Pricing Strategy for Used and Refurbished Devices

Supervised Learning - Pam Lozano

November 17, 2023

Contents / Agenda

- Data Information
- Business Problem Overview and Solution Approach
- Data Background and Contents
- EDA Results
- Data Preprocessing
- Model Building - Linear Regression
- Testing the assumptions of linear regression model
- Model Performance Evaluation
- Executive Summary

Objective

Develop a pricing strategy for used and refurbished phones and tablets by building a price prediction model and identifying key influencing factors.

Key Focus Areas:

1. Create a linear regression model to predict used and refurbished devices prices, based on various features.
2. Examine factors like brand, OS, camera quality, memory, that influence pricing.
3. Understand market trends and how factors like 4G/5G, release year, and usage affect pricing.
4. Investigate eco-friendly practices and encourage used device sales to reduce electronic waste and extend device lifecycles.

Data Information

The data contains information regarding the interaction of users in both groups with the two versions of the landing page.

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 new device of the same model in euros
normalized_used_price	Normalized price of used/refurbished device in euros

Business Problem Overview

- ReCell, a cell phone company, is trying to figure out how to best price used and refurbished devices for maximum profit and competitiveness in the market. .
- The company must address a thorough analysis of the factors that influence the pricing of these devices.
- The company must effectively position their products in the market based on market dynamics and consumer preferences.
- The company needs to explore ways to reduce waste and be more eco-friendly.
- The goal is to dominate the used and refurbished device market while maximizing profit and sustainability.

Solution Approach

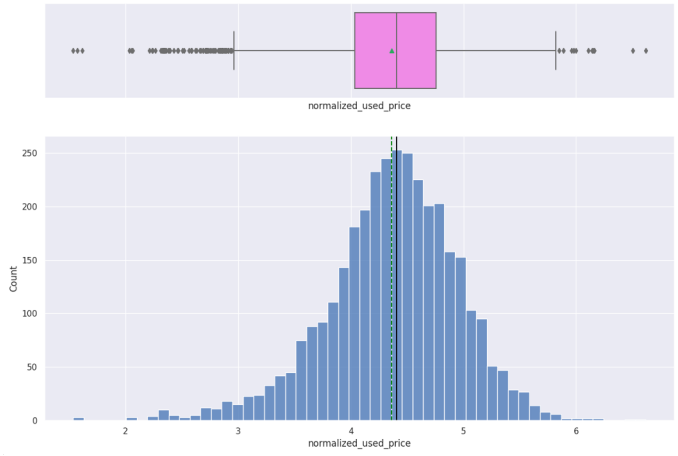
- **Data Cleaning:** Address missing values, duplicates, and outliers to begin preprocessing the data.
- **Exploratory Data Analysis (EDA):** Conduct EDA to explore and identify key attributes influencing pricing.
- **Feature Selection and Engineering:** Identify the most relevant features that impact device pricing and engineer new features if needed.
- **Machine Learning Model :** Choose various machine learning models for price prediction.
- **Model Training and Validation:** Train the selected model on the dataset and validate its performance using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared to ensure it accurately predicts prices.
- **Competitive Analysis:** Use data analytics to gain insights into competitor's pricing strategies and market positioning.
- **Recommendations:** Based on insights, provide actionable recommendations for improving the landing page and increasing conversions.

Data Background and Contents

- Our dataset comprises 3454 rows and 15 columns, providing a comprehensive foundation for our analysis.
- There are a mix of data types in our dataset: integers, objects, and floats, contributing to the diversity of information available for our analysis.
- We've identified some missing values in our dataset, warranting careful consideration in our analysis and data preprocessing steps.
- The average normalized used price is 4.36 euros, serving as a central reference point in understanding the pricing dynamics of our dataset.
- The median normalized used price is 4.41 euros, providing a measure of central tendency in our dataset.
- The lowest normalized used price in our dataset is 1.54 euros, highlighting the range of prices within our analyzed data.
- The highest normalized used price in our dataset is 6.62 euros, representing the upper limit within our analyzed pricing data.

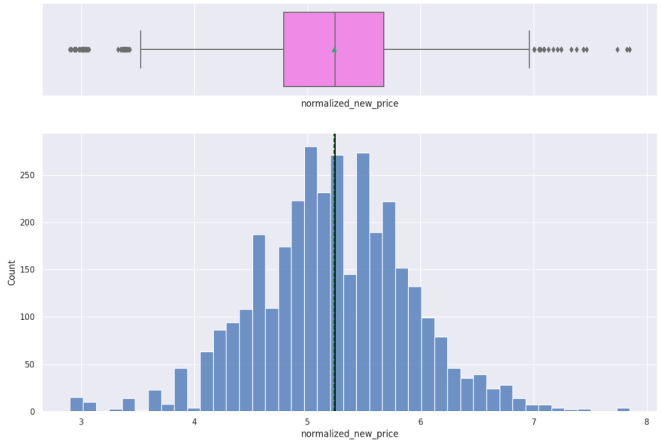
Exploratory Data Analysis

Univariate Analysis - Normalized Used Price



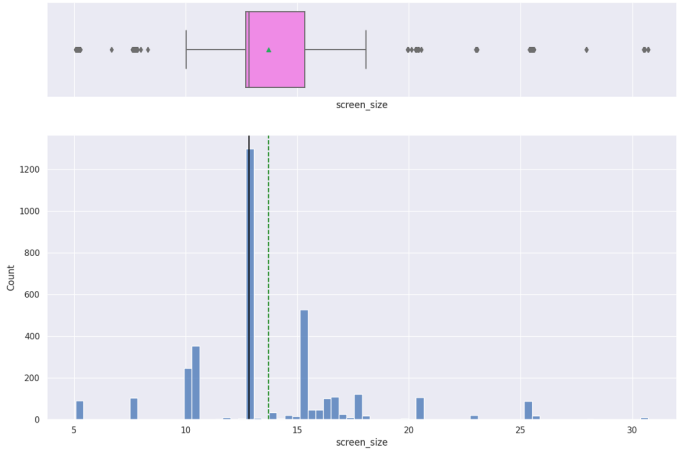
- We observe outliers on both tails in this distribution, indicating potential extreme values.
- The distribution for normalized used price is left-skewed with a median price of 4.41 euros.

Univariate Analysis - Normalized New Price



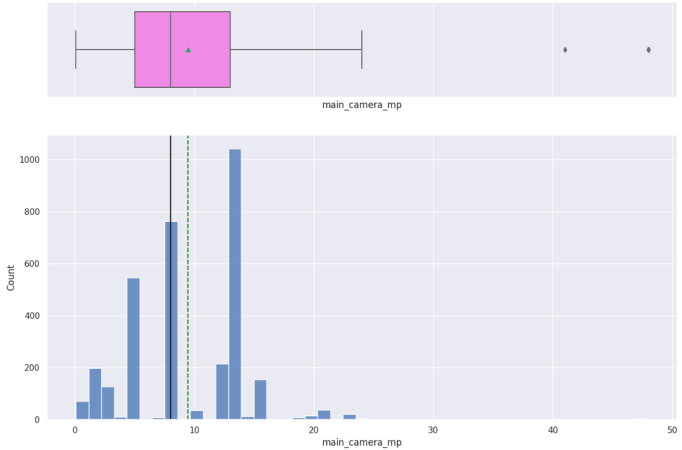
- We've identified several outliers in this distribution, suggesting potential unusual data points.
- The median normalized new price is 5.23 euros.

Univariate Analysis - Screen Size



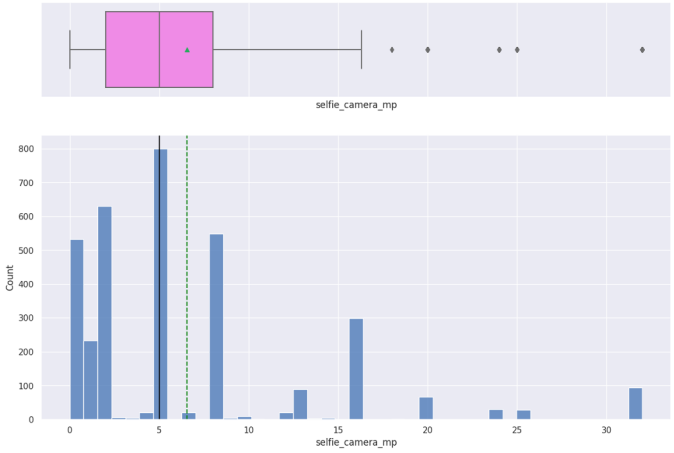
- We've identified some outliers in this distribution, suggesting potential deviations from the overall pattern
- The average screen size is 13.71 cm.

Univariate Analysis - Main Camera MP



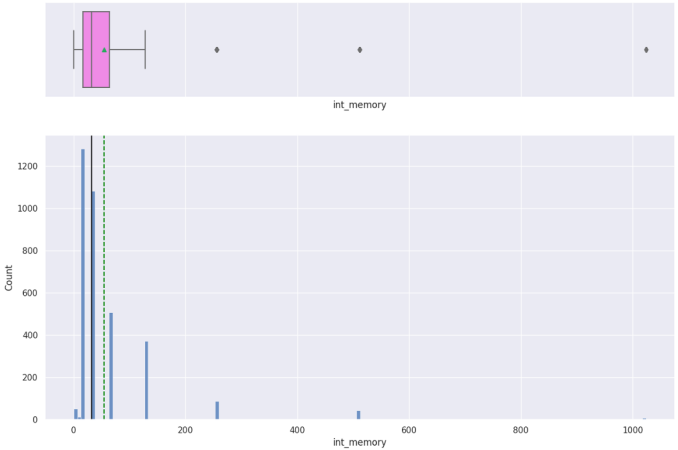
- The distribution for main camera MP is right-skewed with a median of 8 megapixels.
- The majority distribution for main camera MP falls under 25 megapixels

Univariate Analysis - Selfie Camera MP



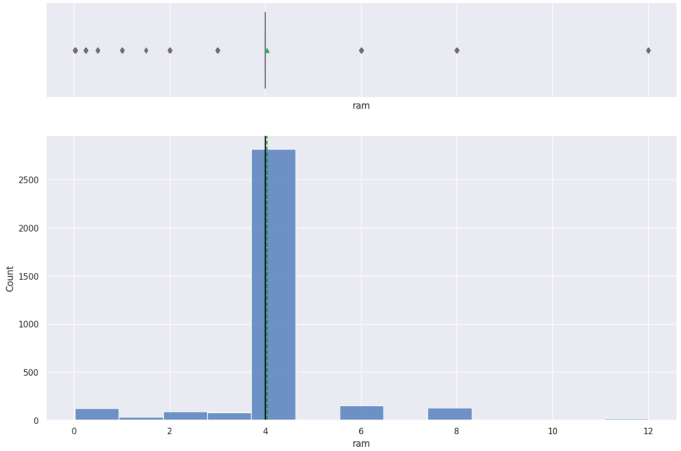
- We've identified a few outliers in this distribution, indicating potential anomalies in our dataset.
- The distribution for selfie camera MP is heavily right-skewed with a median of 5 megapixels.
- We observe spikes in selfie camera MP, particularly around 17 and 33 megapixels.

Univariate Analysis - Int Memory



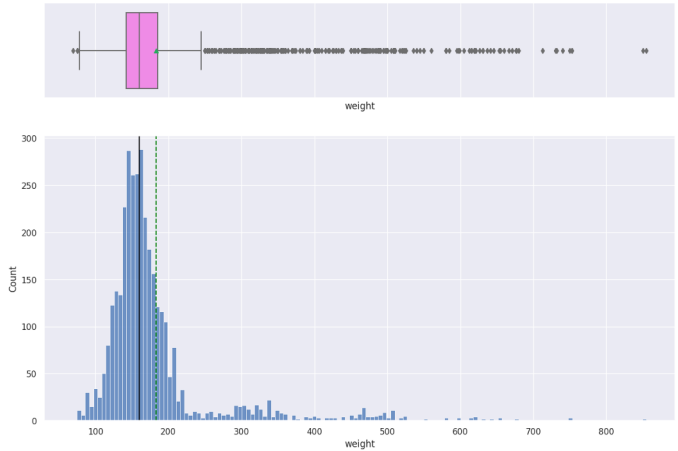
- The distribution for int memory is right skewed.
- The median int memory is 31 GB.

Univariate Analysis - Ram



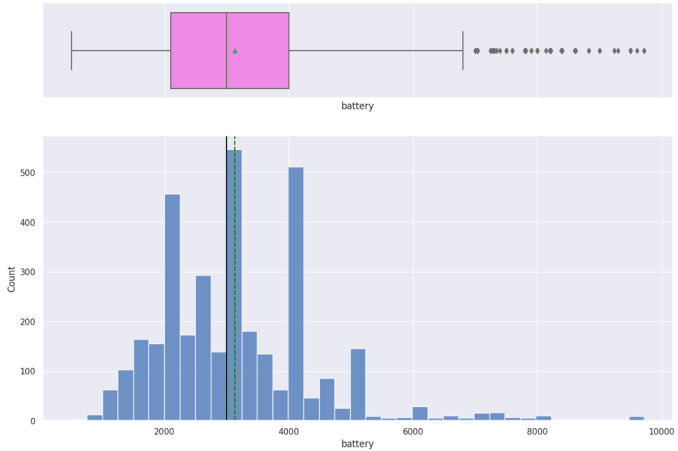
- Several outliers have appeared, suggesting potential unusual values or anomalies.
- The median RAM value in this distribution is 4GB.

Univariate Analysis - Weight



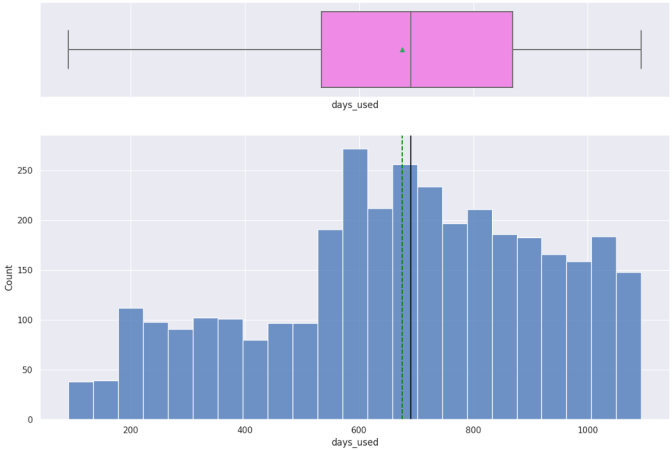
- There are several outliers in this distribution.
- The distribution for weight is heavily right-skewed with a median of 160 grams.

Univariate Analysis - Battery



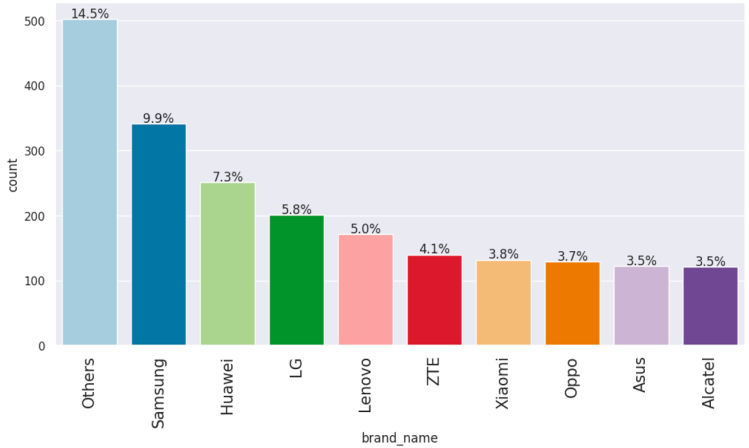
- There are several outliers in this distribution.
- The distribution for battery is heavily right-skewed with a median of 3000 mAh.

Univariate Analysis - Days Used



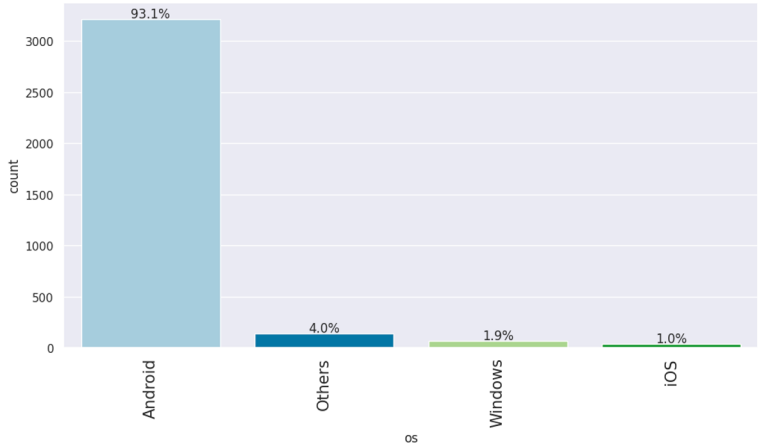
- There are no outliers in this distribution.
- The distribution is skewed to the left, with most values centered around roughly 691 days used.

Univariate Analysis - Brand Name



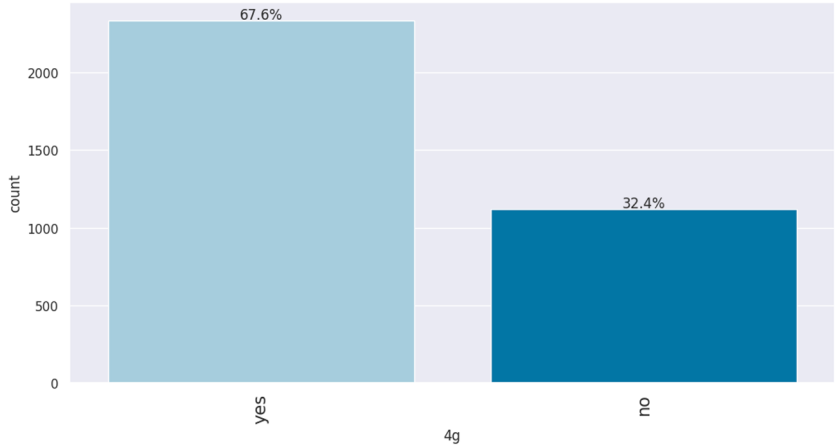
- Most phones in our dataset belong to less common brand names and are categorized as others.

Univariate Analysis - OS



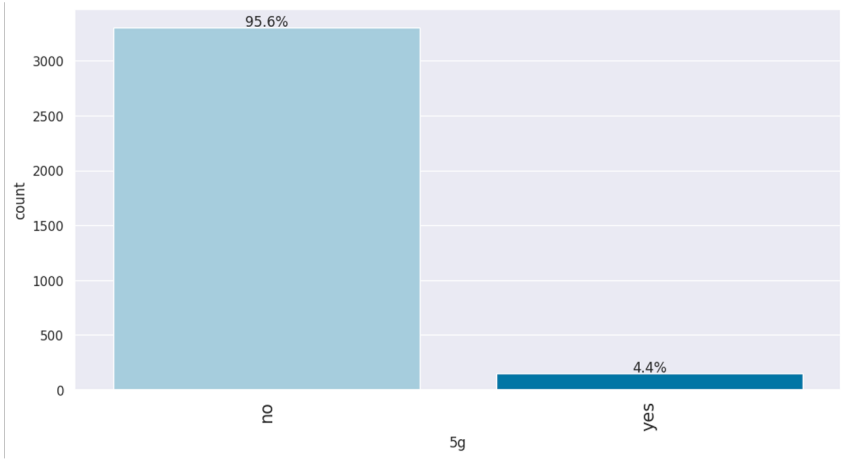
- Most users operate on Android OS, while the smallest proportion prefers iOS.

Univariate Analysis - 4G



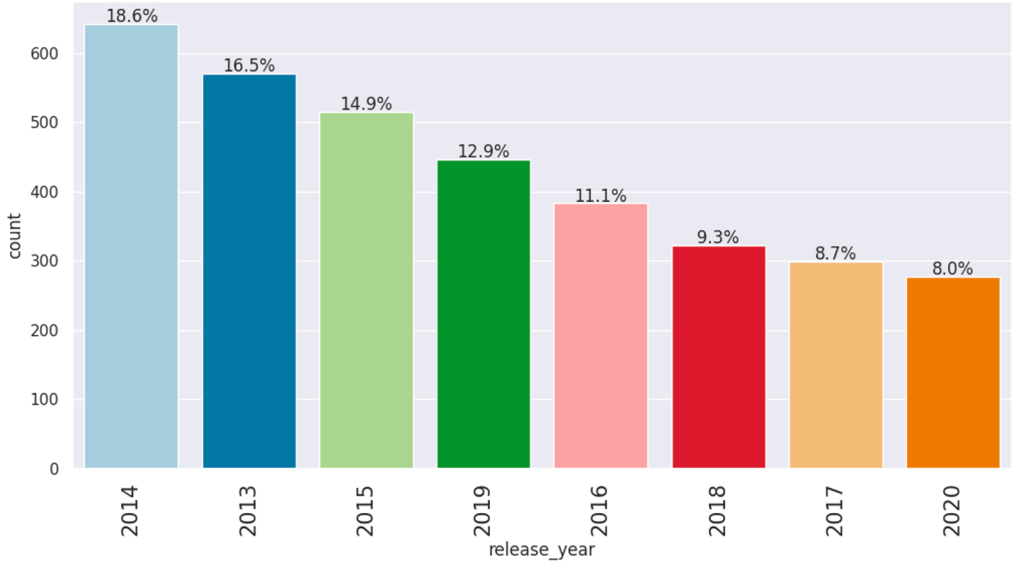
- Approximately two-thirds of the buyers prefer 4G.

Univariate Analysis - 5G



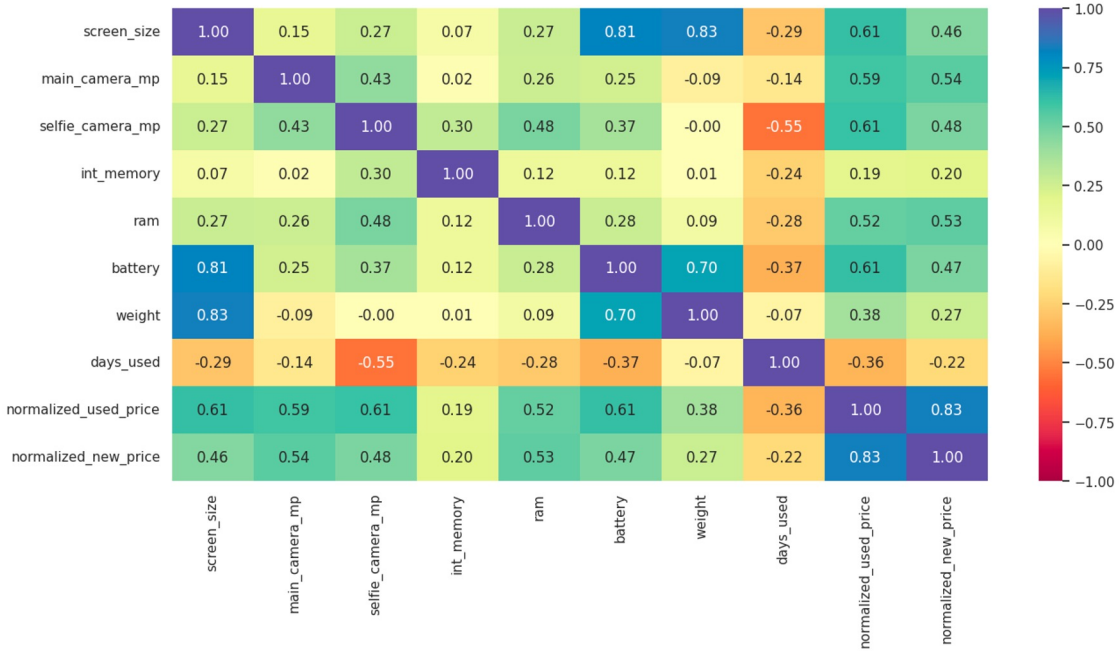
- Most buyers do not prefer 5G technology.

Univariate Analysis - Release Year



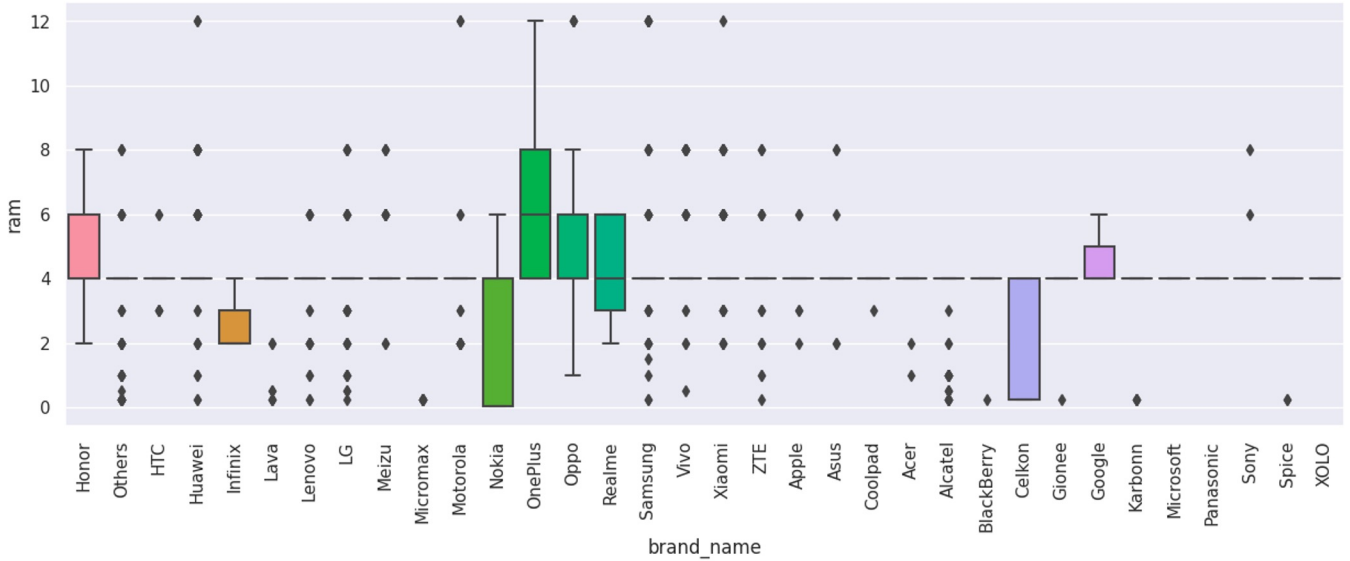
- Over 90% of buyers favor phones released before 2020.

Bivariate Analysis



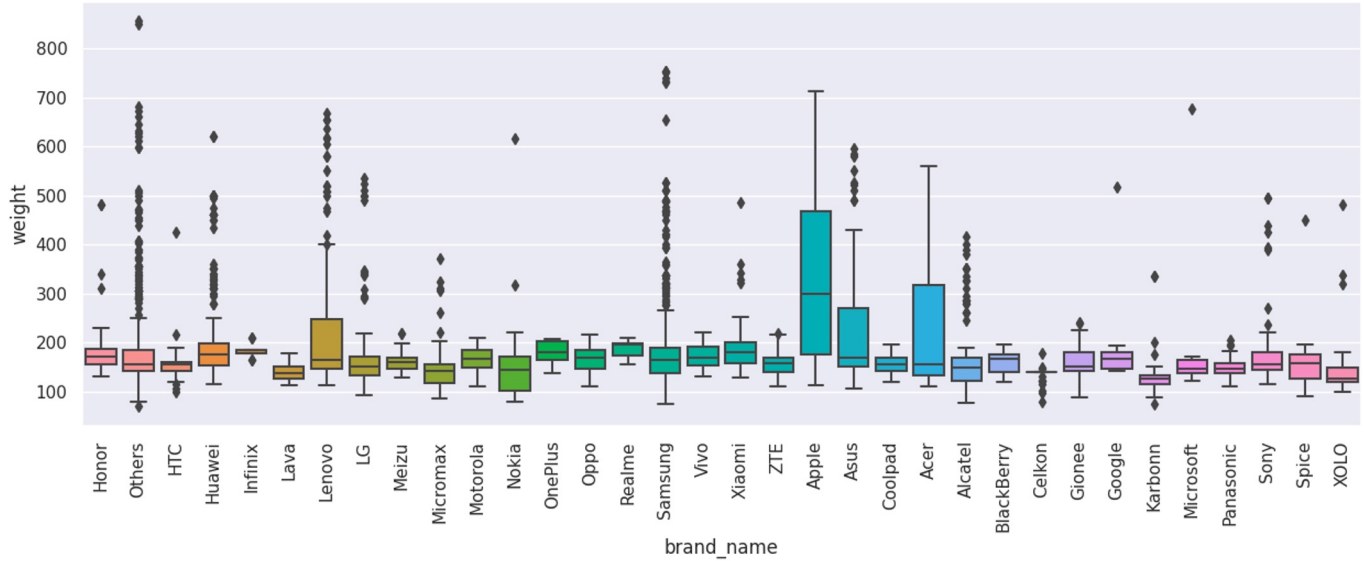
- There's a strong correlation between screen size, weight, and battery in our dataset.
- The normalized used price shows a strong correlation with the normalized new price.
- There is a negative correlation between selfie camera megapixels and the number of days used.

Bivariate Analysis - Brand Name vs Ram



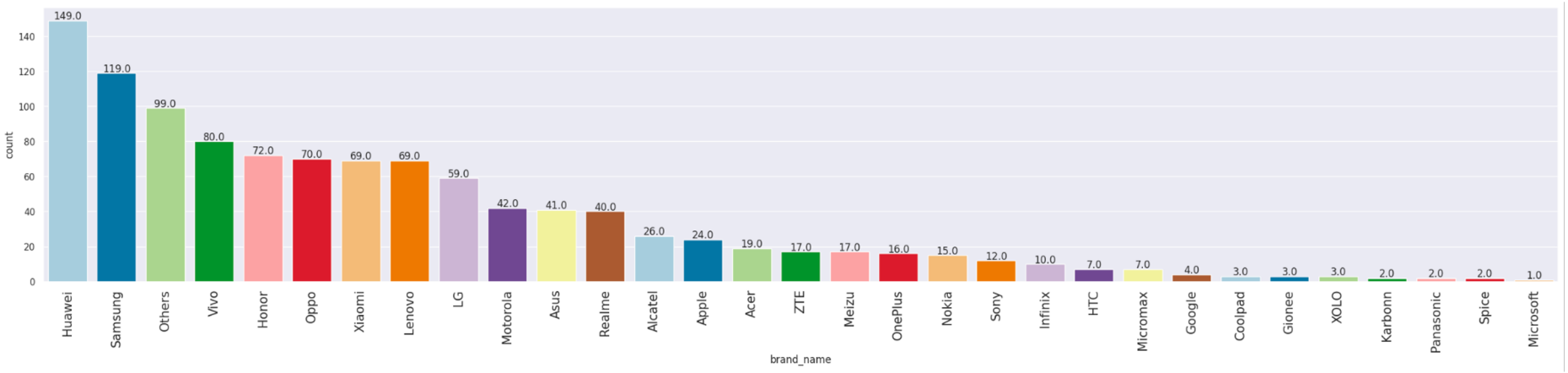
- Infinix, Nokia, and Celkon brands tend to have lower amounts of RAM.
- OnePlus boasts the highest amount of RAM among the observed brands.

Bivariate Analysis - Brand Name vs Weight



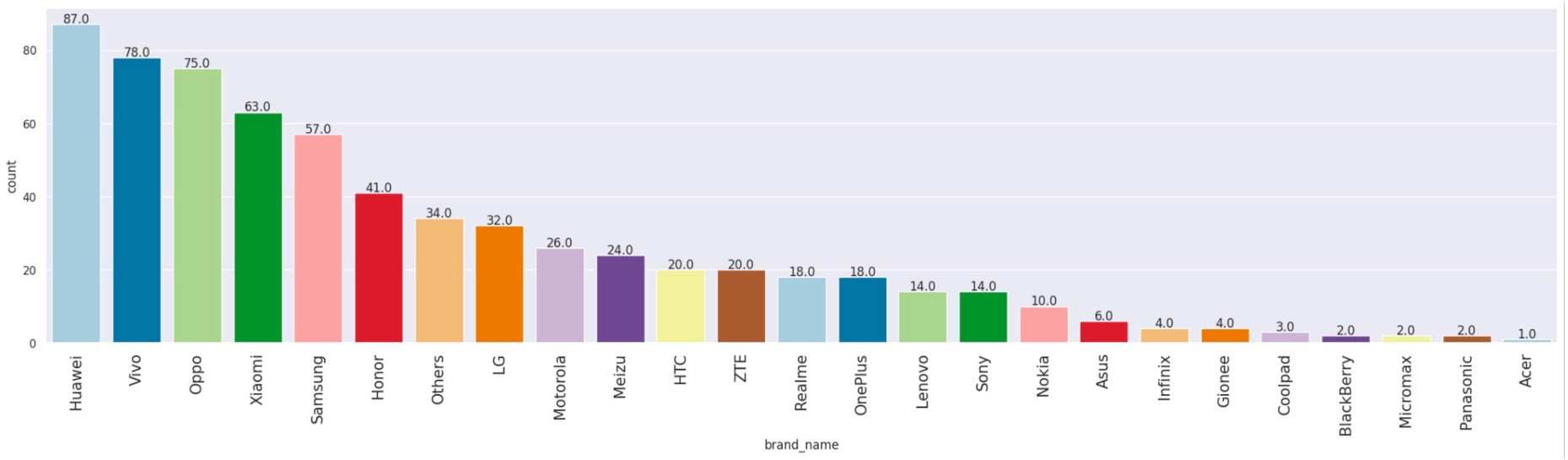
- Acer and Apple brands are characterized by the highest device weight.
- Lava and Karbonn brands exhibit the lowest device weight.

Bivariate Analysis - Screen Size vs Brand Name



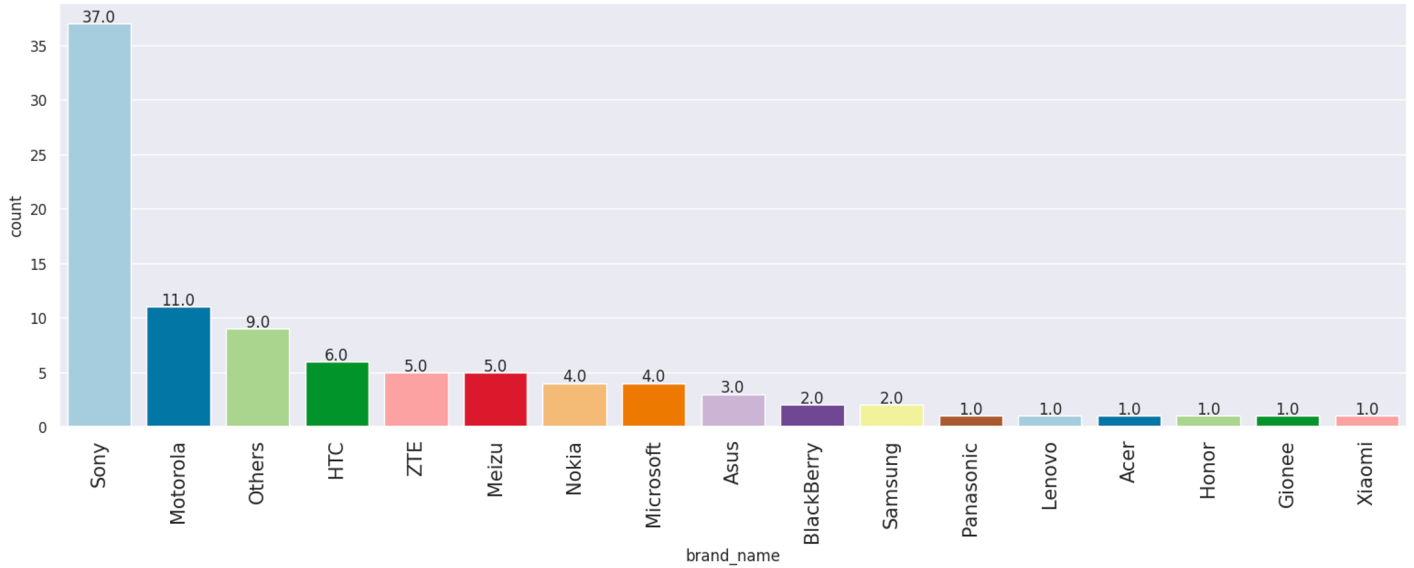
- Huawei and Samsung, being popular brands, are associated with larger screen sizes.
- Spice and Microsoft, less popular brands, have larger screen sizes.

Bivariate Analysis - Selfie Camera vs Brand Name



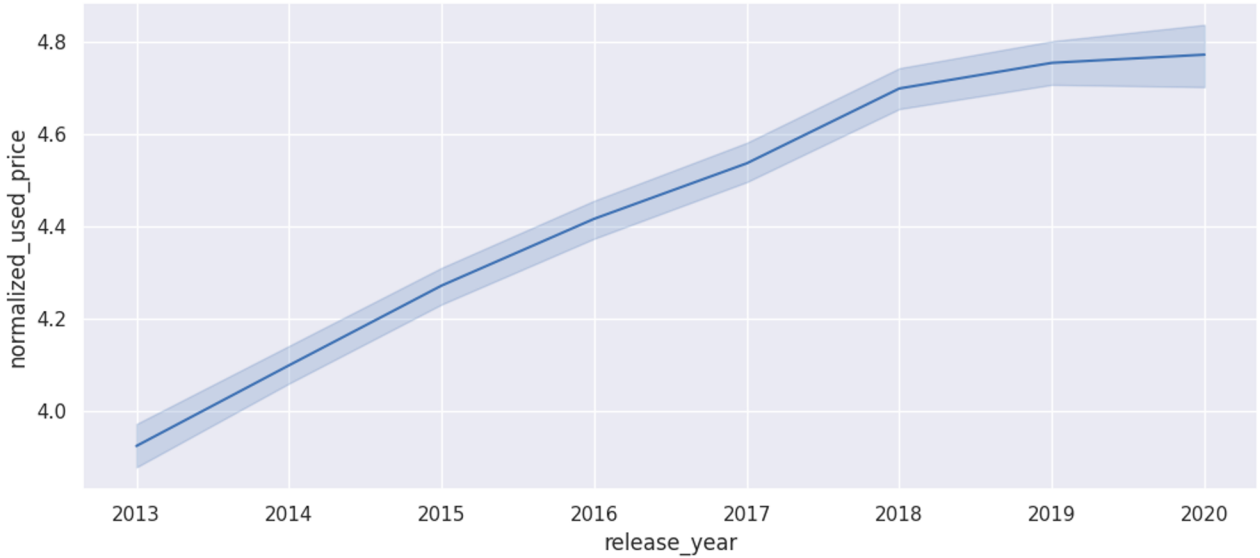
- Huawei and Vivo are the most preferred brands among buyers seeking phones with selfie cameras.
- Panasonic and Acer are the less preferred brands among buyers seeking phones with selfie cameras.

Bivariate Analysis - Main Camera vs Brand Name



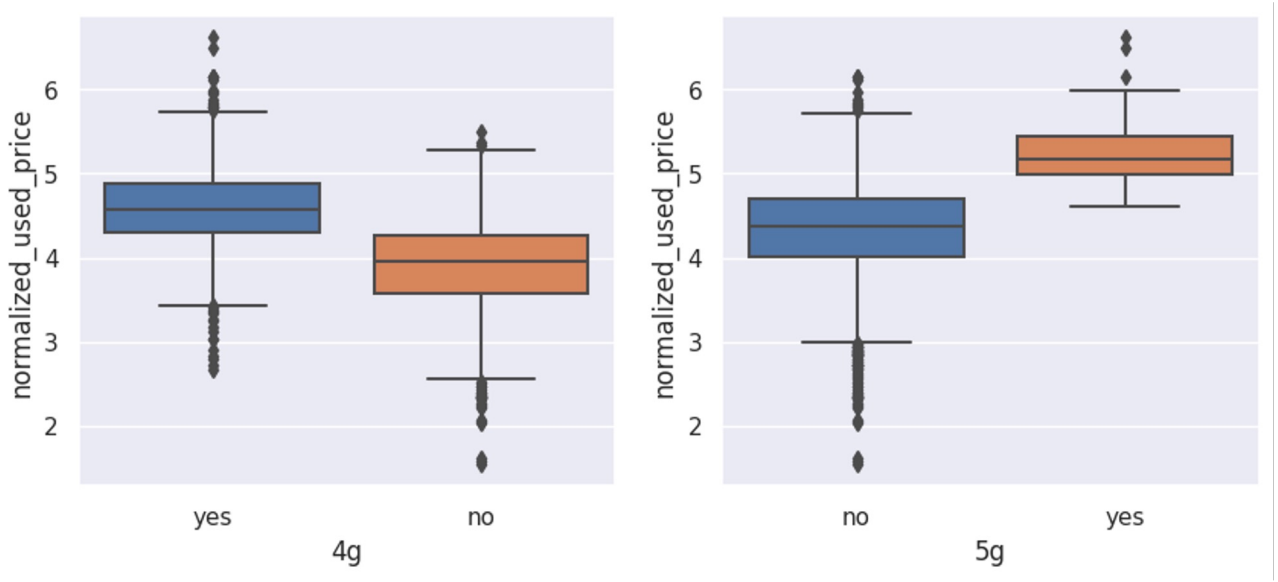
- Sony stands out as the preferred brand for phones with the main camera.
- Xiaomi is the least preferred brand for phones with the main camera.

Bivariate Analysis - Release Year vs Normalized Used Price



- There's a trend of newer phones having higher prices, ranging from 2013 to 2020.

Bivariate Analysis - Normalized Used Price vs 4G and 5G



- The price for 5G phones is higher than that for 4G phones.

Data Preprocessing

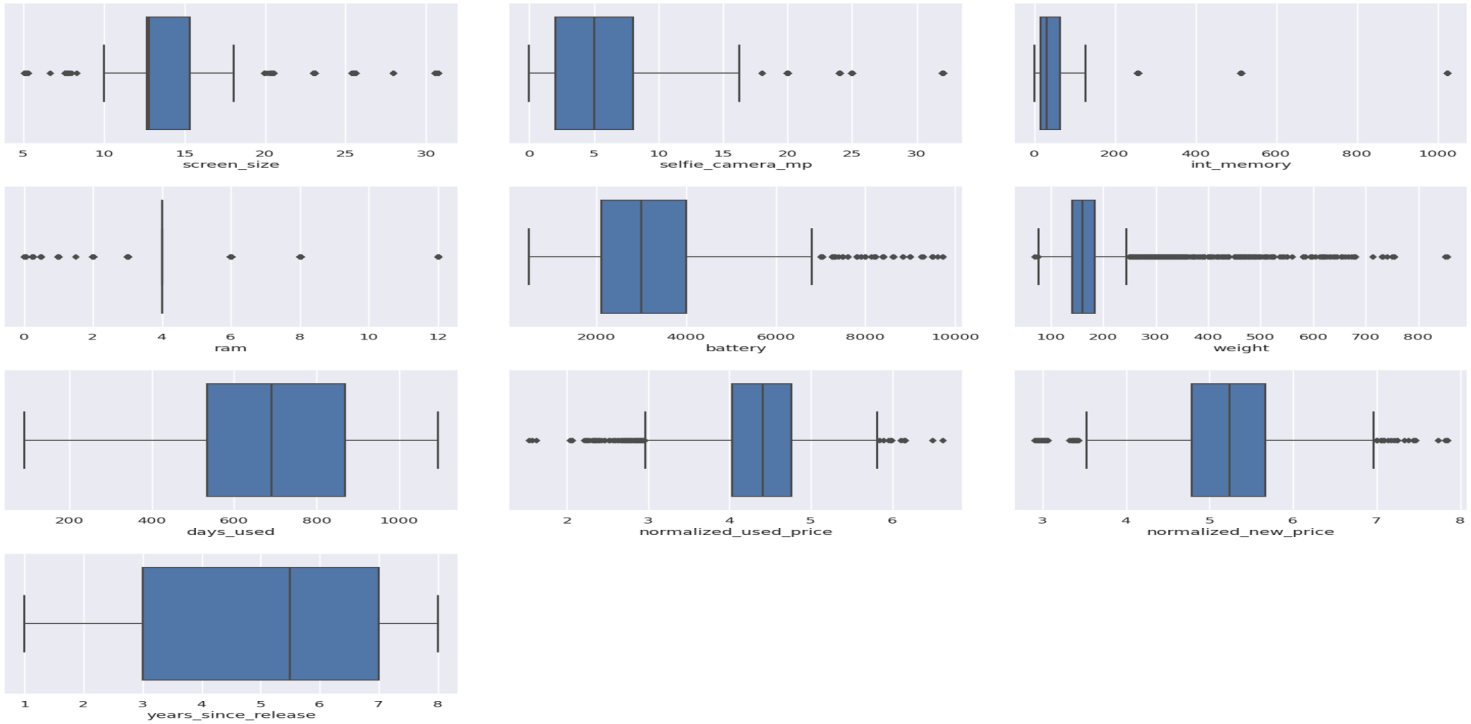
Data Preprocessing - Missing Value Imputation

- No duplicate data was found in our dataset.
- Missing value treatment
    - For the variables, main_camera_mp, selfie_camera_mp, int_memory, ram, battery, and weight, we will impute the missing values with median by grouping the data on release_year and brand_name.
    - For the variables, main_camera_mp, selfie_camera_mp, battery, and weight, we will impute the missing values with median by grouping the date on brand_name.
    - We will fill the remaining missing values in the main_camera_mp column by the column median.

Data Preprocessing - Feature Engineering

- Feature Engineering
  - We will create a new column years_since_release from the release_year column.
  - We will consider the year of data collection, 2021, as the baseline.
  - We will drop the release_year column.

Data Preprocessing - Outlier Check



- We've identified several outliers in the data, but no treatment will be applied.

Data Preprocessing - Data Preparation for Modeling

We aim to predict the normalized price of used devices. To achieve this:

a. We'll encode categorical features.
b. The data will be split into train and test sets for model evaluation.
c. A Linear Regression model will be built using the train data, and we'll assess its performance.

- After creating the dummy variables, there are 5 rows and 89 columns.
- The data is divided into a 70:30 ratio for training and testing purposes.
  - The training data consists of 2417 rows.
  - The testing data consists of 1037 rows in test data.

Model Building - Linear Regression

Model Building - Model Performance Check

- We will be using the OLS model with the train and test data to get the parameters of the linear regression model.
- The values from the training data show the R- squared is .85, and the adj R-Squared is .84. This is good since the values are close.
- The values from the testing data show the R- squared is .84, and the adj R-Squared is .83. This is good since the values are close.
- Training Data

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.226107	0.17736	0.849942	0.844202	4.247918

- Testing Data

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.239404	0.185272	0.841094	0.82616	4.495925

Testing the Assumptions

Test for Multicollinearity

- We will use the VIF to test for multicollinearity.
    - There was only one column that showed a VIF score > 5.
    - The screen_size column showed a VIF of 8.07, which is moderate multicollinearity, and will be dropped because the score is > 5.

- We'll check p-values. If they're > .05, we'll consider dropping those predictors. A p-value above 0.05 suggests the variable doesn't significantly affect the target.
    - The highest p-value is 0.040 for 'main_camera_mp_10.0.' We won't drop any predictors since all p-values are < 0.05, signifying their significance.

Test for Multicollinearity cont'd

- Checking the OLS model with the updated data.
- Training Set

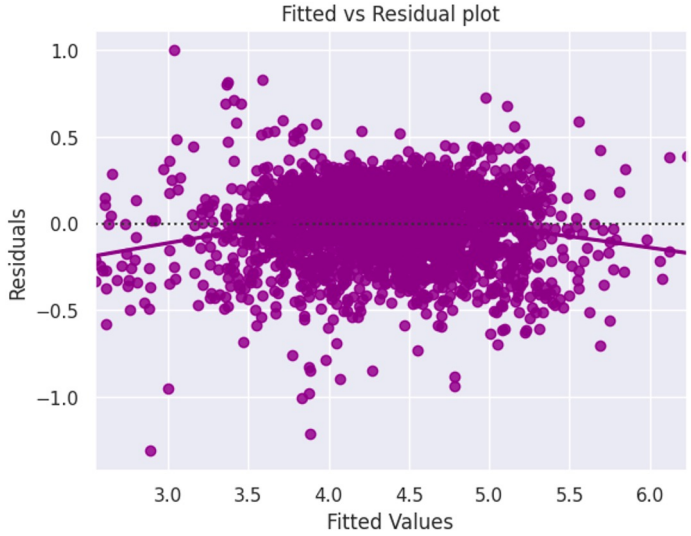
	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.22812	0.179195	0.847258	0.845532	4.289722

- Testing Set

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.238456	0.184816	0.842349	0.83813	4.488782

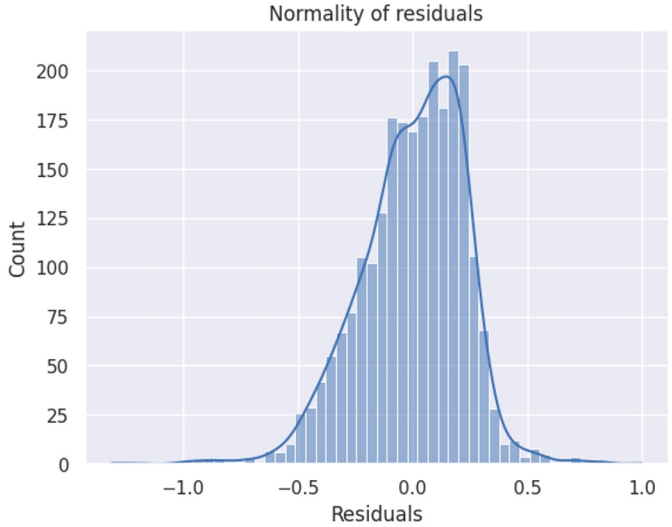
- The values in in both the training and testing set are comparable, thus indicating that the model is not overfitting.

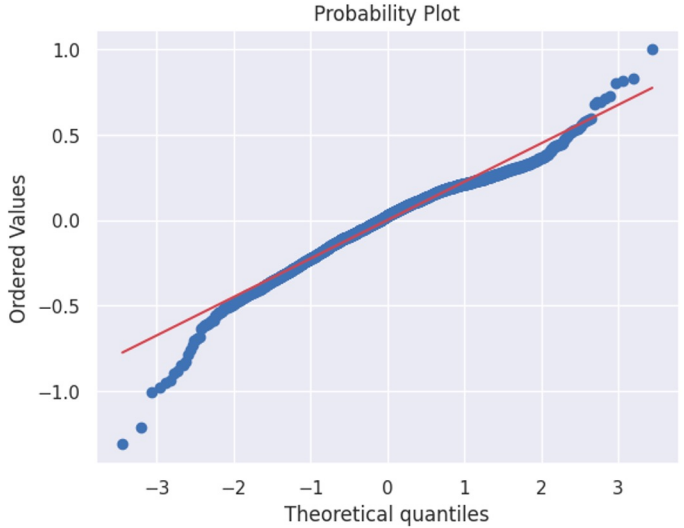
Test for Linearity and Independence



- The scatter plot shows the distribution of residuals (errors) vs fitted valued (predicted values).
- The graph doesn't show a linear pattern, indicating a non-linear and independent relationship.

Test for Normality





- The histogram displays a bell shape with some skewness.
- The line mostly resembles a normal distribution, except for the tails.
- The Shapiro-Wilk test indicates a p-value > 0.05, allowing us to assume normality.

Test for Homoscedasticity

- We use the Goldfeld-Quandt test to assess homoscedasticity.
- The p-value is 0.000005, which is less than 0.05, indicating statistical significance.
- The p-value is greater than 0.05, confirming that the residuals exhibit homoscedasticity, satisfying the assumption.

Model Performance Evaluation

Model Performance Summary

- With all linear regression assumptions met, the final model can be analyzed, and used for prediction.
- Training Set

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.22812	0.179195	0.847258	0.845532	4.289722

- Testing Set

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.238456	0.184816	0.842349	0.83813	4.488782

- Our model accounts for around 85% of the variation in the data.
- Both the train and test RMSE and MAE are low and comparable, indicating no overfitting.
- Our model predicts within 4.5% of the normalized used price, as indicated by the MAPE on the test set.

Model Performance Summary

There are key variables in our OLS model that we could look at that significantly influences the outcomes, analysis and predictions.

1. The 95% confidence interval for 'screen_size' is [0.018, 0.030], providing a range in which we can be reasonably confident about the true impact of screen size on our target variable.

1. The p-value for battery is 0.034, indicating that it is statistically significant in predicting price.

1. The coefficient for weight 0.0011, representing its impact on our target variable, normalized used price.

Executive Summary

Executive Summary - Conclusions

- Model Fit and Significance:
  - Adjusted R-squared of 0.845 indicates a good model fit.
  - Constant coefficient of 1.74 sets a baseline for the dependent variable.
- Variable Selection:
  - No variables were dropped as all p-values are below 0.05.
- Model Accuracy and Overfitting:
  - Low RMSE and MAE values (both 0.18) indicate accurate predictions.
  - Train and test metrics are comparable, suggesting no overfitting.
- Variable Importance and Multicollinearity:
  - 'screen_size' was dropped due to a VIF of 8, addressing multicollinearity.
  - Marginal improvement in adjusted R-squared after dropping variables.
- Assessment of Linearity:
  - Graph reflects no linear pattern, suggesting non-linear and independent relationships.

Executive Summary - Conclusions continued

- Statistical Significance:
  - Overall model is statistically significant (p-value: 0.000005).
- Model Interpretability:
  - Model explains about 85% of the variation in the data.
- Prediction Accuracy:
  - Low MAPE on the test set (within 4.5% of the normalized used price) indicates effective predictions.
- Overall Model Evaluation:
  - Model (olsmodel_final) is deemed good for both prediction and inference purposes, considering statistical significance, accurate predictions, and interpretability.

Executive Summary - Recommendations

- Improve Camera Features:
  - Enhance features like camera quality ('main_camera_mp_10.0') to attract more customers.
- Promote Sustainability:
  - Emphasize the environmental benefits of purchasing used devices in marketing campaigns.
- Stay Competitive:
  - Regularly analyze competitors to adapt pricing strategies and stay competitive in the market.
- Educate Customers:
  - Create customer-friendly materials explaining factors influencing device pricing, with a focus on important features like camera quality.
- Optimize Supply Chain:
  - Continuously improve the supply chain to reduce operational costs and enable more competitive pricing.
- Highlight Predictive Capability:
    - Design marketing campaigns emphasizing the model's ability to predict device pricing accurately.

These recommendations can contribute to ReCell's success in the used and refurbished device market.