RECELL (Case Study)

Project 3: Supervised Learning

Date: November 11th, 2022.

By: Ijeoma Ejem

Contents / Agenda

- Executive Summary pg. 3 - 5.
- Business Problem and Solution Overview pg. 6.
- Data Overview pg. 7.
- EDA Results pg. 8 - 32.
- Data Processing Overview pg. 33 - 34.
- Model Building Overview pg. 35 - 36.
- Linear Assumption Checks pg. 37 - 38.
- Final Model Summary pg. 39.

Executive Summary

	Conclusion	Recommendation
1	Most phone brands appear to have a similar median RAM size of 4GB. The brands with the highest RAM size (including outliers) are OnePlus, Huawei, Motorola, Oppo, Samsung, Xiaomi.	This is useful for customers to know. ReCell can create more awareness on its homepage by highlighting phones from the listed brands that have large RAM and educating customers about what that means for them.
2	As the power of a phone battery increases the weight also tends to increase. Brands with battery power of 9000 mAh and above are 'Others', Lenovo, Samsung, Apple and Google.	The listed brands should be advertised to people who typically need longer lasting batteries on their devices, like; frequent travelers.
3	The top 5 brands with selfie camera megapixel counts over 8 are; Huawei (87 phones), Vivo (78 phones), Oppo (75 phones), Xiaomi (63 phones) and Samsung (57 phones).	Mobile devices from these brands should be compared with other features they offer and then selected for marketing to people who take lots of selfies .eg. Content creators, Influencers, Gen Z’s, Millennials.etc.

	Conclusion	Recommendation
4	The top 5 brands with rear camera megapixel counts over 16 are; Sony (37 phones), Motorola (11 phones), 'Others' (9 phones), HTC (6 phones) and ZTE (5 phones)	Mobile devices from the listed brands should be compared with other features they offer and then selected for marketing to people who usually take lots of high quality pictures .eg. Content creators, photographers.etc.
5	The more recent the release of a phone, the more expensive it is likely to be. This is the same across both new and used phones.	Newly released phones/tablets should also be marketed more frequently as they sell at higher prices. ReCell can also build budget calculators and other customer friendly tools onto its website to help customers find newest releases and deals within their budget, based on their preferences.
6	Based on high correlations between screen size and battery, people who enjoy larger screens for the purpose of entertainment will usually have the added benefit of more battery power. The phone with the largest screen size also has the highest mAh	There are also many phones with large screen sizes and lesser battery strength so people who are looking to buy phones mainly for entertainment should be recommended phones with longer lasting batteries on the ReCell platform.

	Conclusion	Recommendation
7	The median price of used phones/tablets tend to be higher if they have 4G and 5G capabilities than those without. Used phones with 5G capabilities tend to have a higher median price than those with 4G capabilities.	Devices with 5g capabilities sell at a higher price and should be pushed more in marketing, to emphasize the benefits of newer features.
8	Top attributes identified in the data with the highest degree of correlation are; Normalized used price and normalized new price (83%), Screen size and weight (83%), Screen size and battery (81%), Weight and battery (70%).	As these attributes are generally top concerns for customers. I recommend creating preset categories or sophisticated selection tools that will help customers make the best choices based on factors most important to them. This will reduce incidents of returns and negative experiences.
9		We will need demographic, geographic and psychographic data as there are more factors that can affect the choices ReCell customers will make. We will be able to create models that make predictions that will drive up demand and sales.

Business Problem and Solution Overview

Problem:

- 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.
- ReCell aims to tap into the potential of this market by analyzing its available data, building a linear regression model to predict the price of used phones/tablets and identifying factors that significantly influence it.

Solution:

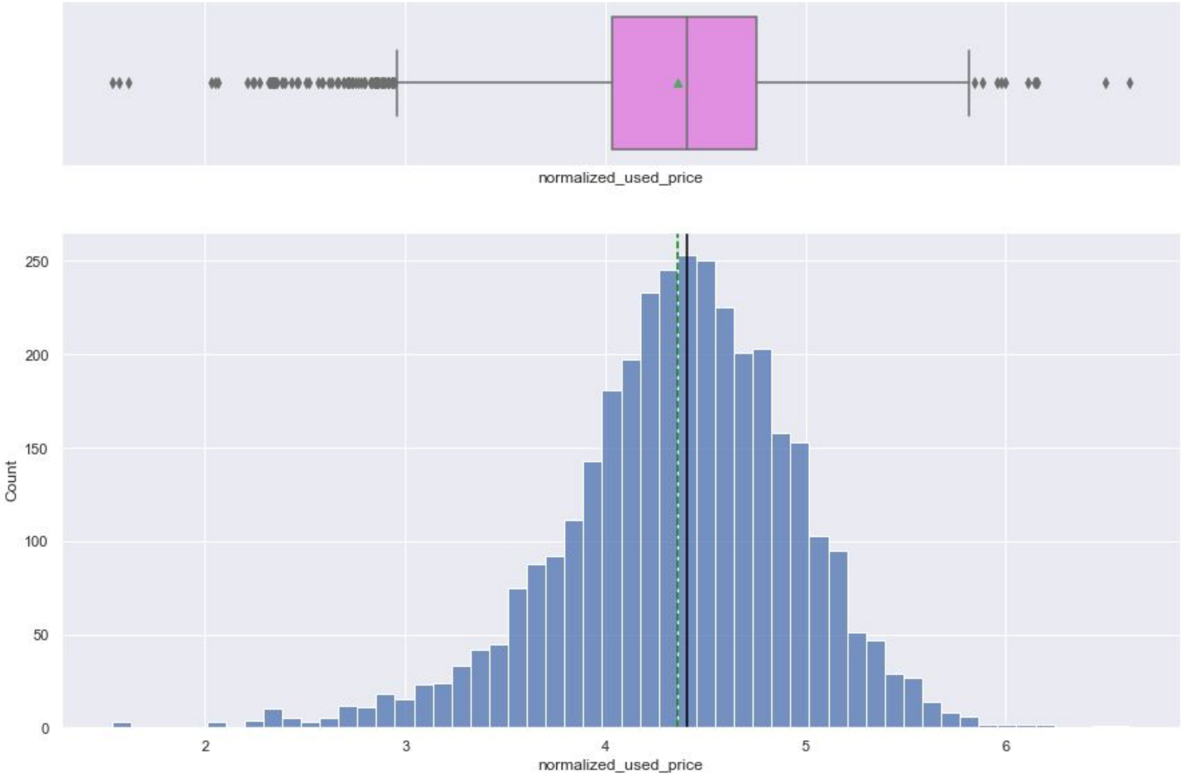
- Using exploratory data analysis (EDA) to find significant influences or patterns among dependent and independent variables.
- Building and testing a linear regression model to predict the price of used phones/tablets.
- Analyzing results from the data and providing business insights and recommendations to help take the company to the next level.

Data Overview

- ❖ There are 3454 rows (observations) and 15 columns (attributes) in total.
- ❖ The attributes we are analyzing and building a model around have 4 categorical variables and 11 numerical variables.
- ❖ From the statistical summary of the combined numerical and categorical data, the following is deduced;
  - The average used price for a phone/tablet is 4.36 euros, while the average new price is 5.23 euros.
  - The average release year for phones/tablets is 2015.
  - The average number of days used is 674 days.
- ❖ There are 202 missing values and 0 duplicate values in the data.

EDA Results

UNIVARIATE ANALYSIS

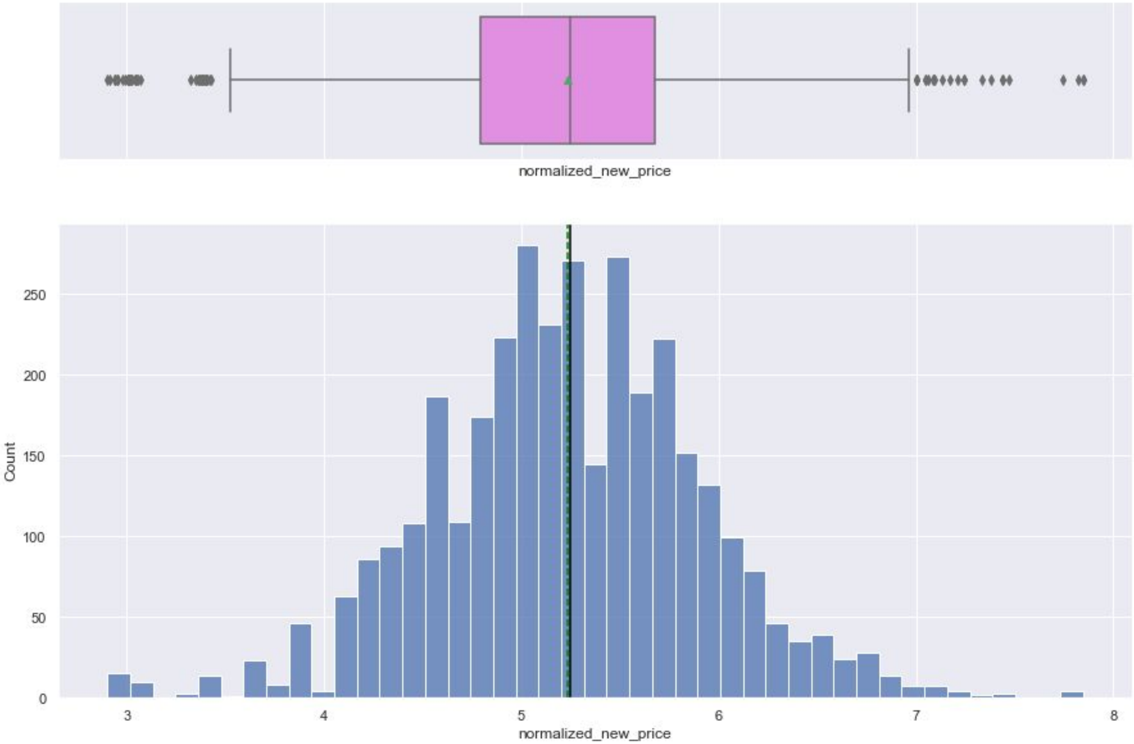


Normalized Used Price

The price of used phone/tablets has been normalized to fit a normal distribution. There are high numbers of outliers of both tails of the plot. The mean and median of normalized used phone price are both around 4.4 euros.

EDA Results

UNIVARIATE ANALYSIS

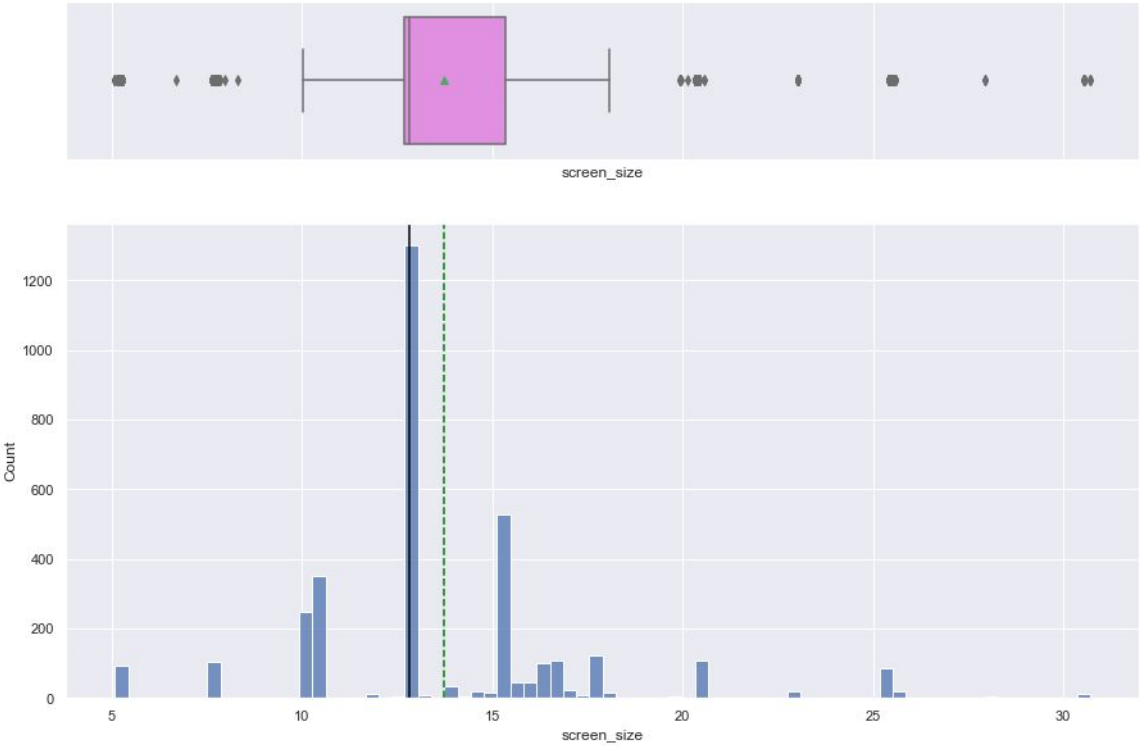


Normalized New Price

Similar to used phone price, the price of new phones/tablets has been normalized to fit a normal distribution. There are numerous outliers on both tails of the plot. The mean and median of normalized used phone price are both around 5.3 euros.

EDA Results

UNIVARIATE ANALYSIS

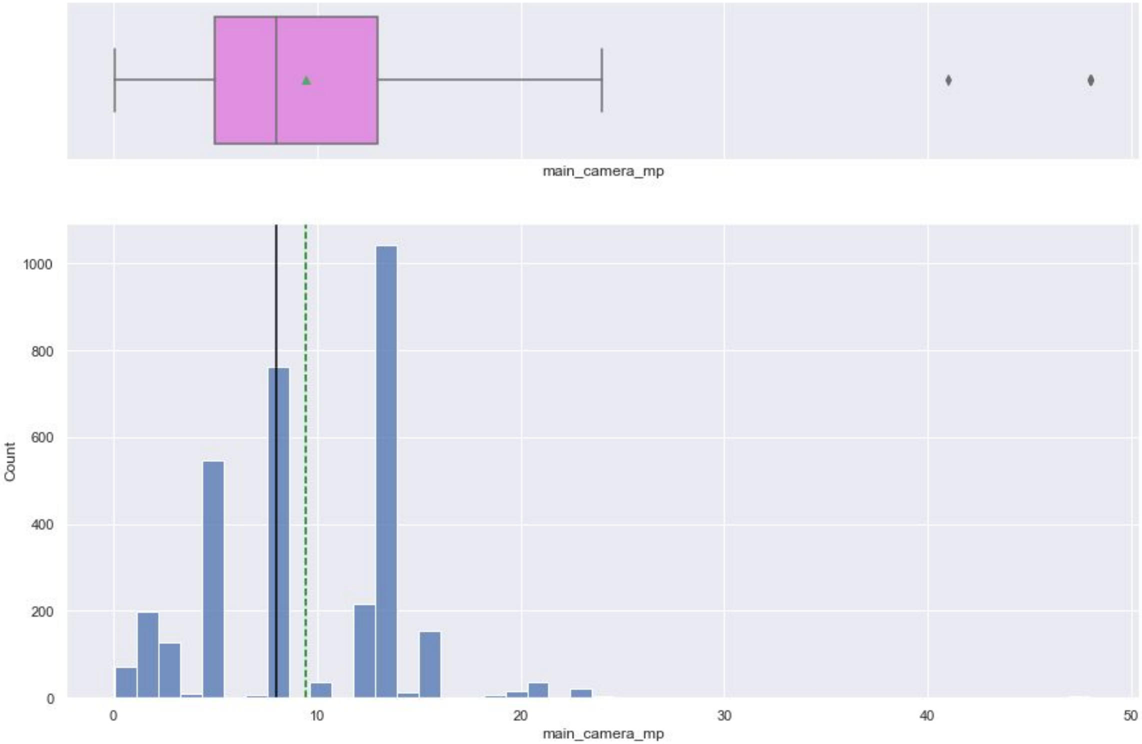


Screen Size

This distribution is right skewed with outliers on both tails. It has a median of about 13 cm and a mean of about 14 cm.

EDA Results

UNIVARIATE ANALYSIS

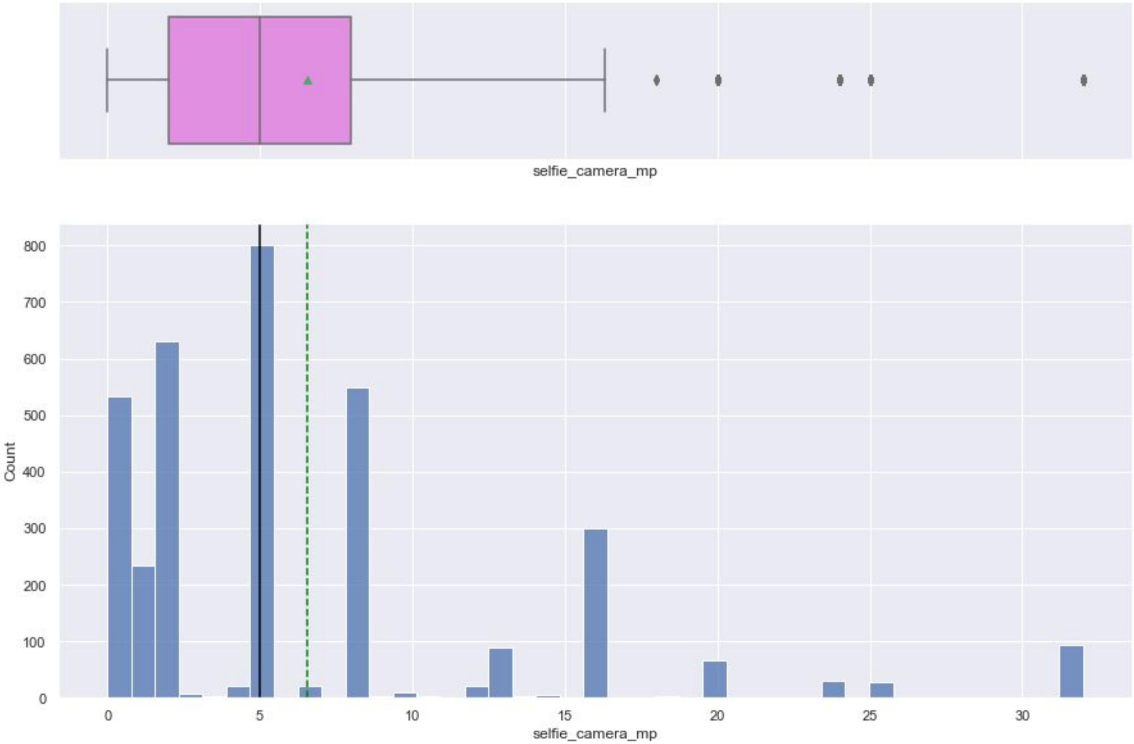


Main Camera Megapixel

This distribution is right skewed with few outliers on the right tail. It has a median of about 8 megapixels and a mean of about 9 megapixels.

EDA Results

UNIVARIATE ANALYSIS

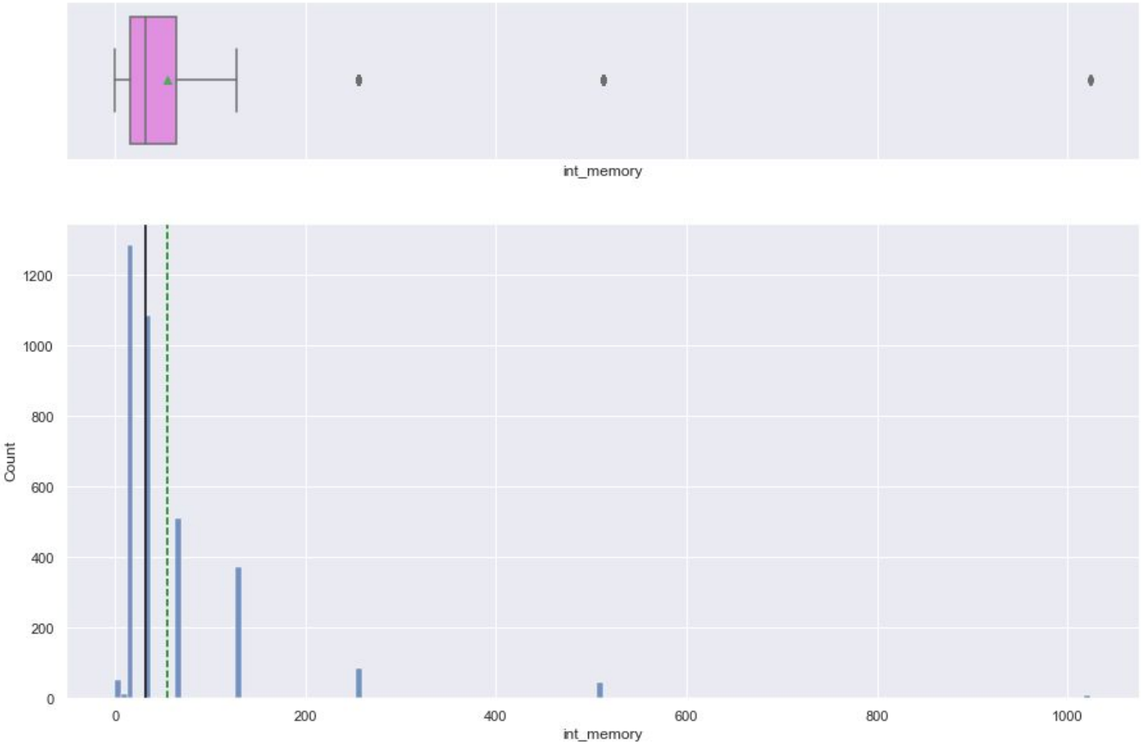


Selfie Camera Megapixel

This distribution is right skewed with some outliers on the right tail. It has a median of about 5 megapixels and a mean of about 6 megapixels.

EDA Results

UNIVARIATE ANALYSIS

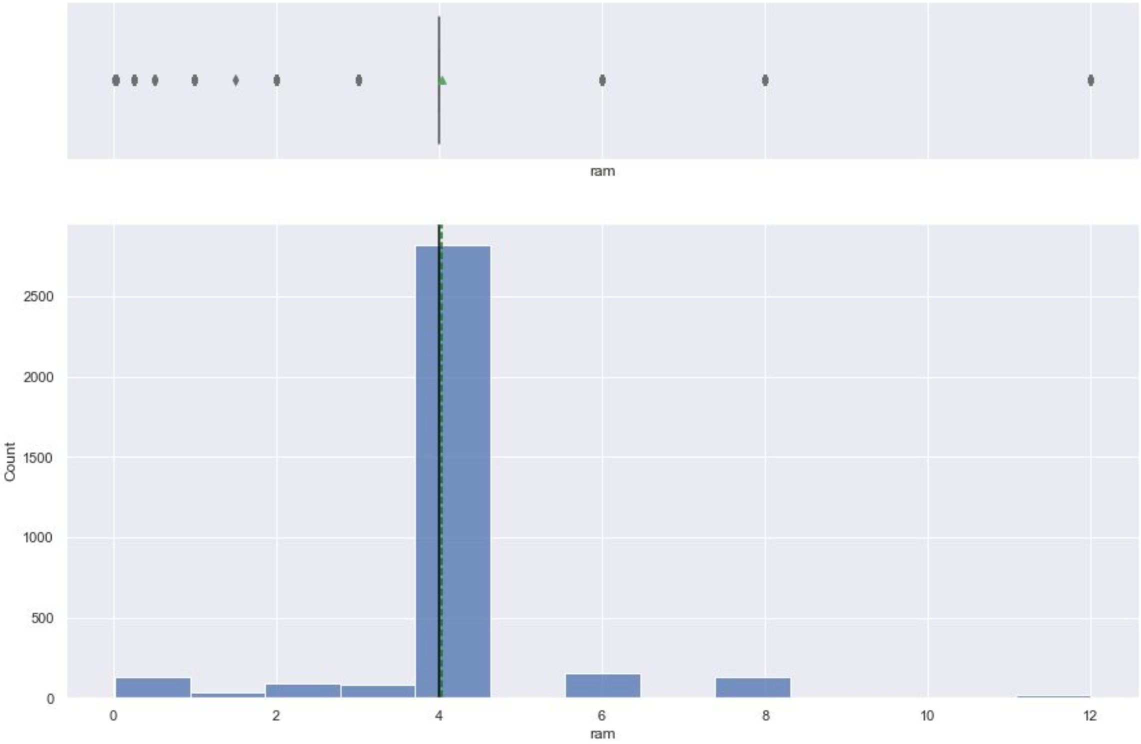


Internal Memory (ROM)

This distribution is right skewed with some spaced out outliers on the right tail. The mean and median of the internal memory fall between 30-50GB.

EDA Results

UNIVARIATE ANALYSIS

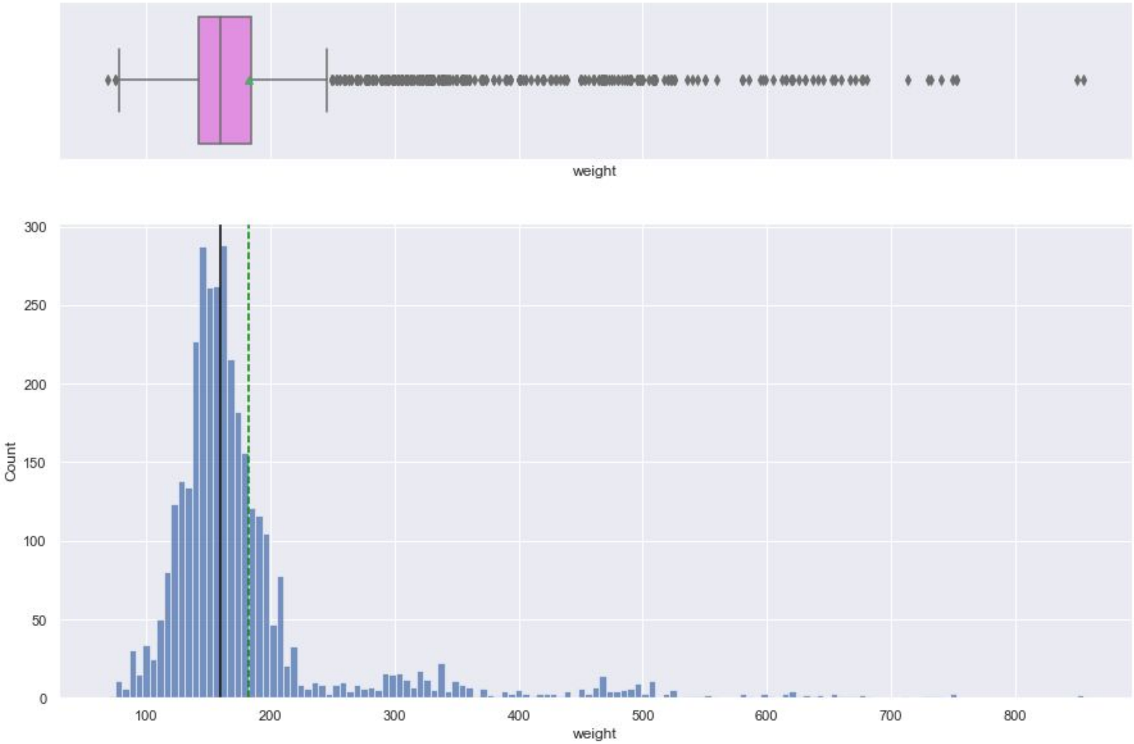


RAM

The mean and median of RAM are 4GB.

EDA Results

UNIVARIATE ANALYSIS

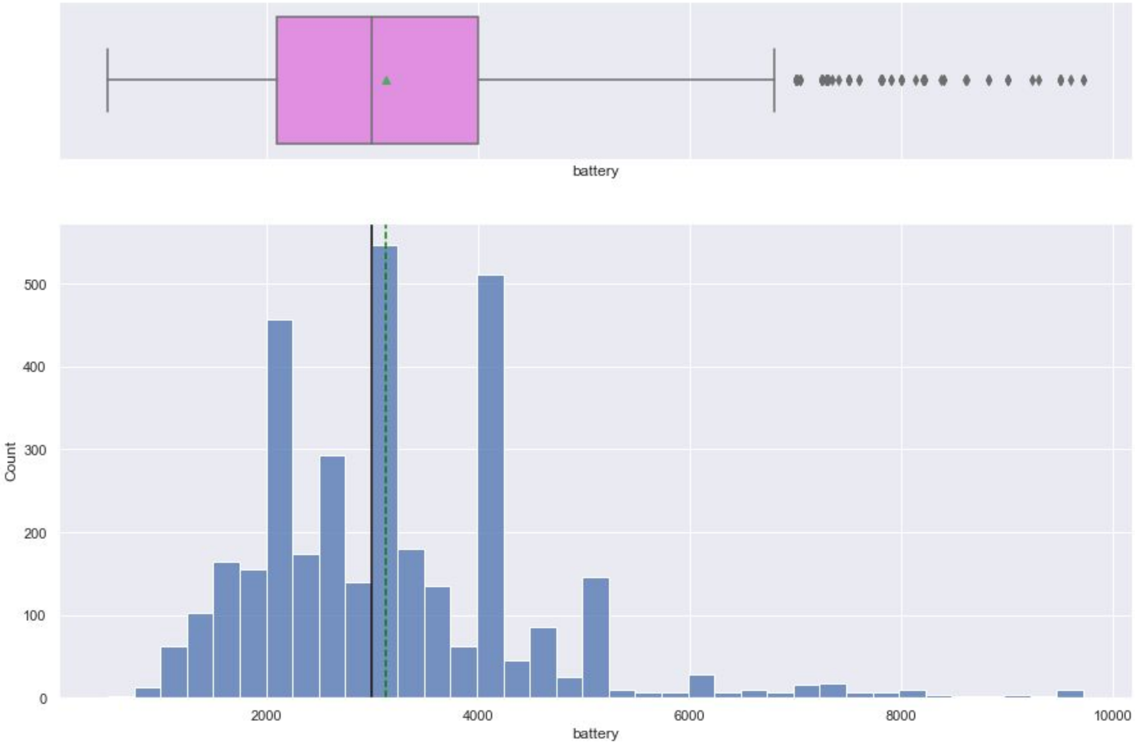


Weight

This distribution is right skewed with few outliers on the left tail and highly dense outliers on the right tail. The median weight of a device is about 160g and the mean is about 180g.

EDA Results

UNIVARIATE ANALYSIS

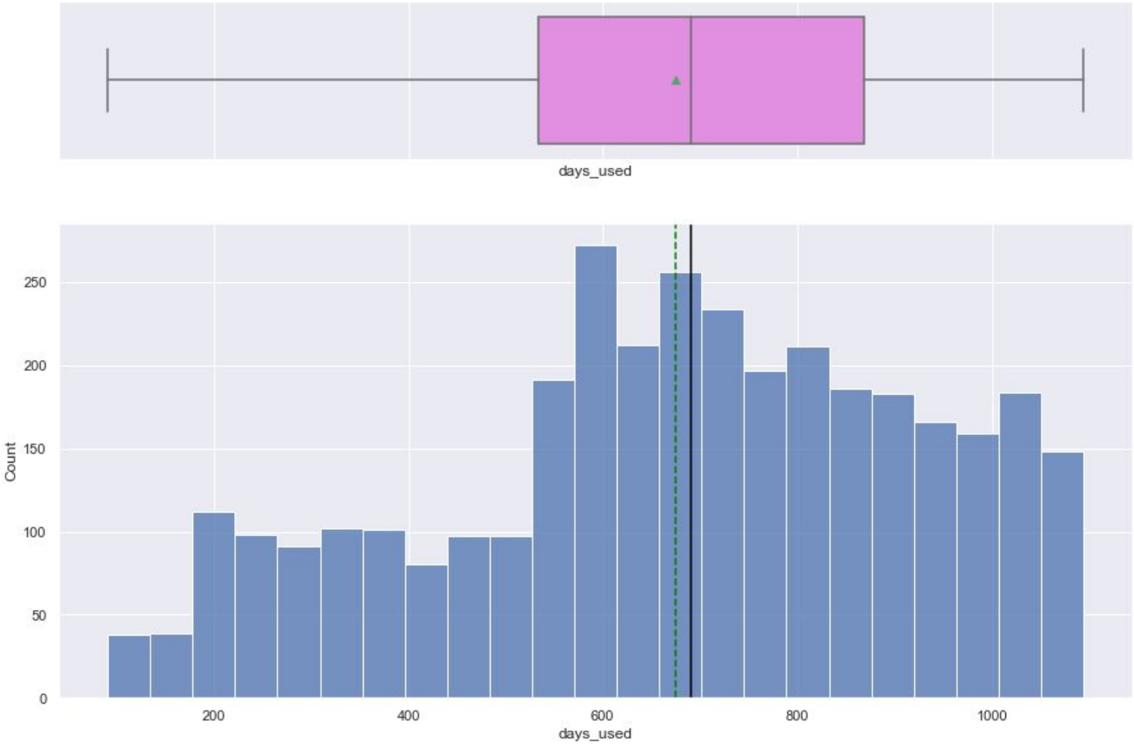


Battery

This distribution is right skewed with all outliers on the right tail. It has a median of about 3,000 mAh and a mean of about 3,150 mAh.

EDA Results

UNIVARIATE ANALYSIS

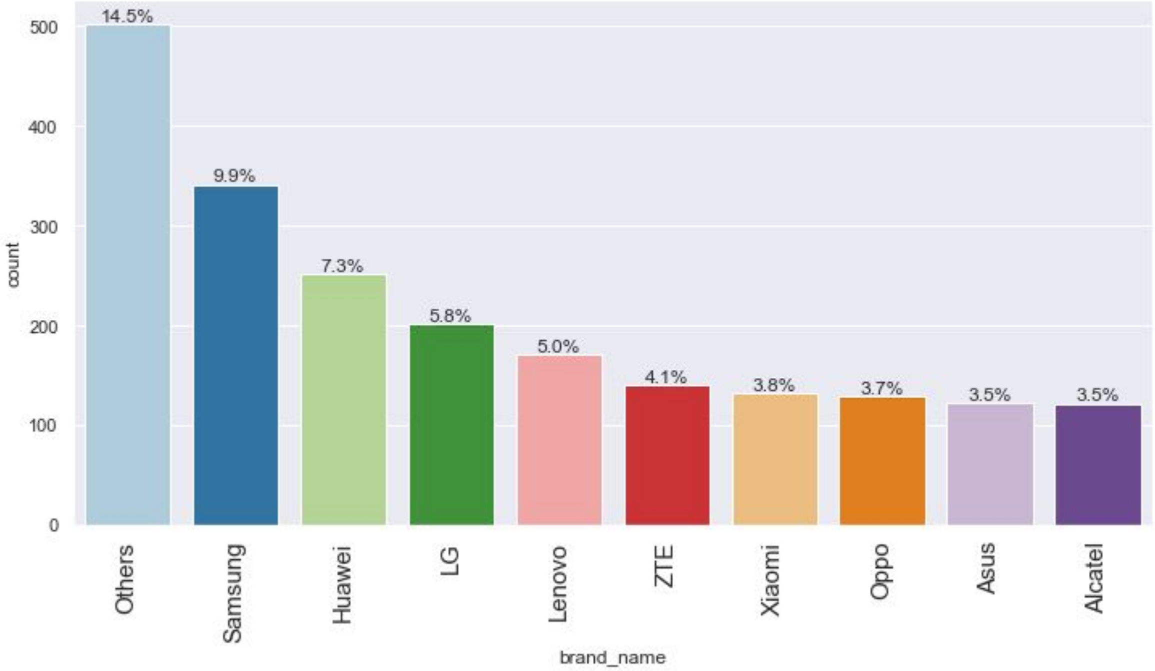


Days Used

This distribution is left skewed with no outliers. It has a mean and median ranging between 650 and 700 days.

EDA Results

UNIVARIATE ANALYSIS

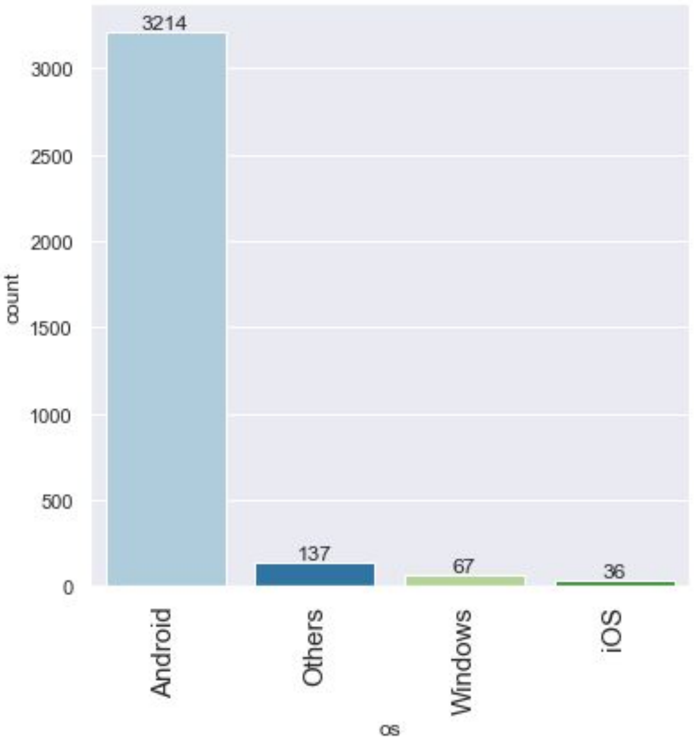


Brand Name

The highest count of phones by brand name in the data are categorized under 'Others' at 14.5%, followed by Samsung at 9.9% and then Huawei at 7.3%.

EDA Results

UNIVARIATE ANALYSIS

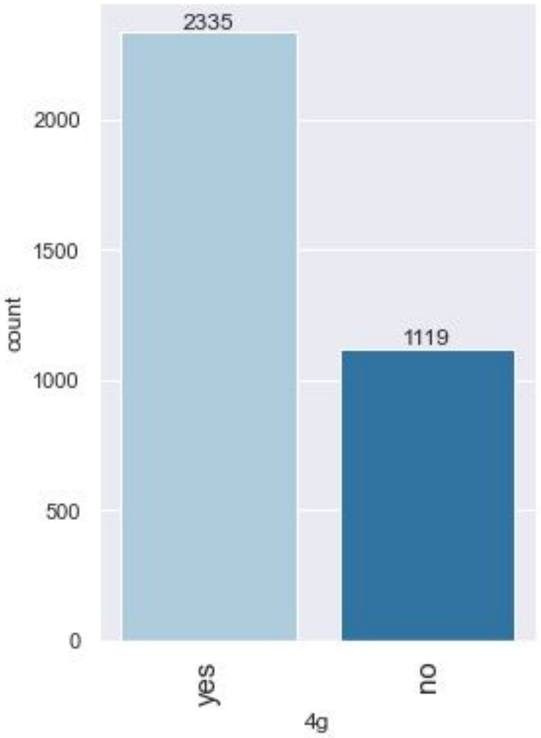


OS

The highest count of phones by OS in the data are Androids at 3214, followed by 'Others' at 137, then Windows at 67 and lastly iOS at 36 phones.

EDA Results

UNIVARIATE ANALYSIS

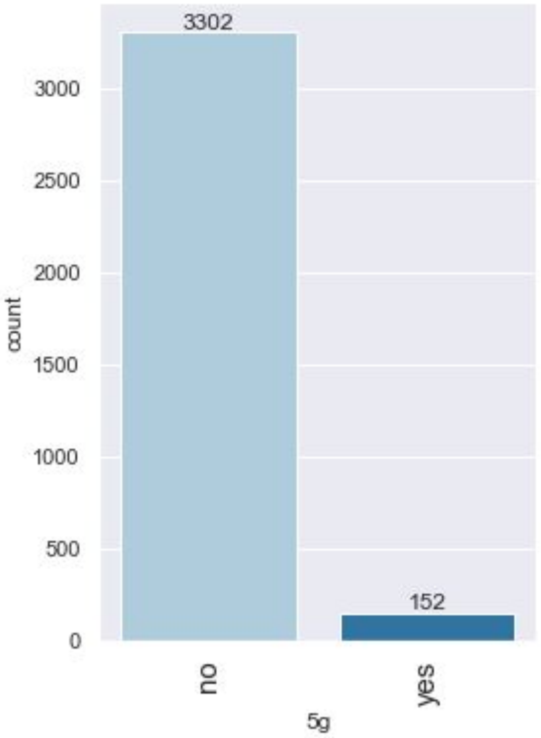


4G

2335 phones in the data have 4G capabilities, while 1119 phones do not.

EDA Results

UNIVARIATE ANALYSIS

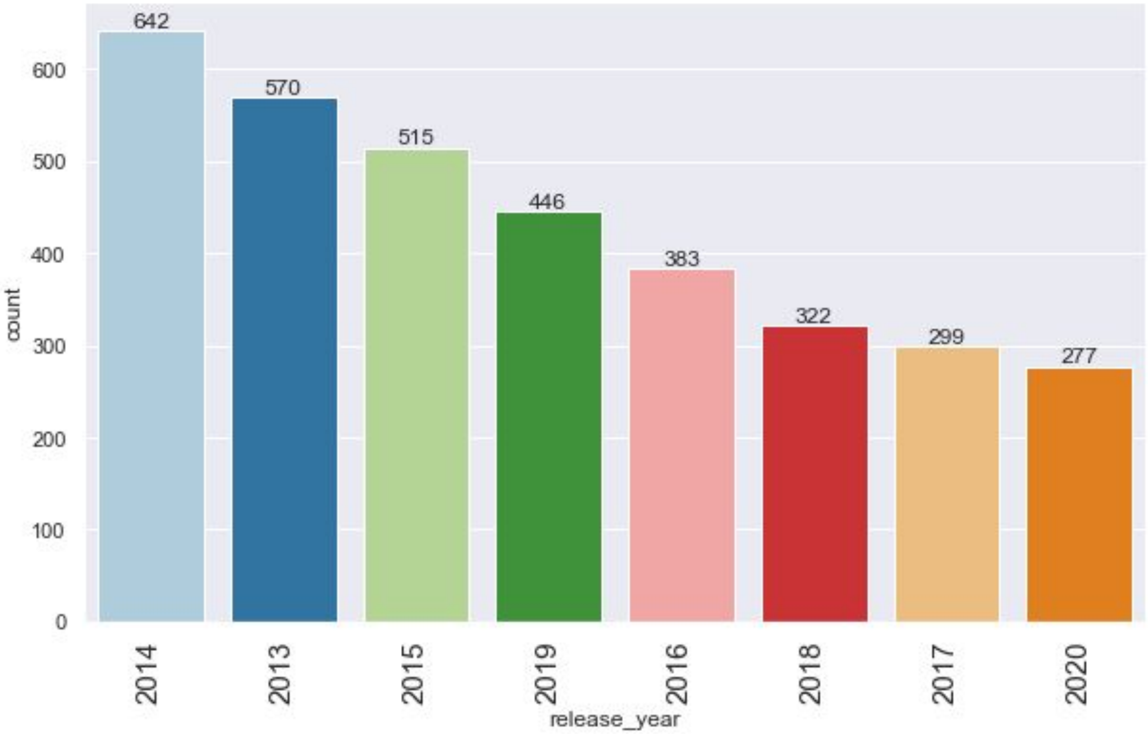


5G

152 phones in the data have 5G capabilities, while 3302 phones do not.

EDA Results

UNIVARIATE ANALYSIS



Release Year

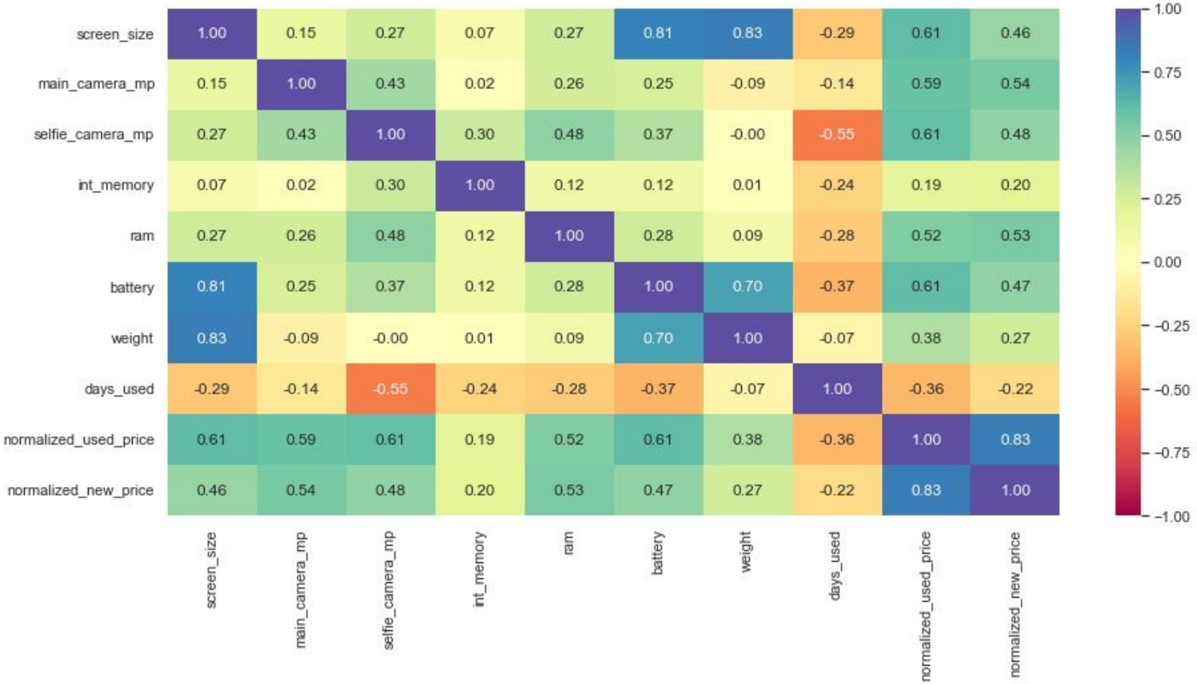
Most phones in the data were released between 2013-2015. 2020 is the year with the lowest count of released phones at a count of 277. 2014 is the year with the highest count of phones released with a count of 642.

EDA Results

BIVARIATE ANALYSIS

Heatmap Plot

The columns with the highest correlation in the data are: Normalized used price and normalized new price (83%), Screen size and weight (83%), Screen size and battery (81%) and Battery and Weight (70%).



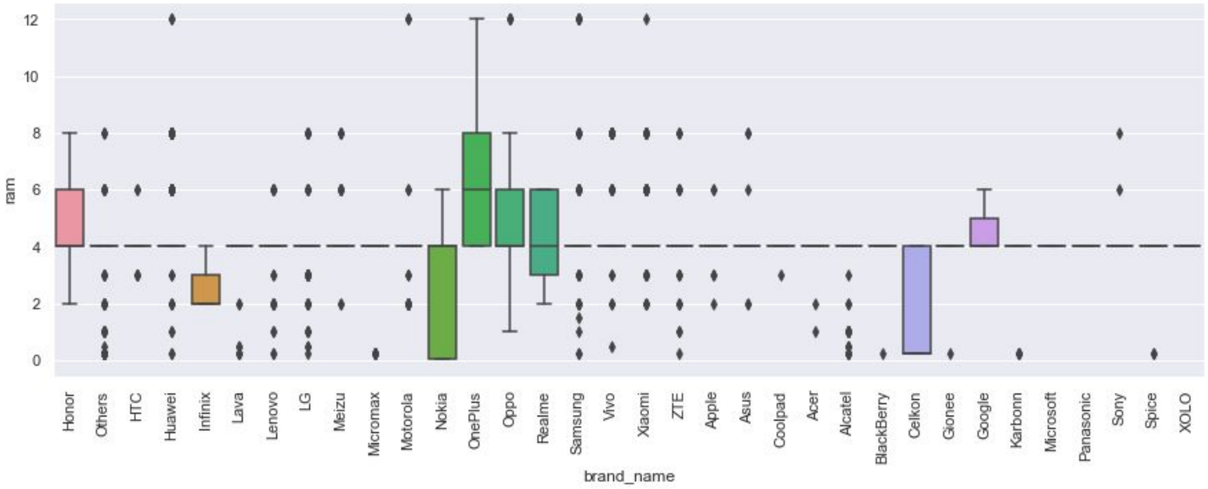
	screen_size	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	days_used	normalized_used_price	normalized_new_price
screen_size	1.00	0.15	0.27	0.07	0.27	0.81	0.83	-0.29	0.61	0.46
main_camera_mp	0.15	1.00	0.43	0.02	0.26	0.25	-0.09	-0.14	0.59	0.54
selfie_camera_mp	0.27	0.43	1.00	0.30	0.48	0.37	-0.00	-0.55	0.61	0.48
int_memory	0.07	0.02	0.30	1.00	0.12	0.12	0.01	-0.24	0.19	0.20
ram	0.27	0.26	0.48	0.12	1.00	0.28	0.09	-0.28	0.52	0.53
battery	0.81	0.25	0.37	0.12	0.28	1.00	0.70	-0.37	0.61	0.47
weight	0.83	-0.09	-0.00	0.01	0.09	0.70	1.00	-0.07	0.38	0.27
days_used	-0.29	-0.14	-0.55	-0.24	-0.28	-0.37	-0.07	1.00	-0.36	-0.22
normalized_used_price	0.61	0.59	0.61	0.19	0.52	0.61	0.38	-0.36	1.00	0.83
normalized_new_price	0.46	0.54	0.48	0.20	0.53	0.47	0.27	-0.22	0.83	1.00

EDA Results

BIVARIATE ANALYSIS

The amount of RAM is important for the smooth functioning of a device. We analyzed how the amount of RAM varies across brands.

Most brands seem to have similar median RAM size of 4GB. The brands with the highest RAM size on their devices (including outliers) are OnePlus, Huawei, Motorola, Oppo, Samsung, Xiaomi.



EDA Results

BIVARIATE ANALYSIS

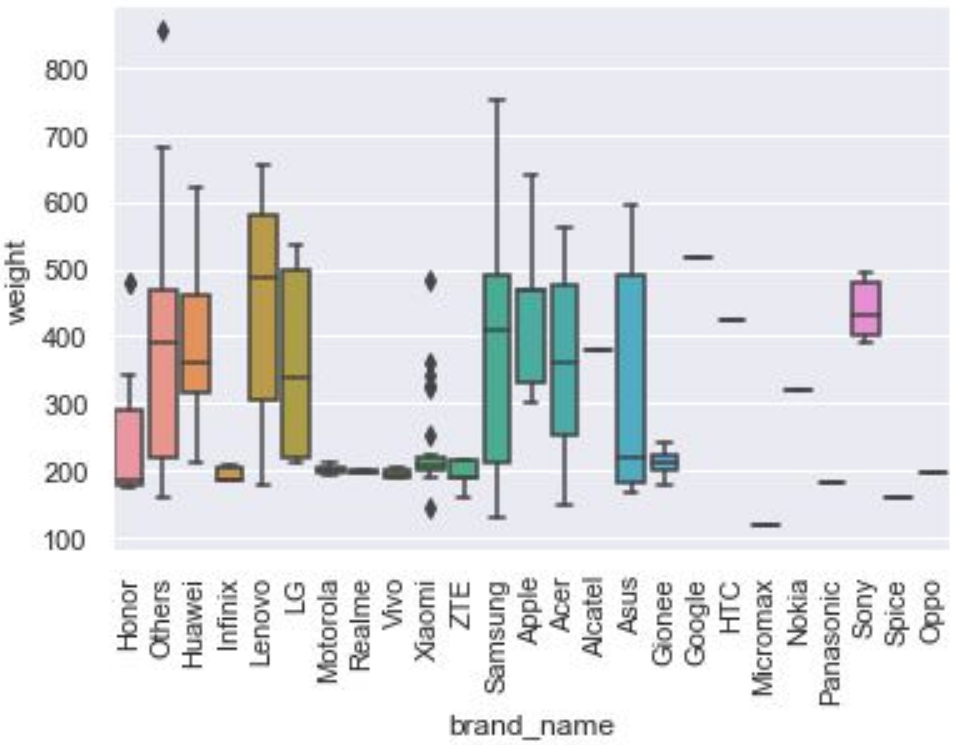
People who travel frequently require devices with large batteries to run through the day. But large battery often increases weight, making it feel uncomfortable in the hands. We created a new dataframe of only those devices which offer a large battery and analyze.

There are 341 rows and 15 columns in the new dataset created for phone batteries over 4,500 mAh.

There is a high correlation between battery and weight of device. In the heatmap we created, the correlation between the two is 70%. The scatterplot for battery and weight shows a positive correlation, as the battery power increases the weight also tends to increase.

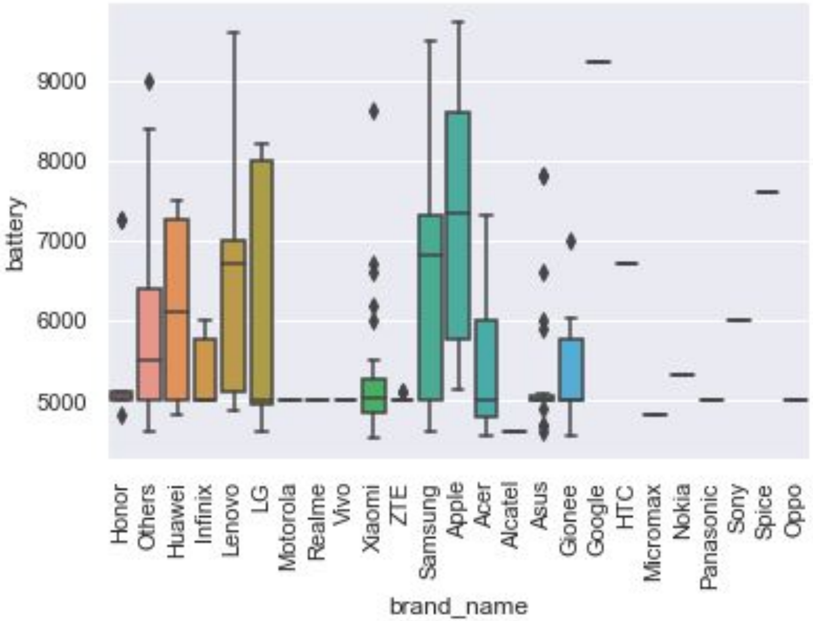
Brands with phone weight over 600 grams are 'Others', Huawei, Lenovo, Samsung and Apple.

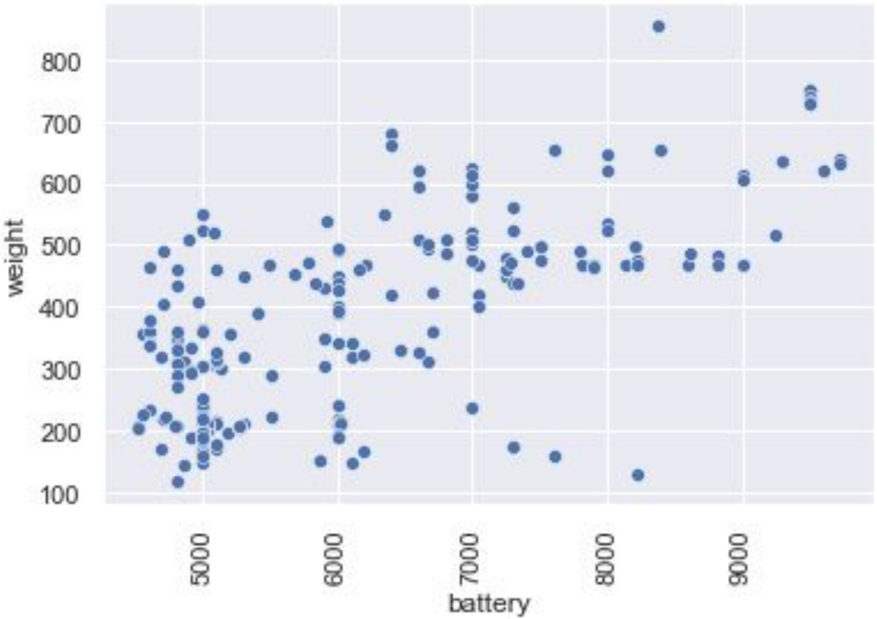
Brands with battery power of 9,000 mAh and above are 'Others', Lenovo, Samsung, Apple and Google.



EDA Results

BIVARIATE ANALYSIS





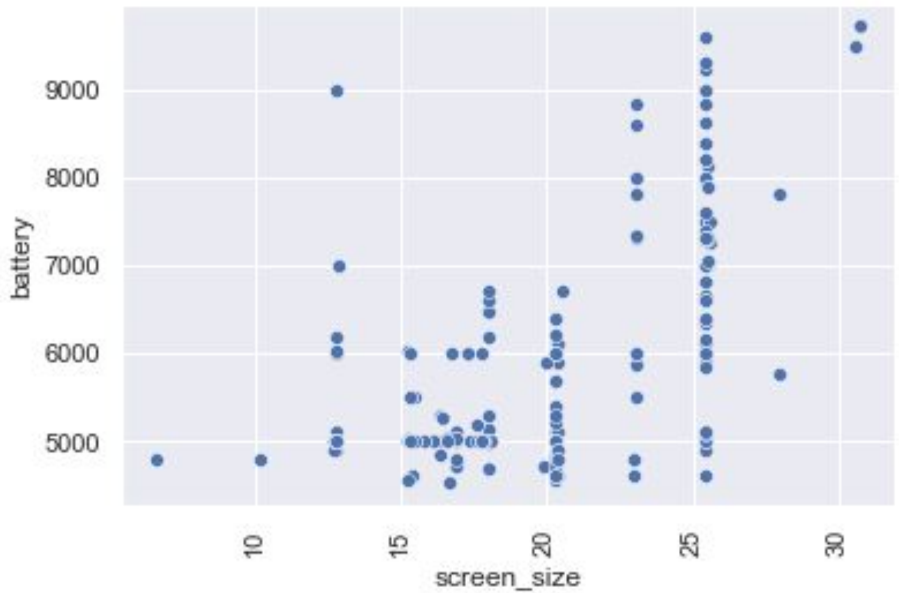
EDA Results

BIVARIATE ANALYSIS

People who buy phones and tablets primarily for entertainment purposes prefer a large screen as they offer a better viewing experience. We created a new dataframe of only those devices which are suitable for such people and analyzed.

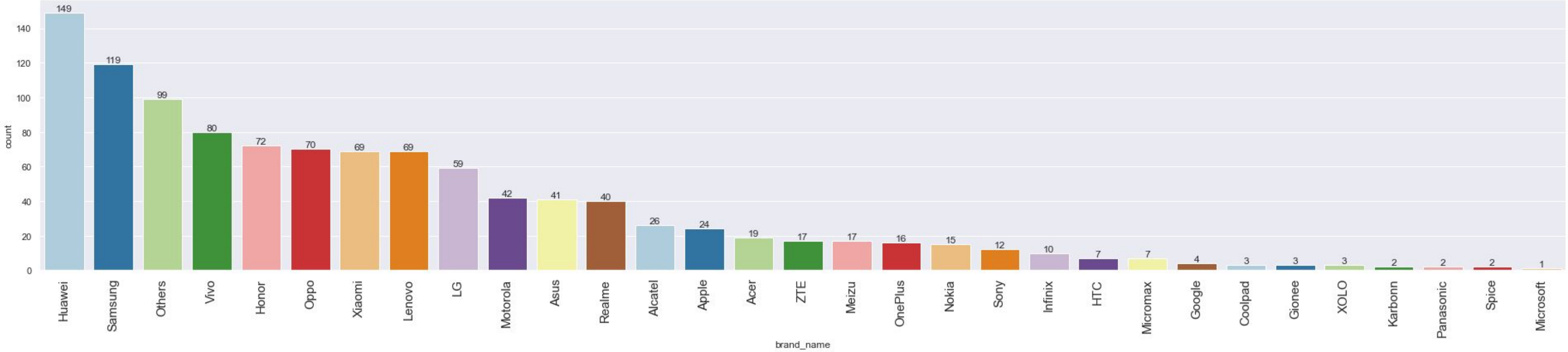
There are 1099 rows and 15 columns in the dataset created for phone screen sizes over 15.24cm.

In the heatmap we created, the correlation between screen size and battery was 83%, which is quite high. This means that people who enjoy larger screens for the purpose of entertainment will usually have the added benefit of more battery power. However, there are many phones with large screen sizes and less battery power. The scatterplot for battery and screen size shows that the device with the largest screen size will also have the highest battery strength.



EDA Results

BIVARIATE ANALYSIS



The top 5 brands with the highest count of phones/tablets with screen sizes over 15.24cm are; Huawei (149 phones), Samsung (119 phones), 'Others' (99 phones), Vivo (80 phones) and Honor (72 phones).

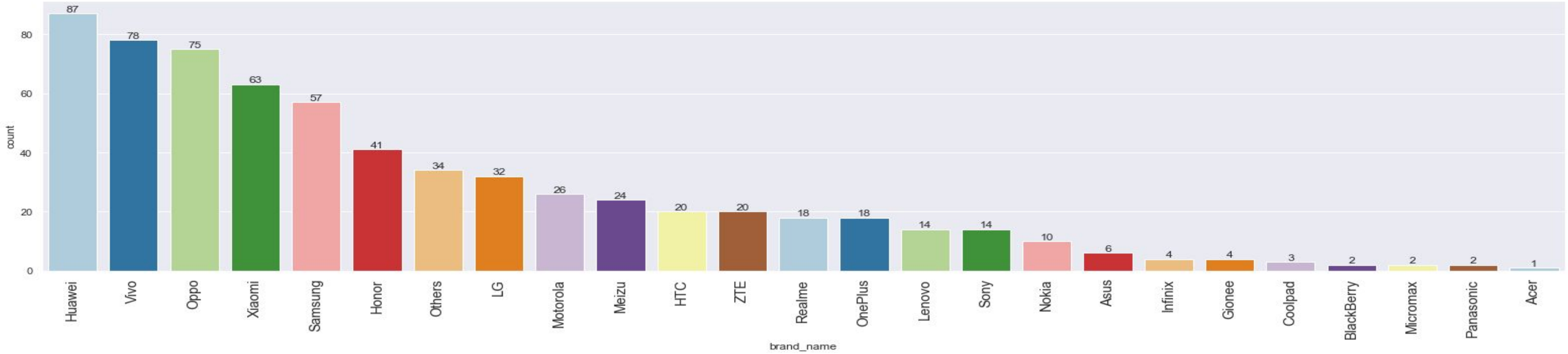
EDA Results

BIVARIATE ANALYSIS

Everyone likes a good camera to capture their favorite moments with loved ones. Some customers specifically look for good front cameras to click cool selfies. We created a new dataframe of only those devices which are suitable for this customer segment and analyzed.

There are 655 rows and 15 columns in the new dataset created for selfie camera megapixels over 8 megapixels.

The top 5 brands with selfie camera megapixel counts over 8 are; Huawei (87 phones), Vivo (78 phones), Oppo (75 phones), Xiaomi (63 phones) and Samsung (57 phones).



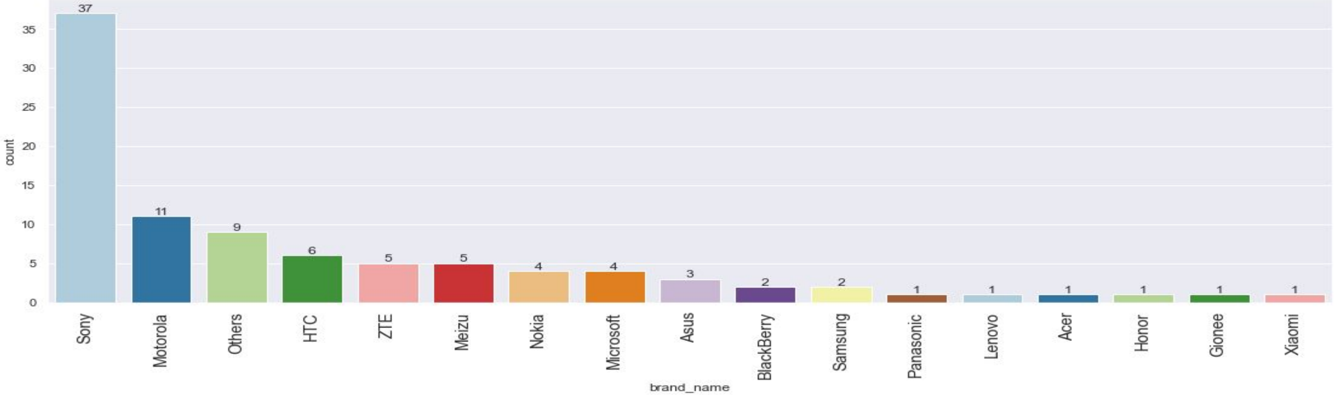
EDA Results

BIVARIATE ANALYSIS

We did a similar analysis for rear cameras.

There are 94 rows and 15 columns in the new dataset created for main (rear) camera megapixels over 16 megapixels.

The top 5 brands with main camera megapixel counts over 16 are; Sony (37 phones), Motorola (11 phones), 'Others' (9 phones), HTC (6 phones) and ZTE (5 phones).

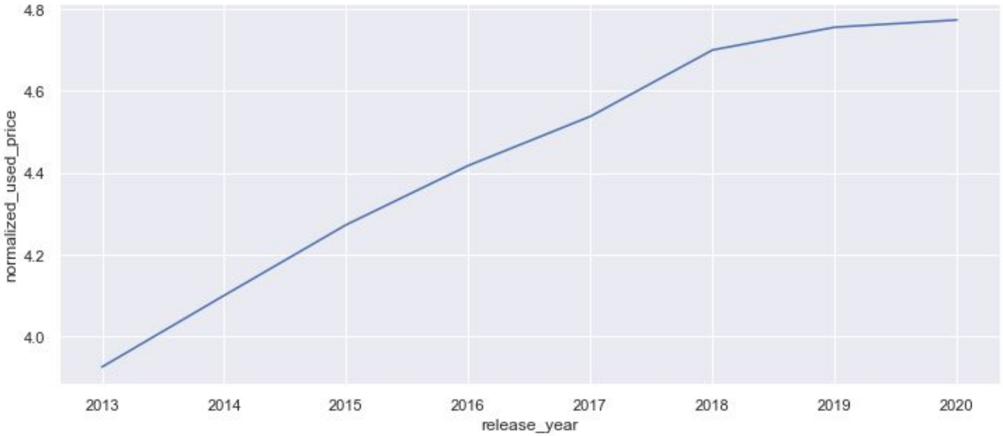


EDA Results

BIVARIATE ANALYSIS

We analyzed how the price of used devices varies across the years.

This line plot shows a positive correlation between normalized used price and release year. Understandably, as new phones are released the more expensive they are likely to be. This is likely to be the same across both new and used phones.

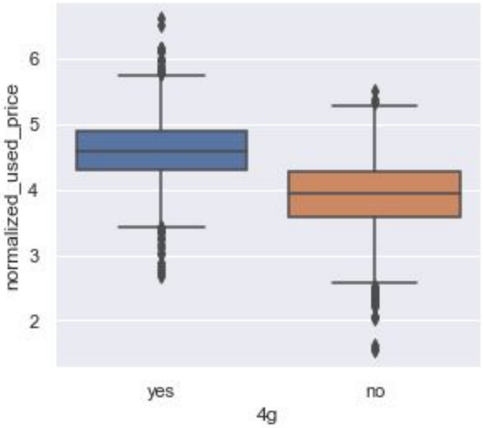


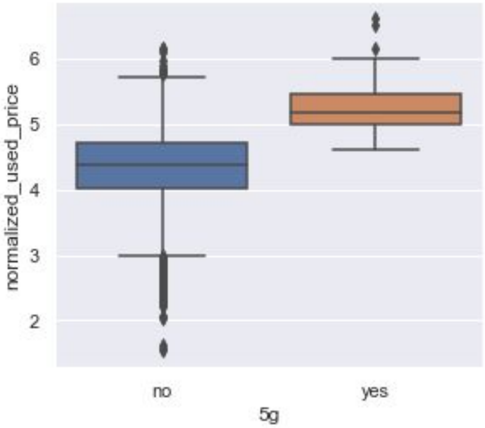
EDA Results

BIVARIATE ANALYSIS

We analyzed how the prices vary for used phones and tablets offering 4G and 5G networks.

The boxplot shows that the median price of used phones tends to be higher if they have 4G and 5G capabilities than those without. The median price of a used phones with 4G capabilities is about 4.6 euros, while the median price of used phones with 5G capabilities is higher at about 5.2 euros. While the median price of a used phones without 4G capabilities is about 4 euros, and the median price of used phones without 5G capabilities is about 4.5 euros.





Data Processing Overview

Data Preprocessing:

We imputed the missing values in the data with median values, grouped by ‘brand_name’.

Feature Engineering:

We created a column for 'years since release', which shows the length of time (in years) since the original release date of the phones/tablets in the dataset. We found that the average number of years since release is 5 years. The median value is 5.5 years with a standard deviation of 2.3 years. The minimum number of years since release is 1 year and the maximum is 8 years.

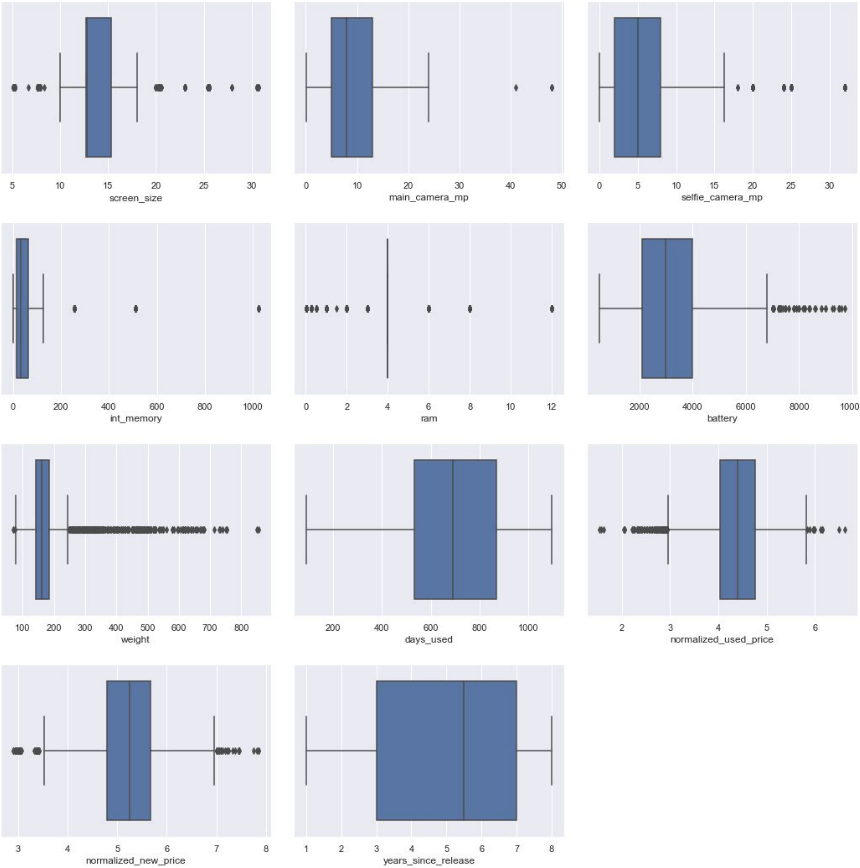
Data Preparation for Modeling:

We encoded the categorical columns/features, then split the data into ‘training’ and ‘testing’ data in order to train and test the built model. We split the data into 70:30; Number of rows in training data (70%) = 2417, Number of rows in test data (30%) = 1037.

Data Processing Overview

Outlier Check:

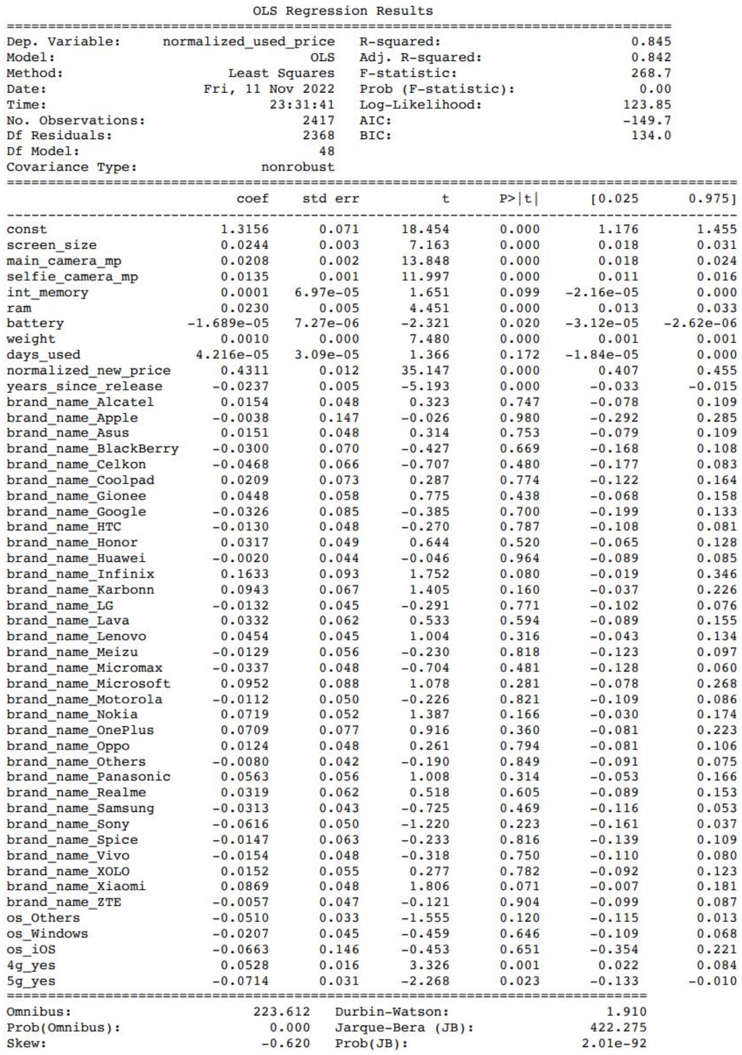
All columns with the exception of days used and years since release have outliers. Weight appears to have the densest and highest number of outliers, while internal memory (ROM) and main camera mp appear to have the least number of outliers in the dataset. It is likely not safe to treat these outliers as they seem to hold a lot of information that would otherwise be lost if we were to eliminate them.



Model Building Overview

Initial Model Performance Evaluation:

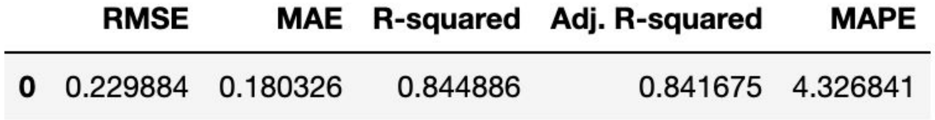
The training *R*2 is 84.4%, indicating that the model explains 84% of the variation in the train data. The training and test model are comparable so there is no underfitting. MAE and RMSE on the train and test sets are also comparable, which shows that the model is not overfitting. MAE indicates that our current model is able to predict used price of phones within a mean error of 0.18 euros on the test data. MAPE on the test data suggests we can predict within 4.5% of the used price.



```
                            OLS Regression Results
==================================================================================
Dep. Variable:     normalized_used_price   R-squared:                       0.845
Model:                               OLS   Adj. R-squared:                  0.842
Method:                    Least Squares   F-statistic:                     268.7
Date:                   Fri, 11 Nov 2022   Prob (F-statistic):               0.00
Time:                           23:31:41   Log-Likelihood:                 123.85
No. Observations:                   2417   AIC:                            -149.7
Df Residuals:                       2368   BIC:                             134.0
Df Model:                             48
Covariance Type:               nonrobust
==========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                       1.3156      0.071     18.454      0.000       1.176       1.455
screen_size                 0.0244      0.003      7.163      0.000       0.018       0.031
main_camera_mp              0.0208      0.002     13.848      0.000       0.018       0.024
selfie_camera_mp            0.0135      0.001     11.997      0.000       0.011       0.016
int_memory                  0.0001   6.97e-05      1.651      0.099   -2.16e-05       0.000
ram                         0.0230      0.005      4.451      0.000       0.013       0.033
battery                 -1.689e-05   7.27e-06     -2.321      0.020   -3.12e-05   -2.62e-06
weight                      0.0010      0.000      7.480      0.000       0.001       0.001
days_used                4.216e-05   3.09e-05      1.366      0.172   -1.84e-05       0.000
normalized_new_price        0.4311      0.012     35.147      0.000       0.407       0.455
years_since_release        -0.0237      0.005     -5.193      0.000      -0.033      -0.015
brand_name_Alcatel          0.0154      0.048      0.323      0.747      -0.078       0.109
brand_name_Apple           -0.0038      0.147     -0.026      0.980      -0.292       0.285
brand_name_Asus             0.0151      0.048      0.314      0.753      -0.079       0.109
brand_name_BlackBerry      -0.0300      0.070     -0.427      0.669      -0.168       0.108
brand_name_Celkon          -0.0468      0.066     -0.707      0.480      -0.177       0.083
brand_name_Coolpad          0.0209      0.073      0.287      0.774      -0.122       0.164
brand_name_Gionee           0.0448      0.058      0.775      0.438      -0.068       0.158
brand_name_Google          -0.0326      0.085     -0.385      0.700      -0.199       0.133
brand_name_HTC             -0.0130      0.048     -0.270      0.787      -0.108       0.081
brand_name_Honor            0.0317      0.049      0.644      0.520      -0.065       0.128
brand_name_Huawei          -0.0020      0.044     -0.046      0.964      -0.089       0.085
brand_name_Infinix          0.1633      0.093      1.752      0.080      -0.019       0.346
brand_name_Karbonn          0.0943      0.067      1.405      0.160      -0.037       0.226
brand_name_LG              -0.0132      0.045     -0.291      0.771      -0.102       0.076
brand_name_Lava             0.0332      0.062      0.533      0.594      -0.089       0.155
brand_name_Lenovo           0.0454      0.045      1.004      0.316      -0.043       0.134
brand_name_Meizu           -0.0129      0.056     -0.230      0.818      -0.123       0.097
brand_name_Micromax        -0.0337      0.048     -0.704      0.481      -0.128       0.060
brand_name_Microsoft        0.0952      0.088      1.078      0.281      -0.078       0.268
brand_name_Motorola        -0.0112      0.050     -0.226      0.821      -0.109       0.086
brand_name_Nokia            0.0719      0.052      1.387      0.166      -0.030       0.174
brand_name_OnePlus          0.0709      0.077      0.916      0.360      -0.081       0.223
brand_name_Oppo             0.0124      0.048      0.261      0.794      -0.081       0.106
brand_name_Others          -0.0080      0.042     -0.190      0.849      -0.091       0.075
brand_name_Panasonic        0.0563      0.056      1.008      0.314      -0.053       0.166
brand_name_Realme           0.0319      0.062      0.518      0.605      -0.089       0.153
brand_name_Samsung         -0.0313      0.043     -0.725      0.469      -0.116       0.053
brand_name_Sony            -0.0616      0.050     -1.220      0.223      -0.161       0.037
brand_name_Spice           -0.0147      0.063     -0.233      0.816      -0.139       0.109
brand_name_Vivo            -0.0154      0.048     -0.318      0.750      -0.110       0.080
brand_name_XOLO             0.0152      0.055      0.277      0.782      -0.092       0.123
brand_name_Xiaomi           0.0869      0.048      1.806      0.071      -0.007       0.181
brand_name_ZTE             -0.0057      0.047     -0.121      0.904      -0.099       0.087
os_Others                  -0.0510      0.033     -1.555      0.120      -0.115       0.013
os_Windows                 -0.0207      0.045     -0.459      0.646      -0.109       0.068
os_iOS                     -0.0663      0.146     -0.453      0.651      -0.354       0.221
4g_yes                      0.0528      0.016      3.326      0.001       0.022       0.084
5g_yes                     -0.0714      0.031     -2.268      0.023      -0.133      -0.010
==============================================================================
Omnibus:                      223.612   Durbin-Watson:                   1.910
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              422.275
Skew:                          -0.620   Prob(JB):                     2.01e-92
```

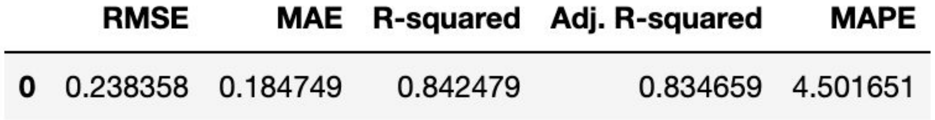
Model Building Overview

Training Data Performance Check:



	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.229884	0.180326	0.844886	0.841675	4.326841

Testing Data Performance Check:



	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.238358	0.184749	0.842479	0.834659	4.501651

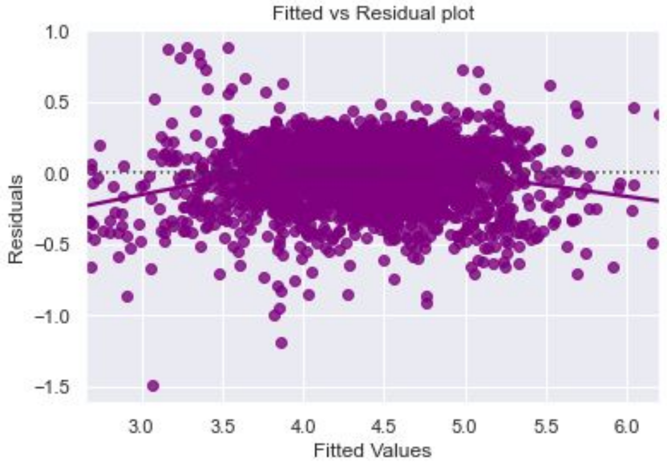
Checking Linear Assumptions

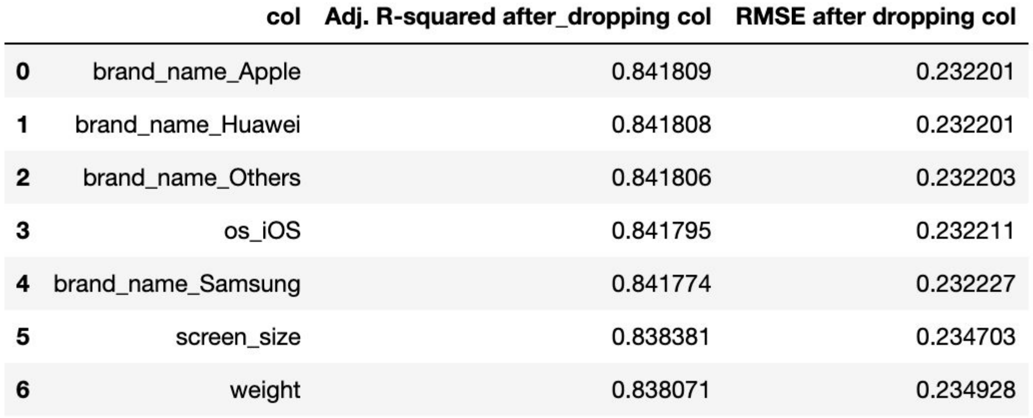
Test for Multicollinearity:

Some variables show moderate to high multicollinearity (7 initially identified with VIF scores above 5) so we proceeded to drop values one after the other using a loop until our p-values for all variables were below 0.05.

Test for Linearity and Independence:

No visible patterns were found in this test, we concluded that our model is linear and independent. Hence, the assumption of linearity and independence in the data is satisfied.





	col	Adj. R-squared after_dropping col	RMSE after dropping col
0	brand_name_Apple	0.841809	0.232201
1	brand_name_Huawei	0.841808	0.232201
2	brand_name_Others	0.841806	0.232203
3	os_iOS	0.841795	0.232211
4	brand_name_Samsung	0.841774	0.232227
5	screen_size	0.838381	0.234703
6	weight	0.838071	0.234928

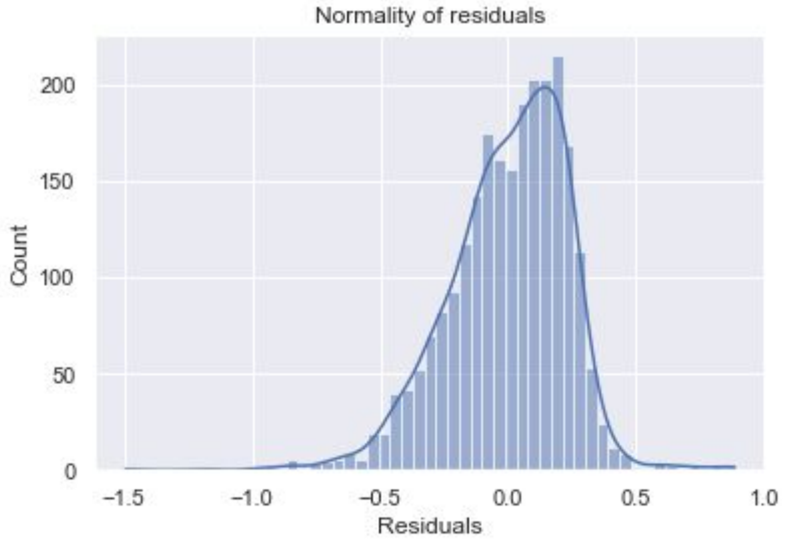
Checking Linear Assumptions

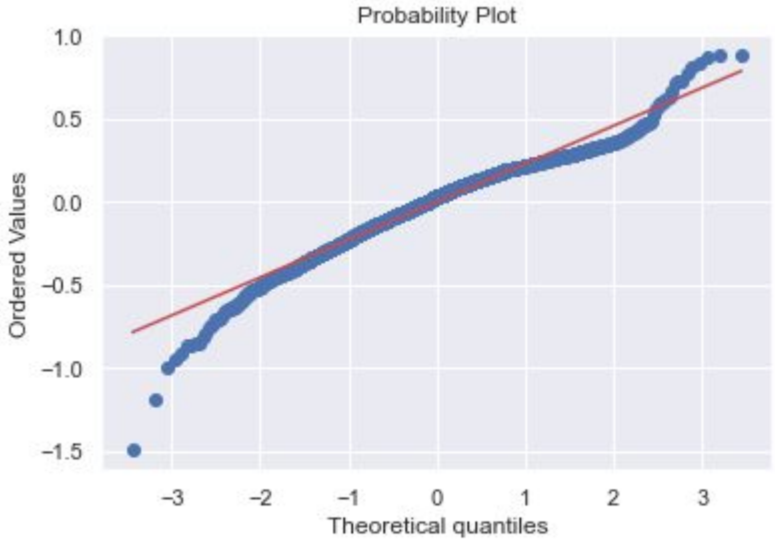
Test for Homoscedasticity:

Our p-value is 0.229. Since p-value > 0.05, we concluded that the residuals are homoscedastic and that the assumption is satisfied.

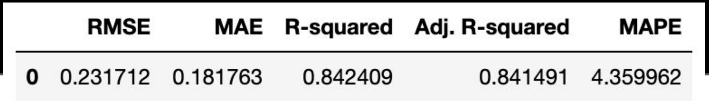
Test for Normality:

The Shapiro-Wilk test proves the p-value of 2.13 to be greater than 0.05. The plotted distribution is also relatively close to being normal. As a result, we will conclude that the assumption is satisfied.





Final Model Summary

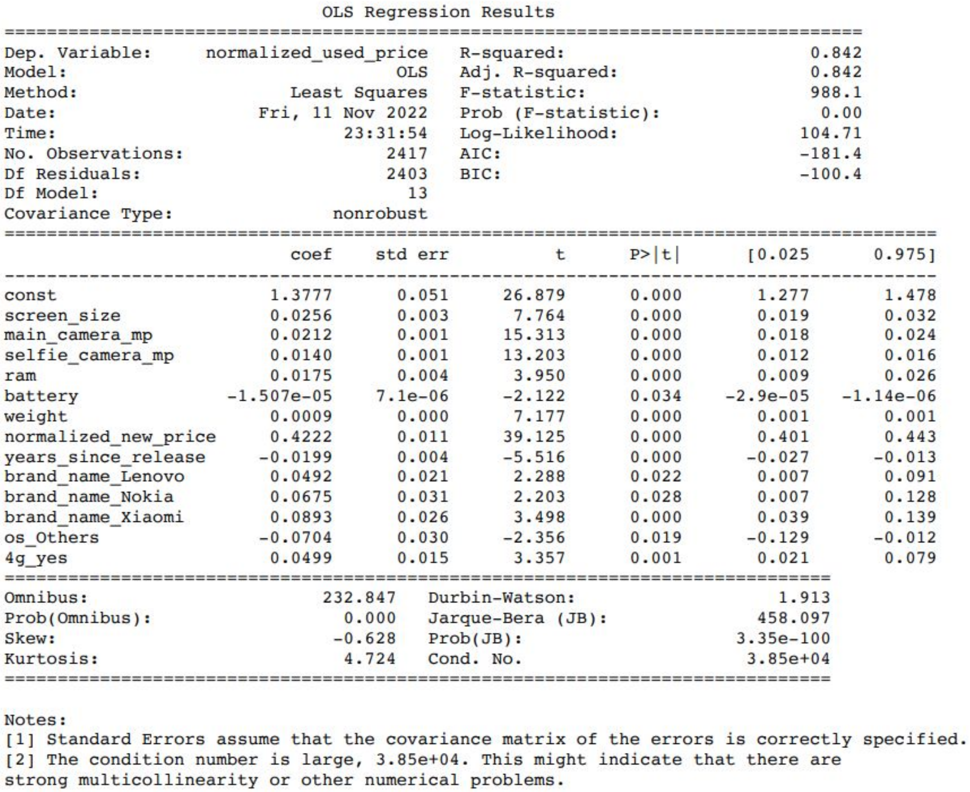


	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.231712	0.181763	0.842409	0.841491	4.359962

Final Model Performance Evaluation:

The training R2 is still 84%, indicating that the model explains 84% of the variation in the trained data. MAE indicates that our final model is able to predict used price of phones within a mean error of 0.18 euros on the test data. MAPE on the test data suggests we can predict within 4.35% of the used price of phones/tablets.

The training and testing data returned the exact same values in the performance check. These values have minor visible changes from the initial model performance check. No overfitting or underfitting is seen in our final model. Since all assumptions are fulfilled, we have concluded that we are satisfied with our final linear regression model.



```
                            OLS Regression Results
================================================================================
Dep. Variable:     normalized_used_price   R-squared:                      0.842
Model:                               OLS   Adj. R-squared:                 0.842
Method:                    Least Squares   F-statistic:                    988.1
Date:                   Fri, 11 Nov 2022   Prob (F-statistic):              0.00
Time:                           23:31:54   Log-Likelihood:                104.71
No. Observations:                   2417   AIC:                           -181.4
Df Residuals:                       2403   BIC:                           -100.4
Df Model:                             13
Covariance Type:               nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                     1.3777      0.051     26.879      0.000       1.277       1.478
screen_size               0.0256      0.003      7.764      0.000       0.019       0.032
main_camera_mp            0.0212      0.001     15.313      0.000       0.018       0.024
selfie_camera_mp          0.0140      0.001     13.203      0.000       0.012       0.016
ram                       0.0175      0.004      3.950      0.000       0.009       0.026
battery               -1.507e-05    7.1e-06     -2.122      0.034    -2.9e-05   -1.14e-06
weight                    0.0009      0.000      7.177      0.000       0.001       0.001
normalized_new_price      0.4222      0.011     39.125      0.000       0.401       0.443
years_since_release      -0.0199      0.004     -5.516      0.000      -0.027      -0.013
brand_name_Lenovo         0.0492      0.021      2.288      0.022       0.007       0.091
brand_name_Nokia          0.0675      0.031      2.203      0.028       0.007       0.128
brand_name_Xiaomi         0.0893      0.026      3.498      0.000       0.039       0.139
os_Others                -0.0704      0.030     -2.356      0.019      -0.129      -0.012
4g_yes                    0.0499      0.015      3.357      0.001       0.021       0.079
==============================================================================
Omnibus:                      232.847   Durbin-Watson:                   1.913
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              458.097
Skew:                          -0.628   Prob(JB):                    3.35e-100
Kurtosis:                       4.724   Cond. No.                     3.85e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.85e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

THE END :)