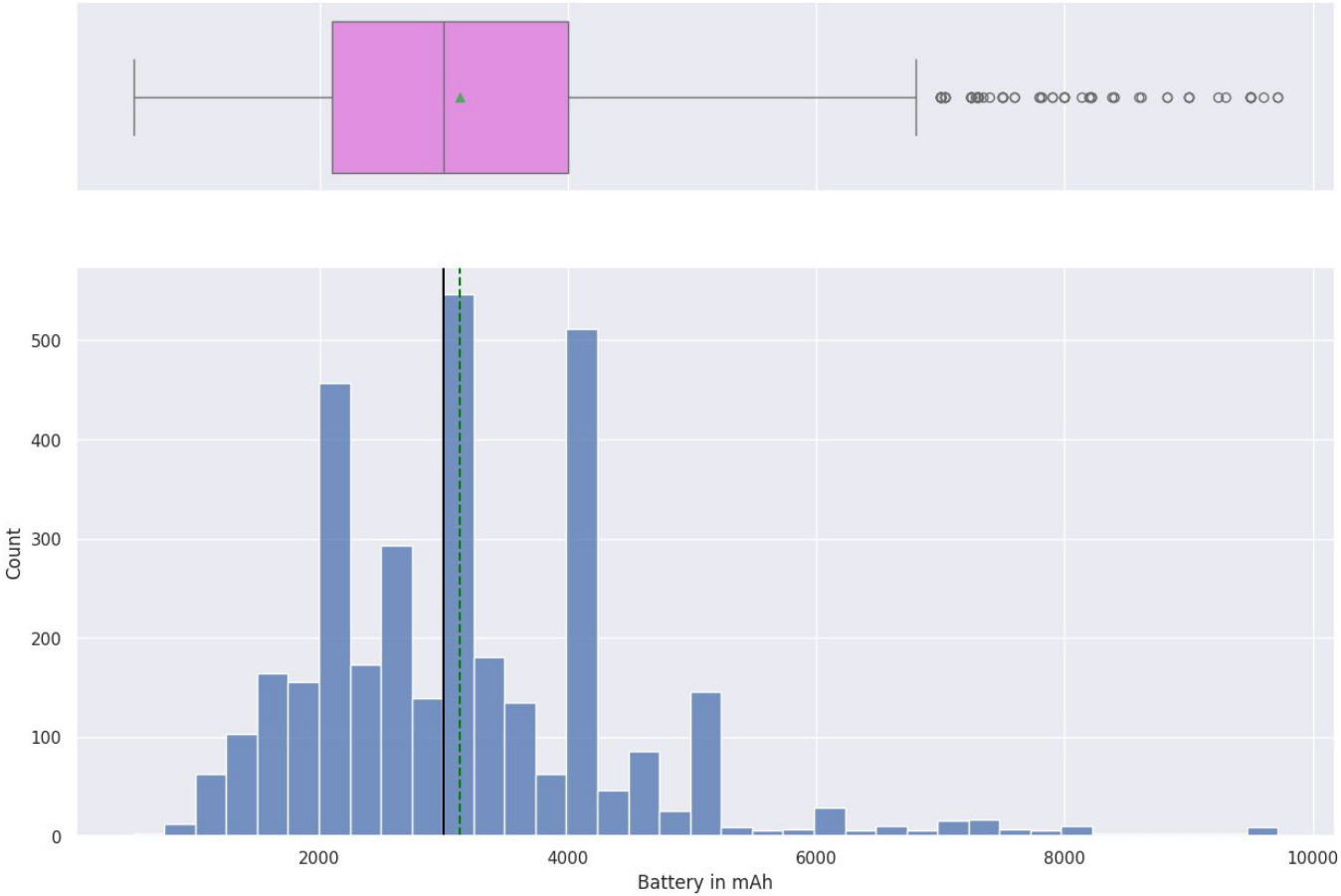




Observation:

- Most of the devices weigh between 100 grams and 200 grams.
- There are still some devices that way more than 200 grams, these are most likely bigger phones or tablets.

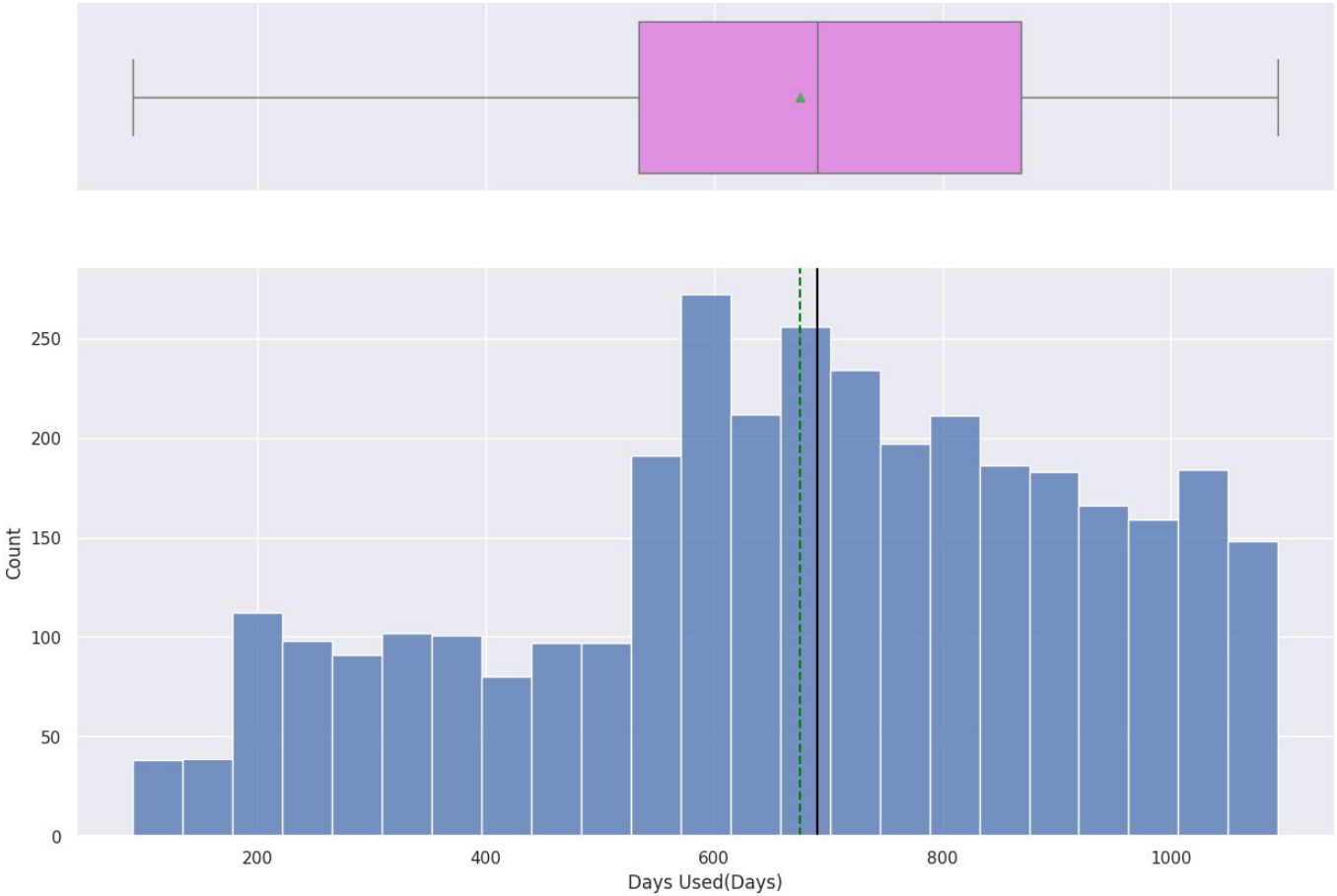
Battery



Observation:

- Most devices have a battery life of about 2000 mAh, 3000 mAh, & 4000 mAh.

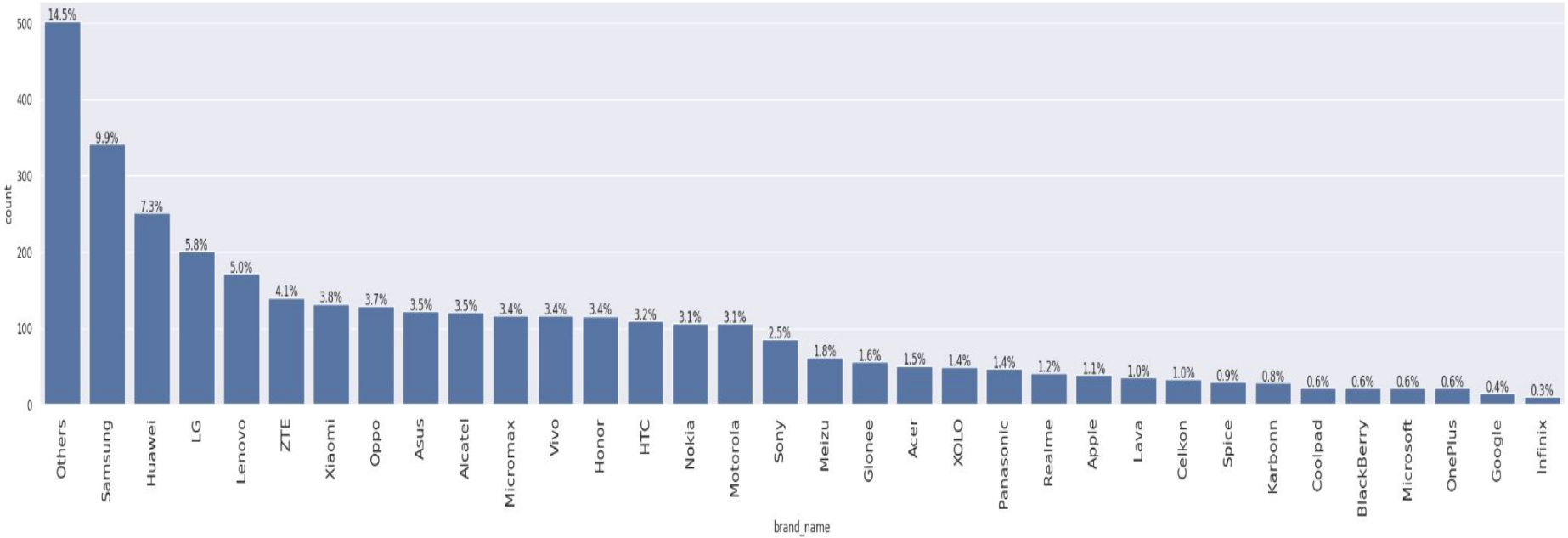
Days Used



Observation

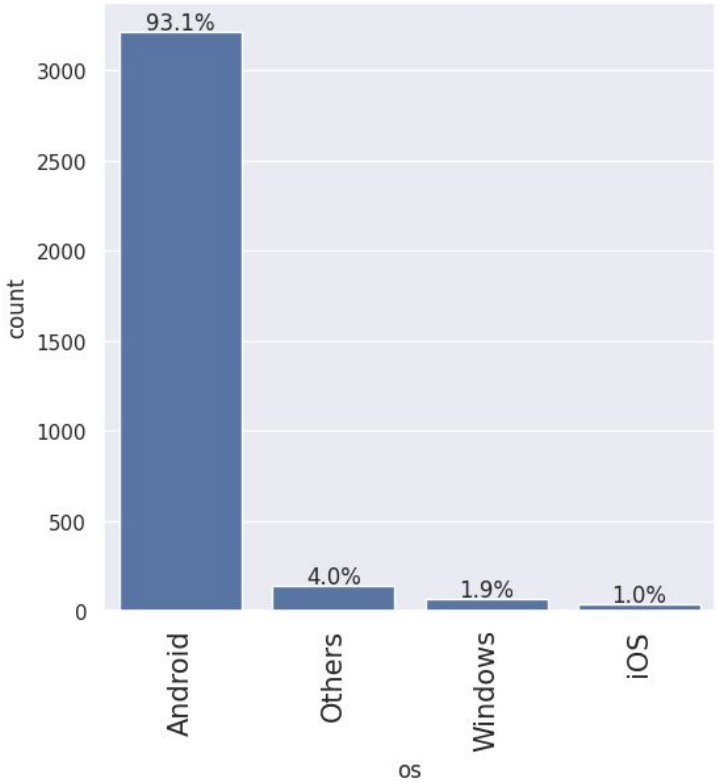
- Majority of the devices have been used for 600 days.

Brand Name



Observations:
14.5% of the devices are brands other than the named brands in the data. Samsung is the second most common, making up 9.9% of the devices.

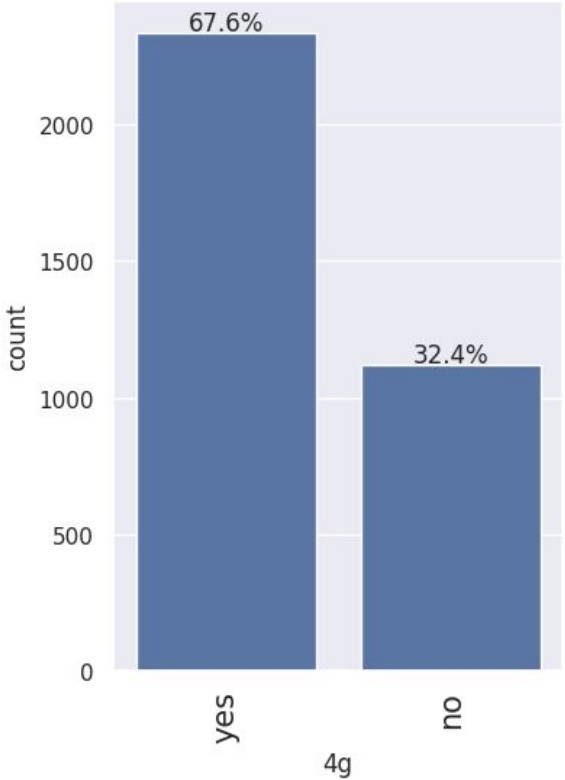
Operating System

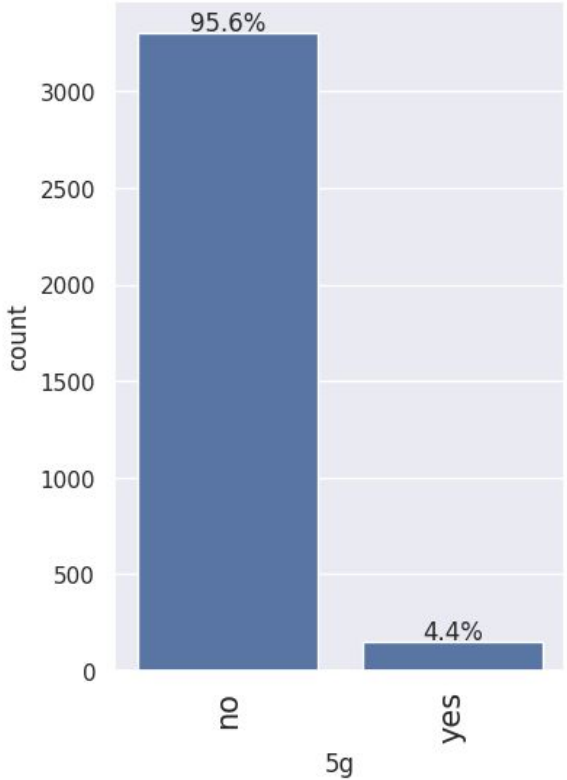


Observation:

- Most of the devices have an Android operating system.
- Only 1% of the devices are IOS; Apples.

4G & 5G





Observation:

- 67% of the devices have 4G and 4.4% have 5G.

Release Year

Observations:

- The devices released in 2014 were the most refurbished.

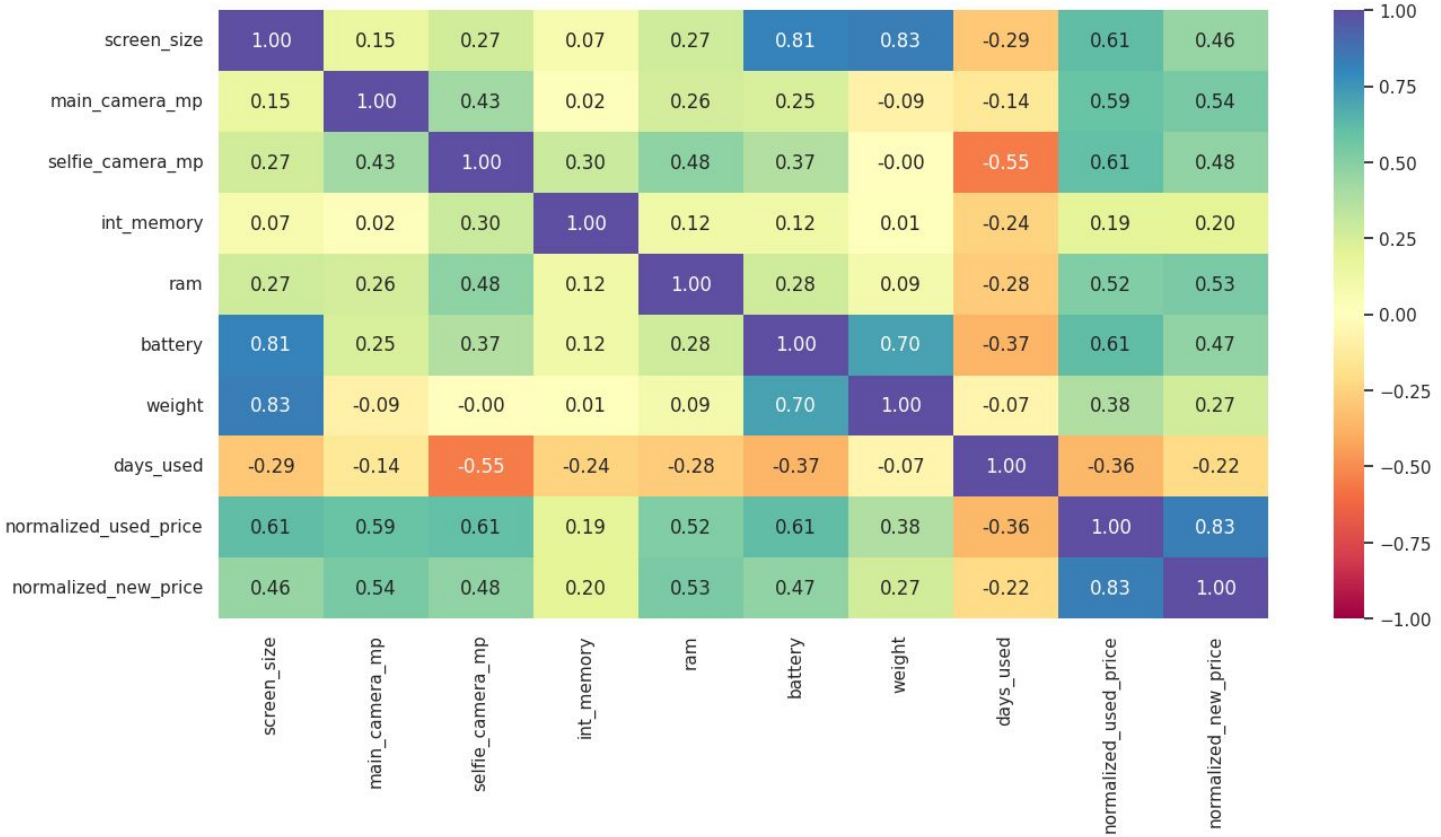
EDA Results - Bivariate Analysis

Heat Map

Observation:

Based on the heat map, it can be said that used price and new price are highlighly correlated.

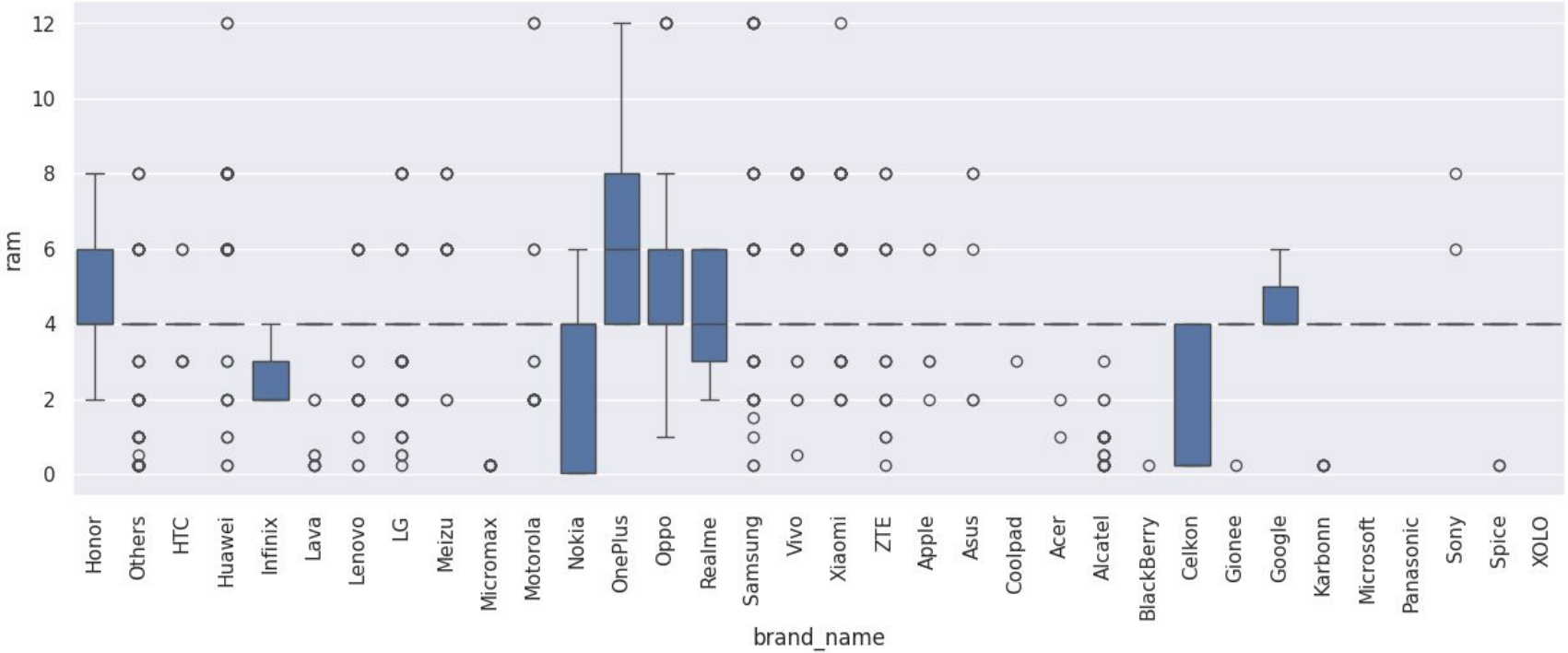
Screen size, battery, and weight are correlated. This gives an idea of the features of the devices.



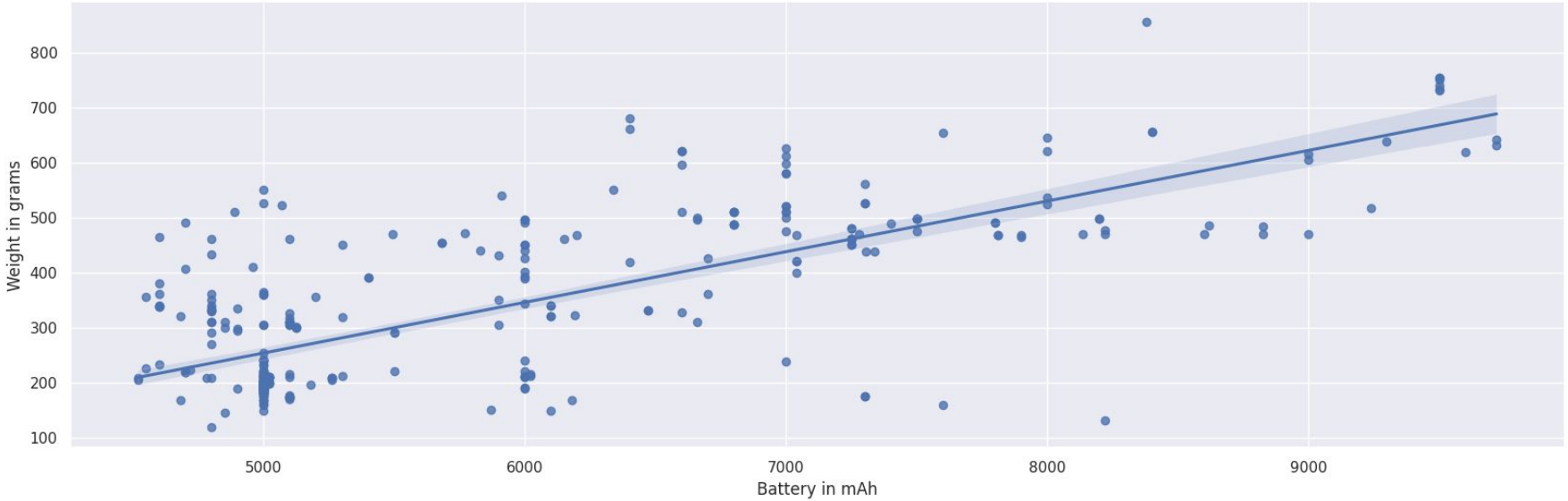
Ram & Brand Name

Observations:

- A wide variety of brands show minimum and maximum values of 4Gb in the box plot, with outliers at 2Gb, 3Gb, 6Gb, 8Gb, and 12Gb.



Devices with Larger Batteries & Weight



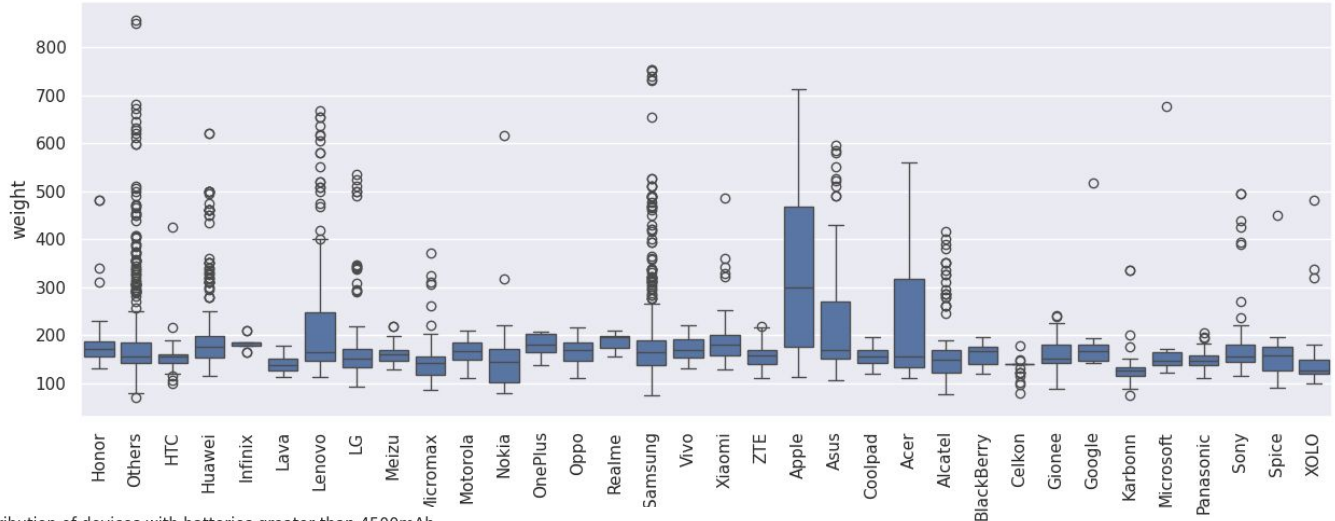
Observation:

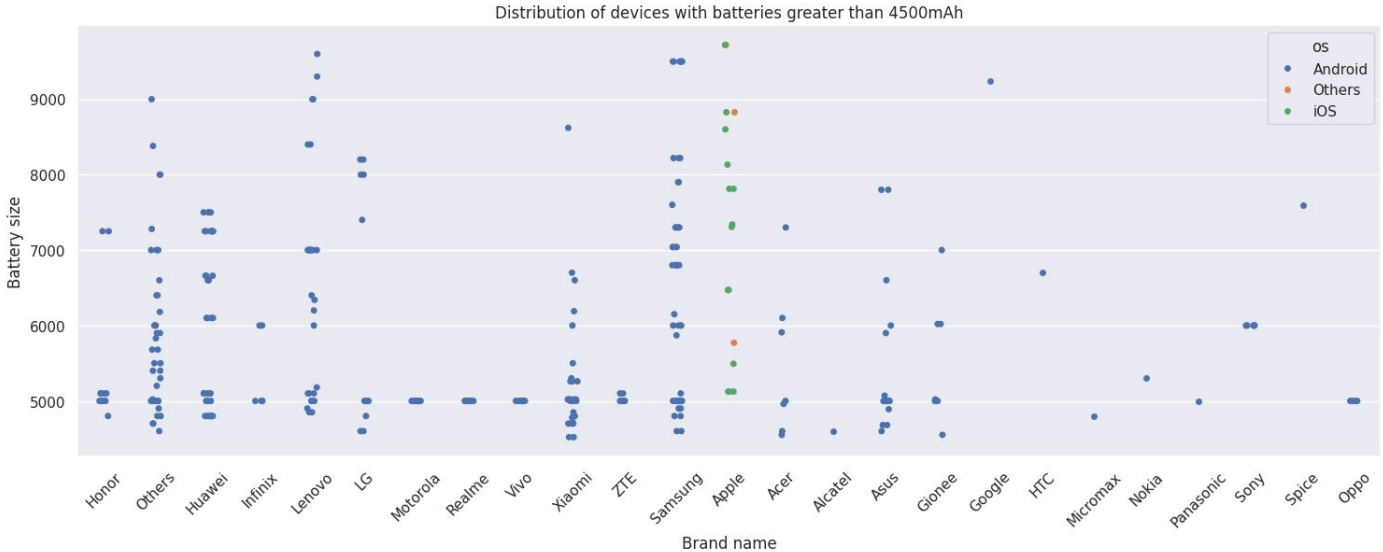
Linear correlation factor between large batteries(4500mAh)and phone's weight is 0.76

The correlation between weight and battery size is higher when batteries are larger than 4500 mAh.

Brand Name, Weight & Battery Size

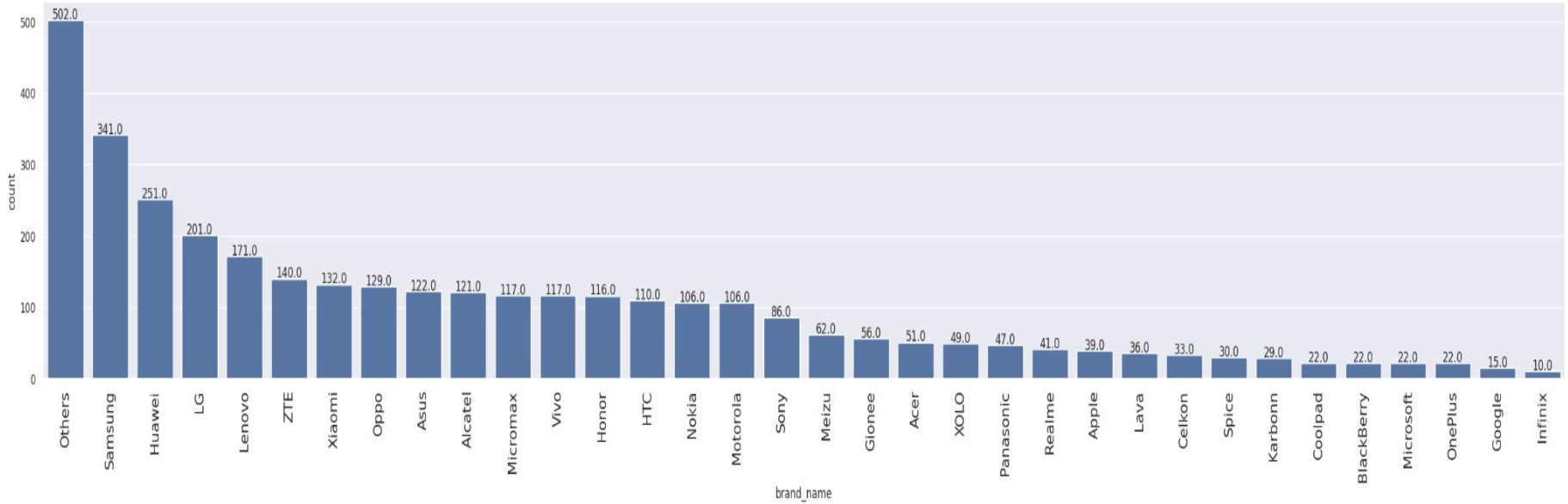
- It seems that apple devices weigh more on average, most likely due to a bigger battery capacity.





Devices with Larger Screens

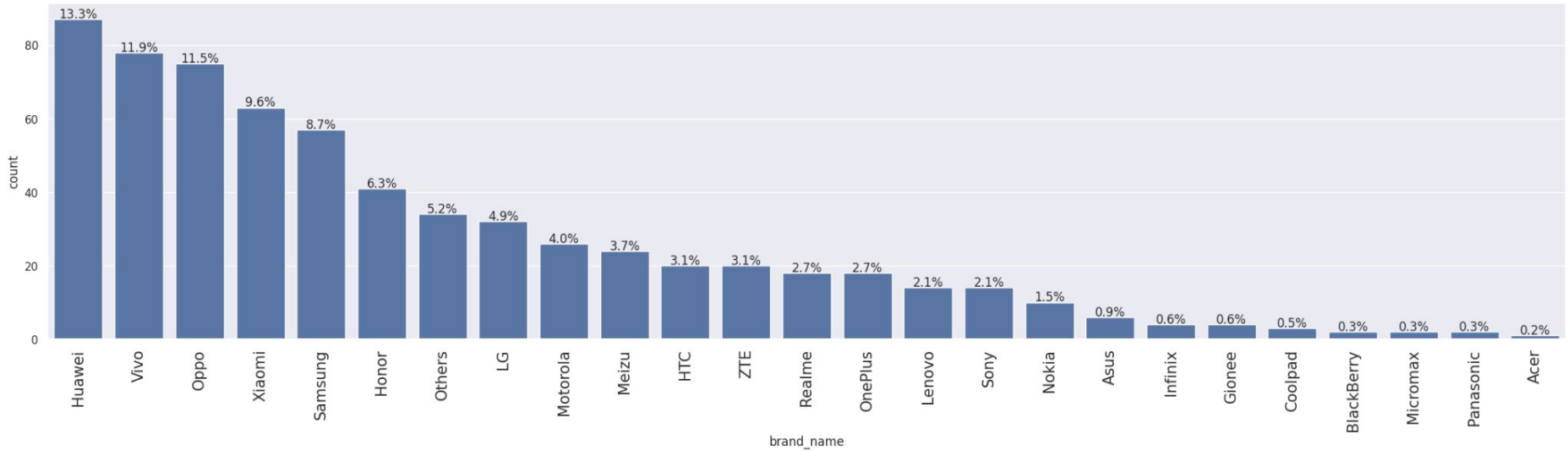
Other brands and Samsung have the most devices with a larger screen size. Apple devices, which have been seen to be a competitor in the other categories show to actually have smaller screens.



Selfie Cameras

Huawei offers the most devices with better front camera resolution. Vivo and Oppo are close competitors.

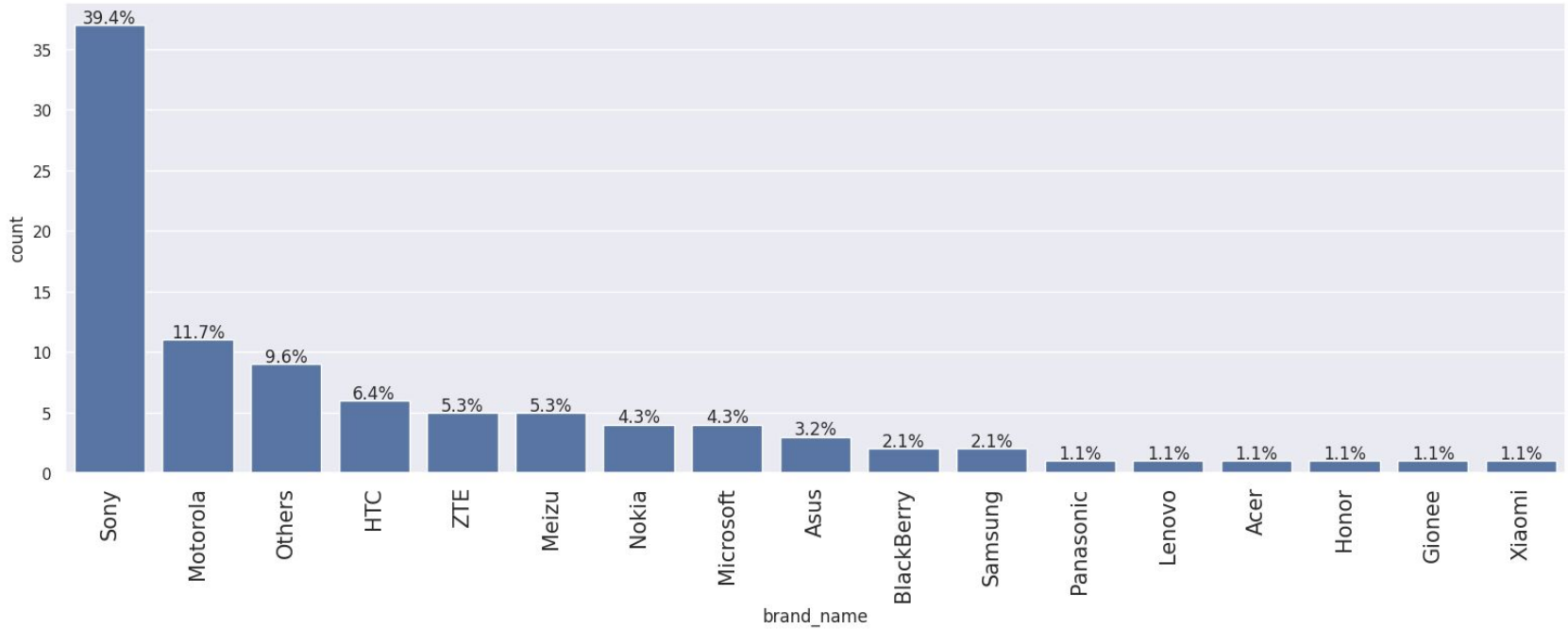
Acer has very few devices with a good selfie camera.



Rear Cameras

Sony has the most devices with better rear cameras. Around 39% have a resolution higher than 16MP. Surprisingly, Samsung and Xiaomi do not have many devices with this characteristic.

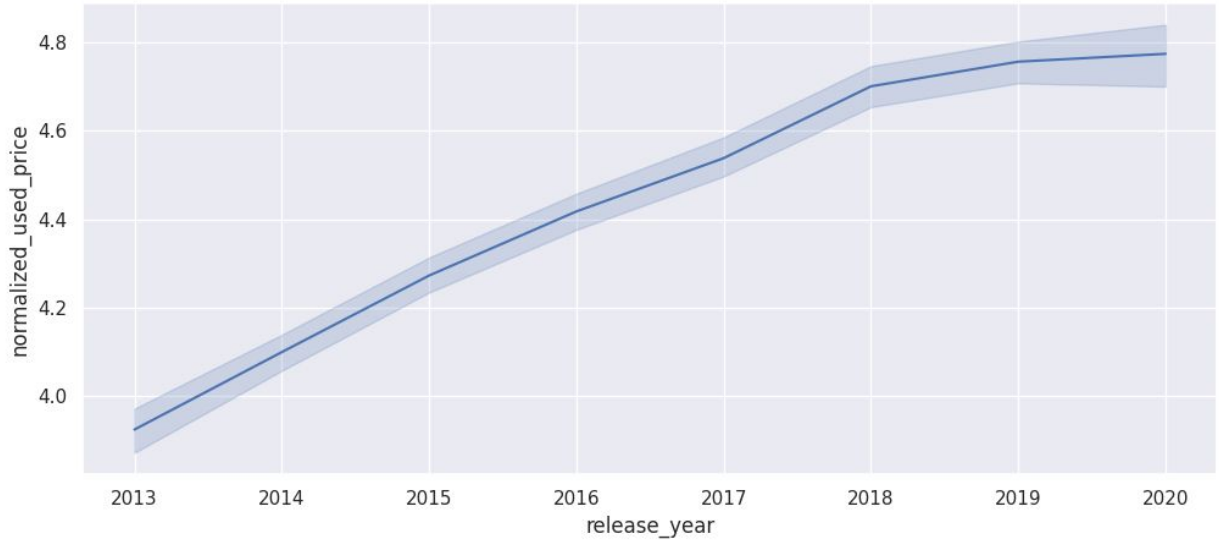
Acer is again among the devices with lower performing cameras.



Used Devices Across the Years

As release year increases, the used price increases.

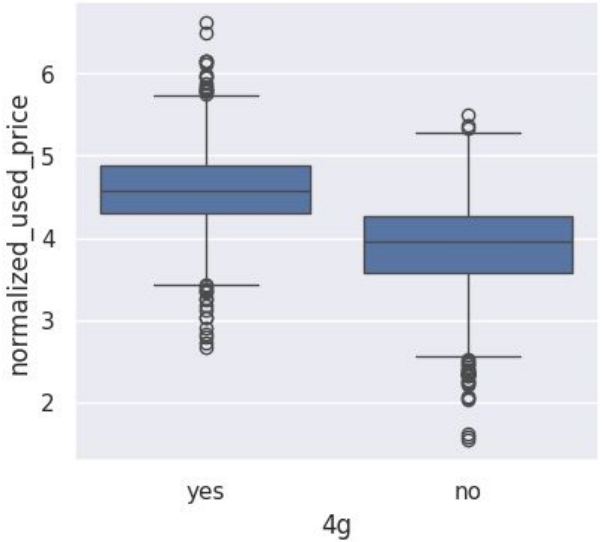
The newer the phone the more expensive the used phone will be.

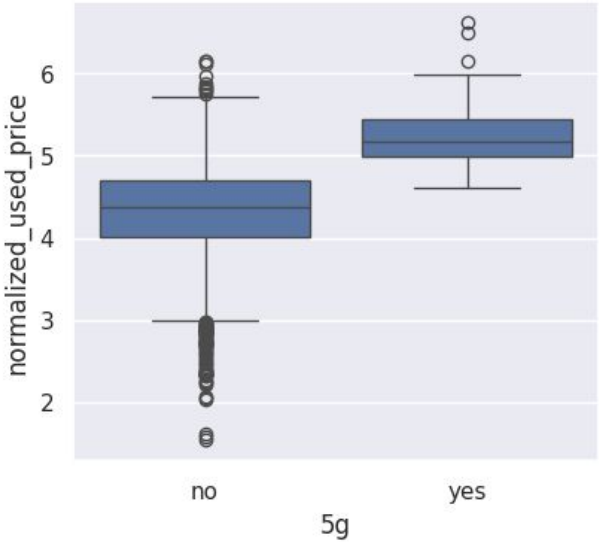


Pricing for Used Phones and Tablets with 4G and 5G.

Devices with 5g have a higher used price. Most likely because these devices are more desirable in the market.

However, 4g devices are not too far b behind.





Data Preprocessing

Missing Value Imputation

Main camera, weight, battery, memory, ram & selfie camera all have missing values.

The second table shows that the missing values have been treated.

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

dtype: int64

	0
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

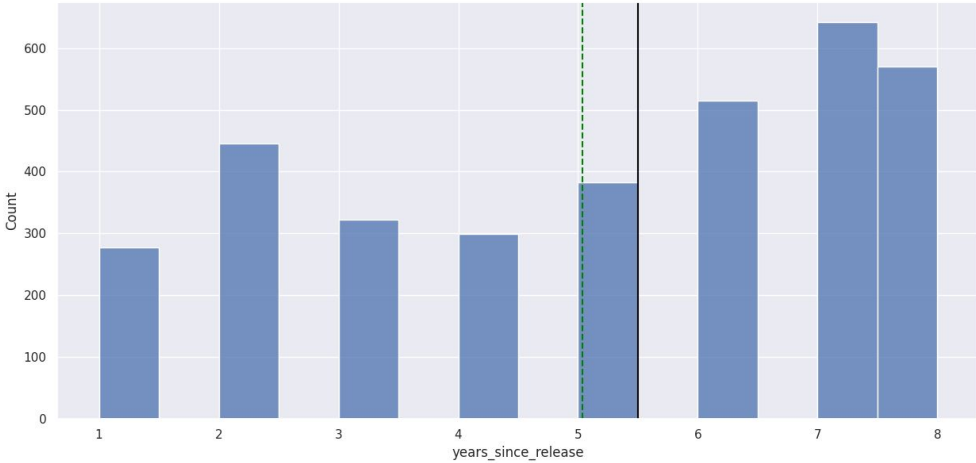
dtype: int64

Feature Engineering

Creating a new column called “years_since_release” to ensure accurate analysis of the data

	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

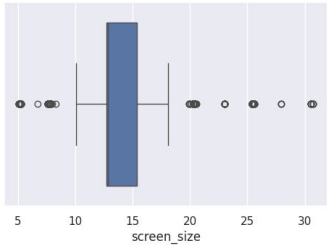
dtype: float64

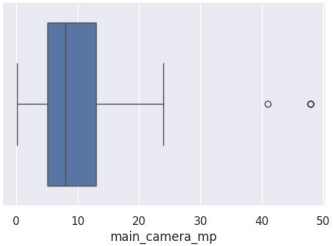


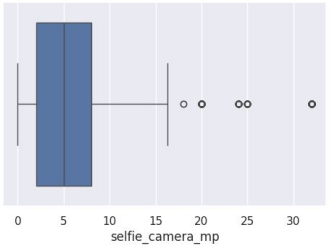
Outlier Check

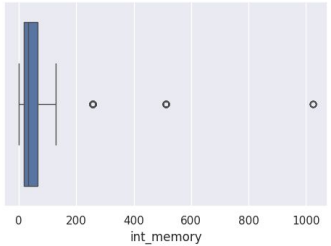
After performing an outlier detection test it is apparent that there are a lot of outliers.

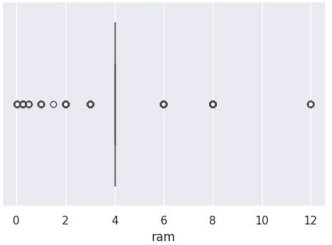
However, these outliers should not be treated as removing these may lead to biased testing that does not represent the sample accurately.

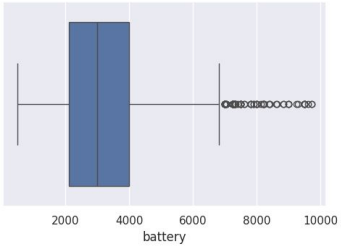


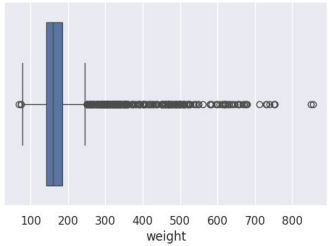


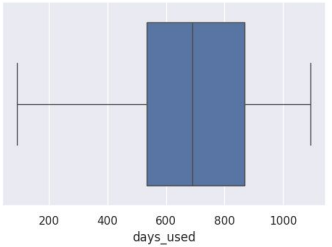


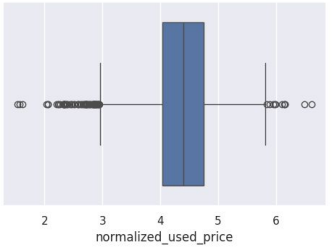


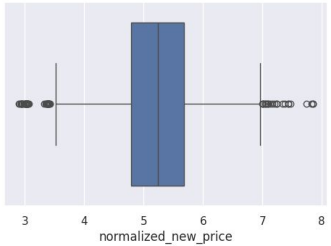


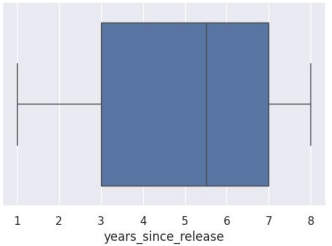












Data Preparation for Modeling

- In this section the data was prepared for the ML model.
- Categorical values were encoded to work in the model
- Data was also split into train and test sets with a 70:30 ratio
- Finally, intercept for train and test variables was added

Dropping Used Price

```
  brand_name       os  screen_size   4g   5g  main_camera_mp  \
0      Honor  Android        14.50  yes   no            13.0
1      Honor  Android        17.30  yes  yes            13.0
2      Honor  Android        16.69  yes  yes            13.0
3      Honor  Android        25.50  yes  yes            13.0
4      Honor  Android        15.32  yes   no            13.0

   selfie_camera_mp  int_memory  ram  battery  weight  days_used  \
0               5.0        64.0  3.0   3020.0   146.0        127
1              16.0       128.0  8.0   4300.0   213.0        325
2               8.0       128.0  8.0   4200.0   213.0        162
3               8.0        64.0  6.0   7250.0   480.0        345
4               8.0        64.0  3.0   5000.0   185.0        293

   normalized_new_price  years_since_release
0              4.715100                    1
1              5.519018                    1
2              5.884631                    1
3              5.630961                    1
4              4.947837                    1

0    4.307572
1    5.162097
2    5.111084
3    5.135387
4    4.389995
Name: normalized_used_price, dtype: float64
```

Dummy Variables (does not show full table)

	const	brand_name	os	screen_size	4g	5g	main_camera_mp	selfie_
0	1.0	Honor	Android	14.50	yes	no	13.0	
1	1.0	Honor	Android	17.30	yes	yes	13.0	
2	1.0	Honor	Android	16.69	yes	yes	13.0	
3	1.0	Honor	Android	25.50	yes	yes	13.0	
4	1.0	Honor	Android	15.32	yes	no	13.0	

Model Performance Summary

Linear regression

Model explains 84% of the variance in the training set.

10 predictors

Based on these results the model will be good for prediction,

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.840
Method:	Least Squares	F-statistic:	1267.
Date:	Fri, 11 Oct 2024	Prob (F-statistic):	0.00
Time:	02:59:03	Log-Likelihood:	89.188
No. Observations:	2417	AIC:	-156.4
Df Residuals:	2406	BIC:	-92.68
Df Model:	10		
Covariance Type:	nonrobust		

Train and Test Sets

Mean absolute error in training set is about 0.184 and RMSE is 0.233. MAE and RMSE on test test are about 0.184 and 0.238.

The train and test sets are comparable, which shows that the model is not overfitting

Training Performance

cell output actions

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.233205	0.183583	0.840372	0.839375	4.409285

Test Performance

cell output actions

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.237467	0.18389	0.843654	0.841357	4.491875

Test for Multicollinearity

Dropping p-values

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.840
Method:	Least Squares	F-statistic:	1582.
Date:	Fri, 11 Oct 2024	Prob (F-statistic):	0.00
Time:	03:12:31	Log-Likelihood:	87.706
No. Observations:	2417	AIC:	-157.4
Df Residuals:	2408	BIC:	-105.3
Df Model:	8		
Covariance Type:	nonrobust		

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.23732	0.183571	0.843847	0.842479	4.480411

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.233348	0.18355	0.840176	0.839579	4.407851

Testing & Removing Multicollinearity

VIF after dropping Screen Size

	feature	VIF
0	const	98.993556
1	screen_size	6.097696
2	main_camera_mp	1.838534
3	selfie_camera_mp	2.538478
4	int_memory	1.204239
5	ram	1.569363
6	battery	3.758687
7	weight	5.518991
8	days_used	2.486067
9	normalized_new_price	2.321351
10	years_since_release	3.798380

	feature	VIF
0	const	98.993556
1	brand_name	NaN
2	os	NaN
3	screen_size	6.097696
4	4g	NaN
5	5g	NaN
6	main_camera_mp	1.838534
7	selfie_camera_mp	2.538478
8	int_memory	1.204239
9	ram	1.569363
10	battery	3.758687
11	weight	5.518991
12	days_used	2.486067
13	normalized_new_price	2.321351
14	years_since_release	3.798380

	col	Adj. R-squared after_dropping col	RMSE after dropping col
0	const	0.996236	0.270348
1	weight	0.836964	0.235730
2	screen_size	0.834146	0.237759

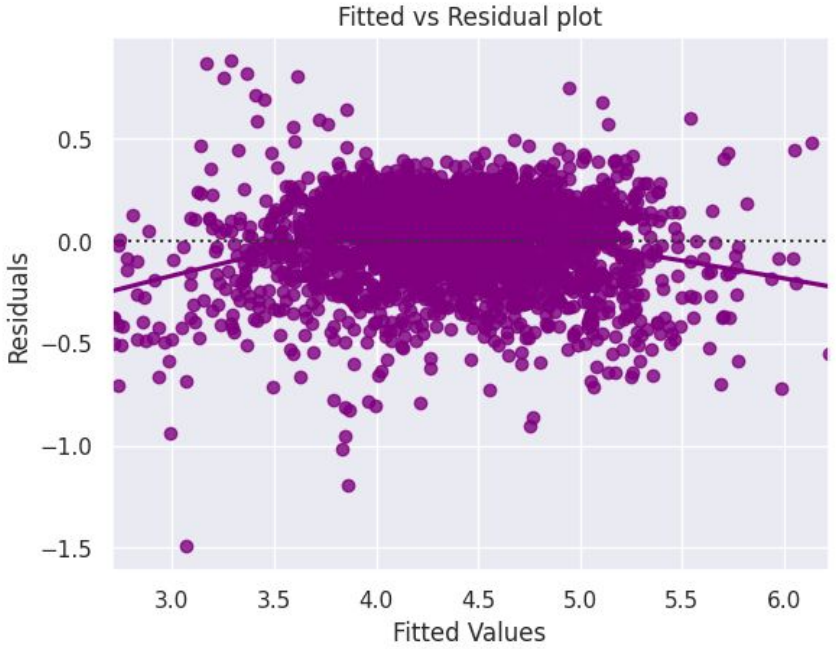
Multicollinearity Continued

- Most predictors showed to have high variance inflation, indicating that the model did have multicollinearity issue.
- To solve this, high p values were dropped
- This resulted in a better model without multicollinearity

Test for linearity and independence

As seen in the residual plot there is no pattern. Therefore, there is independence among the variables.

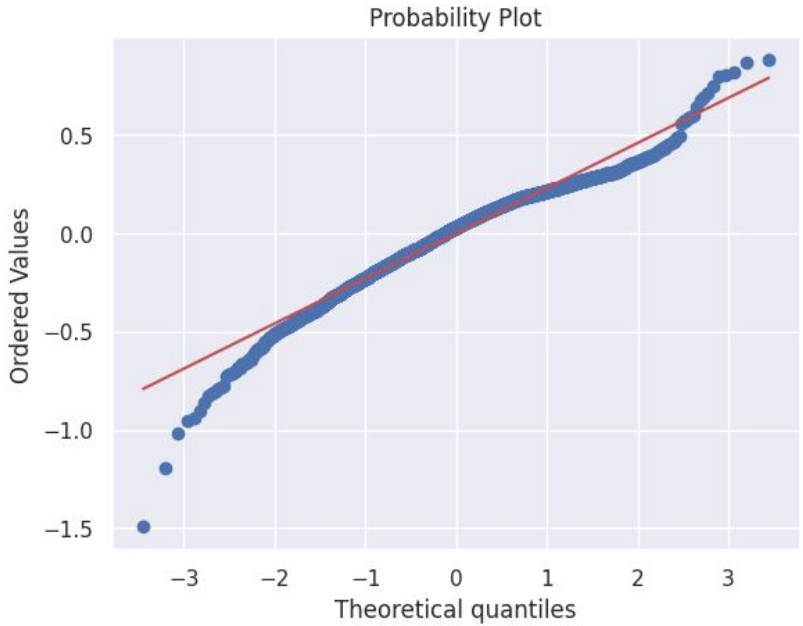
Based on this, the model is linear and and residuals are independent.

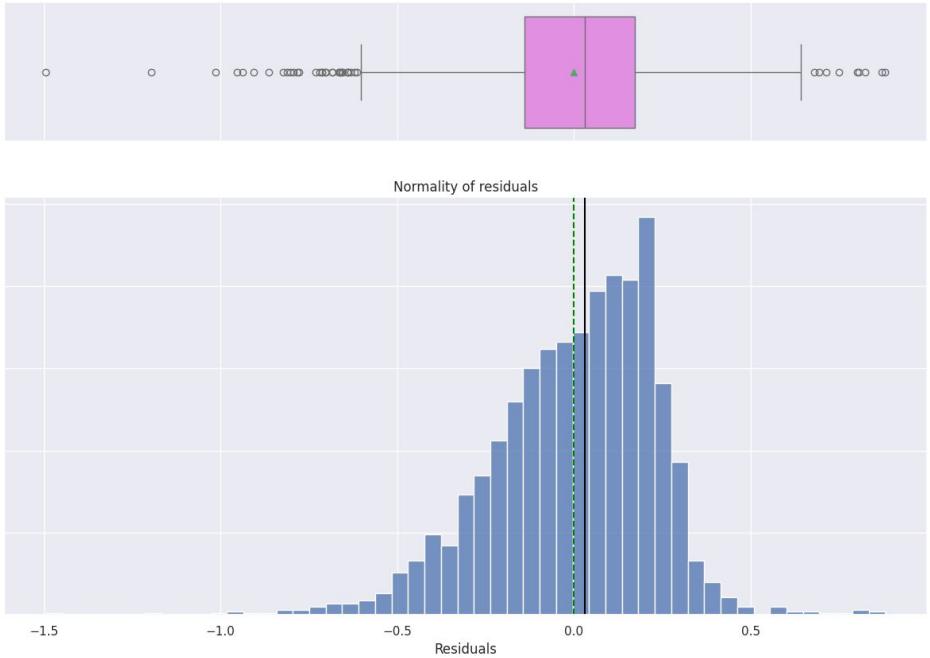


Test for Normality

Based on the probability plot and histogram, the residuals are normally distributed.

Therefore, we can accept normality and the models passes the test for normality






```
stats.shapiro(df_pred['Residuals'].values) ## Complete the code to apply the Shapiro-Wilks test
```

```
ShapiroResult(statistic=0.9678086423937371, pvalue=7.634328509630319e-23)
```

Test for Homoscedasticity

P value is greater than 0.05.

We can determine that the residuals are homoscedastic.

```
import statsmodels.stats.api as sms
from statsmodels.compat import lzip

name = ["F statistic", "p-value"]
test = sms.het_goldfeldquandt(df_pred['Residuals'].values,olsmodel2.model.exog)
lzip(name, test)
```

```
[('F statistic', 1.0595652195489762), ('p-value', 0.15825251352781292)]
```

Final Model Summary

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.840
Method:	Least Squares	F-statistic:	1582.
Date:	Fri, 11 Oct 2024	Prob (F-statistic):	0.00
Time:	03:16:06	Log-Likelihood:	87.706
No. Observations:	2417	AIC:	-157.4
Df Residuals:	2408	BIC:	-105.3
Df Model:	8		
Covariance Type:	nonrobust		

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.23732	0.183571	0.843847	0.842479	4.480411

Training Performance

ode cell
l+M B

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.233348	0.18355	0.840176	0.839579	4.407851

Conclusion & Business Recommendations

Business Recommendations and Insights

- Significant factors influencing used phone prices include selfie_camera, screen_size, and int_memory, with higher values leading to increased prices. New_price also has a strong positive correlation the used price.
- Expensive brands tend to have more refurbished phones with larger screens and better selfie cameras, while cheaper brands have fewer such phones. Features like 5G contribute positively to prices, while factors like weight, battery size, RAM, and Android OS are insignificant for prediction.
- ReCell should focus on phones with better selfie cameras, bigger screens, and more storage, as these drive up used phone prices and appeal to buyers.
- ReCell should also Invest in high-end brands with better features, and not focus on lower-end models that don’t add as much value to the resale price.
- In the future it would be beneficial to gather more data on ReCell’s customers. This could include characteristics such as age, sex, income, etc. Moreover, ReCell should collect data on the refurbishment cost to identify which models are most cost-effective to refurbish and resell.