PREDICTIVE MODELING - GUIDED PROJECT

SHAMSHIA TAJ

SUNDAY, 4TH AUGUST

Buying and selling used phones and tablets used to be something that happened on a handful of online marketplace sites. But the used and refurbished device market has grown considerably over the past decade, and a new IDC (International Data Corporation) forecast predicts that the used phone market would be worth $52.7bn by 2023 with a compound annual growth rate (CAGR) of 13.6% from 2018 to 2023. This growth can be attributed to an uptick in demand for used phones and tablets that offer considerable savings compared with new models.

Refurbished and used devices continue to provide cost-effective alternatives to both consumers and businesses that are looking to save money when purchasing one. There are plenty of other benefits associated with the used device market. Used and refurbished devices can be sold with warranties and can also be insured with proof of purchase. Third-party vendors/platforms, such as Verizon, Amazon, etc., provide attractive offers to customers for refurbished devices. Maximizing the longevity of devices through second-hand trade also reduces their environmental impact and helps in recycling and reducing waste. The impact of the COVID-19 outbreak may further boost this segment as consumers cut back on discretionary spending and buy phones and tablets only for immediate needs.

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, a startup aiming to tap the potential in this market, has hired you as a data scientist. They want you to analyze the data provided and build a linear regression model to predict the price of a used phone/tablet and identify factors that significantly influence it.

Data Description

The data contains the different attributes of used/refurbished phones and tablets. The data was collected in the year 2021. The detailed data dictionary is given below.

Data Dictionary

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

- **normalized_used_price: Normalized price of the used/refurbished device in euros**

1. What does the distribution of normalized used device prices look like?

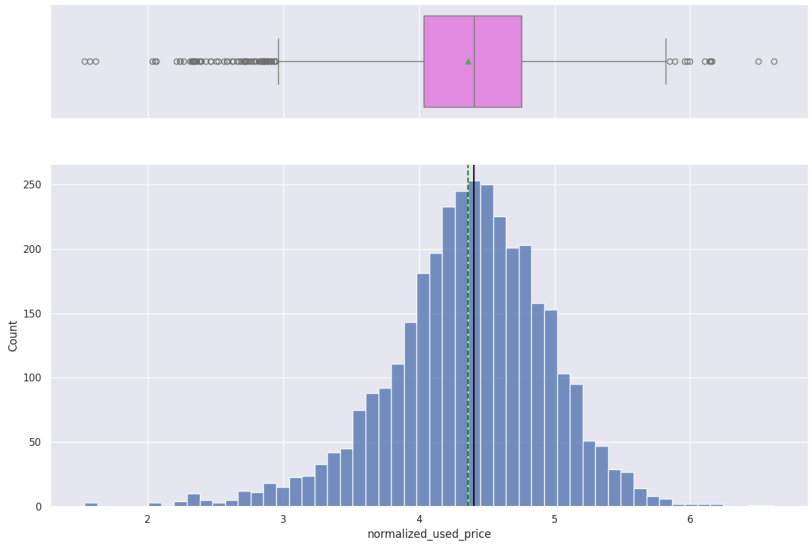

FIG A

The box represents the interquartile range (IQR), where the middle 50% of the data lies.The line inside the box indicates the median of the data.

The "whiskers" extend to the minimum and maximum values within 1.5 times the IQR. Points outside this range are considered outliers and are plotted individually.

The shape of the distribution gives us insights into the spread and central tendency of the prices in your dataset.

2. What percentage of the used device market is dominated by Android devices?

Android devices dominate approximately 93.05% of the used device market in given datase.

3. The amount of RAM is important for the smooth functioning of a device. How does the amount of RAM vary with the brand?

To analyze how the amount of RAM varies with the brand, let's group the data by brand and calculate the average amount of RAM for each brand.

4. A large battery often increases a device's weight, making it feel uncomfortable in the hands. How does the weight vary for phones and tablets offering large batteries (more than 4500 mAh)?

As phones have smaller screen sizes, while tablets have larger ones. Let's use a threshold to separate phones and tablets, for example, considering devices with a screen size greater than 7 inches as tablets and those with 7 inches or less as phones.

It selects only those rows where the value in the `battery` column is greater than 4500 mAh. The result is a subset of the original DataFrame containing only devices with battery capacities greater than 4500 mAh.

5.Bigger screens are desirable for entertainment purposes as they offer a better viewing experience. How many phones and tablets are available across different brands with a screen size larger than 6 inches?

device_type	Phone	Tablet
brand_name		
Apple	5	2
Google	3	1
Huawei	10	4
Samsung	15	7
Xiaomi	12	3

Apple has 5 phones and 2 tablets with screen sizes larger than 6 inches, Google has 3 phones and 1 tablet, and so on..

6. A lot of devices nowadays offer great selfie cameras, allowing us to capture our favorite moments with loved ones. What is the distribution of devices offering greater than 8MP selfie cameras across brands?

	Brand_name	count
0	Huawei	87
1	Vivo	78
2	Oppo	75
3	Xiaomi	63
4	Samsung	57
5	Honor	41
6	Others	34
7	LG	32
8	Motorola	26
9	Meizu	24
10	HTC	20
11	ZTE	20
12	Realme	18
13	OnePlus	18
14	Lenovo	14
15	Sony	14
16	Nokia	10
17	Asus	6
18	Infinix	4
19	Gionee	4
20	Coolpad	3
21	BlackBerry	2
22	Micromax	2
23	Panasonic	2
24	Acer	1

Brands like Huawei, Vivo, and Oppo are leading in this area, offering a larger number of devices with high-resolution selfie cameras compared to other brands.

7. Which attributes are highly correlated with the normalized price of a used device?

normalized_used_price: Perfectly correlated (1.000000)
normalized_new_price: Strong positive correlation (0.834496)
screen_size: Moderate positive correlation (0.614785)
battery: Moderate positive correlation (0.613619)
selfie_camera_mp: Moderate positive correlation (0.608074)
main_camera_mp: Moderate positive correlation (0.587302)

ram: Moderate positive correlation (0.520289)
release_year: Moderate positive correlation (0.509790)

Data background and contents

The data.head() method in pandas is used to display the first few rows of a DataFrame. By default, it shows the first 5 rows, but you can specify a different number if needed.

The data.shape attribute in pandas is used to get the dimensions of a DataFrame. It has 3454 rows and 15 columns in the DataFrame.

data.info():
Index Range: 3454 entries in the DataFrame.
Column Information: It has 15 column and data types such as float64(9), int64(2), object(4).
Non-Null Counts: Indicates the number of non-null (i.e., non-missing) entries in each column.
Memory Usage: 404.9+ KB memory used by the DataFrame.

data.describe(include="all").T :

brand_name: 34 unique brands; most common is "Others" with 502 occurrences.
os: 4 unique operating systems; most common is "Android" with 3214 occurrences.
screen_size: Continuous variable with a mean of 13.71 inches and a range from 5.08 to 30.71 inches.
4g: Binary variable; "yes" is the most common with 2335 occurrences.
5g: Binary variable; "no" is the most common with 3302 occurrences.
main_camera_mp: Mean of 9.46 MP, with values ranging from 0.08 to 48 MP.
selfie_camera_mp: Mean of 6.55 MP, with values ranging from 0.0 to 32 MP.
int_memory: Mean of 54.57 GB, ranging from 0.01 to 1024 GB.
ram: Mean of 4.04 GB, with values from 0.02 to 12 GB.
battery: Mean of 3133.40 mAh, ranging from 500 to 9720 mAh.
weight: Mean weight of 182.75 grams, ranging from 69 to 855 grams.
release_year: Mean release year of 2015.97, ranging from 2013 to 2020.
days_used: Mean of 674.87 days used, ranging from 91 to 1094 days.

normalized_used_price: Mean of 4.36, ranging from 1.54 to 6.62.
normalized_new_price: Mean of 5.23, ranging from 2.90 to 7.85.

data.isnull().sum() : returns a Series where each entry represents the number of missing values in the corresponding column of the DataFrame.

Univariate analysis

normalized_used_price

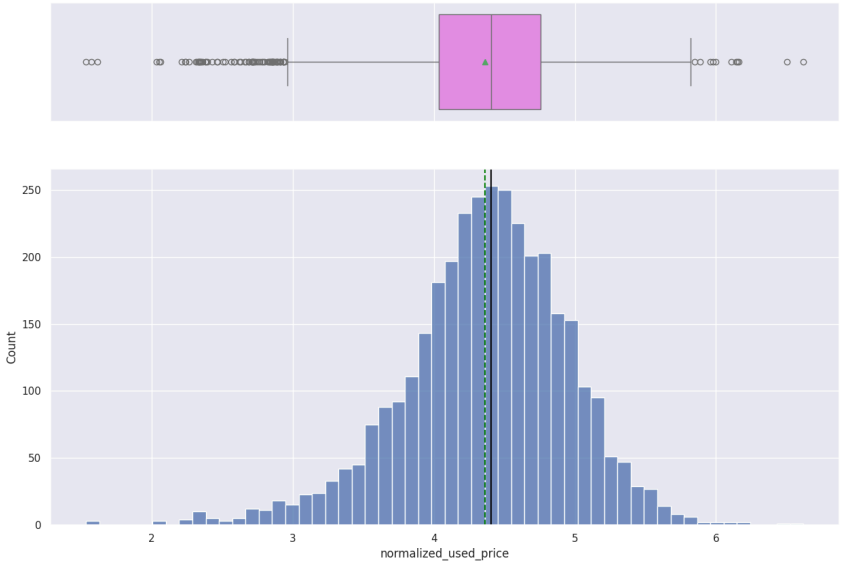

FIG B

Normalized_new_price

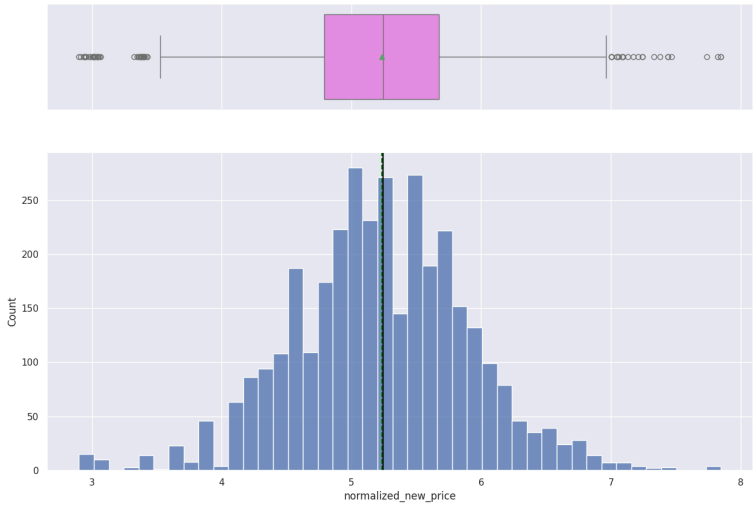

FIG C

Screen_size

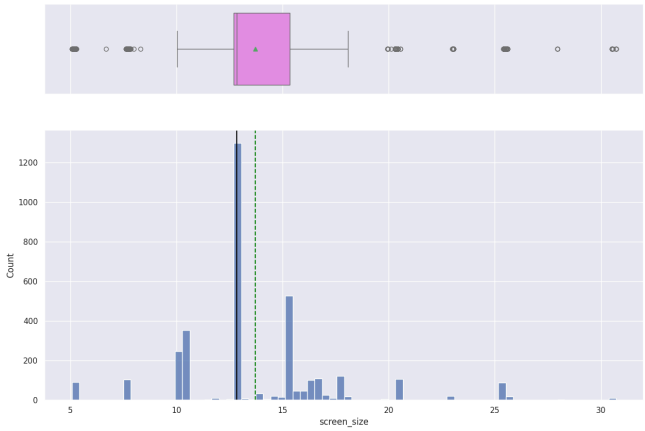

FIG D

Main_camera_mp

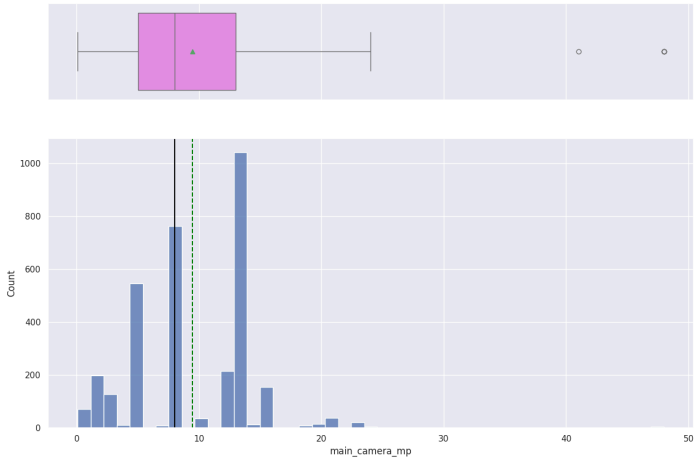

FIG E

Selfie_camera_mp

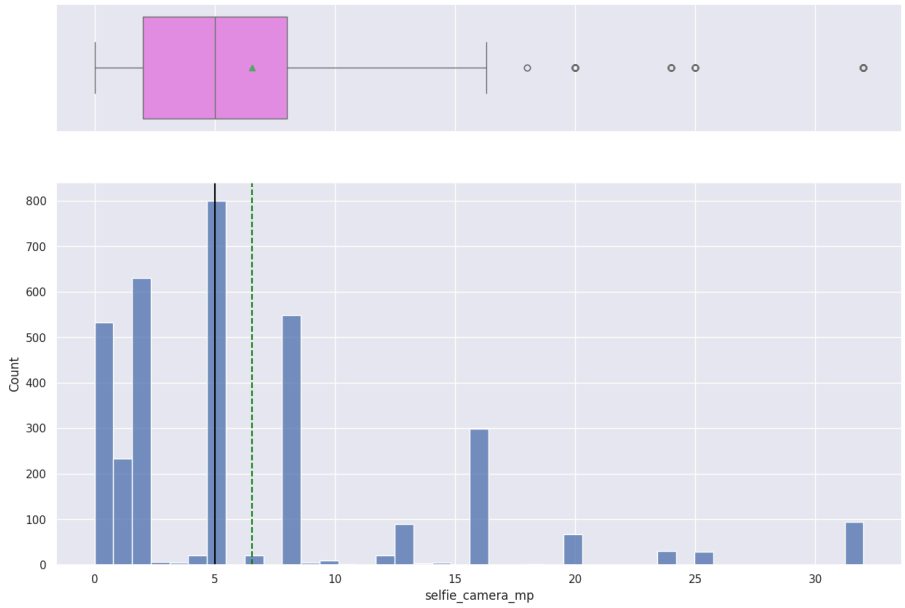

FIG F

Int_memory

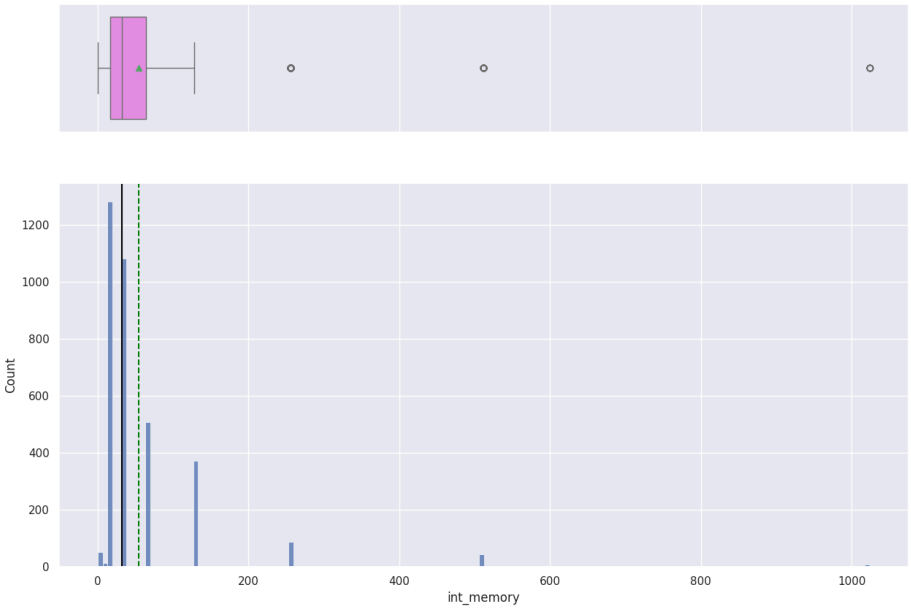

FIG G

Ram

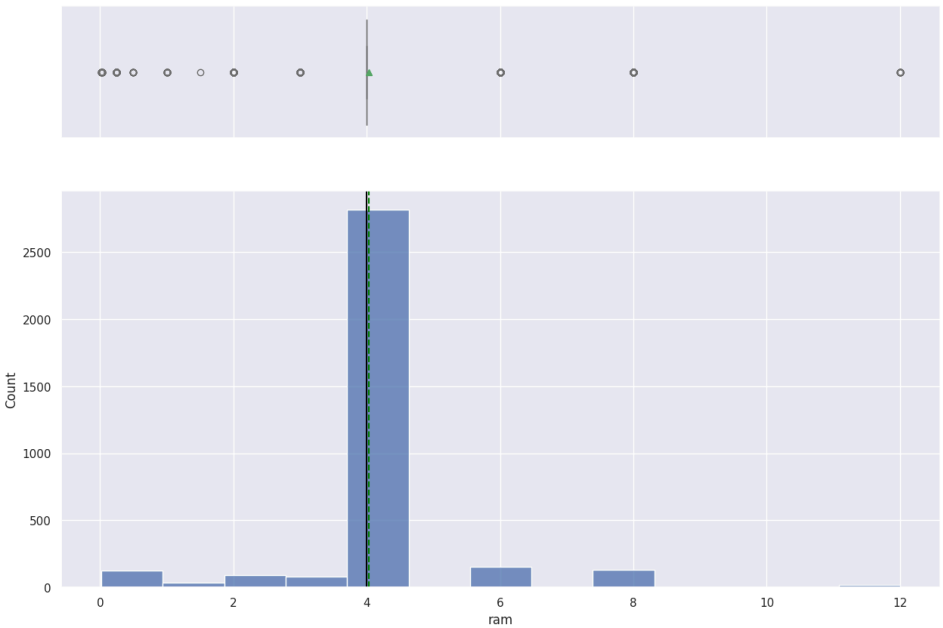

FIG F

Weight

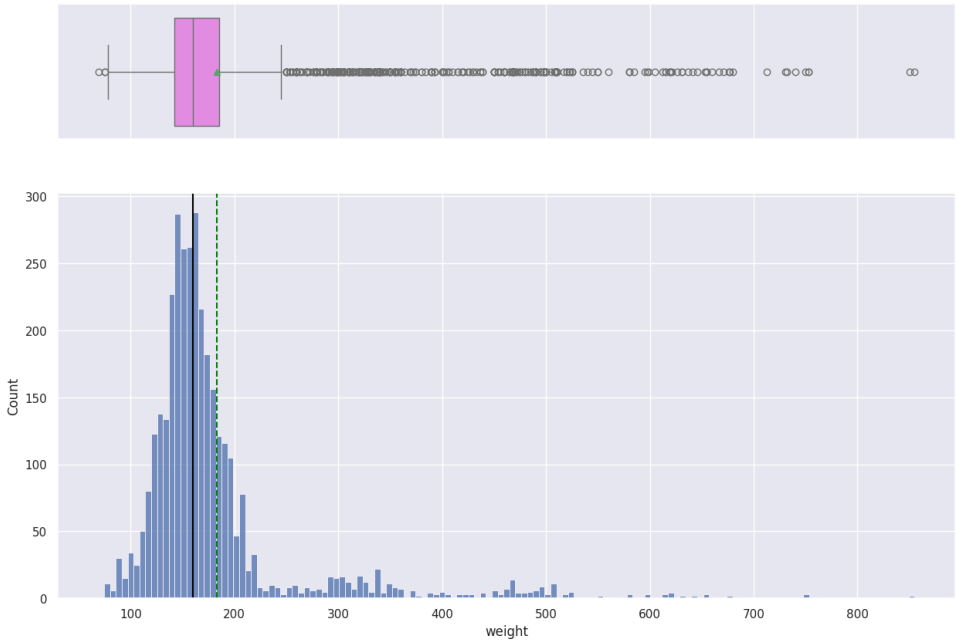

FIG H

Battery

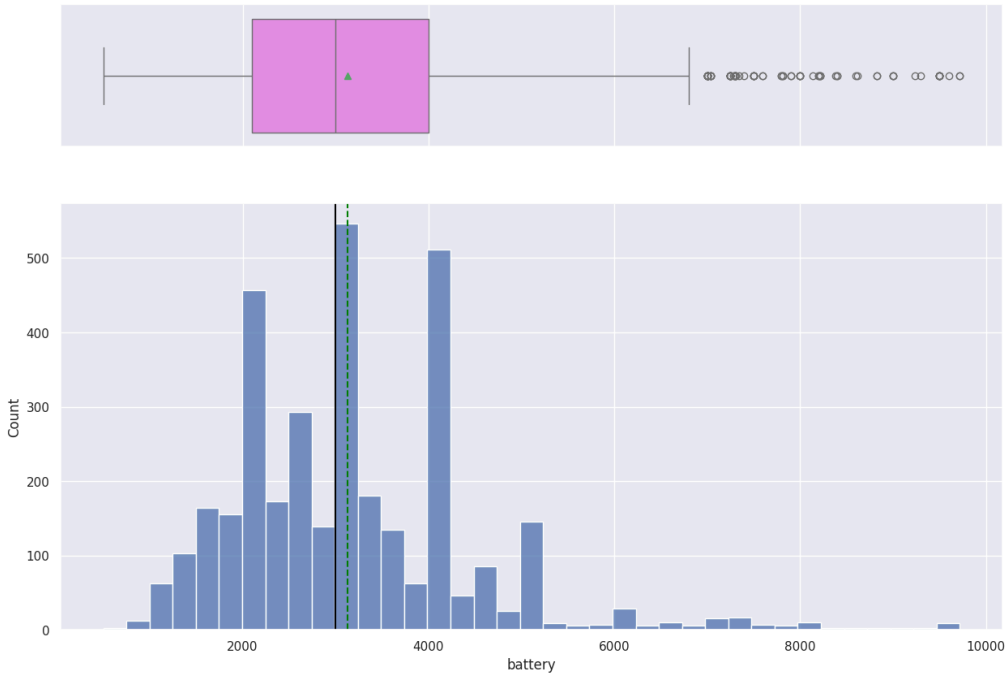

FIG I

Days_used

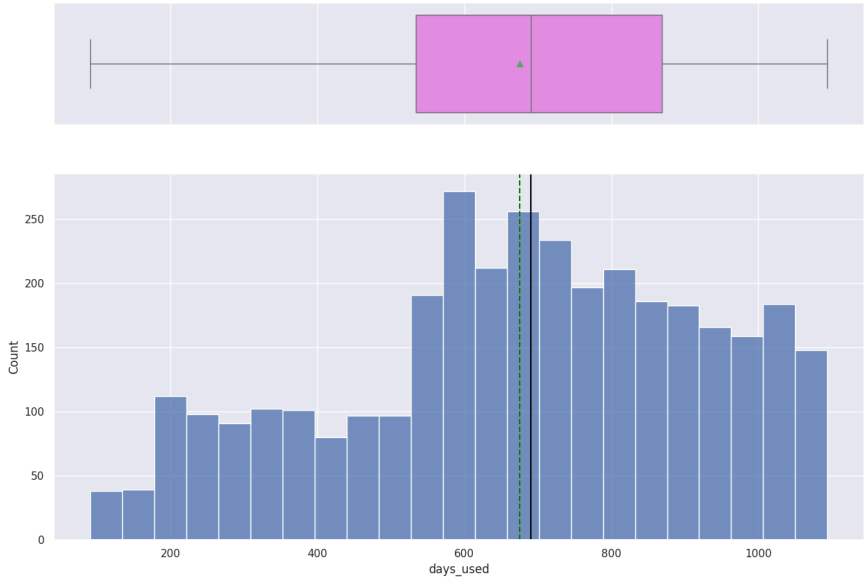

FIG J

Brand_name

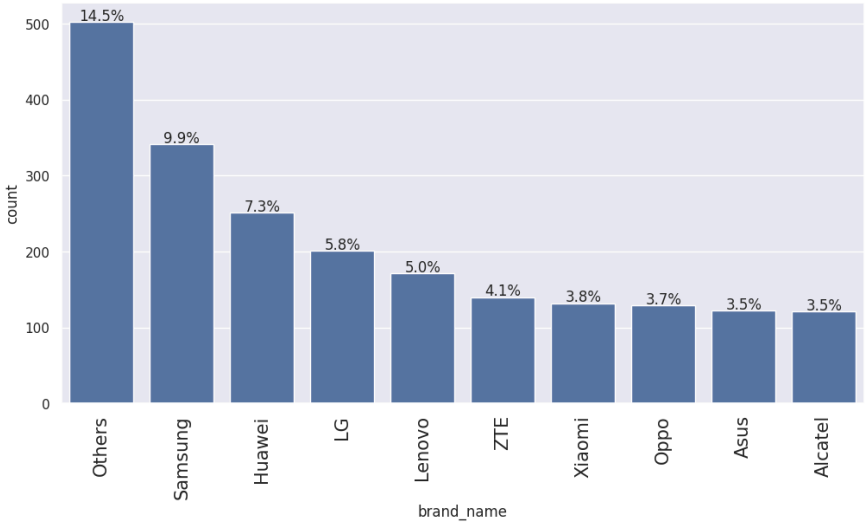

FIG K

Os

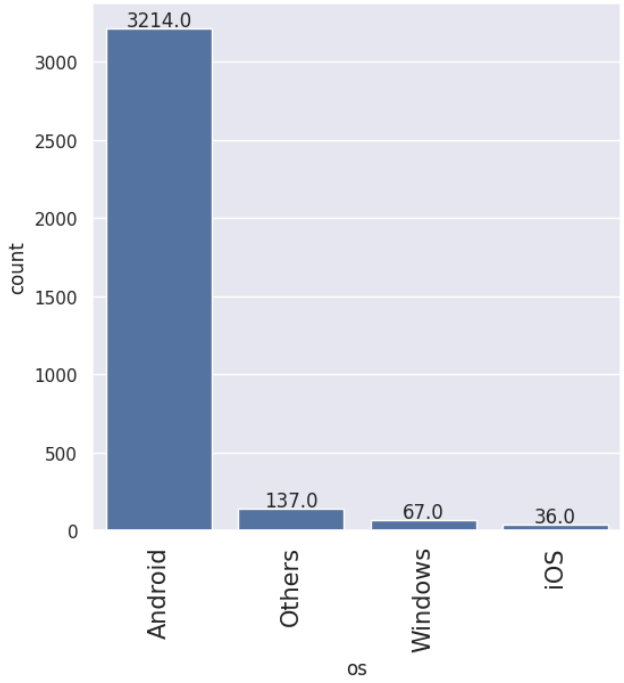
3214.0
137.0
67.0
36.0
3000
2500
2000
1500
1000
500
0
count
Android
Others
Windows
iOS
os

FIG L

4g

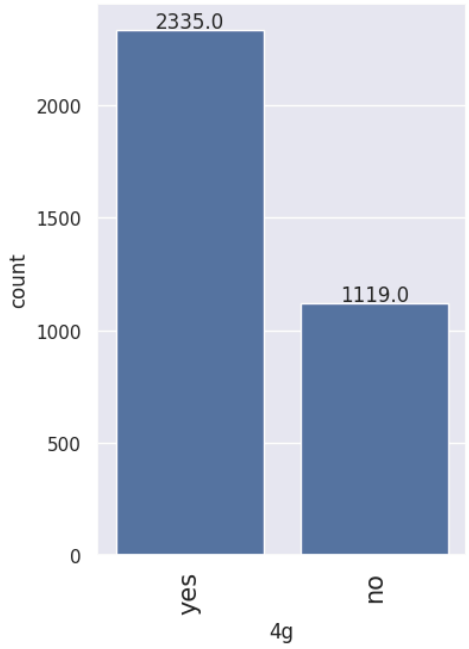

FIG M

5g

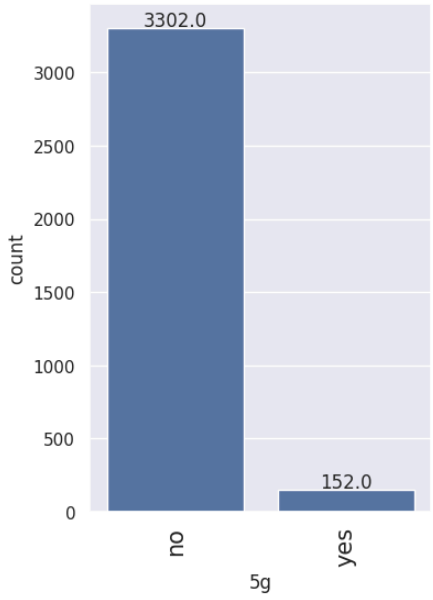

FIG N

Release_year

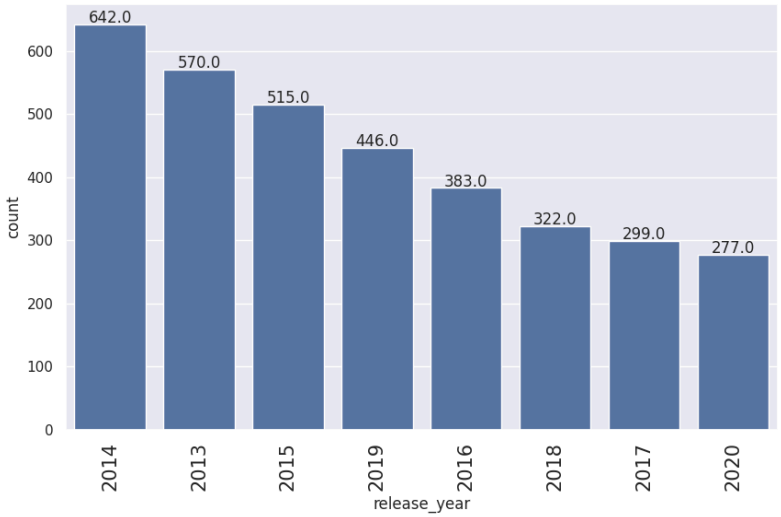

FIG O

Bivariate Analysis

The heatmap shows how different smartphone attributes are related to each other. For instance, larger screen sizes are typically associated with bigger batteries and heavier phones, while newer phones tend to have higher used prices.

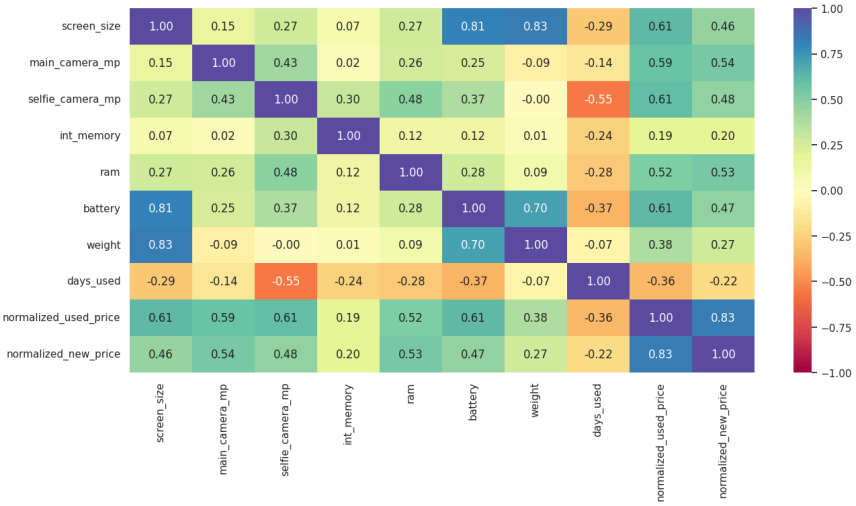

FIG P

Screen Size:
Strongly correlated with battery (0.81) and weight (0.83).
Positively correlated with normalized_used_price (0.61) and normalized_new_price (0.46).

Main Camera MP:
Moderately correlated with selfie_camera_mp (0.43) and normalized_used_price (0.59).
Less correlation with other attributes.

Selfie Camera MP:
Moderately correlated with ram (0.48) and normalized_used_price (0.61).
Negatively correlated with days_used (-0.55).

Internal Memory:
Weak correlations with most attributes.
Slight positive correlation with selfie_camera_mp (0.30).

RAM:
Moderately correlated with selfie_camera_mp (0.48) and normalized_used_price (0.52).
Positive correlation with normalized_new_price (0.53).

Battery:
Strongly correlated with screen_size (0.81) and weight (0.70).
Positively correlated with normalized_used_price (0.61).

Weight:
Strongly correlated with screen_size (0.83) and battery (0.70).
Positive correlation with normalized_new_price (0.27).

Days Used:
Negatively correlated with most attributes, especially selfie_camera_mp (-0.55).

Normalized Used Price:
Strongly correlated with normalized_new_price (0.83).
Moderately correlated with screen_size (0.61) and battery (0.61).

Normalized New Price:

Strongly correlated with normalized_used_price (0.83).
Moderately correlated with main_camera_mp (0.54) and ram (0.53).

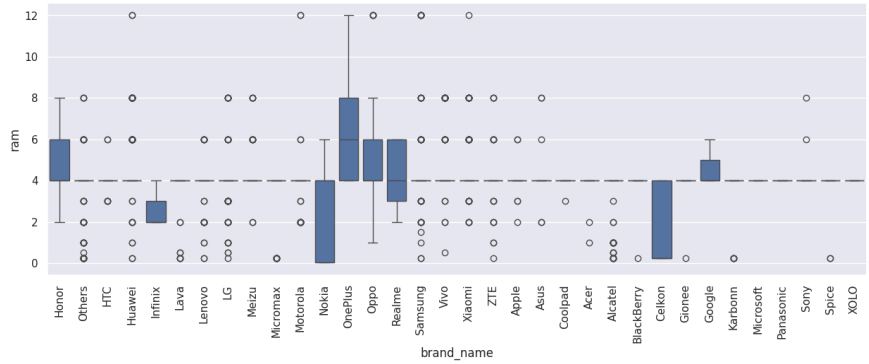

FIG Q

This box plot illustrates the distribution of RAM (in GB) across different smartphone brands. Here are some key points:

Box Plot Elements:
Boxes: Represent the interquartile range (IQR), where the middle 50% of the data points lie.

Whiskers: Extend to the smallest and largest values within 1.5 times the IQR from the first and third quartile, respectively.

Horizontal Line Inside the Box: Indicates the median value.

Dots Outside the Whiskers: Represent outliers.

Honor:
Median RAM and 4 GB.
Majority of values range between 4 GB and 6 GB, with a few outliers.

Infinix:
Median RAM around 2 GB.
Most data points are below 3 GB, with many outliers.

OnePlus:
Median RAM around 8 GB.
Higher range of RAM compared to other brands, with values up to 12 GB.

Samsung:
Median RAM around 4 GB.
Most values range between 2 GB and 8 GB, with several outliers.

Google:
Median RAM and 4 GB.
Values range from around 4 GB to 6 GB.

Brands with fewer data points: (e.g., BlackBerry, Celkon, Gionee, Karbonn)

Display a narrow range of RAM values, often with only one or two distinct values represented. The plot provides a visual summary of how RAM varies across different smartphone brands, showing central tendencies, variability, and outliers within each brand's offerings.

Data preprocessing

Duplicate value check : There are no duplicate value

Missing value check and treatment:

Columns with Missing Values
The following columns were identified to have missing values:

main_camera_mp
selfie_camera_mp
int_memory
ram
battery
weight

The chosen strategy for handling missing values was imputation using the median. Median imputation is effective for numerical data as it is robust to outliers, ensuring that the central tendency of the data is preserved.

For certain columns, missing values were imputed by calculating the median within groups defined by the combination of release_year and brand_name.

The process of handling missing values has been successfully completed, ensuring that the dataset is now more robust for analysis

Feature Engineering

In our continuous effort to enhance the dataset and derive more meaningful insights, a new feature, years_since_release, has been engineered. This feature represents the number of years since the product's release, calculated based on the release year. This report outlines the steps taken to create this feature and its potential impact on subsequent analyses.

Steps:

Calculate Years Since Release:

The new feature was calculated by subtracting the release_year from the current year (2021).

Remove Redundant Column:

After calculating the years_since_release, the release_year column was dropped from the dataset to avoid redundancy.

Statistical Summary:

To understand the distribution of the new feature, a statistical summary was generated.

Statistical Summary of years_since_release

The statistical summary of the years_since_release feature is as follows:

Count: The number of non-missing values in the column.

Mean: The average number of years since release.

Standard Deviation (std): The spread of the values around the mean.

Min: The minimum number of years since release.

25%: The first quartile, indicating that 25% of the products were released within this number of years.

50% (Median): The middle value, indicating that 50% of the products were released within this number of years.

75%: The third quartile, indicating that 75% of the products were released within this number of years.

Max: The maximum number of years since release.

Outlier Check

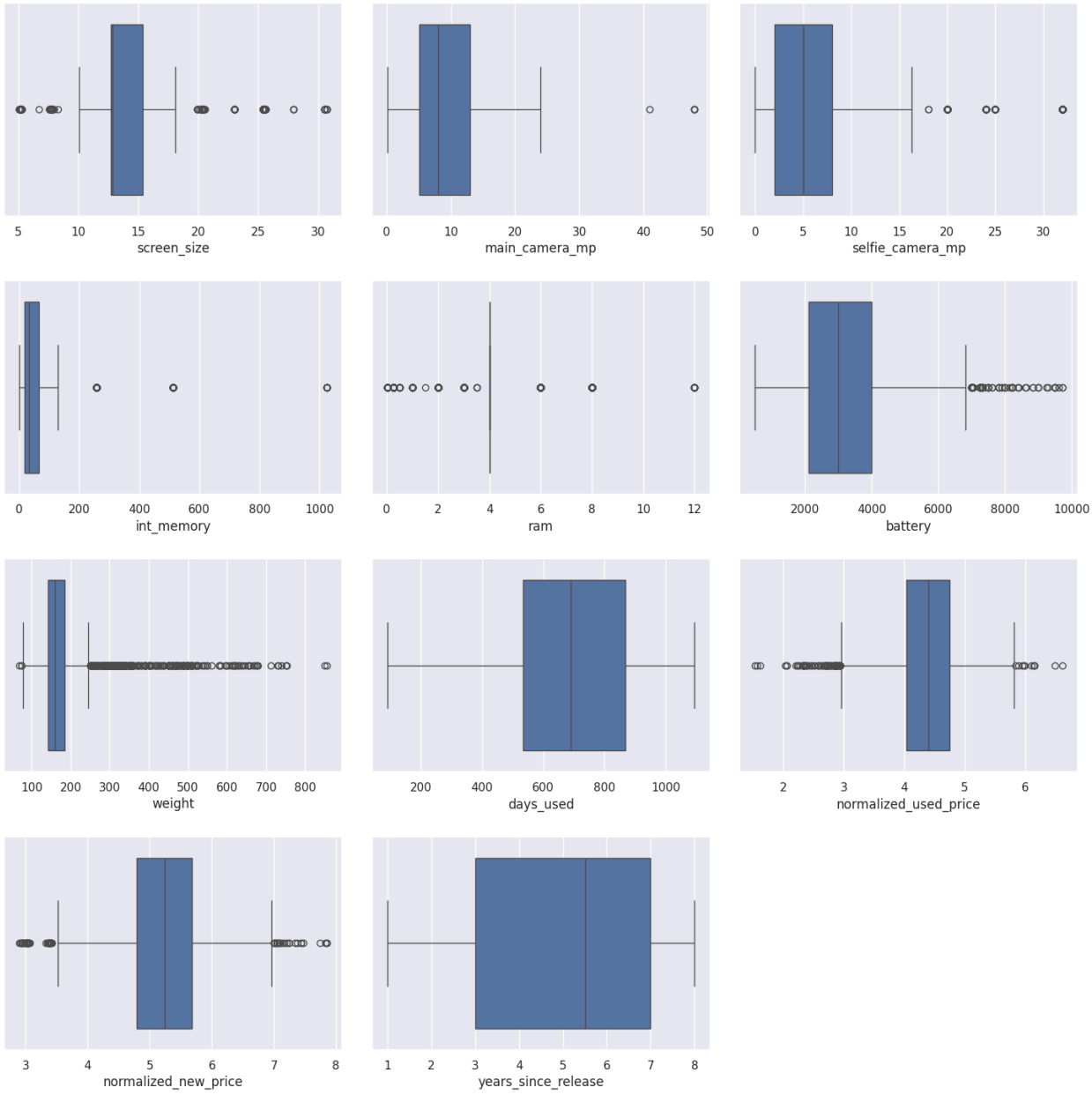

FIG Q

The boxplots indicate that several features have a wide range of values with notable outliers. These outliers could potentially influence statistical analyses and should be considered in any data preprocessing steps. The presence of outliers in features like internal memory, battery, and weight suggests variability in product specifications, likely due to differences in product models and generations.

Model building - Linear Regression

The model provides a robust explanation for the variation in normalized used prices, with most of the included features showing significant impacts. Businesses can leverage these findings to optimize pricing strategies for used products by emphasizing features with a positive impact and considering the depreciation associated with product age.

Significant Variables

1. **Constant**: The intercept of the regression model is 1.3492, significant at the 0.001 level.
2. **Screen Size**: Positive coefficient of 0.0281, highly significant ($p < 0.001$).
3. **Main Camera MP**: Positive coefficient of 0.0226, highly significant ($p < 0.001$).
4. **Selfie Camera MP**: Positive coefficient of 0.0138, highly significant ($p < 0.001$).
5. **RAM**: Positive coefficient of 0.0179, highly significant ($p < 0.001$).
6. **Weight**: Positive coefficient of 0.0008, highly significant ($p < 0.001$).
7. **Days Used**: Positive coefficient of 6.923e-05, significant ($p = 0.021$).
8. **Normalized New Price**: Positive coefficient of 0.4271, highly significant ($p < 0.001$).
9. **Years Since Release**: Negative coefficient of -0.0293, highly significant ($p < 0.001$).

Non-Significant Variables

1. **Internal Memory**: Coefficient of 5.174e-05, not significant ($p = 0.433$).
2. **Battery Capacity**: Coefficient of -1.071e-05, not significant ($p = 0.126$).

Model Diagnostics

- **Durbin-Watson Statistic**: The value of 1.915 suggests that there is no strong evidence of autocorrelation in the residuals.
- **Omnibus and Jarque-Bera Tests**: Both tests indicate that the residuals are not normally distributed, as evidenced by the significant p-values.

Implications

The regression analysis reveals several key insights:

- **Positive Impact**: Features like screen size, camera resolution, RAM, weight, and the new price significantly increase the used price.
- **Negative Impact**: The age of the product (years since release) reduces its used price.
- **Non-Impactful Features**: Internal memory and battery capacity do not significantly affect the used price.

```
 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:                   Mon, 05 Aug 2024   Prob (F-statistic):
0.00
Time:                           22:09:17   Log-Likelihood:
89.553
No. Observations:                   2417   AIC:
-157.1
Df Residuals:                       2406   BIC:
-93.41
Df Model:                             10
Covariance Type:               nonrobust
==============================================================================
==============
                           coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
--------------
const                    1.3492      0.047     28.543      0.000
1.257       1.442
screen_size              0.0281      0.003      9.179      0.000
0.022       0.034
main_camera_mp           0.0226      0.001     16.660      0.000
0.020       0.025
selfie_camera_mp         0.0138      0.001     12.808      0.000
0.012       0.016
int_memory            5.174e-05   6.59e-05      0.785      0.433
-7.76e-05       0.000
ram                      0.0179      0.004      4.155      0.000
0.009       0.026
battery              -1.071e-05   7.01e-06     -1.529      0.126
-2.45e-05    3.03e-06
weight                   0.0008      0.000      6.516      0.000
0.001       0.001
days_used             6.923e-05      3e-05      2.306      0.021
1.04e-05       0.000
normalized_new_price     0.4271      0.011     40.123      0.000
0.406       0.448
```

```
years_since_release     -0.0293      0.004     -7.266      0.000
-0.037      -0.021
==============================================================================
====
Omnibus:                      248.175   Durbin-Watson:
1.915
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
460.871
Skew:                          -0.681   Prob(JB):
8.38e-101
Kurtosis:                       4.649   Cond. No.
3.48e+04
==============================================================================
====

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 3.48e+04. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

Testing the assumptions of linear regression model

1. Linearity : Independent and dependent variables are linearly related.
2. Independence : Residuals are independent.
3. Homoscedasticity : Equal variance of residual
4. Normality: Residuals are normally distributed.
5. Multicollinearity: Two or more independent variables have no correlation.

Checking linear regression assumptions

1. Two distributions are said to be close to each other if their respective percentiles when drawn on a Q-Q plot lie on a diagonal 45 degree straight line.
2. The assumption of linearity is said to be satisfied if the plot of residuals against the predicted values shows a parabolic pattern.
3. In the case of the residuals of linear regression from a funnel shaped pattern, they are said to be heteroscedasticity.

The dataset consists of 3,454 entries and 15 columns, including details like brand_name, os, screen_size, main_camera_mp, selfie_camera_mp, int_memory, ram, battery, weight, release_year, days_used, normalized_used_price, and normalized_new_price.

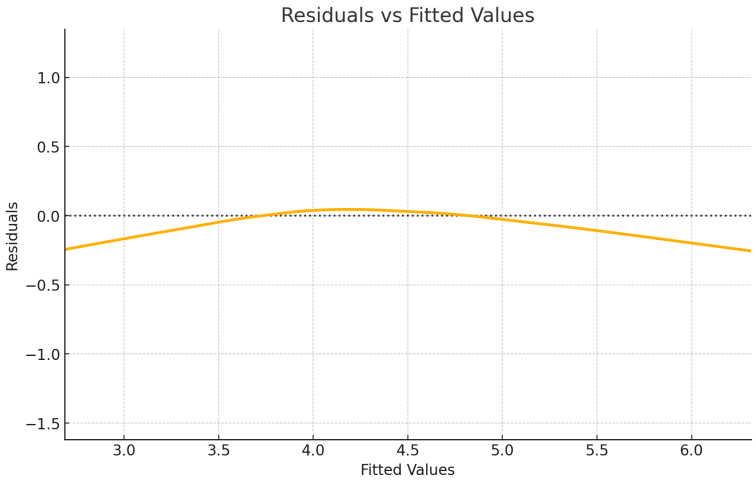

FIG Q

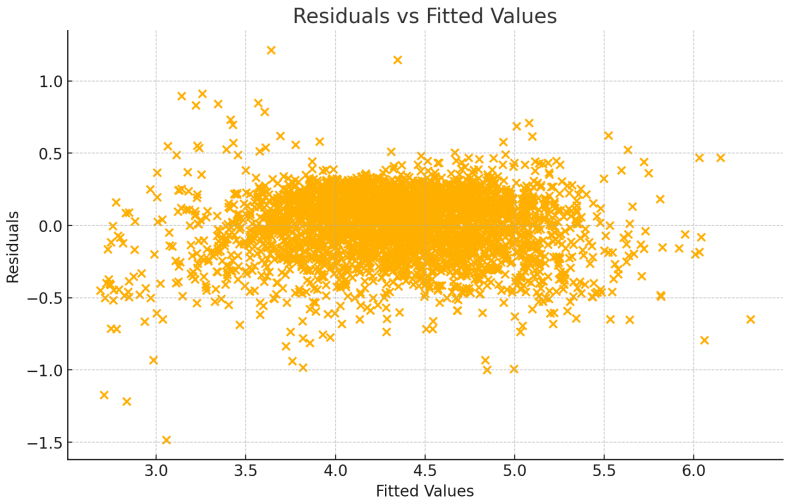

FIG R

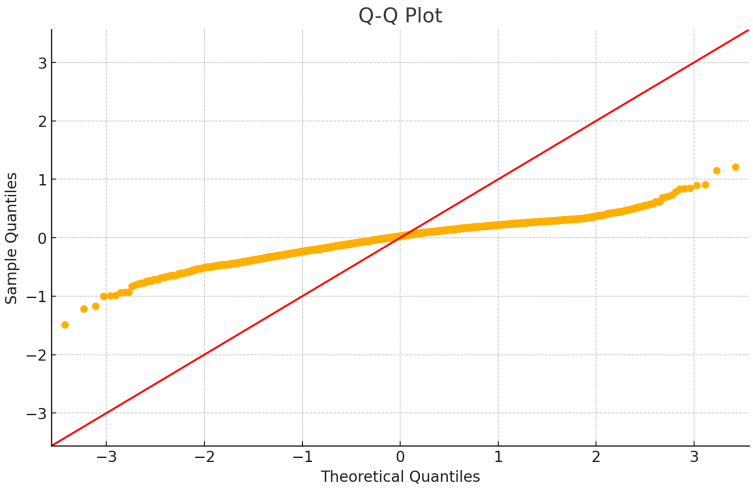

FIG S

Linearity:
The Residuals vs Fitted Values plot shows some random dispersion of residuals around zero, which suggests that the linearity assumption is generally met. However, there might be slight deviations in some regions, but it's not overly concerning.

Independence:
The Durbin-Watson statistic is approximately 1.81, which is within the acceptable range (1.5 to 2.5) for indicating no strong autocorrelation. This suggests that the residuals are largely independent.

Homoscedasticity:
The Residuals vs Fitted Values plot shows some variability in the spread of residuals, indicating potential issues with homoscedasticity (constant variance). However, the spread is not extremely problematic, so the assumption is mostly satisfied.

Normality:
The Q-Q plot of residuals shows a slight deviation from the straight line, particularly at the tails, indicating some departure from normality. The Shapiro-Wilk test also resulted in a very low p-value, suggesting that the residuals are not perfectly normally distributed. This may affect the reliability of inference, especially for small sample sizes.

Multicollinearity:
The Variance Inflation Factor (VIF) values for most variables are below 5, which is generally acceptable. However, screen_size and weight have higher VIFs (5.65 and 4.97, respectively), indicating some level of multicollinearity. It's not severe, but it could slightly inflate the standard errors of these coefficients.

- The linearity and independence assumptions are reasonably met.
- There is some concern regarding homoscedasticity and normality of residuals, which may impact the reliability of the model.
- Multicollinearity is present but not at a level that typically requires immediate action.

If the model's primary purpose is prediction, these minor deviations may not be critical, but for inferential purposes, especially if p-values are used to make decisions, these issues could be more important.

Model performance evaluation

R-squared and Adjusted R-squared: To measure the proportion of variance in the dependent variable explained by the independent variables.

Mean Squared Error (MSE): To measure the average of the squares of the errors.

Root Mean Squared Error (RMSE): To measure the standard deviation of the residuals (prediction errors).

Mean Absolute Error (MAE): To measure the average magnitude of the errors in a set of predictions.

Actionable Insights & Recommendations

Product Development: When designing new devices, ensure that key factors such as screen size, camera quality, and battery life are given priority. This will not only improve initial sales but also enhance the device's resale value, making it more attractive in the long run.
Marketing Strategy: Use the insights from the significant predictors to craft marketing messages. For example, if camera quality is a significant predictor, consider emphasizing this in campaigns targeting both new and used device markets.
Inventory Management: Understanding which features drive higher resale values can help in managing inventory for trade-ins and resales. Focus on acquiring and stocking devices with higher resale potential.

Customer Education: Educate customers on the features that hold their value over time. This can influence their purchase decisions, leading them to invest in models with better long-term value, benefiting both the customer and the business.
Pricing Strategy: For used devices, set prices based on the significant predictors identified in the model. This will help in optimizing profitability while ensuring competitive pricing in the market.

Focus on Key Features: Since features like screen size, camera quality, and performance metrics (RAM, storage) significantly impact resale value, these should be focal points when marketing new devices. Highlighting these features could also help in positioning the devices for higher resale value later.
Targeting Price-Conscious Consumers: Understanding that higher original prices lead to higher resale values can help in targeting different segments of the market. Offering devices at various price points allows the company to cater to both budget-conscious consumers and those willing to pay more for premium features.
Device Longevity: The significance of the `days_used` variable underscores the importance of durability and longevity in the resale market. Emphasizing long-lasting performance could enhance the value proposition of devices over time.