PREDICTIVE MODELLING

GUIDED

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ReCell PROBLEM STATEMENT

The market for used and refurbished devices has grown significantly over the past decade with predictive market value of $52.7 billion by 2023. ReCell is a startup which aims to leverage machine learning to develop a dynamic pricing strategy for used and refurbished devices. This report presents an analysis and the development of the linear regression model to predict the price of the used devices.

DATA DESCRIPTION

- Brand_name: Name of the manufacturing brand
- Os: Operating system
- Screen_size: size of the screen in cm
- 4g: Whether 4g is available
- 5g: Whether 5g is available
- Main_camera_mp: Resolution of the rear camera in megapixels
- Selfie_camera_mp: Resolution of the front camera in megapixels
- Int_memory: Internal memory in GB
- ram: RAM in GB
- Battery: Battery capacity in mAh
- Weight: Weight in grams
- Release_year: Year the device model was released
- Days_used: Number of days the 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

- There are 3454 rows and 15 columns in the dataset.
- There are 9 numerical variables and 4 objects variable in the dataset.
- There are no duplicate values in the dataset.
- The dataset contains complete information with no missing values.

EDA (Exploratory Data Analysis)

UNIVARIATE ANALYSIS

1. Distribution & Boxplot for 'normalized_used_price'

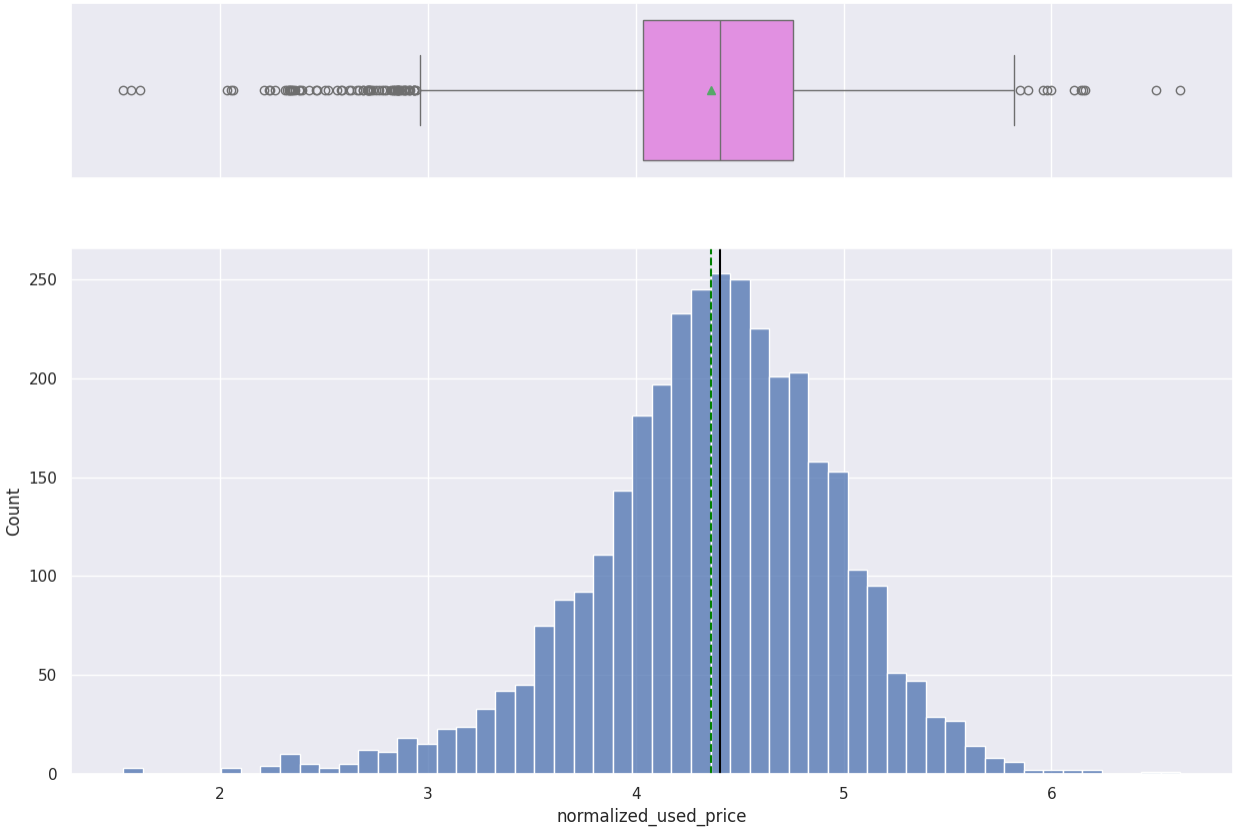

FIGURE 1

Insights:

- The histogram shows that the plot is normally distributed.

- The distribution is slightly right skewed. This suggests that the most of the used device prices are centered around mean with few higher outliers.
- The peak of the distribution is around the mean value suggesting a higher concentration of devices around this range.
- The boxplot shows the presence of the several outliers on both the lower and higher ends of the price range.
- The mean and the median are very close showing the distribution is symmetric though the mean is slightly higher than the median.
- The whiskers of the boxplot extend to the right, showing more variability in the higher used prices.
- The IQR is narrow, showing that most of the device price are close around the median.

2. Distribution & Boxplot for normalized_new_price:

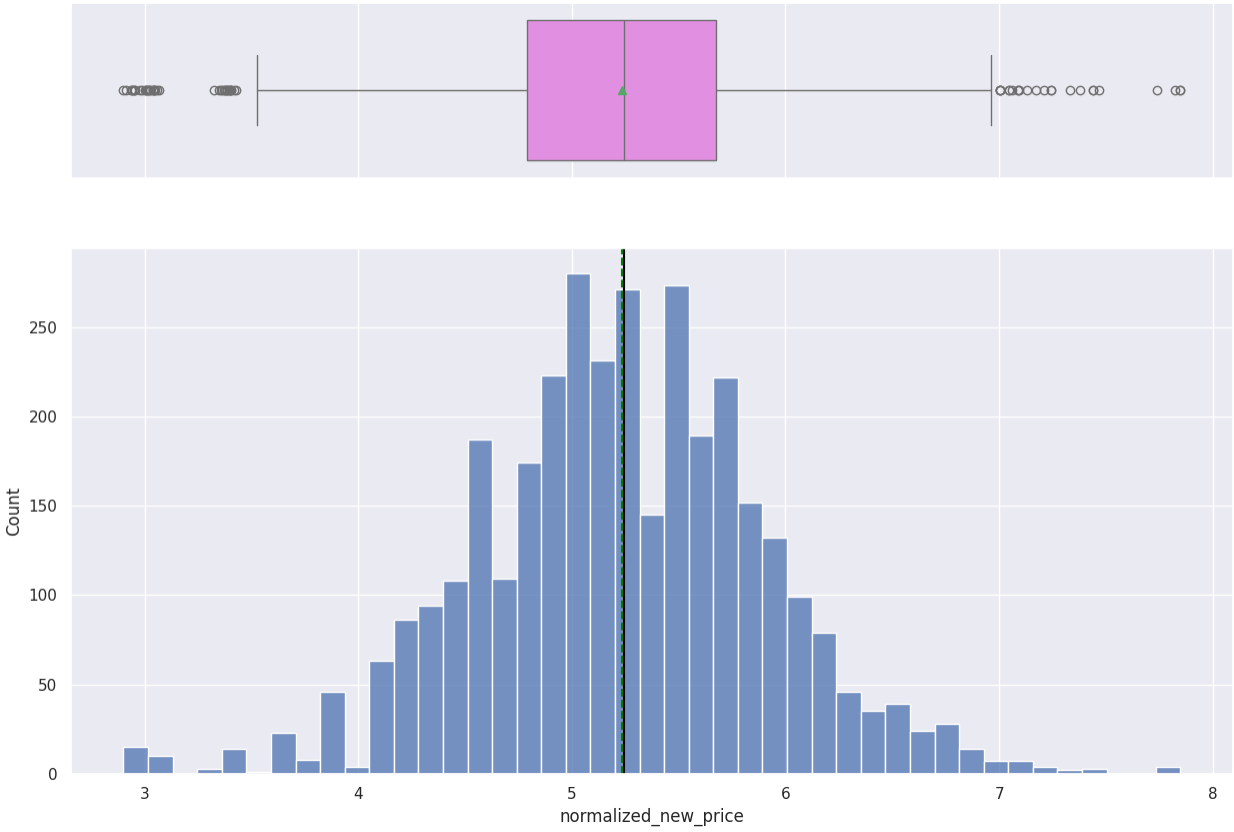

FIGURE 2

Insights:

- The histogram shows roughly a normal distribution with slight right skew.
- The peak of the distribution is around the mean value.

- The mean & median values are very close showing symmetric distribution. This indicates that most of the new devices are priced around the mean value with few expensive models.
- Outliers on the higher end represents high new end devices.
- The boxplot shows few outliers compared to the used prices, showing that the new device prices are more consistent.
- The IQR is wider compared to the used prices, showing broader spread of the new device prices.
- The whiskers of the boxplot are longer showing more variability in the pricing of the new devices.

3. Histogram & Boxplot for ‘screen_size’

IGURE

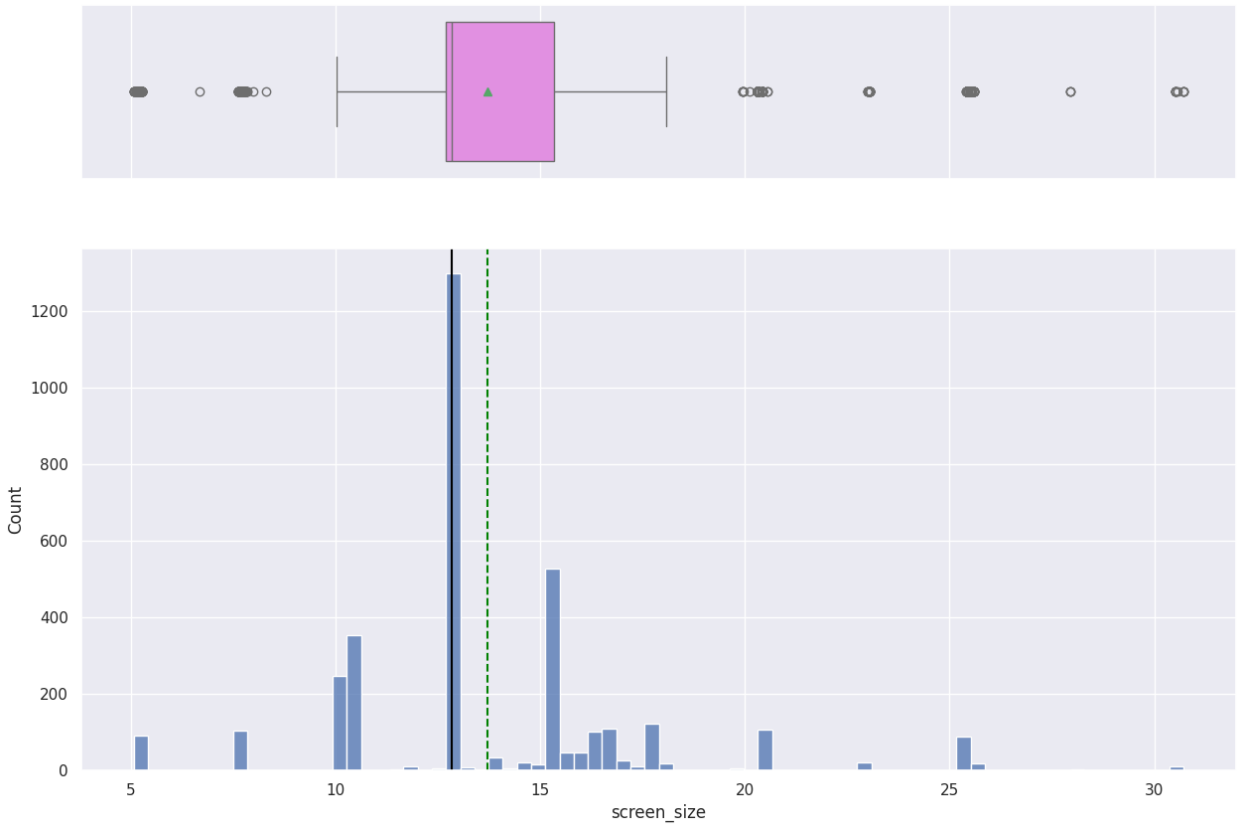

FIGURE 3

Insights:

- There are several outliers present on both the upper and lower ends. Devices that have screen size less than 8cm and greater than 17cm are considered to be outliers.
- The ‘screen_size’ variable shows high concentration with a sharp peak at 13cm.
- The median is approx. 13 cm as shown in the boxplot.
- The mean is slightly higher than the median indicating a slight right skew.

- The histogram indicates that the distribution is right skewed showing that there are more devices with the small screen sizes and fewer with larger screen size.

4. Boxplot & Histogram for ‘main_camera_mp’

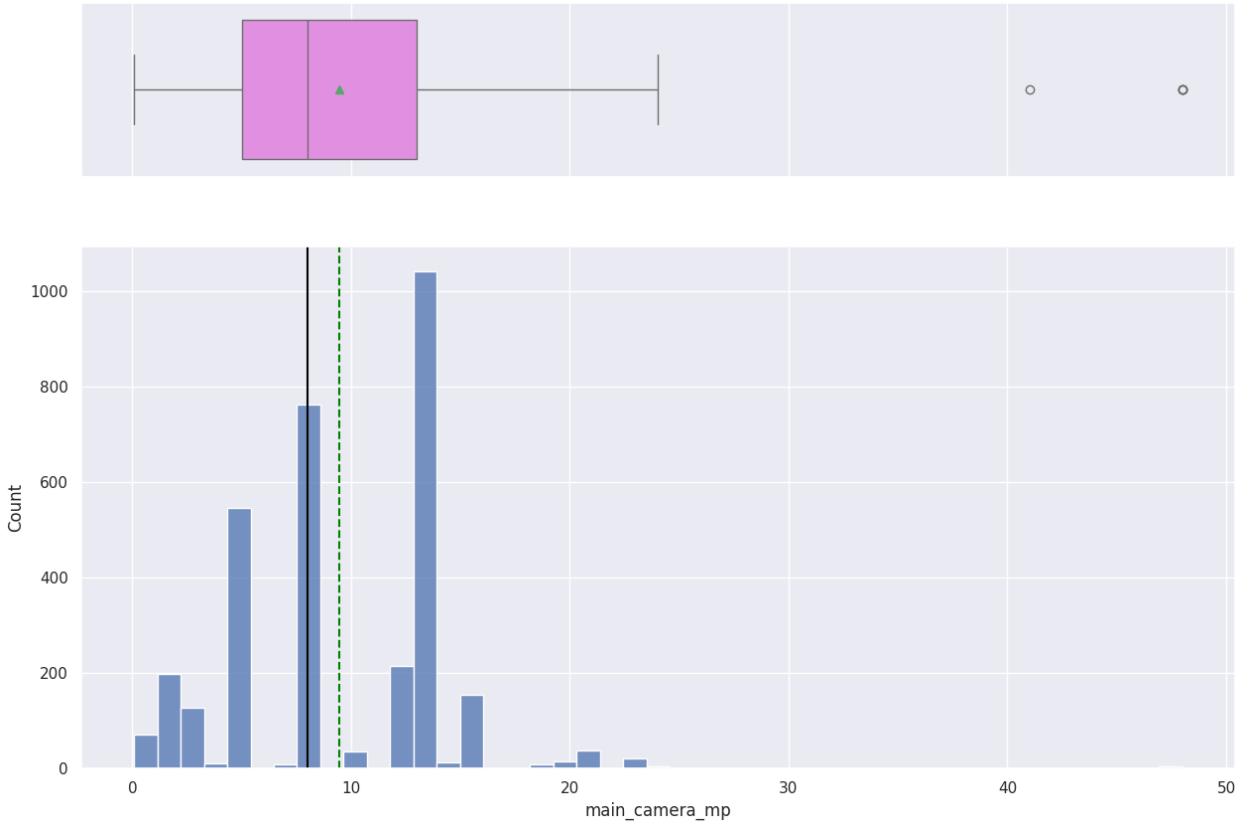

FIGURE 4

Insights:

- There are several outliers present mostly at the upper end, with some devices having camera resolution up to 50MP.
- The ‘main_camera_mp’ variable has high concentration around 10MP.

- The median is around 12MP.
- The mean is slightly higher than the median suggesting a right skew.
- The histogram shows right skewness with majority of devices having camera resolutions between 8MP - 12MP.

5. Boxplot & Histogram for ‘selfie_camera_mp’

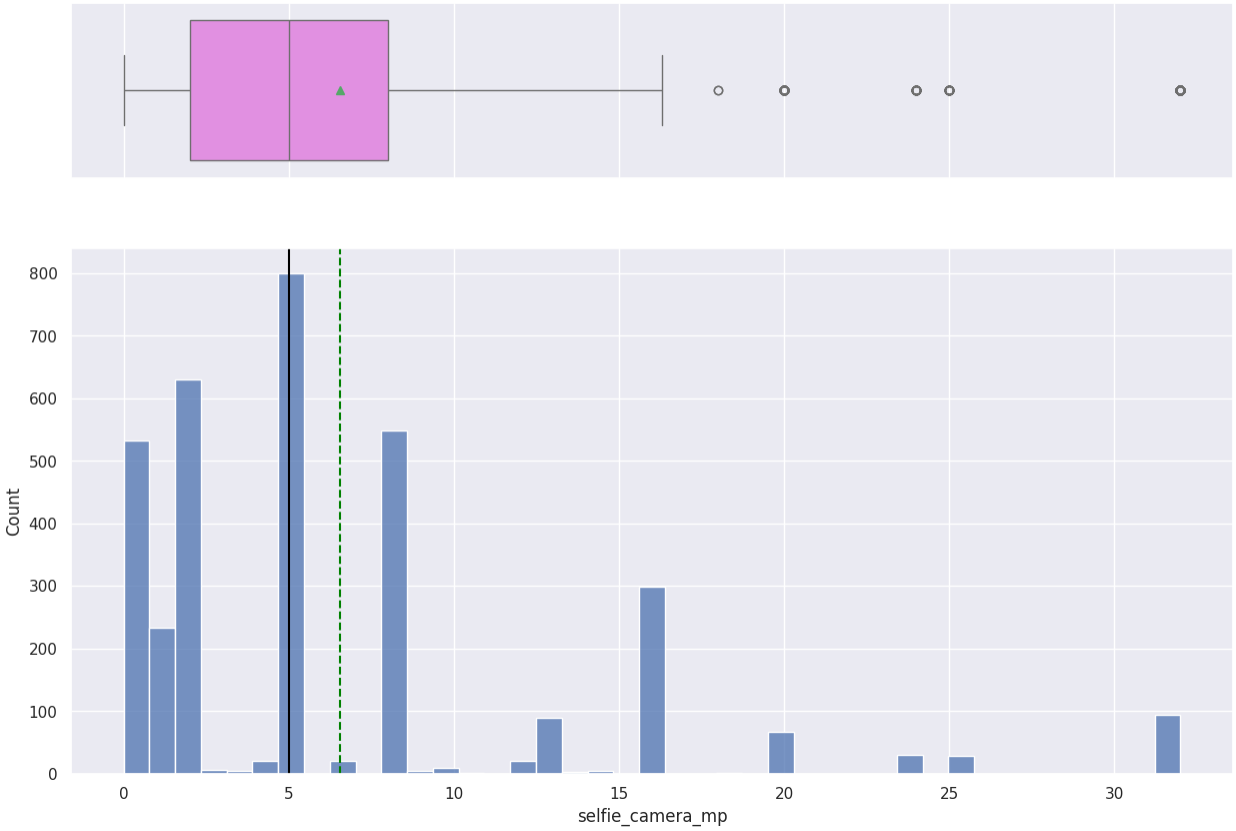

FIGURE 5

Insights:

- The median is around 5MP
- There are outliers present above 20MP

- From histogram it can be seen that most phones have selfie camera in the range of 0 to 10MP
- Few phones have selfie cameras greater than 20MP
- The mean is around 5MP similar to the median.

6. Boxplot & Histogram for ‘int_memory’

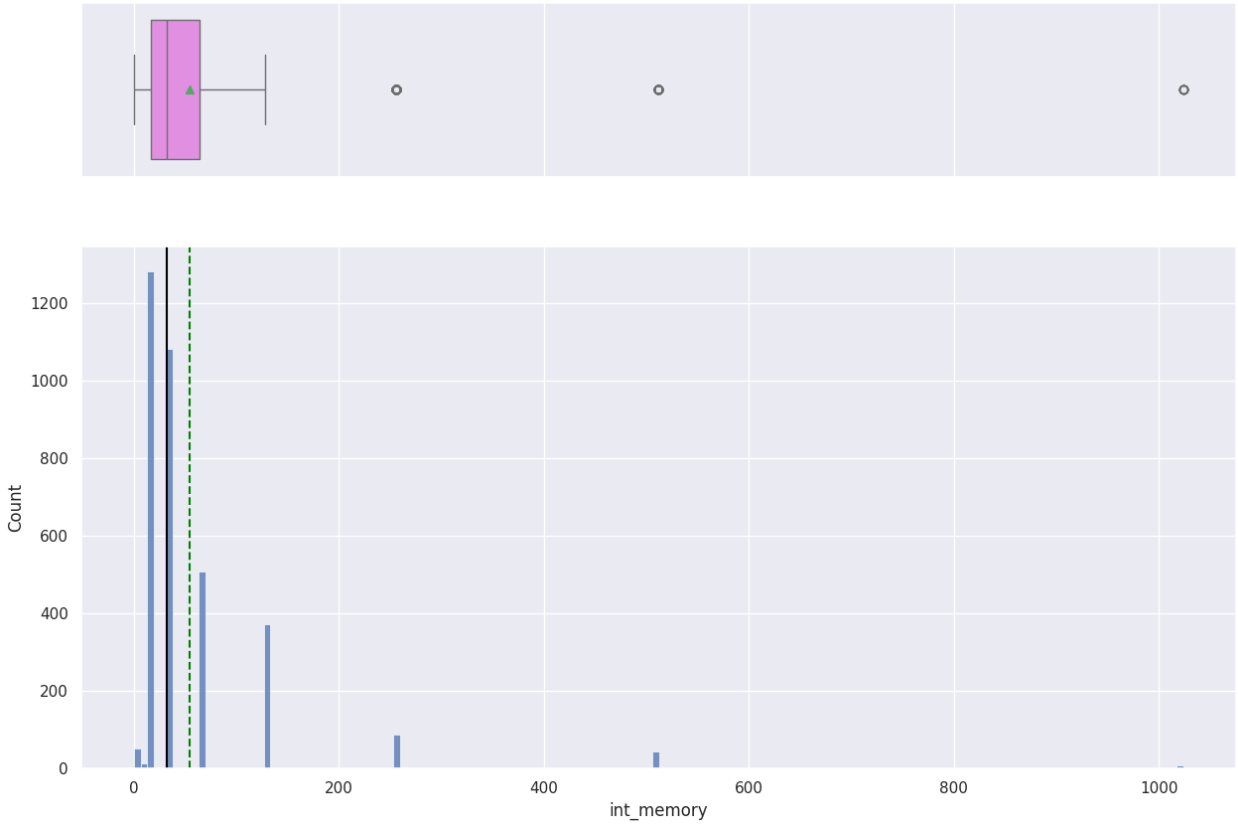

FIGURE 6

Insights:

- The median is around 16 to 32GB
- It can be seen in the boxplot that there are significant outliers above 128GB
- From histogram, it can be seen that most phones have internal memory between 0 – 64GB
- The mean is slightly higher than the median.
- Very few phones are there with internal memory above 256 GB

7. Boxplot & Histogram for 'ram'

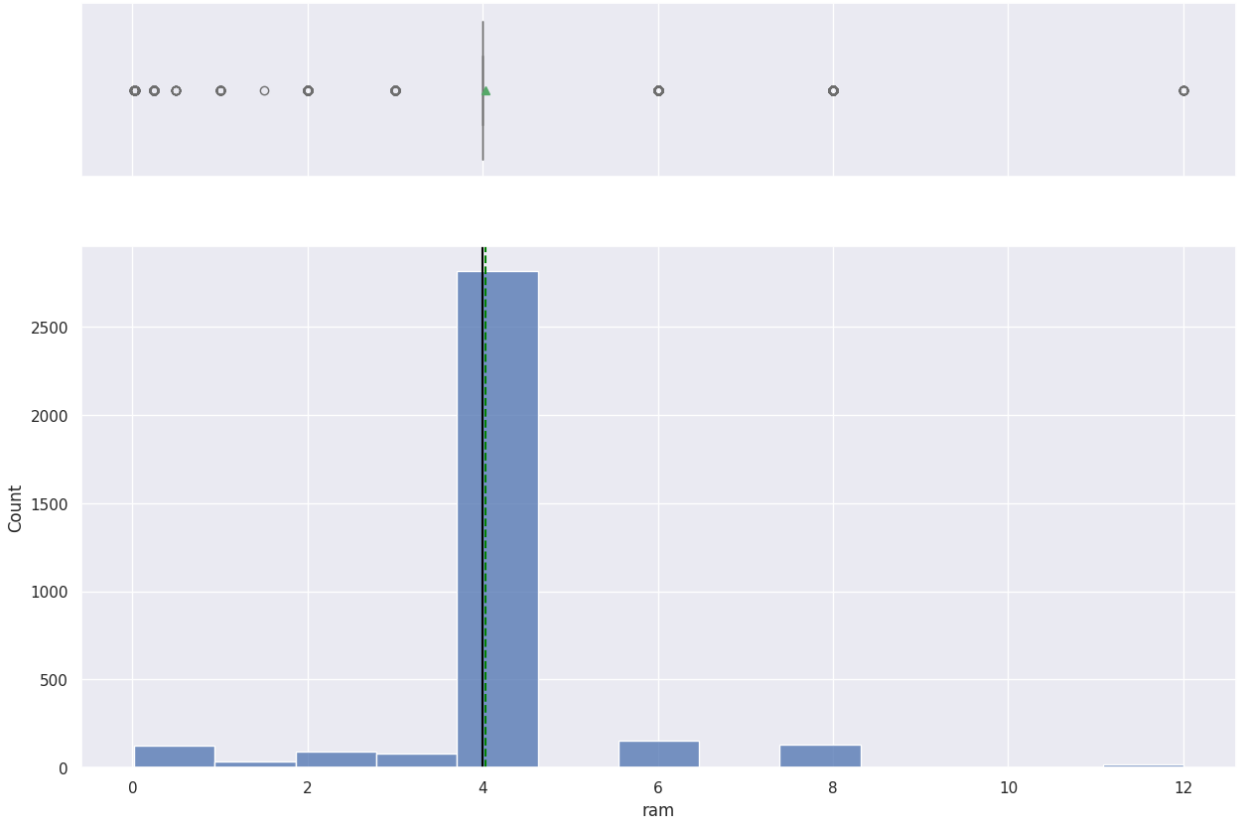

FIGURE 7

Insights:

- The median is around 4GB
- The outliers can be seen above 8GB
- From the histogram, we can see that few phones have RAM higher than 8GB
- Mean and median are close suggesting a symmetric distribution.

8. Boxplot & Histogram for ‘weight’:

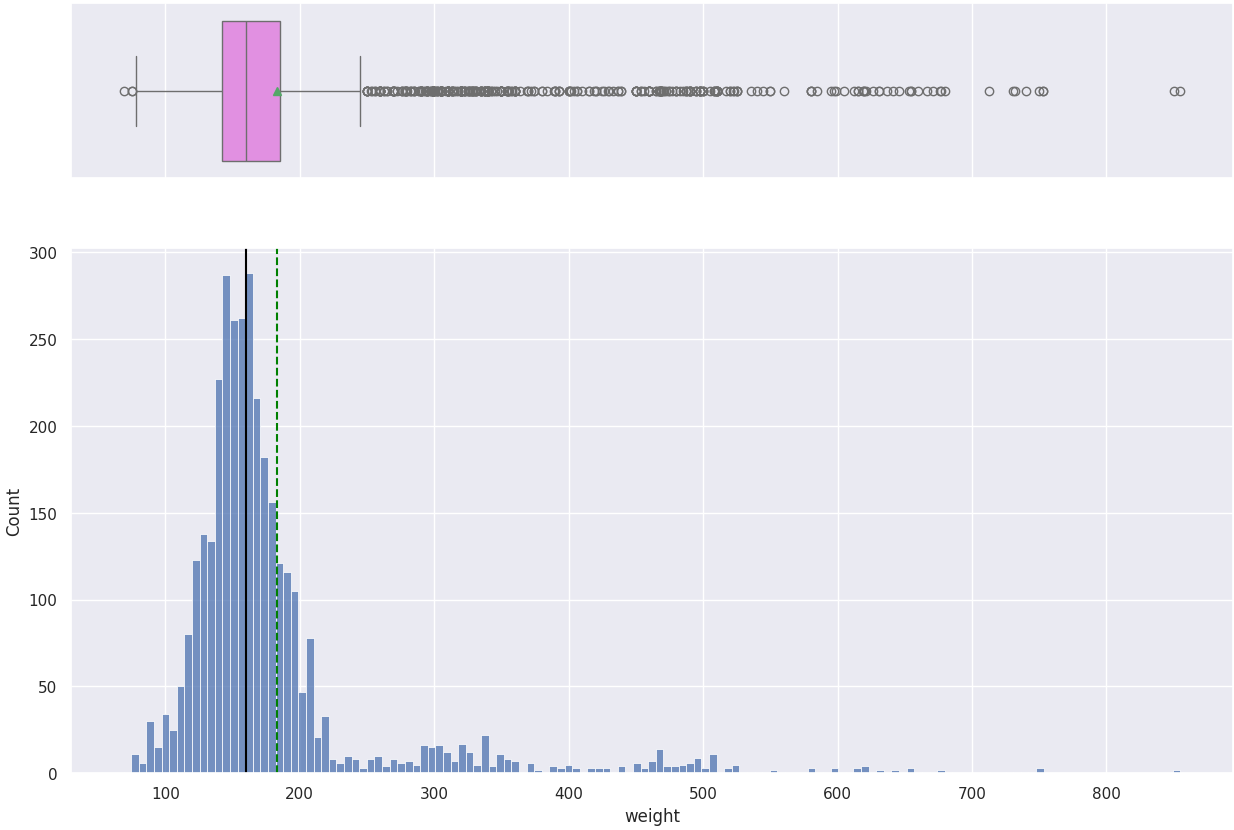

FIGURE 8

Insights:

- The IQR is narrow as compared to the range of data.
- There are many outliers present in the plot, meaning that there are significant number of instances with weights much higher than the median.
- The plot shows that the distribution of ‘weight’ variable is highly skewed to the right showing a long tail of higher values
- Histogram shows high frequency of lower weights with a peak around 100 – 150
- The mean is higher than the median showing positive skewness.
- Most of the data is below 200 with a long tail extending towards higher values.

9. Boxplot & Histogram for ‘battery’:

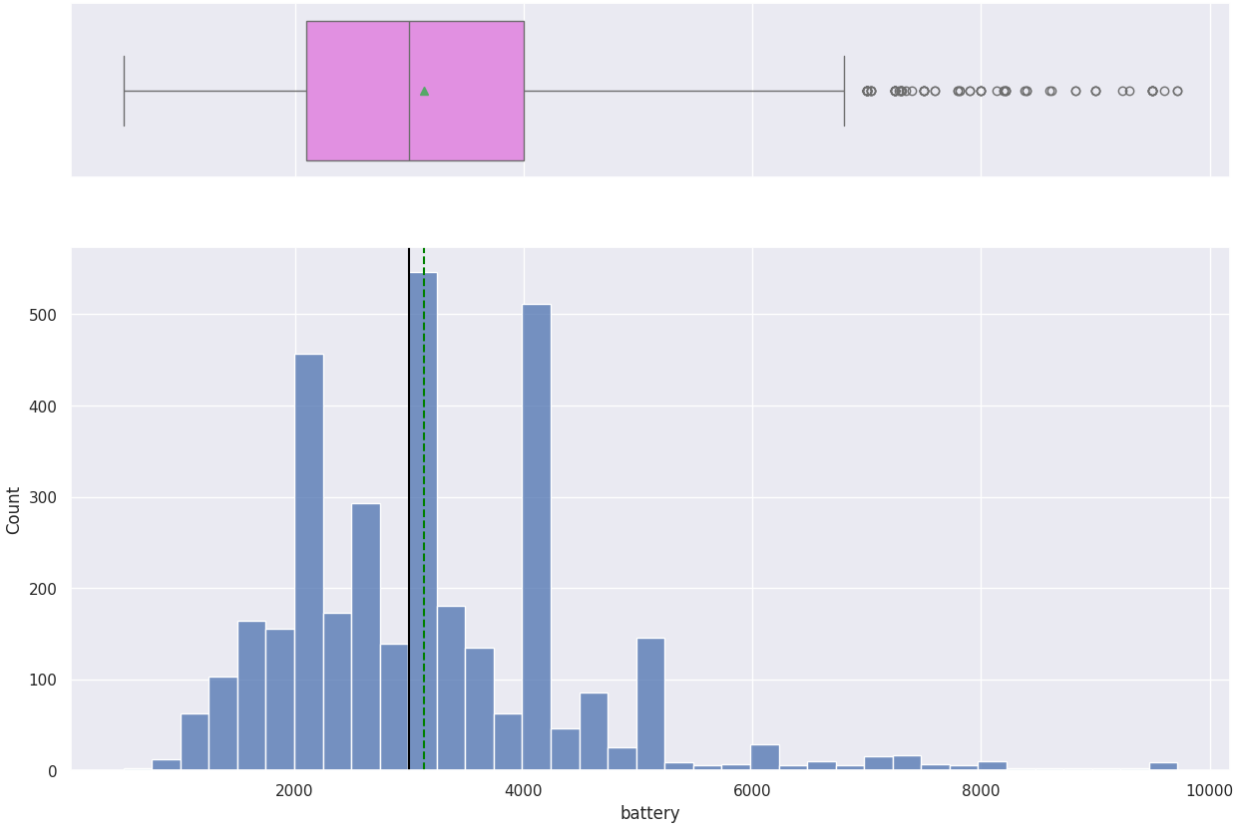

FIGURE 9

Insights:

- The IQR is wider showing more variability within 50% of the data.
- The 'battery' variable shows symmetric distribution with fewer outliers.
- Though there are some outliers on the higher end.
- Histogram shows two peaks indicating a bimodal distribution suggesting presence of two different groups or types of batteries.
- The first peak is around 2000, and second peak around 4000.

- The mean and median are very close to each other showing a relatively symmetric distribution.

10. Boxplot & Histogram for ‘days_used’:

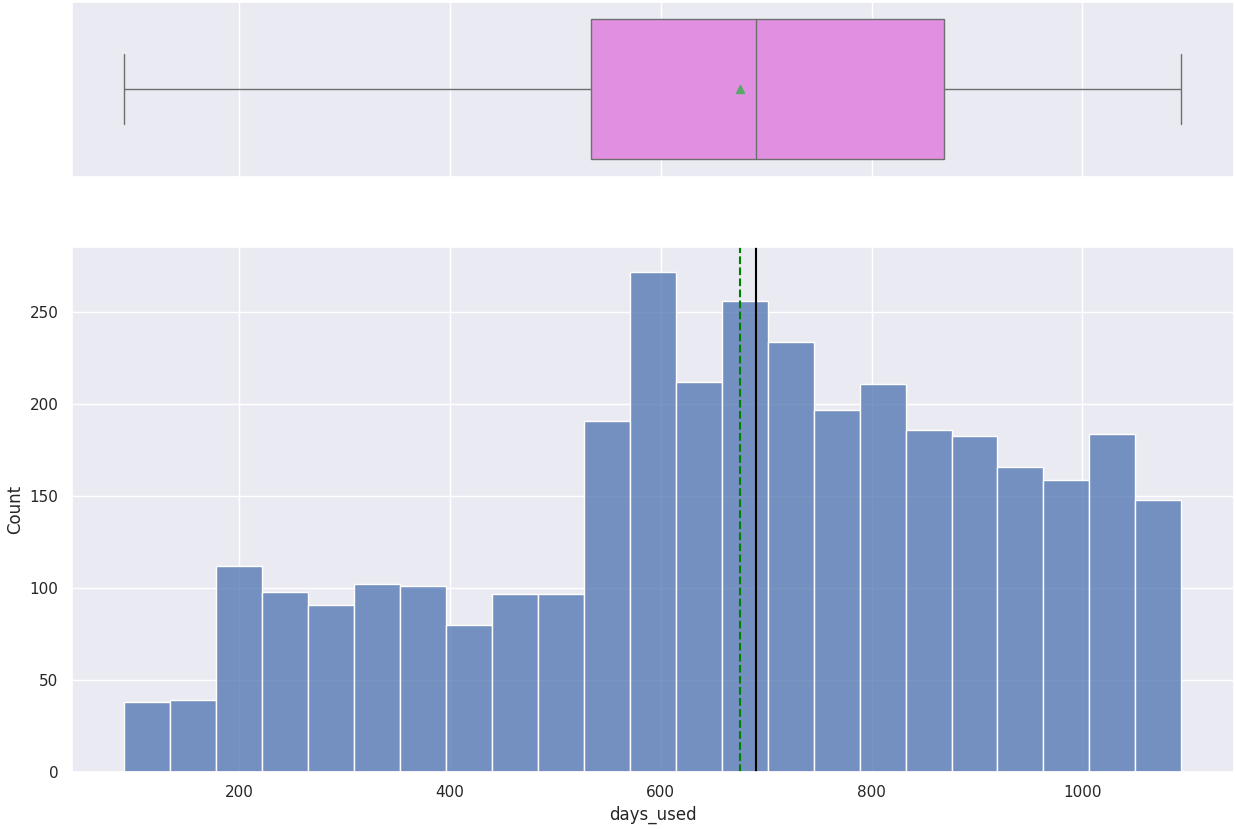

FIGURE 10

Insights:

- The distribution of ‘days_used’ is symmetric with a median close to the center.
- There are few outliers present in the plot.

- The IQR is wide suggesting a substantial range in number of days devices have been used.
- The histogram shows a uniform distribution with multiple peaks.
- The data is spread across a range of values.
- The mean is slightly higher than the median suggesting a positive skew.

BIVARIATE ANALYSIS

★ Heatmap of Correlation Matrix:

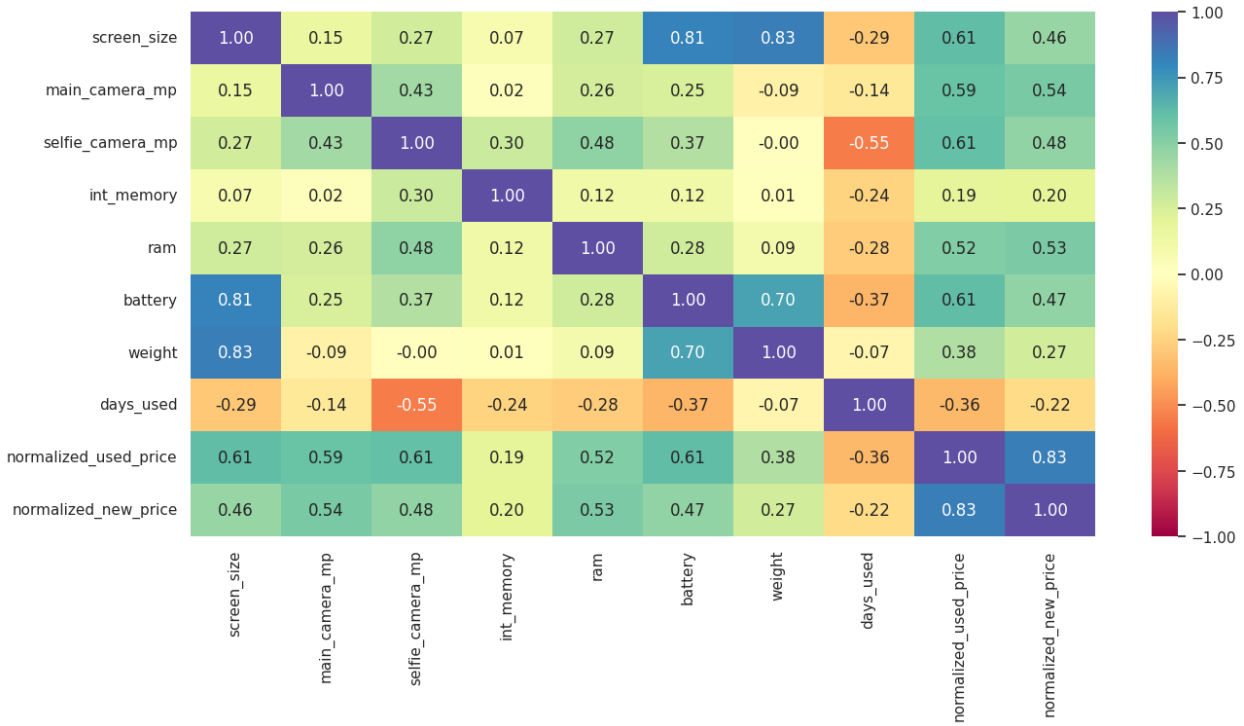

Figure 11

Insights:

- ‘normalized_new_price’ & ‘normalized_used_price’ have a high positive correlation. This shows that the price of used device is related to the price of new device.
- ‘ram’ & ‘int_memory’ shows a positive correlation showing that devices with more RAM also have higher internal memory.
- ‘battery’ shows moderate positive correlation with ‘weight’ suggesting that devices with large batteries are heavier.
- ‘main_camera_mp’ & ‘selfie_camera_mp’ have a moderate positive correlation suggesting that devices with better rear camera have better front cameras too.
- ‘days_used’ shows no correlation with other variables showing that the number of days of a device used does not strongly relate to other features.
- ‘screen_size’ has low correlation with other features showing that screen size varies independently.
- ‘days_used’ and ‘normalized_used_price’ have low to negative correlation suggesting that the longer a device is used the lower its resale price tends to be.

The amount of RAM is important for the smooth functioning of a device. Let's see how the amount of RAM varies across brands.

❖ **Boxplot of ‘brand_name’ vs ‘ram’:**

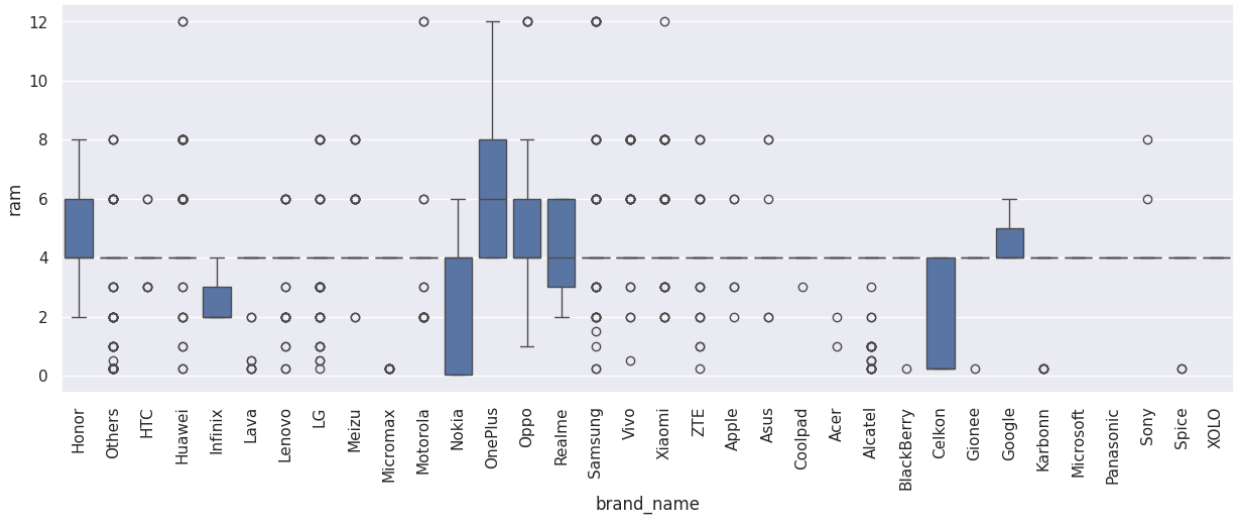

FIGURE 12

Insights:

- Some brands have wide range of RAM options while others have more uniform range.
- Devices ranging from low RAM capacities to some of the highest, Samsung shows the highest variability.
- Brands like Apple & Samsung shows high variability showing these brands offer devices with wide range of RAM capacities.
- Nokia & LG shows less variability showing limited range of RAM capacities.
- LG shows narrower range and lower median.
- There are noticeable outliers in several brands. Brands like Huawei and Sony have outliers suggesting few high-end devices have much higher RAM than their regular range.
- Apple and OnePlus have higher median RAM compared to Nokia and LG.

- This suggests that certain brands offer devices with higher RAM to target performance-oriented users.

People who travel frequently require devices with large batteries to run through the day. But large batteries often increase weight, making it feel uncomfortable in the hands.

After creating a new data frame of only those devices which offer a large battery and analyzing, here are some insights below:

❖ **Boxplot of 'weight' vs 'brand_name':**

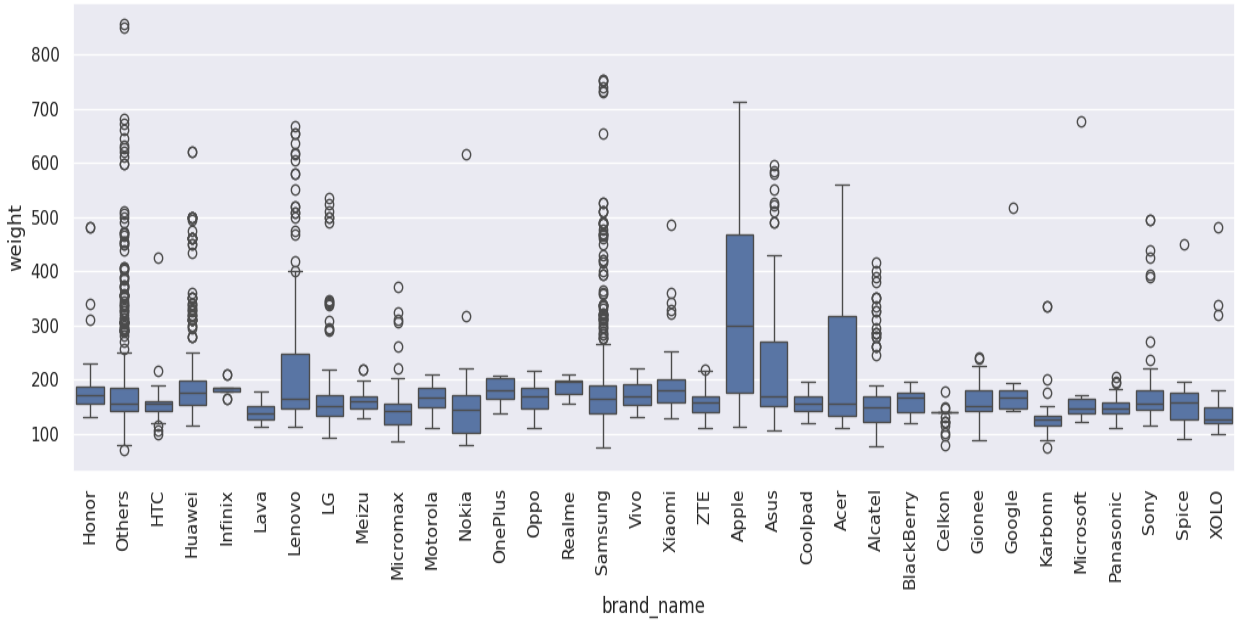

FIGURE 13

There are 341 rows containing batteries greater than 4500mAh

Insights:

- Apple, Acer have the highest median weights suggesting that their products are heavier as compared to other brands.
- Brands like Motorola, Micromax, Nokia, OnePlus, Oppo, Samsung, Vivo and Xiaomi have similar median weights.
- Microsoft, Google, Panasonic shows smaller IQR indicating more consistent weight.
- Coolpad, Apple, Lava, shows variability in weights showing large IQR and several outliers.
- Brands like Lava, Samsung, Huawei, Apple and others have several outliers present suggesting wide range of product weights.
- Whereas brands such as Sony, Microsoft, Google and XOLO are showing fewer outliers and smaller IQR showing higher consistency in product weights.

People who buy phones and tablets primarily for entertainment purposes prefer a large screen as they offer a better viewing experience.

After creating a new data frame of only those devices which are suitable for such people and analyzing, below are the insights:

- There are 1099 rows which contains devices which have screen size greater than 6 inches.
- This shows that there are quite a lot of people who prefer a large screen for better viewing experience.

Everyone likes a good camera to capture their favorite moments with their loved ones. Some customers specifically look for good front cameras to click cool selfies. Let's create a new data frame of only those devices which are suitable of such people and analyzed.

- There are 655 rows out of total, suggesting people who specifically look a good front camera of 8MP to click cool pictures.

After doing the similar analysis for the rear cameras, setting the threshold higher at 16MP, we conclude that:

- 94 columns are there out of the complete data who prefer the main camera to be greater than 16MP.

❖ **Line plot of 'days_used' vs 'release_year':**

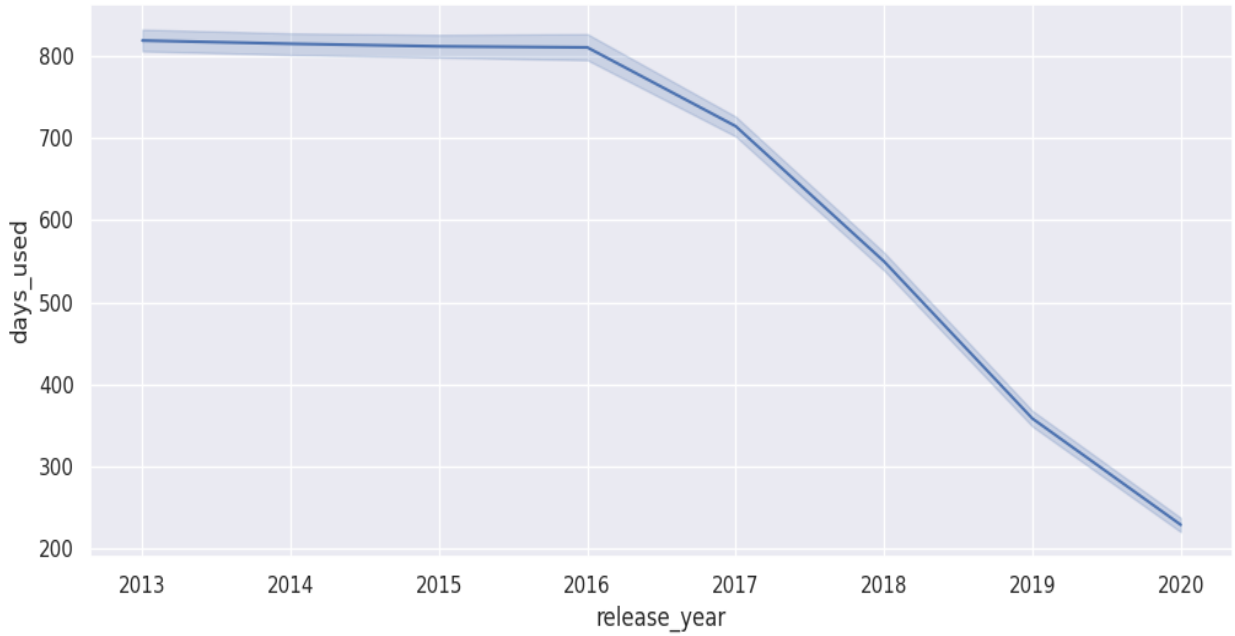

FIGURE 14

Insights:

- Products that were released in 2013 were used for around 800 days on an average, whereas those released in year 2020 were used for approximately 200 days only.
- The plot shows that there is a decline in number of days used as the release year progresses from 2013 to 2020.
- This shows that the newer products have a shorter lifespan/ usage duration due to faster technology advancement or change in consumer behavior.

❖ **Boxplot of 'normalized_used_price' vs '4G & 5G availability':**

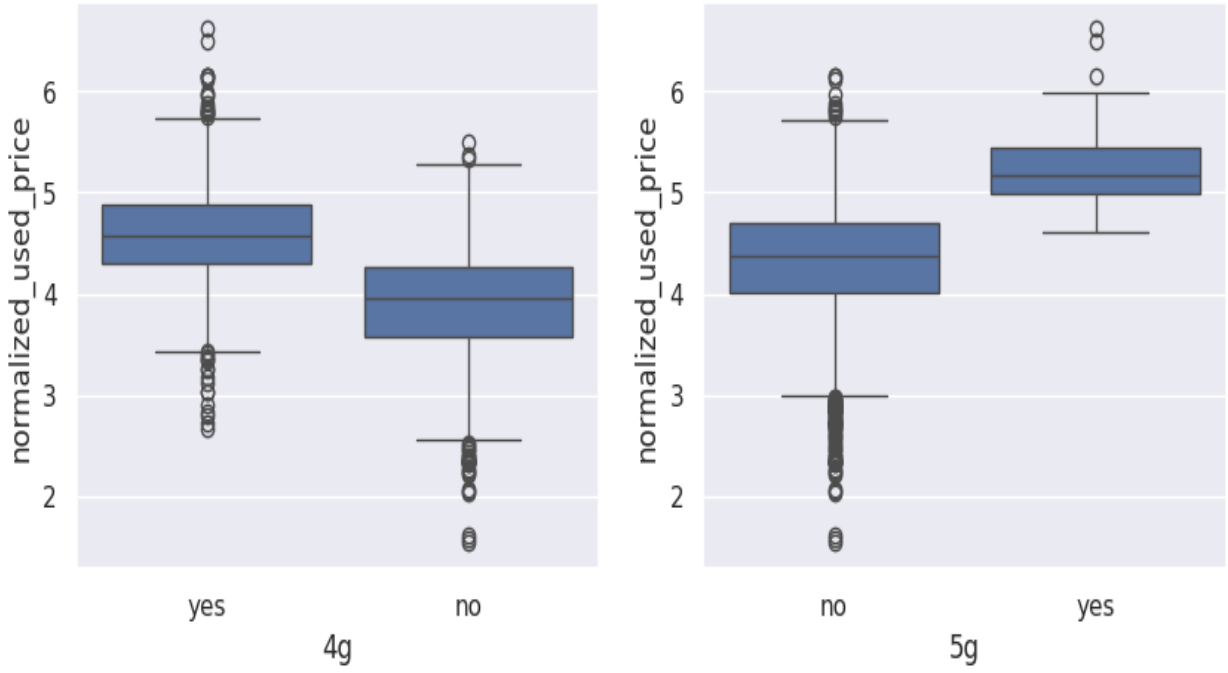

FIGURE 15

Insights:

- Products with 4g availability have higher normalized used price as compared to those without 4g.
- There are quite high value outliers in the 4g category.
- The median of ‘normalized used price’ is higher for 4g enabled products.
- The median price is higher for 5g enabled products.
- The IQR of 5g is also upward showing high value placed on these products in the used market.
- Both 4g and 5g availability have positive impact on the normalized used price.

- The outliers present in the ‘Yes’ categories for both 4g and 5g shows that some of the high-end models gain more value in the used market.

DATA PREPROCESSING

Missing value imputation

We will impute the missing values in the data by column medians grouped by ‘release_year’ & ‘brand_name’

Firstly, we will create a copy of the original data set and then check for missing values in all the columns.

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

TABLE 1

After filling the missing values in the ‘main_camera_mp’ column by column median.

brand_name	0
os	0
screen_size	0
4g	0
5g	0
main_camera_mp	0
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

TABLE 2

FEATURE ENGINEERING

After creating a new column ‘years_since_release’ from the release column

We will drop the ‘release_year’ column and get the mean, min, std, 25%, 50%,75% value.

count	3454.000000
mean	5.034742

std	2.298455
min	1.000000
25%	3.000000
50%	5.500000
75%	7.000000
max	8.000000

TABLE 3

- There is 3454 data points in the data set
- The data is moderately centered around the mean of 5.3
- The standard deviation is about 2.30
- The smallest value in the dataset is 1
- 25% of the data is less than or equal to 3
- The median value is around 5.5, suggesting that the data is slightly skewed to the left since the mean is slightly lower than the median.
- 75% of the data is less than or equal to 7, the percentiles indicate that most of the data points lie between 2 and 7.
- The largest value in the dataset is 8

OUTLIER CHECK:

❖ **Boxplot of Predictors:**

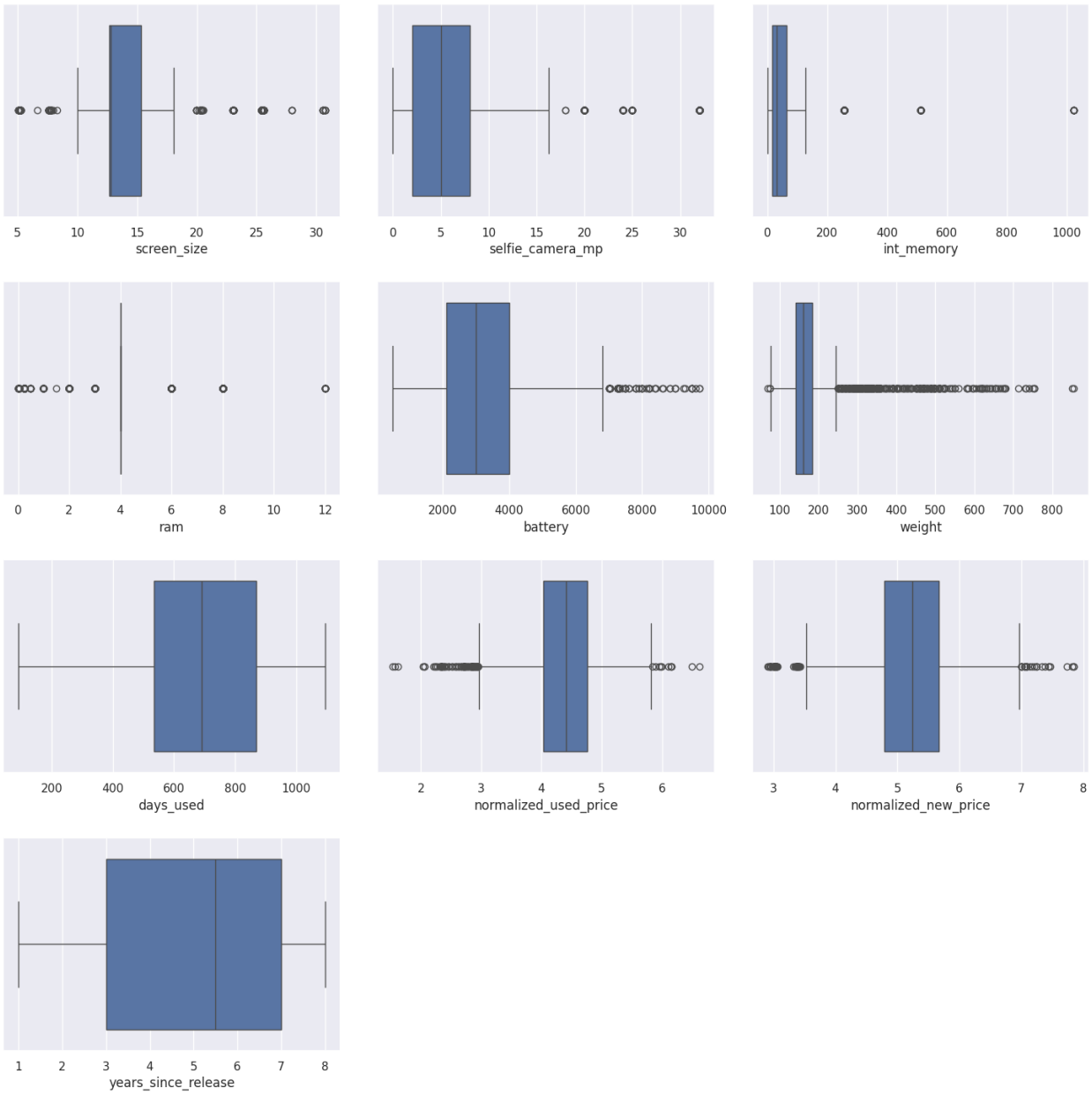

FIGURE 16

Insights:

- **Screen size:**

1. The ‘screen size’ is mostly centered around the mean with few outliers on the higher end.

2. There are some extreme values showing that some devices have larger screen size.

- **Selfie camera MP:**

1. The distribution is tight around the median with some higher values.

2. There are some outliers present, indicating higher megapixel counts.

- **Internal Memory:**

1. The distribution is skewed with many outliers showing higher internal memory.

2. There are notable outliers suggesting higher internal memory.

- **RAM:**

1. The distribution is tight around the median with few outliers.

2. The outliers on the higher end signifies devices with more RAM.

- **Battery:**

1. The distribution is centered around the median with a wide range.

2. There are numerous outliers present on the higher end showing devices with larger battery capacities.

- **Weight:**

1. The distribution has tight central mass with many outliers.
2. There are numerous outliers on the higher end suggesting heavier devices.

- **Days Used:**

1. The distribution is uniform across the range with some outliers.
2. There are some outliers present showing that the devices that has been used notably longer or shorter duration.

- **Normalized Used Price:**

1. The distribution is tight around the median.
2. There are few outliers present showing variations in the ‘Normalized used price’.

- **Normalized New Price:**

1. The distribution is centered around the median.
2. There are few outliers present showing variations in the ‘Normalized new price’.

- **Years Since Release:**

1. The distribution is uniform across the range.
2. There are no major outliers present, showing consistent years since release across the devices.

DATA PREPARATION FOR MODELLING:

We will split the data into train and test to be able to evaluate the model that we build on the train data.

After that we will build a linear regression model using the train data and check its performance.

After splitting the data in 70:30 ratio for train to test data,

Number of rows in train data – 2417

Number of rows in test data – 1037

MODEL BUILDING – LINEAR REGRESSION:

MODEL BUILDING – LINEAR REGRESSION

Statistic	Value
Dependent Variable	normalized_used_price
Model	OLS
R-squared	0.849
Adjusted R-squared	0.846
F-statistic	276.8
Prob (F-statistic)	0.00
Log-Likelihood	123.88
No. Observations	2417
AIC	-149.8
BIC	134.0
Df Residuals	2368
Df Model	48
Covariance Type	nonrobust
Coefficients	
const	-51.6529 (9.213)
screen_size	0.0292 (0.004)
main_camera_mp	0.0228 (0.002)
selfie_camera_mp	0.0117 (0.001)
int_memory	0.0002 (6.77e-05)
ram	0.0313 (0.005)
battery	-1.543e-05 (7.34e-06)
weight	0.0008 (0.000)
release_year	0.0263 (0.005)
days_used	2.753e-05 (3.07e-05)
normalized_new_price	0.4105 (0.012)
brand_name_Alcatel	-0.0804 (0.050)
brand_name_Apple	-0.0495 (0.148)
brand_name_Asus	0.0100 (0.049)
brand_name_BlackBerry	0.0258 (0.072)
brand_name_Celkon	-0.2345 (0.068)
brand_name_Coolpad	-0.0218 (0.071)
brand_name_Gionee	-0.0129 (0.059)
brand_name_Google	-0.1206 (0.083)
brand_name_HTC	-0.0393 (0.050)
brand_name_Honor	-0.0537 (0.051)
brand_name_Huawei	-0.0640 (0.046)
brand_name_Infinix	0.0933 (0.113)
brand_name_Karbonn	-0.0573 (0.068)

brand_name_LG	-0.0629 (0.047)
brand_name_Lava	-0.0235 (0.063)
brand_name_Lenovo	-0.0370 (0.047)
brand_name_Meizu	-0.0682 (0.056)
brand_name_Micromax	-0.0643 (0.049)
brand_name_Microsoft	0.0717 (0.082)
brand_name_Motorola	-0.0567 (0.051)
brand_name_Nokia	0.0314 (0.052)
brand_name_OnePlus	0.0126 (0.073)
brand_name_Oppo	-0.0243 (0.049)
brand_name_Others	-0.0684 (0.044)
brand_name_Panasonic	-0.0432 (0.062)
brand_name_Realme	0.0243 (0.063)
brand_name_Samsung	-0.0633 (0.045)
brand_name_Sony	-0.0667 (0.053)
brand_name_Spice	-0.0348 (0.068)
brand_name_Vivo	-0.0585 (0.050)
brand_name_XOLO	-0.0835 (0.057)
brand_name_Xiaomi	0.0381 (0.050)
brand_name_ZTE	-0.0437 (0.048)
os_Others	-0.0621 (0.033)
os_Windows	-0.0320 (0.043)
os_iOS	-0.0085 (0.148)
4g_yes	0.0392 (0.016)
5g_yes	-0.0682 (0.032)

TABLE 4

- CHECKING MODEL PERFORMANCE ON THE TRAIN SET (70% data)

RMSE	MAE	R-Squared	Adj. R-Squared	MAPE
0.229882	0.178347	0.845596	0.845596	4.289704

TABLE 5

♦ CHECKING MODEL PERFORMANCE ON THE TRAIN SET (30% data)

RMSE	MAE	R-Squared	Adj. R-Squared	MAPE
0.229882	0.178347	0.845596	0.845596	4.289704

TABLE 6

★ CHECKING LINEAR REGRESSION ASSUMPTIONS:

1. NO MULTI COLLINEARITY

Calculating VIF for each feature:

	Feature	VIF
0	A	inf
1	B	inf
2	C	inf

TABLE 7

To remove multi collinearity:

	Col.	R- squared after dropping column	RMSE after dropping column
0	RAM	0.842401	0.233942
1	Scrren_size	0.841297	0.235508

TABLE 8

VIF after dropping the constant:

0	const	3.803475e+06
1	screen_size	8.257756e+00
2	main_camera_mp	2.315276e+00
3	selfie_camera_mp	2.867771e+00
4	int_memory	1.362418e+00
5	ram	2.249443e+00
6	battery	4.055476e+00
7	weight	6.401901e+00
8	release_year	4.865021e+00
9	days_used	2.588858e+00
10	normalized_new_price	3.229074e+00
11	brand_name_Alcatel	3.458630e+00
12	brand_name_Apple	1.119205e+01
13	brand_name_Asus	3.651826e+00
14	brand_name_BlackBerry	1.623496e+00
15	brand_name_Celkon	1.873162e+00
16	brand_name_Coolpad	1.574731e+00
17	brand_name_Gionee	2.076759e+00

18	brand_name_Google	1.388277e+00
19	brand_name_HTC	3.460206e+00
20	brand_name_Honor	3.564128e+00
21	brand_name_Huawei	6.397324e+00
22	brand_name_Infinix	1.191453e+00
23	brand_name_Karbonn	1.628548e+00
24	brand_name_LG	5.353728e+00
25	brand_name_Lava	1.826101e+00
26	brand_name_Lenovo	4.705647e+00
27	brand_name_Meizu	2.411039e+00
28	brand_name_Micromax	3.779588e+00
29	brand_name_Microsoft	2.092823e+00
30	brand_name_Motorola	3.480051e+00
31	brand_name_Nokia	3.745567e+00
32	brand_name_OnePlus	1.586089e+00
33	brand_name_Oppo	4.286517e+00
34	brand_name_Others	1.083418e+01
35	brand_name_Panasonic	1.890929e+00
36	brand_name_Realme	1.966203e+00
37	brand_name_Samsung	8.013236e+00
38	brand_name_Sony	2.883271e+00
39	brand_name_Spice	1.638448e+00
40	brand_name_Vivo	3.732548e+00
41	brand_name_XOLO	2.160226e+00
42	brand_name_Xiaomi	4.079134e+00
43	brand_name_ZTE	4.342056e+00
44	os_Others	1.878983e+00
45	os_Windows	1.740535e+00
46	os_iOS	1.003527e+01
47	4g_yes	2.553476e+00
48	5g_yes	1.826907e+00

TABLE 9

- Build a model, check the p-values of the variables, and drop the column with the highest p-value.
- Create a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value.
- Repeat the above two steps till there are no columns with p-value > 0.05.

Now we will check the model performance on the train set to see the difference.

RMSE	MAE	R-Squared	Adj. R-Squared	MAPE
0.740676	0.554162	-0.570391	-0.571692	12.730602

TABLE 10

Now we will check the model performance on the test set to see the difference.

RMSE	MAE	R-Squared	Adj. R-Squared	MAPE
0.740676	0.554162	-0.570391	-0.571692	12.730602

TABLE 11

2. TEST FOR LINEARITY & INDEPENDENCE:

We will test for linearity and independence by making a plot of fitted values vs residuals and checking for the patterns.

	ACTUAL VALUES	FITTED VALUES	RESIDUALS
1744	4.261975	3.555571	0.706405
3141	4.175156	3.545977	0.629180
1233	4.117410	4.033358	0.084052
3046	3.782597	3.536383	0.246215
2649	3.981922	3.837638	0.144284

TABLE 12

- **Plot of 'Fitted values' vs 'Residuals':**

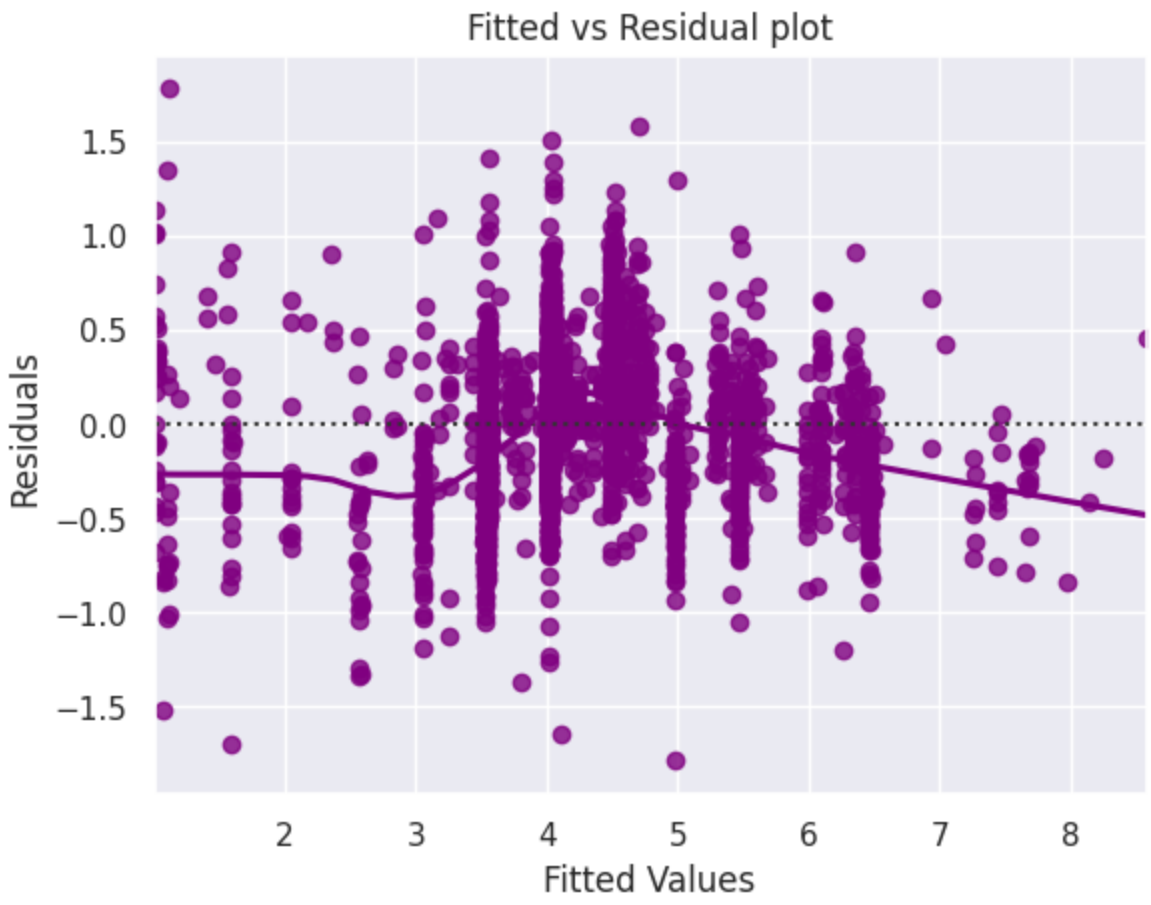

FIGURE 17

Insights:

- The plot shows that the residuals are randomly scattered around the y – axis. This shows that the model captures the underlying relationship in the data well with no clear patterns left in the residuals.
- The lowness smoothing line is flat around zero indicating that the model does not have biased and has captured the main structure of data.

- The slight deviation at the higher fitted values shows that there are some areas where the model's prediction is less accurate.
- There are outliers present in the residuals, most points lying close to zero line, showing that there are no data points with intense influence on the model.
- The 'residuals' spread remains constant across different fitted values, indicating there is no significant heteroscedasticity. This means that the variance of the residuals is same across all levels of fitted values.

  However slight deviation at the higher fitted values might be worth investigating further to ensure the models accuracy across all ranges of data.

3. TEST FOR NORMALITY

We will test for normality by checking the distribution of residuals, by checking the Q-Q plot of residuals, and by using the Shapiro-Wilk test.

❖ **Distribution of the Residuals Plot:**

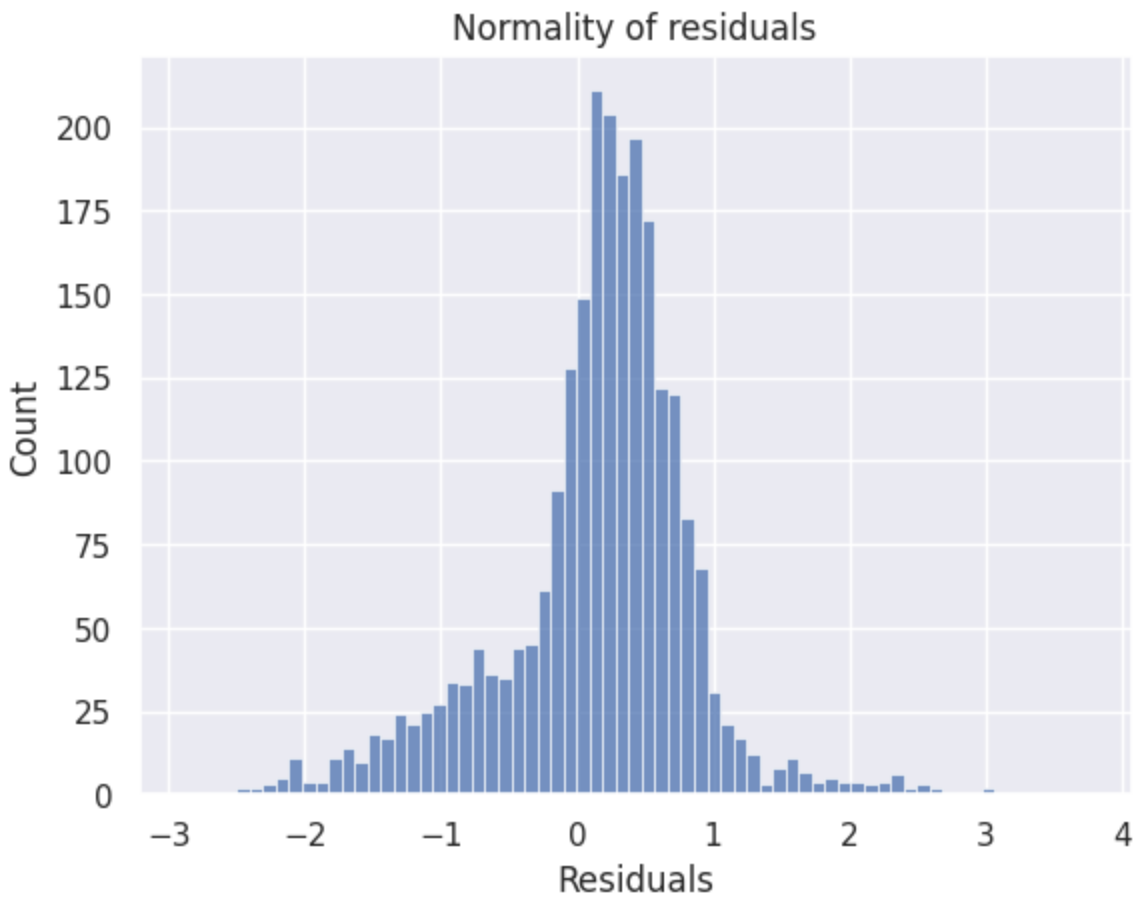

FIGURE 18

❖ **Probability Plot (Q-Q Plot):**

If the p-value of the Shapiro-Wilk test is greater than 0.05, we can say the residuals are normally distributed.

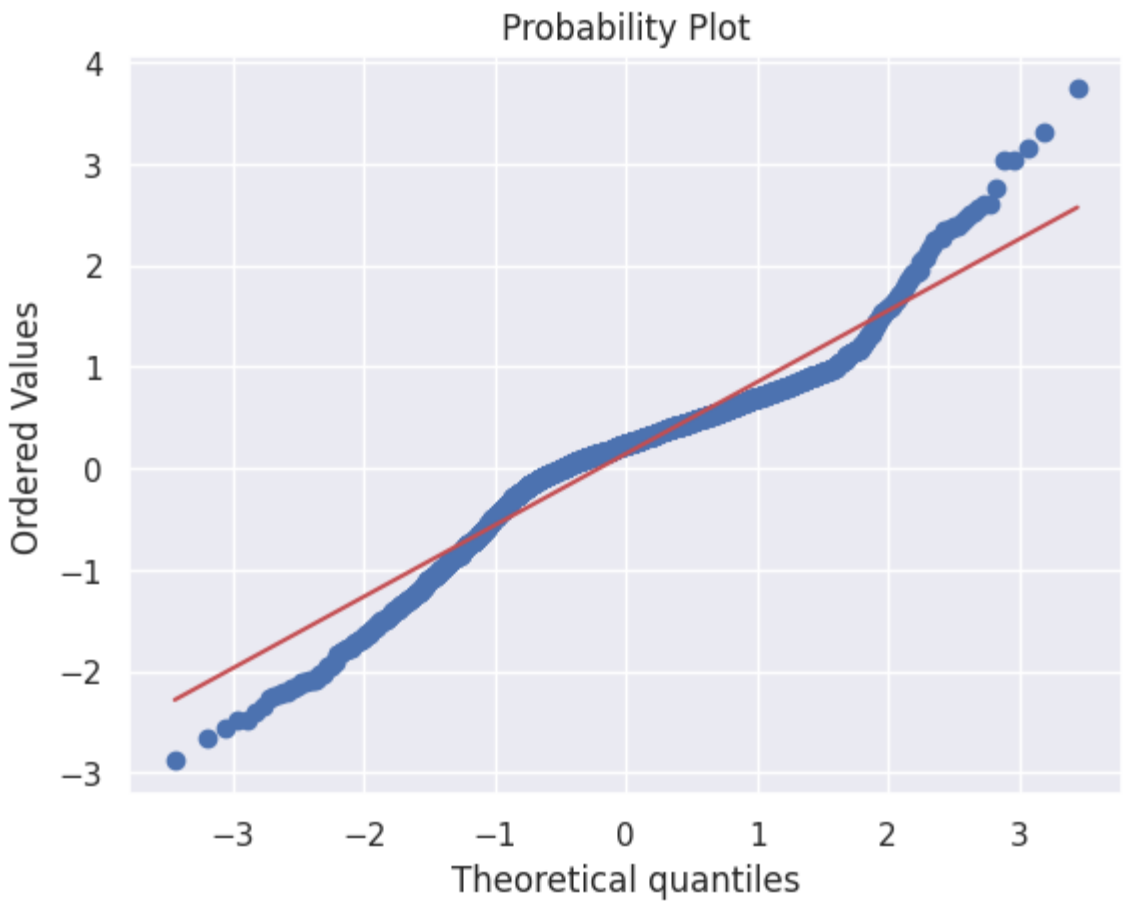

FIGURE 19

The histogram of residuals and the Q-Q Plot provides insights into the normality of the residuals.

Insights on Histogram of Residuals:

- The high frequency of the residuals is centered around zero which is expected in a well fitted model.
- The histogram appears to be symmetric around zero indicating the residuals are normally distributed.

- The tails of the distribution extend equally in both the directions from the mean, supporting the normality assumptions.

Insights on Q – Q Plot:

- There are few points that deviate from the line at the extremes, indicating presence of outliers.
- The residuals align with 45 degrees line indicating a normal distribution. However, there are slight deviation at the tails showing some deviation from the normality.

After performing the Shapiro – Wilks test, we get p - value of

P-value= 3.664893092417428e-30

Since the p- value is less than 0.05, we can conclude that the residuals are not normally distributed.

4. TEST FOR HOMOSCEDASTICITY

We will test for homoscedasticity by using the Goldfredquandt Test.

If we get the p- value > 0.05, we can say the residuals are homoscedastic otherwise they are heteroscedastic.

After performing the Goldfredquandt test we get,

the p- value as **0.99999**

Since the p- value is greater than 0.05 we can conclude that the residuals are homoscedastic.

FINAL MODEL SUMMARY:

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared (uncentered):	0.972
Model:	OLS	Adj. R-squared (uncentered):	0.972
Method:	Least Squares	F-statistic:	4.143e+04
Date:	Tue, 25 Jun 2024	Prob (F-statistic):	0.00
Time:	06:18:33	Log-Likelihood:	-2704.0
No. Observations:	2417	AIC:	5412.
Df Residuals:	2415	BIC:	5424.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
screen_size	0.1919	0.003	63.253	0.000	0.186	0.198
ram	0.3929	0.010	38.724	0.000	0.373	0.413

Omnibus:	200.715	Durbin-Watson:	1.916
Prob(Omnibus):	0.000	Jarque-Bera (JB):	587.697
Skew:	-0.431	Prob(JB):	2.42e-128
Kurtosis:	5.256	Cond. No.	10.3

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

TABLE 13

Final OLS Regression Analysis:

The OLS model was fitted to predict the ‘normalizes used price’ using predictors ‘screen size’ & ‘RAM’

- Model fit:

R- Squared – 0.972

Adj. R- Squared – 0.972

→These values suggests that the model explains approximately 97.2% of the variance in the 'normalizes used price' showing a very strong fit.

- F- Statistic: 4.143e+04 (p- value 0.00)

→ High F- Statistic and corresponding p- value suggests that the model is statistically important and that the predictors collectively have noteworthy effect on 'normalized used price'

- Skewness: -0.431

→ The residuals are slightly skewed to the left.

- Coefficient Analysis:

Screen size: Coeff 0.1919

→ For each unit increase in screen size, the normalized used price increases by 0.1919 holding other factors constant. The effect is highly significant.

RAM: Coeff 0.3929

→ For each unit increase in RAM, the normalized used price increases by 0.3929 the effect is also highly significant.

- The high R- Squared value indicates that the model is robust and explains that a large proportion of the variance is in the normalized used price.

- Both screen size and RAM are important predictors of the normalized used price, with RAM having larger impact.

- The residuals are not perfectly normally distributed.

➢ Now we will check the model performance on the final model train set to see the difference.

RMSE	MAE	R-Squared	Adj. R-Squared	MAPE
0.740676	0.554162	-0.570391	-0.571692	12.730602

TABLE 14

The model performance on the final model test set:

RMSE	MAE	R-Squared	Adj. R-Squared	MAPE
0.740676	0.554162	-0.570391	-0.571692	12.730602

TABLE 15

SUMMARY:

- The analysis of the dataset including the residual distribution, Q-Q Plot, Boxplot for various predictors has yielded valuable insights into the factors influencing the used prices of devices.
- The model demonstrates a good fit with residuals approximately normally distributed, though there are some mild deviations and outliers are present.

ACTIONABLE INSIGHTS:

1. **Normality of Residuals:**

- The presence of outliers indicates potential areas for model improvement.
- The residuals are approximately normally distributed signifying that the model's assumptions are mostly met.

2. **Predictors Distribution:**

- Numerous predictors such as internal memory, battery capacity and weight have significant outliers.

- Devices with higher specifications like larger screen size, more RAM, higher battery tend to have higher used prices.

3. Mild deviations at the tails and presence of the outliers need attention.
4. **Boxplot Analysis:**

- Several predictors have notable outliers present.

BUSINESS RECOMMENDATIONS:

1. **Pricing strategy and targeted marketing:**

- Focus marketing efforts on high specification devices, showcasing their USP or features (e.g larger screen size, higher RAM, better battery life) to justify higher used price.
- For devices which has been identified as outliers. (e.g high internal memory, weight) develop exclusive marketing messages or pricing strategies to appeal all segments.

2. **Inventory Optimization:**

- Ensure adequate stock of devices with high demand features such as high RAM and battery capacity as these features tend to attract higher used prices.
- Regularly manage and monitor outlier devices to ensure they do not adversely affect turnover sales.

3. Feature Education:

- Educate customers on the value of high specification features explaining these features enhance user experience and device life.
- Provide tips on device performance maintenance which can positively influence used device prices.

4. Continuous Data Analysis:

- Implementation of a system for continuous data collection and analysis to monitor trends in device specifications and their impact on used price.
- Refine predictive models to improve accuracy, adding up new data and addressing the outliers.

5. Segment by Preference:

- Conduct price sensitivity analysis for various customer segments to optimize pricing strategies.
- Use insights from the analysis to differentiate customers based on their preferences for device specifications allowing more exclusive marketing and sales.

CONCLUSION:

- By implementation of the above recommendations, the company can enhance its pricing, marketing, and inventory control strategies further leading to improved customer satisfaction and increased revenue from the used device sales.
- Regular data analysis and model refinement will help the company to remain competitive in the dynamic used device market.