ReCell

Project 3 – Supervised Learning-Foundations

April 13, 2025

Submitted By:
Alex Kyeremateng Botwe

Contents / Agenda

	Topics	Page No.
	Executive Summary	5
	Business Problem Overview and Solution Approach	10
	EDA Results	14
	Data Preprocessing	37
	Model Performance Summary	52
	Appendix	57

List of Tables

List of Figures

List of Figures

Executive Summary

- There has been a considerable growth over the past decade in the global market in the demand for used and refurbished smartphones and tablets which is driven by both consumers and businesses' demands for cost-effective and sustainable device as an alternative to new devices which are more expensive. The 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. Based on these predictions, ReCell as a startup company is positioning itself to capitalize on this opportunity.
- As part of its go-to-market strategy, ReCell aims to implement a machine learning-based pricing engine to dynamically and accurately price second-hand devices.

Executive Summary (cont.)

- **Business Insights:**
  a. **Duration Strongly Impacts Price:** Devices with longer days on the market show a clear decline in normalized used price

  b. **New Price Is a Strong Driver for Used Price:** The normalized new price has the largest positive impact on normalized used price

  c. **Technical Specs Drive Value Retention:**
     - Devices with better cameras tend to have higher used prices, especially for main cameras
     - Larger batteries drive higher used prices
     - Performance specifications like RAM correlates positively with resale prices
     - Operating System affects the value of the device
     - Network Capabilities like 4G and 5G has an influence on the value of the device

Executive Summary (cont.)

d. **Brand Influence Price Stability:** Brand perception drives resale value as some brands hold their value better.

e. **Newer Devices Retain More Value :** More expensive devices retain a higher portion of their value when resold.

- **Business Recommendations:**

  a. Bundle value enhancing features like better cameras, RAM, and 4G for higher resale value

  b. Leverage brand strength in pricing strategies, especially for brands with high resale reputation.

Executive Summary (cont.)

c. Regularly update models and highlight camera quality and performance in marketing

d. Buy phones with higher RAM, strong camera specs, and mainstream OS for better resale value

e. Avoid lesser-known OS variants or experimental devices if resale is a priority

f. Sell sooner rather than later since years since release significantly reduce value.

g. Focus inventory on phones with better specs, newer models, and recognized brands

h. Price 5G enabled devices carefully since customers may not be seeing the full value yet.

i. Consider bundling accessories or warranties to improve resale potential of older models

Executive Summary (cont.)

- **Conclusions:**
  - The closeness of training and test performance metrics (RMSE, MAE, MAPE) indicates no overfitting, suggesting that the model is robust and generalizes well to unseen data.
  - With R^2 of approximately 0.84, the model is highly reliable for predicting used device prices.
  - This model can therefore be confidently used for predicting normalized used prices.
  - ReCell can thus use this model to set trade-in values, forecast inventory recovery values and help consumers estimate device value over time.

Business Problem Overview and Solution Approach

- **Business Problem Overview:**

The used and refurbished smartphone/tablet market has experienced rapid growth, offering consumers cost-effective alternatives to new devices. With projections indicating a market value of $52.7 billion by 2023, ReCell is looking to strategically positioning itself by leveraging data and ML-driven insights.
One of the key challenges in this space is the lack of a standardized, dynamic pricing model. Pricing used devices is complex, as it depends on numerous factors such as brand, technical specifications, age, and usage patterns. Relying on manual pricing or flat-rate models can lead to lost revenue opportunities or overpricing which will deters buyers.

To address this challenge, ReCell seeks to implement a data-driven ML solution to:

- Predict the fair market value of used and refurbished devices
- Identify key drivers of price depreciation
- Enable scalable and consistent pricing across inventory

Business Problem Overview and Solution Approach (cont.)

- **Solution Approach**

To predict a fair market value and reveal depreciation drivers that affect used device prices, a **a machine learning model** was built using the following steps:

1. **Data Collection & Preparation**

   - Historical device data with the following parameters: RAM, camera MP, weight, Brand, OS, 4G/5G capability, battery, days used, screen size, new price and used price was collected

   - Data cleaning, normalization, and encoding of categorical variables was employed

   - Outlier detection, Feature engineering and Data preparation for modeling were performed

Business Problem Overview and Solution Approach (cont.)

2. **Exploratory Data Analysis (EDA)**

   - Explored feature distributions, correlations, and outliers
   - Identified potential predictors and multicollinearity risks

3. **Model Development & Refinement**

   - Applied Ordinary Least Squares (OLS) regression
   - Iteratively removed high VIF variables to address multicollinearity
   - Refined model using p-value-based backward elimination

Business Problem Overview and Solution Approach (cont.)

4. **Model Evaluation & Validation**

   - Evaluated performance on train and test split using: RMSE, MAE, R^2, Adjusted R^2 and MAPE

   - Achieved high generalizability (Test $R^2 = 0.838$, MAPE $\approx 4.6\%$)

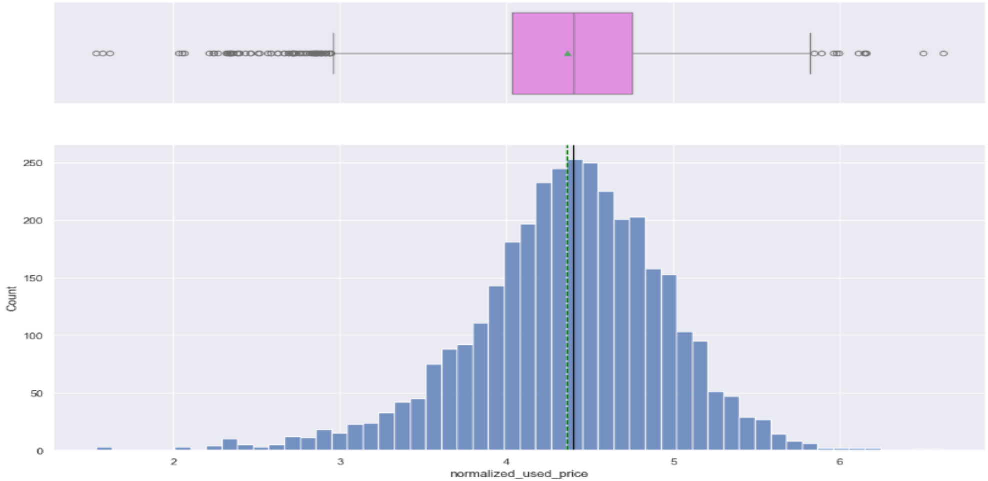
5. **Residual Diagnostics & Assumption Testing**

   - Validated linear regression assumptions: Linearity, independence, homoscedasticity, and normality test

   - Used residual plots, Shapiro-Wilk test

EDA – Univariate Analysis

- **Normalized Used Price**



Observations:

Fig 1: Univariate Analysis of Normalized Used Price

- This is an approximately normal distribution, mostly symmetrical around the mean and median
- There are many outliers on both ends possibly heavily used, outdated, or poor specifications or likely new releases or premium models that haven't depreciated much.
- Most used phones fall into a tight range — centered around the 4.0 to 4.8 mark

Link to Appendix slide on data background check

- Normalized New Price

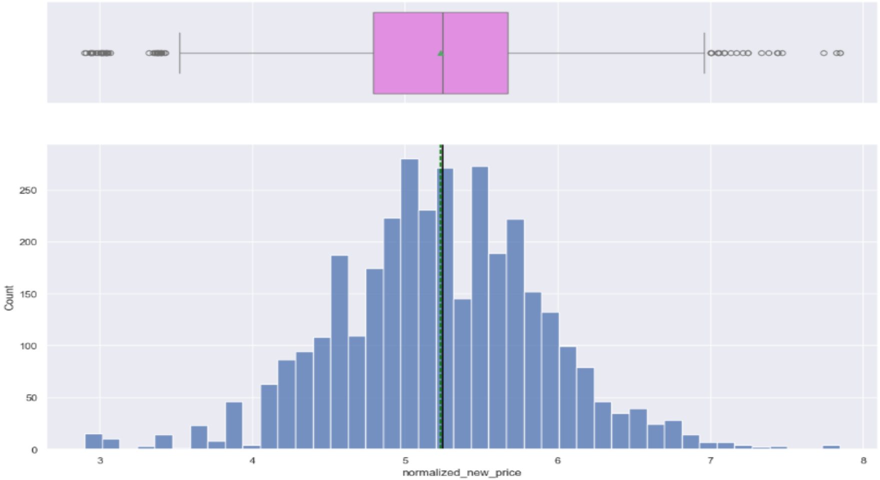


Fig 2: Univariate Analysis of Normalized New Price

Observations:

- This is a slight Right-Skewed Distribution. The data is nearly symmetrical but with a slightly longer right tail. This means there are more expensive outliers than cheap ones for new devices.
- New devices are generally sold around a consistent price point, but some premium models push the distribution upward.
- The market for used devices is broader, with more devices depreciating heavily or priced very low.

Link to Appendix slide on data background check

- Screen Size

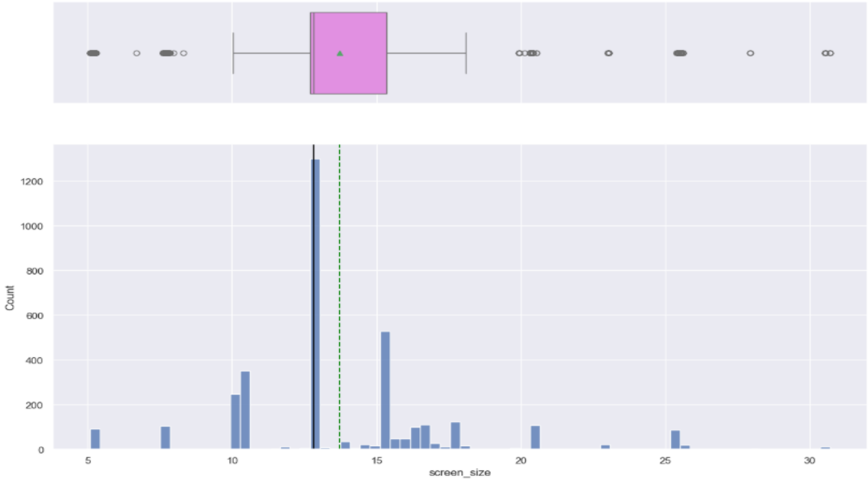


Fig 3: Univariate Analysis of Screen Size

Observations:

- This is a right-skewed distribution with numerous outliers on both ends, especially the high end
- This means most phones have screen sizes clustered around 12–14 inches.
- A few very large devices (maybe tablets or incorrect entries) stretch the distribution.

Link to Appendix slide on data background check

- **Main Camera MP**

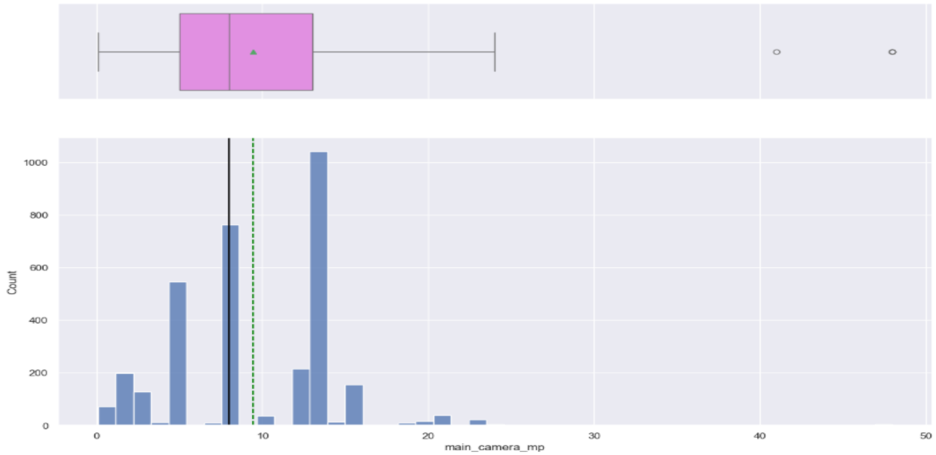


Fig 4: Univariate Analysis of Main Camera MP

Observations:

- This is a positively skewed distribution indication more phones have lower MP cameras, but with a few high-end ones go all the way up to about 48 MP
- There's a small number of phones with very low MP values (0–2 MP) and very high outliers (above 30 MP).

Link to Appendix slide on data background check

- ## Selfie Camera MP

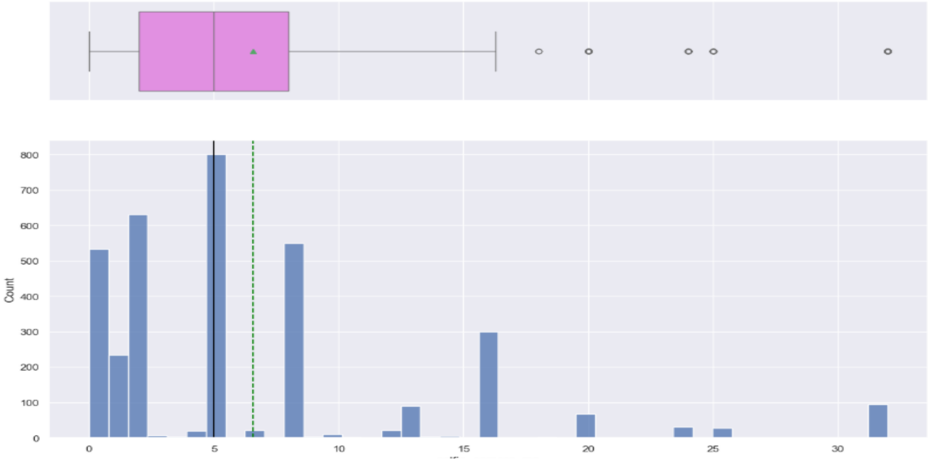


Fig 5: Univariate Analysis of Selfie Camera MP

Observations:

- This is Right-Skewed or Positively Skewed distribution
- Most devices have lower selfie camera megapixels (MP)
- The majority of devices have a selfie camera between 2 MP and 8 MP.
- There is a small but significant segment of premium devices with very high-resolution selfie cameras (above 15 MP)

Link to Appendix slide on data background check

- **Internal Memory**

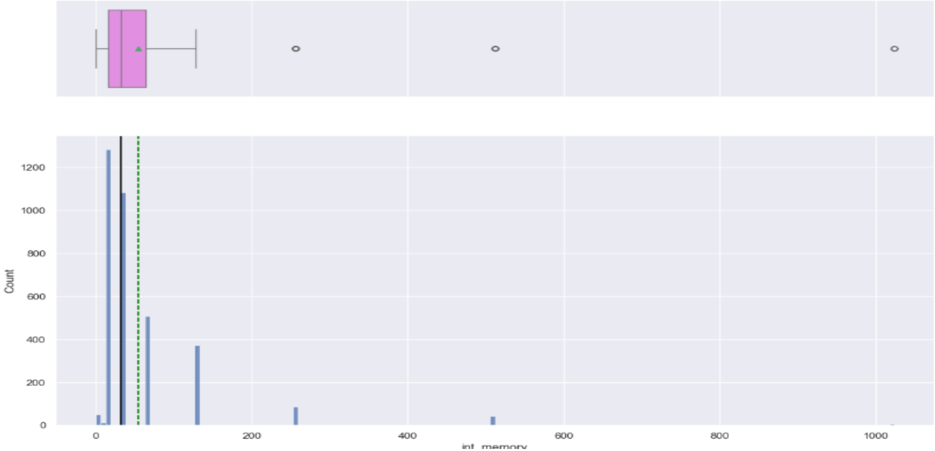


Fig 6: Univariate Analysis of Internal Memory

Observations:

- This is a right-skewed positively skewed distribution.
- Majority of devices have a low int_memory. Long tail to the right indicates a few with very large int_memory

Link to Appendix slide on data background check

- RAM

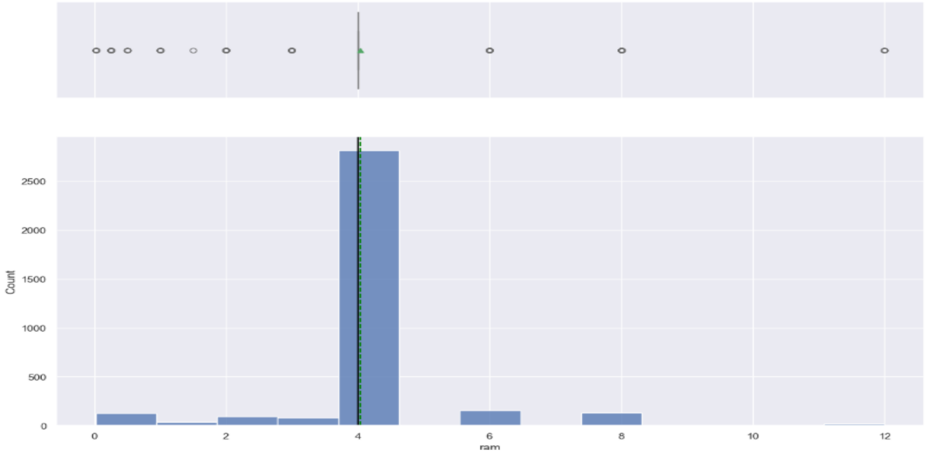


Fig 7: Univariate Analysis of RAM

Observations:

- This is showing a highly left-skewed distribution also called negatively skewed.
- This distribution suggest a higher number of devices have ram size of about 4GB with few on the lower end ram size

Link to Appendix slide on data background check

- Weight

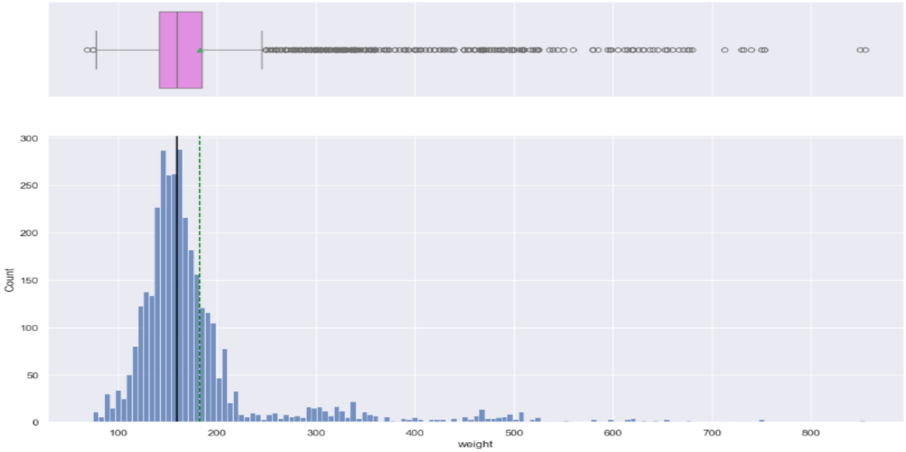


Fig 8: Univariate Analysis of Weight

Observations:

- The distribution is Right-skewed indicating that most phones are relatively lightweight between 100g and 200g.
- The presence of outliers suggests there are a few exceptionally heavy devices

Link to Appendix slide on data background check

- Battery

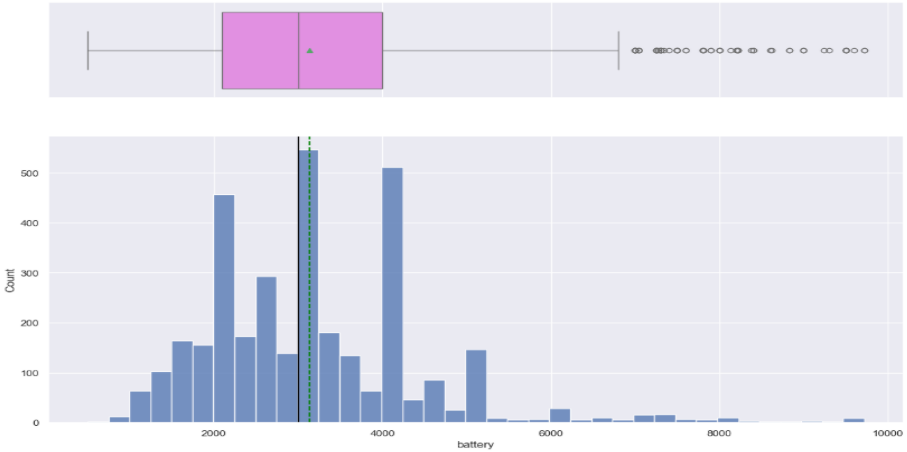


Fig 9: Univariate Analysis of Battery

Observations:

- This is a multimodal and right-skewed distribution with a high concentration of battery capacity of 2100 mAh and another around 4000 mAh.
- There is also a long right tail suggesting a small number of high-capacity devices

Link to Appendix slide on data background check

- Days Used

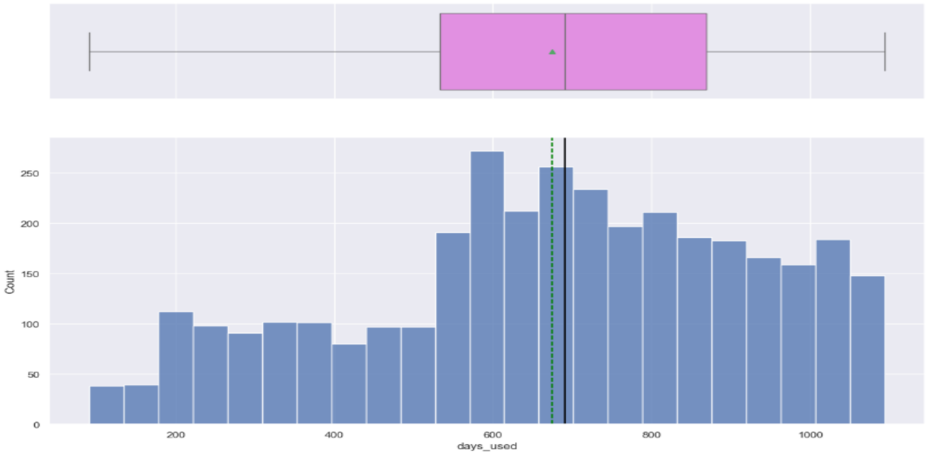


Fig 10: Univariate Analysis of Days Used

Observations:

- This is a roughly uniform with a slightly right-skewed tail distribution
- Most devices have been used between 600 to 1000 days (i.e. roughly 1.5 to 3 years), aligning with typical smartphone lifecycles.
- Some devices are lightly used less than 300 days which could possibly be newer models or underused devices.
- Very few extreme outliers on either end, suggesting a healthy spread and minimal noise

Link to Appendix slide on data background check

- **Brand Name**

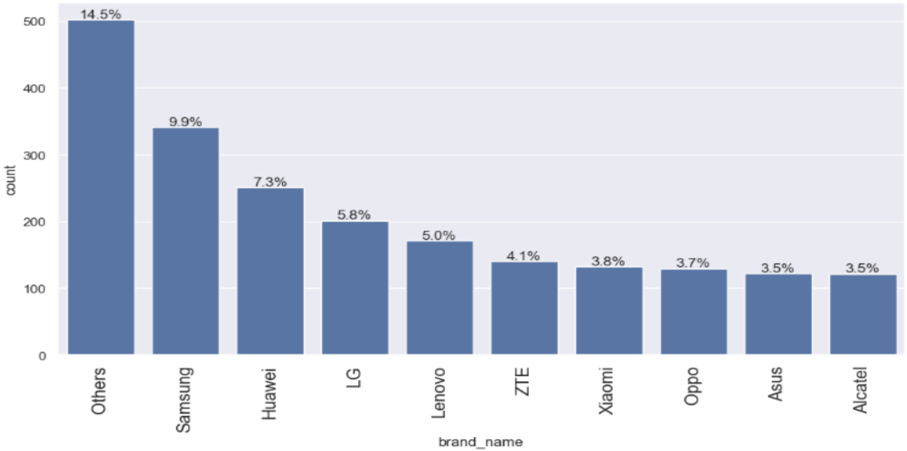


Fig 11: Univariate Analysis of Brand Name

Observations:

- Other brands of device contributes the largest share of 14.5% followed by Samsung showing 9.9% share and Huawei with 7.3%. This suggests that there may be many low-frequency brands grouped into this bucket
- The used device market is highly fragmented beyond the top brands.
- Brands like Xiaomi, Oppo, Asus, and Alcatel have smaller shares (~3.5% to 3.8%). This indicates these brands are relatively less available in the secondary market in your data.

Link to Appendix slide on data background check

- OS

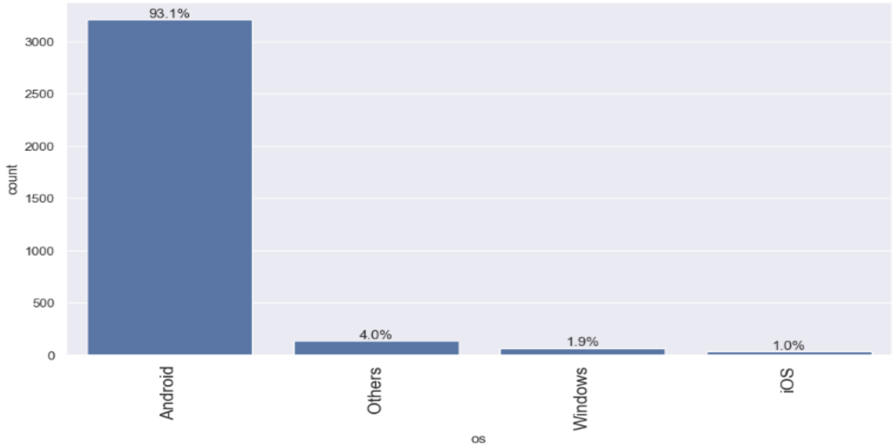


Fig 12: Univariate Analysis of OS

Observations:

- Android is the most popular with a market share of 93.1% with iOS devices only constituting 1.0% of share
- This suggests that the used device market is heavily Android-dominated.

Link to Appendix slide on data background check

- **4g**

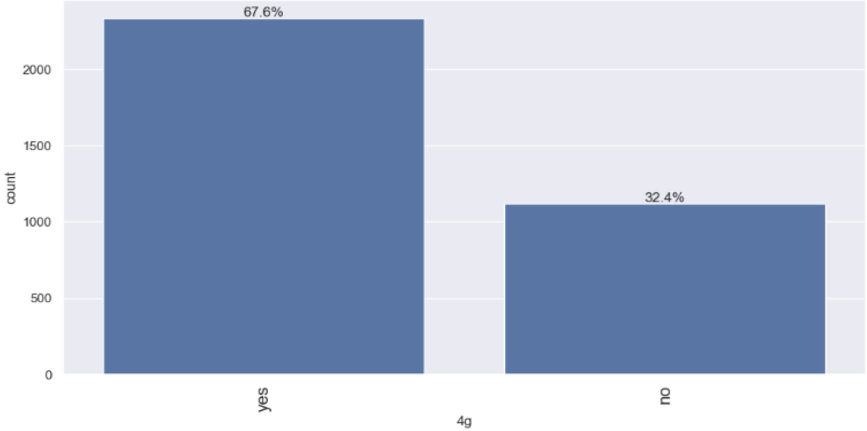


Fig 13: Univariate Analysis 4g

Observations:

- Around 67.6% of the devices support 4G, while 32.4% do not.
- This indicates that 4G support is a common feature among used devices, but a significant portion of devices still lack 4G capability.

Link to Appendix slide on data background check

- 5g

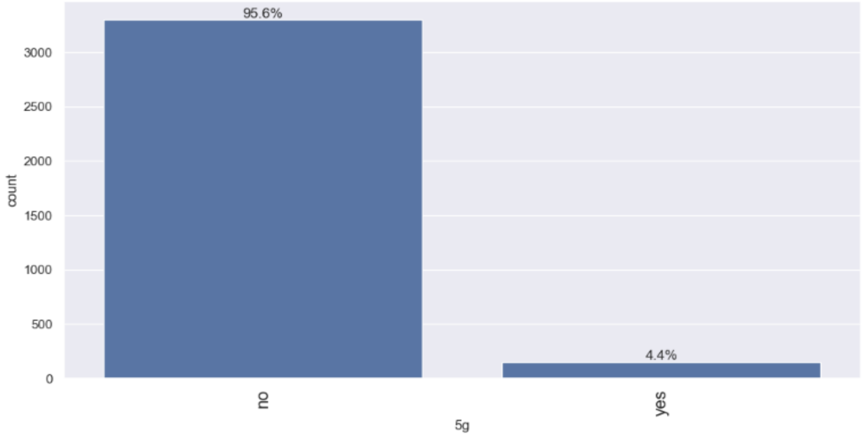


Fig 14: Univariate Analysis of 5g

Observations:

- A significant majority of devices (95.6%) do not support 5G.
- Only 4.4% of the devices in the dataset have 5G capability.

Link to Appendix slide on data background check

- # Release Year

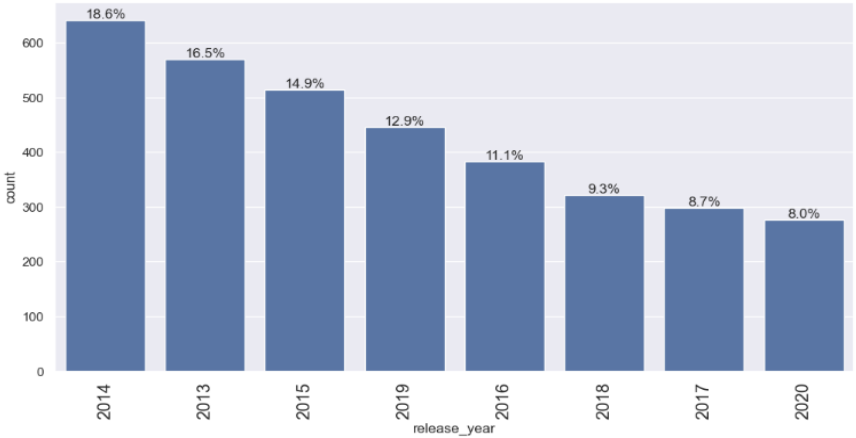


Fig 15: Univariate Analysis of Release Year

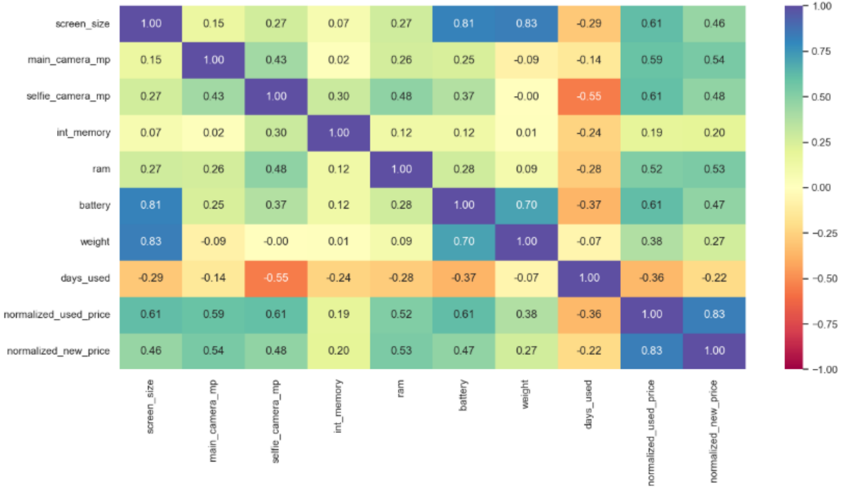
Observations:

- The majority of the devices were released between 2013 and 2015, contributing about 50% of the total devices.
- 2014 has the highest number of devices (18.6%) followed closely by 2013 (16.5%)
- After 2015, the number of devices declines steadily, indicating fewer older devices in the dataset.
- Recent years (2019-2020) have a smaller share (around 8%-13%), possibly due to fewer devices entering the used device market from recent releases.

Link to Appendix slide on data background check

EDA – Bivariate Analysis

- **Correlation Check**



	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

Observations:

Fig 16: Correlation Plot

- Internal Memory is positively correlated to RAM suggesting that Devices with higher internal memory usually come with more RAM.
- Normalized New Price is positively correlated to RAM and Internal Memory indicating that Newer, more expensive devices tend to have better specification with more RAM and storage
- Normalized Used Price also have a positive correlation with Normalized New Price suggesting that Used prices are strongly tied to the original price, although some depreciation is expected.

Link to Appendix slide on data background check

- RAM vs Brand Name

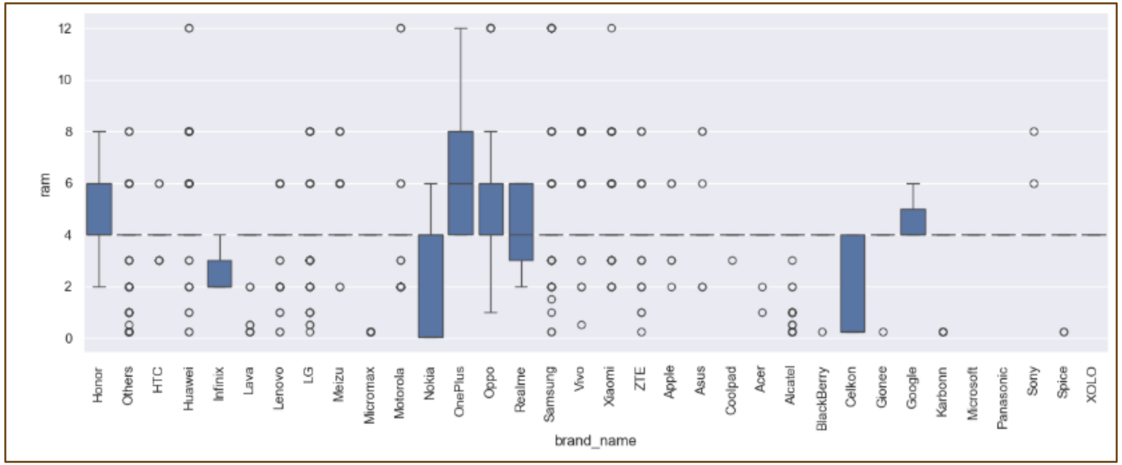


Fig 17: Bivariate Analysis of RAM vs Brand Name

Observations:

- Apple devices have a relatively narrow range of RAM, centered around lower values compared to Android brands.
- Samsung, OnePlus, and Xiaomi generally offer higher RAM configurations with a wider range
- Realme, Vivo, and Oppo also show diverse RAM offerings, suggesting they cover various market segments.
- Brands like Infinix and Tecno tend to cluster around lower RAM values

Link to Appendix slide on data background check

- **Weight vs Brand Name**

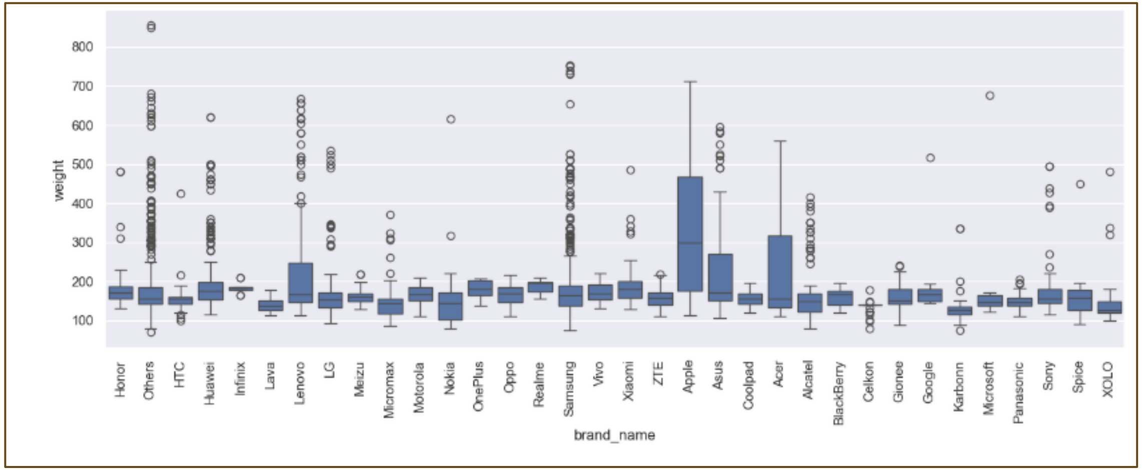


Fig 18: Bivariate Analysis of Weight vs Brand Name

Observations:

- There are 341 devices in the dataset with a battery capacity greater than 4500 mAh
- Samsung, Xiaomi, and Realme offer a good range of high-battery devices with relatively balanced weight
- A few brands like OnePlus and Infinix show some outliers with notably heavy devices
- Lenovo and Huawei show higher median weights for large-battery devices

Link to Appendix slide on data background check

Devices with Large Screen greater than 6inch

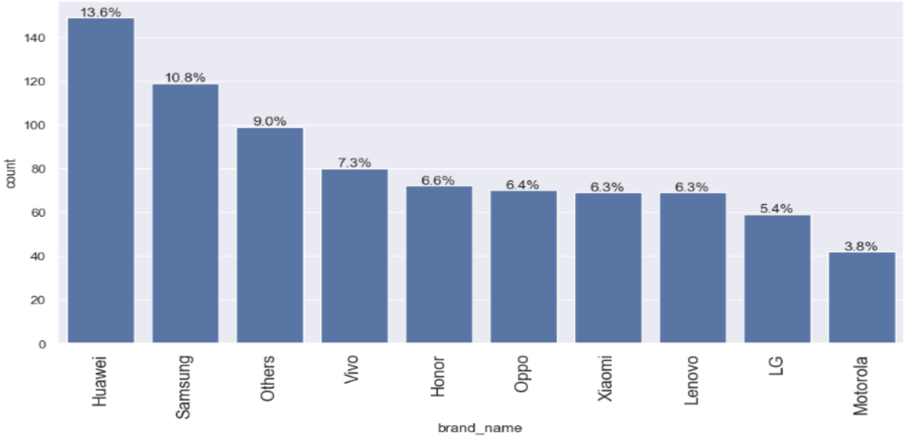


Fig 19: Devices with Large screens greater than 6inch

Observations:

- There are 1,099 devices in the dataset with a large screen greater than 6 inches making them suitable for entertainment like video streaming or gaming.
- Huawei holds the top spot of 13.6%, making up a substantial portion of large-screen offerings.
- Samsung and Vivo are also key players, likely offering a mix of affordable and mid-range models with big displays.
- Honor, Oppo, Xiaomi, and Lenovo are clustered showing moderate emphasis on larger screens.

Link to Appendix slide on data background check

- **Devices with Selfie Camera greater than 8 MP**

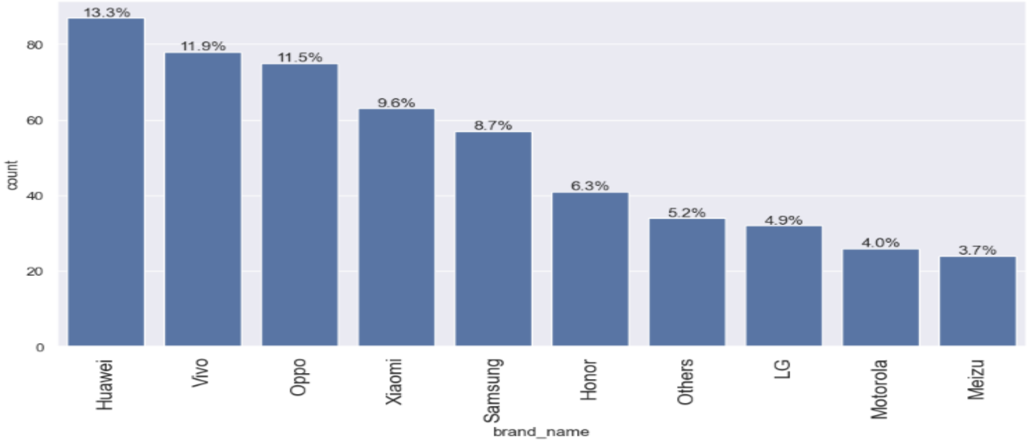


Fig 20: Devices with Selfie Cameras greater than 8MP

Observations:

- There are 655 devices in the dataset with selfie cameras greater than 8 MP, making them attractive options for selfie lovers and social media enthusiasts.
- Huawei is the top contributor in the high-resolution selfie camera segment
- Vivo and Oppo closely follow reinforcing their market position as selfie and photography-focused brands
- Xiaomi and Samsung also show strong representation

[*Link to Appendix slide on data background check*]

- **Devices with Main Camera greater than 16 MP**

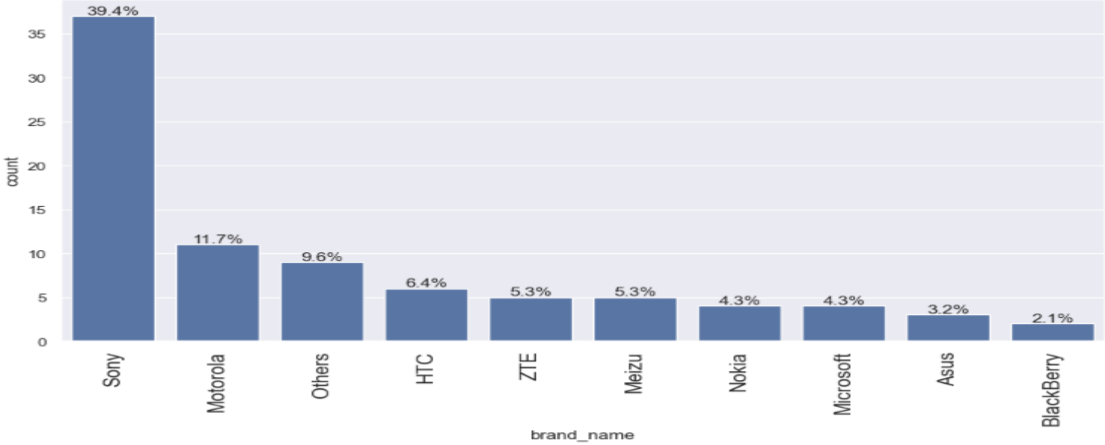


Fig 21: Devices with Main Camera greater than 16MP

Observations:

- There are 94 devices in the dataset with main (rear) cameras greater than 16 MP, indicating a more select group of devices focused on high-resolution photography.
- Sony dominates the segment with 39.4% reflecting its strength in camera sensor technology
- Motorola follows in a distant second with 11.7%
- Brands like HTC, ZTE , Meizu, Nokia, and Microsoft make up smaller shares

Link to Appendix slide on data background check

- **Price of Used Devices vs Release Year**

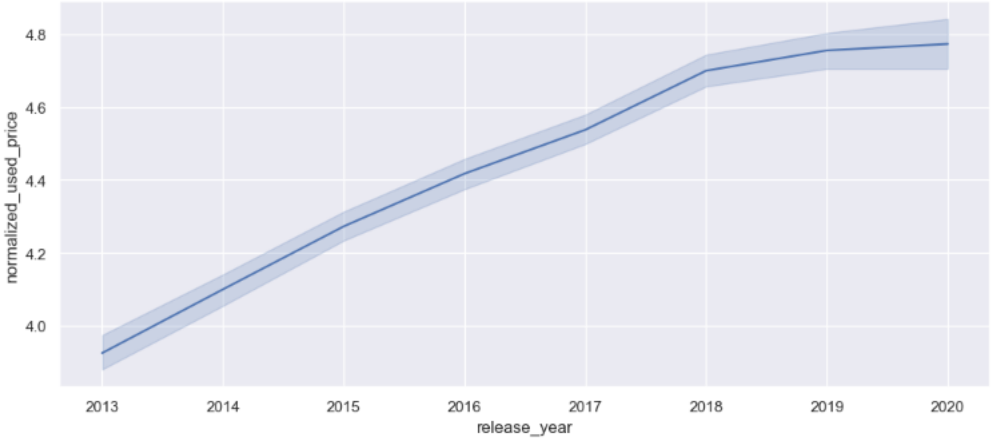


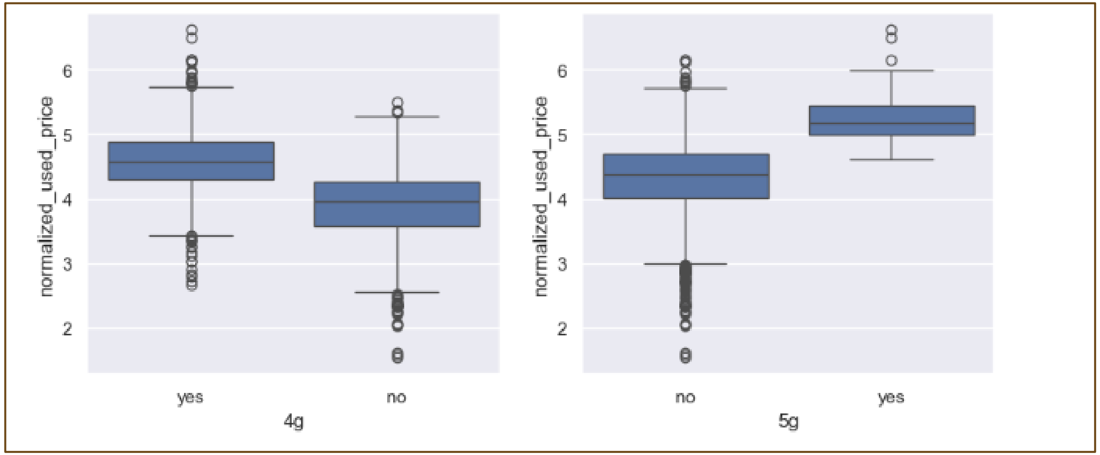
Fig 22: Prices of Device vs Release Year

Observations:

- Newer devices from 2020 to 2022 generally have higher used prices, which is expected as they're closer to their original launch and retain more value.
- There's a sharp decline in used price for devices released before 2018, indicating depreciation over time
- The pricing curve is relatively smooth, showing a typical depreciation trend rather than sudden drops, which suggests a stable second-hand market.

Link to Appendix slide on data background check

- **Price of Used Devices vs Release Year**



Observations:

Fig 23: Devices with Large screens greater than 6inch

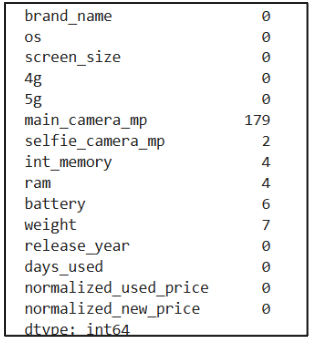
- 4G Support: Most devices support 4G, and they show a wide spread in prices, indicating a variety of models from budget to premium. Devices without 4G tend to have significantly lower used prices, likely due to older technology or limited usability today.
- 5G Support: Devices with 5G support are clearly more expensive on average in the used market. This reflects their newer release dates, better hardware, and growing demand for future-proof connectivity.
- Non-5G devices dominate the dataset but come at more affordable prices.

Link to Appendix slide on data background check

Data Preprocessing

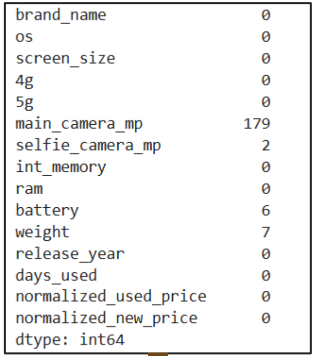
- ## Missing Value Treatment

Missing values in the dataset after checking



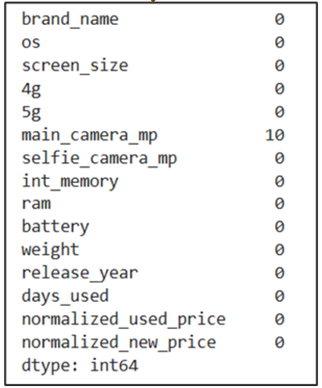
```
brand_name              0
os                      0
screen_size             0
4g                      0
5g                      0
main_camera_mp        179
selfie_camera_mp        2
int_memory              4
ram                     4
battery                 6
weight                  7
release_year            0
days_used               0
normalized_used_price   0
normalized_new_price    0
dtype: int64
```

Imputing missing values in the data by the column medians grouped by release_year and brand_name



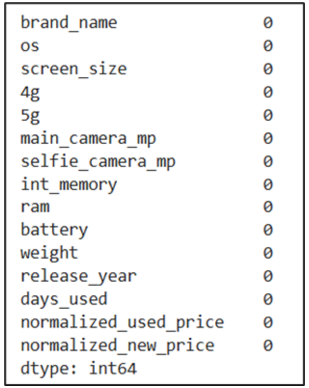
```
brand_name              0
os                      0
screen_size             0
4g                      0
5g                      0
main_camera_mp        179
selfie_camera_mp        2
int_memory              0
ram                     0
battery                 6
weight                  7
release_year            0
days_used               0
normalized_used_price   0
normalized_new_price    0
dtype: int64
```

Imputing missing values in the dataet by the column medians grouped by brand_name



```
brand_name              0
os                      0
screen_size             0
4g                      0
5g                      0
main_camera_mp         10
selfie_camera_mp        0
int_memory              0
ram                     0
battery                 0
weight                  0
release_year            0
days_used               0
normalized_used_price   0
normalized_new_price    0
dtype: int64
```

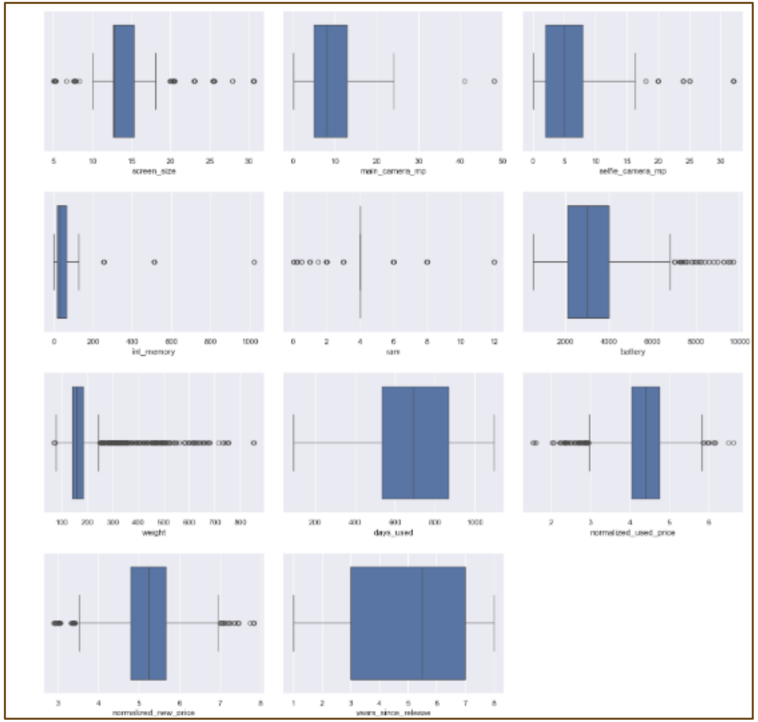
Imputing missing values in the data by the column medians grouped by main_camera_mp



```
brand_name              0
os                      0
screen_size             0
4g                      0
5g                      0
main_camera_mp          0
selfie_camera_mp        0
int_memory              0
ram                     0
battery                 0
weight                  0
release_year            0
days_used               0
normalized_used_price   0
normalized_new_price    0
dtype: int64
```

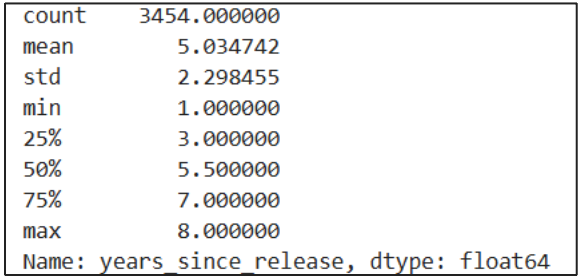
All missing values treated in the dataset

- **Outlier Check**

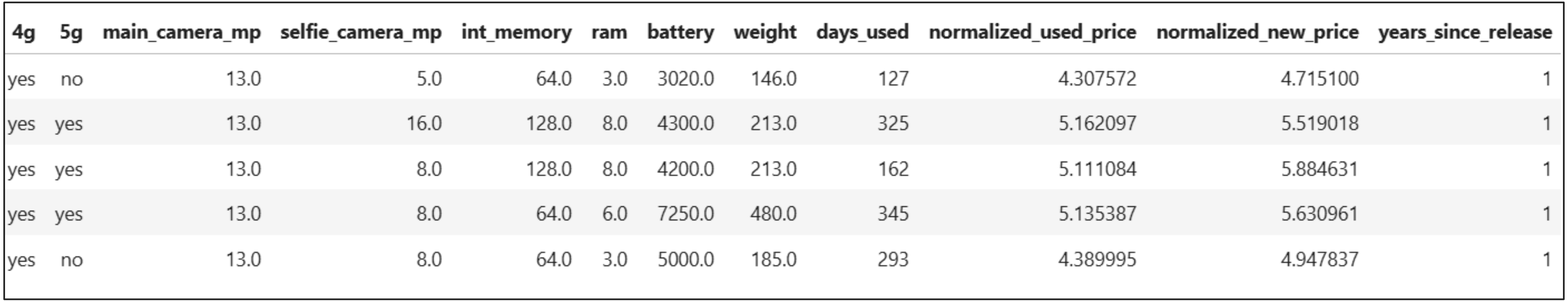


- Outliers at the higher end, likely due to high specification device
- Some possible inconsistencies in ram
- Outliers are mostly at the top, suggesting premium devices with high-megapixel cameras.
- A few extreme values, especially in battery
- Weight also has heavy outliers, possibly due to devices with large screens
- Some large-screen outliers are also detected

- Feature engineering



```
count    3454.000000
mean        5.034742
std         2.298455
min         1.000000
25%         3.000000
50%         5.500000
75%         7.000000
max         8.000000
Name: years_since_release, dtype: float64
```

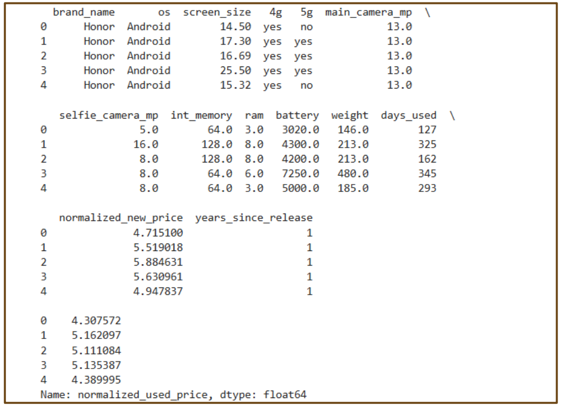


4g	5g	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	days_used	normalized_used_price	normalized_new_price	years_since_release
yes	no	13.0	5.0	64.0	3.0	3020.0	146.0	127	4.307572	4.715100	1
yes	yes	13.0	16.0	128.0	8.0	4300.0	213.0	325	5.162097	5.519018	1
yes	yes	13.0	8.0	128.0	8.0	4200.0	213.0	162	5.111084	5.884631	1
yes	yes	13.0	8.0	64.0	6.0	7250.0	480.0	345	5.135387	5.630961	1
yes	no	13.0	8.0	64.0	3.0	5000.0	185.0	293	4.389995	4.947837	1

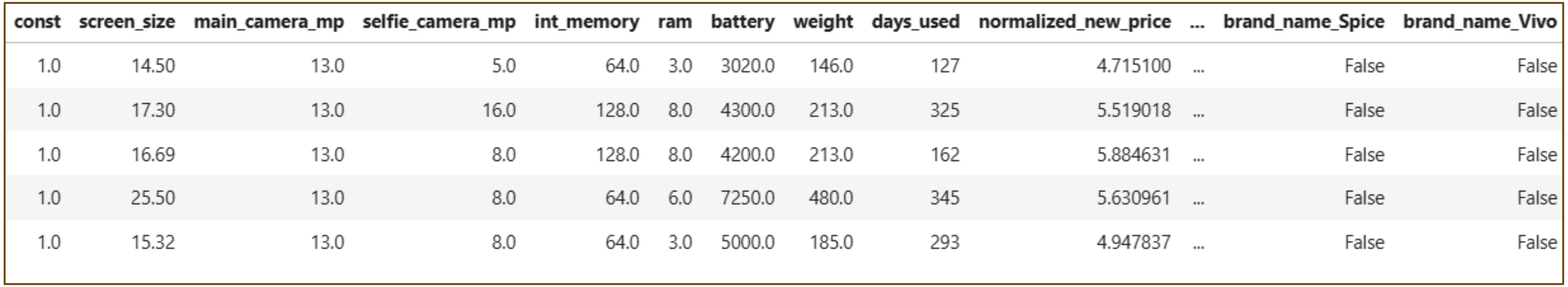
This indicates the new column showing the years_since_release and its is a float64 data type

Data preparation for modeling

Defining Dependent and Independent Variables

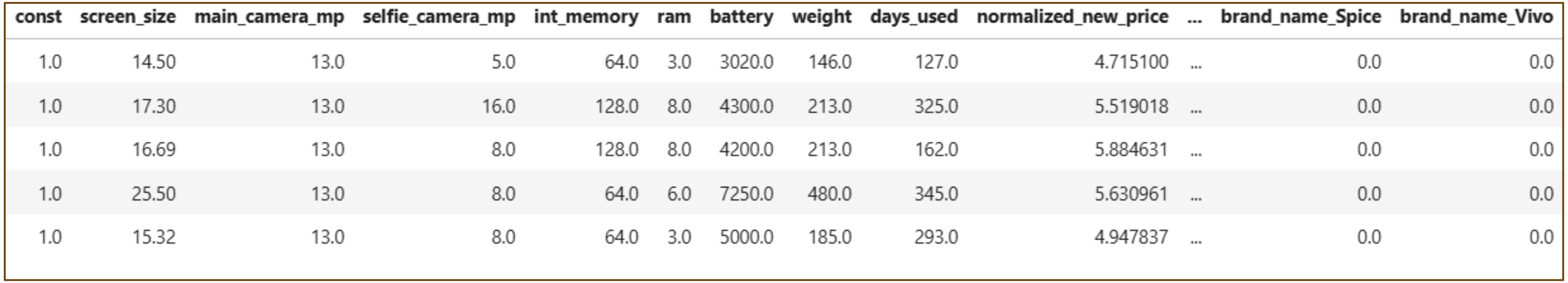


```
  brand_name       os  screen_size   4g   5g  main_camera_mp  \
0      Honor  Android        14.50  yes   no            13.0
1      Honor  Android        17.30  yes  yes            13.0
2      Honor  Android        16.69  yes  yes            13.0
3      Honor  Android        25.50  yes  yes     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

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

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

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



const	screen_size	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	days_used	normalized_new_price	...	brand_name_Spice	brand_name_Vivo
1.0	14.50	13.0	5.0	64.0	3.0	3020.0	146.0	127	4.715100	...	False	False
1.0	17.30	13.0	16.0	128.0	8.0	4300.0	213.0	325	5.519018	...	False	False
1.0	16.69	13.0	8.0	128.0	8.0	4200.0	213.0	162	5.884631	...	False	False
1.0	25.50	13.0	8.0	64.0	6.0	7250.0	480.0	345	5.630961	...	False	False
1.0	15.32	13.0	8.0	64.0	3.0	5000.0	185.0	293	4.947837	...	False	False

Intercept added to data and creating dummies for independent features



const	screen_size	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	days_used	normalized_new_price	...	brand_name_Spice	brand_name_Vivo
1.0	14.50	13.0	5.0	64.0	3.0	3020.0	146.0	127.0	4.715100	...	0.0	0.0
1.0	17.30	13.0	16.0	128.0	8.0	4300.0	213.0	325.0	5.519018	...	0.0	0.0
1.0	16.69	13.0	8.0	128.0	8.0	4200.0	213.0	162.0	5.884631	...	0.0	0.0
1.0	25.50	13.0	8.0	64.0	6.0	7250.0	480.0	345.0	5.630961	...	0.0	0.0
1.0	15.32	13.0	8.0	64.0	3.0	5000.0	185.0	293.0	4.947837	...	0.0	0.0

Input attributes converted into float type for modeling

- **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 Performance Summary

- **Overview of ML model and its parameters**

Objective: To develop a predictive model that accurately estimates the fair market price of used and refurbished devices, enabling ReCell to optimize pricing strategies, enhance revenue, and ensure competitive offerings.

Model Type: Ordinary Least Squares (OLS) Linear Regression

- A supervised learning algorithm used to model the linear relationship between the dependent variable (price) and one or more independent variables (device features).
- The goal is to minimize the residual sum of squares (the difference between observed and predicted values).

Link to Appendix slide on model assumptions

Key Model Parameters:

- **Dependent Variable**: normalized_used_price
- **Independent Variables**: screen_size, main_camera_mp, selfie_camera_mp, int_memory, ram, battery, weight, 4g, 5g, release_year, days_used, normalized_new_price, brand_name, os
- **fit_intercept**

- **Summary of most important factors used by the ML model for prediction**

  - **Normalized New Price**: This is the strongest predictor in the model. A higher new price leads to proportionally higher used price. It indicates a perceived long value and brand positioning

Link to Appendix slide on model assumptions

- **Main Camera MP**: Devices with higher main camera resolution tend to retain value better

- **Selfie Camera MP:** High-resolution selfie cameras also correlate with higher used value and this is important for buyers focused on social media and video calls

- **RAM:** Higher RAM indicates better performance and multitasking capability. They directly influence both new and used market pricing

- **Weight:** Positively correlated with perceived durability or value. May reflect build quality, battery size, or robustness

- **Years Since Release:** Negative impact on price. Newer devices retain more value; depreciation increases with age

Link to Appendix slide on model assumptions

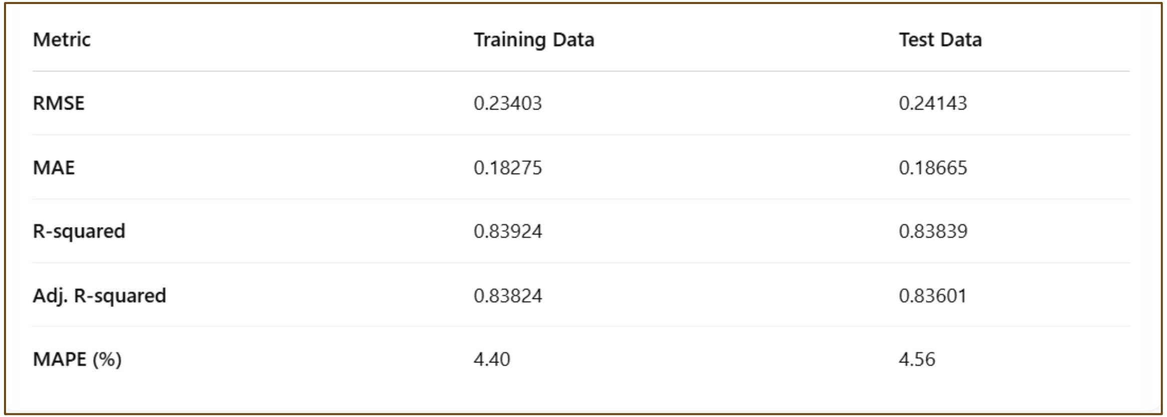
- **Brand Name**: Certain brands may retain best while others show lower used prices.

- **Connectivity:** 5G support is not available for a significant majority of devices where as 4G support is a common feature among used devices.

- **Operating System:** iOS devices show distinct pricing behavior, often priced higher in both new and used market. Android is the most popular OS and negatively impact price

Link to Appendix slide on model assumptions

- **Summary of key performance metrics for training and test data in tabular format for comparison**



Metric	Training Data	Test Data
RMSE	0.23403	0.24143
MAE	0.18275	0.18665
R-squared	0.83924	0.83839
Adj. R-squared	0.83824	0.83601
MAPE (%)	4.40	4.56

- The model performs consistently well across both datasets
- Low RMSE/MAE indicates small prediction errors
- High R^2 shows that the model explains approximately 84% of the variance in used price
- Minimal performance drop on test data implies good generalization with no significant overfitting.

Link to Appendix slide on model assumptions

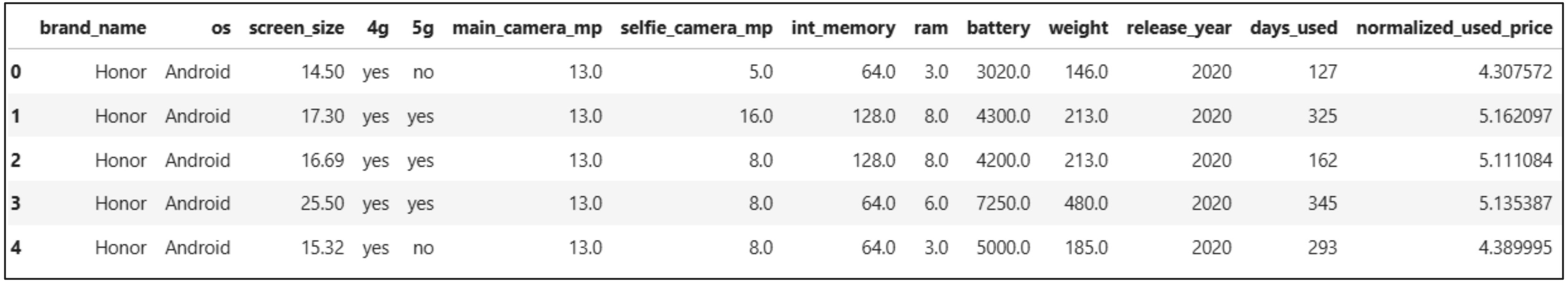
APPENDIX

Data Background and Contents

- **Data Overview**

The dataset consists of 3454 rows and 15 columns, representing data information about different brands and specifications of used and refurbished devices

Devices with higher specifications seems to have higher normalized_used_price wheras those with more days_used have slight decrease in normalized_used_price



	brand_name	os	screen_size	4g	5g	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	release_year	days_used	normalized_used_price
0	Honor	Android	14.50	yes	no	13.0	5.0	64.0	3.0	3020.0	146.0	2020	127	4.307572
1	Honor	Android	17.30	yes	yes	13.0	16.0	128.0	8.0	4300.0	213.0	2020	325	5.162097
2	Honor	Android	16.69	yes	yes	13.0	8.0	128.0	8.0	4200.0	213.0	2020	162	5.111084
3	Honor	Android	25.50	yes	yes	13.0	8.0	64.0	6.0	7250.0	480.0	2020	345	5.135387
4	Honor	Android	15.32	yes	no	13.0	8.0	64.0	3.0	5000.0	185.0	2020	293	4.389995

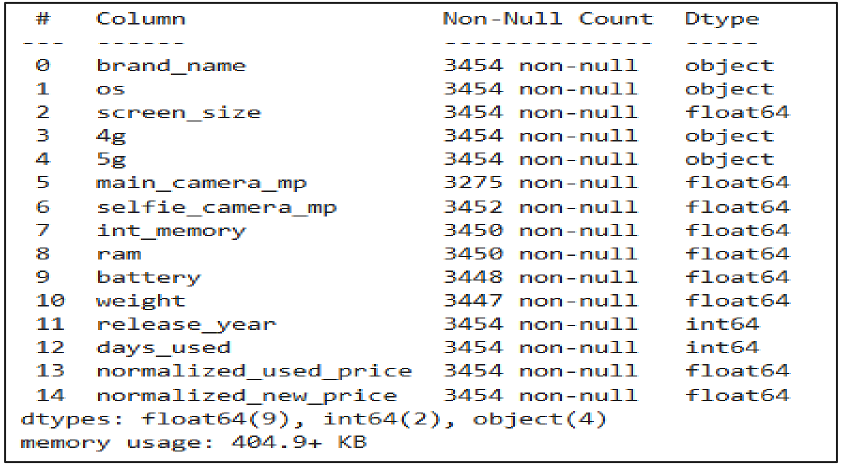
Table 1: Top 5 rows of the Dataset

Data Background and Contents (cont.)

- **Data Background**

The dataset used-device-data was used in the preparation of machine learning-based pricing model to dynamically and accurately price second-hand devices.

- **Data Contents**



```
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   brand_name             3454 non-null   object
 1   os                     3454 non-null   object
 2   screen_size            3454 non-null   float64
 3   4g                     3454 non-null   object
 4   5g                     3454 non-null   object
 5   main_camera_mp         3275 non-null   float64
 6   selfie_camera_mp       3452 non-null   float64
 7   int_memory             3450 non-null   float64
 8   ram                    3450 non-null   float64
 9   battery                3448 non-null   float64
 10  weight                 3447 non-null   float64
 11  release_year           3454 non-null   int64
 12  days_used              3454 non-null   int64
 13  normalized_used_price  3454 non-null   float64
 14  normalized_new_price   3454 non-null   float64
dtypes: float64(9), int64(2), object(4)
memory usage: 404.9+ KB
```

Table 2: Information on the Data Set

Data Background and Contents (cont.)

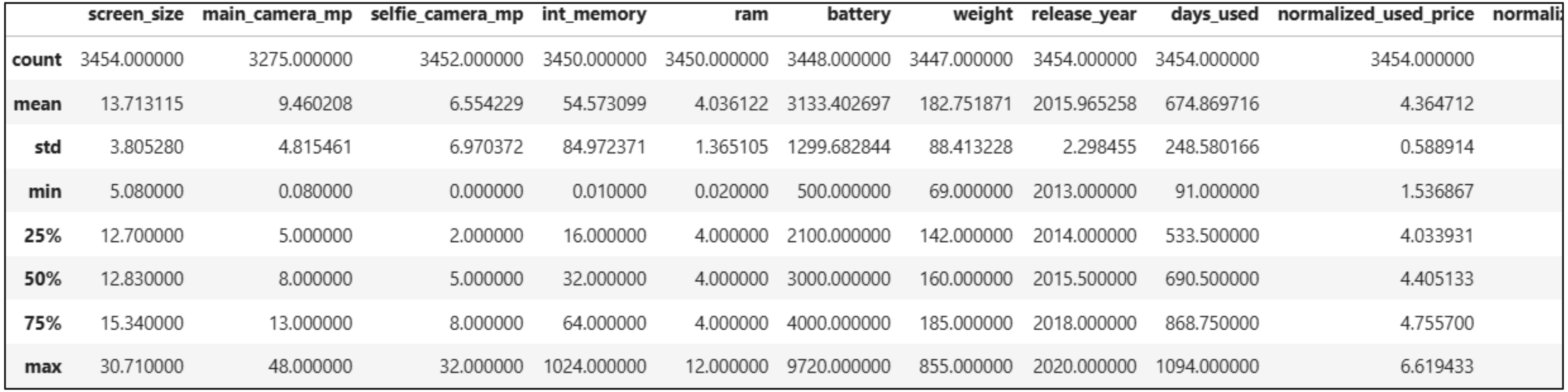
There are three datatypes namely: int64 (2), float64 (9) and object (4) with 11 numeric and 4 strings. The target variable is the normalized_used_price, which is of float type.

- brand_name (*object*): Name of manufacturing brand.
- os (*object*) : OS on which the device runs
- screen_size (*float64*): Size of the screen in cm
- 4g (*object*): Whether 4G is available or not
- 5g (*object*): Whether 5G is available or not
- main_camera_mp (*float64*): Resolution of the rear camera in megapixels
- selfie_camera_mp (*float64*): Resolution of the front camera in megapixels

Data Background and Contents (cont.)

- int_memory (*float64*): Amount of internal memory (ROM) in GB
- ram (*float64*): Amount of RAM in GB
- Battery (*float64*): Energy capacity of the device battery in mAh
- Weight (*float64*): Weight of the device in grams
- release_year (*int64*): Year when the device model was released
- days_used (*int64*): Number of days the used/refurbished device has been used
- normalized_new_price (*float64*): Normalized price of a new device of the same model in euros
- normalized_used_price (*float64*): Normalized price of the used/refurbished device in euros

Data Background and Contents (cont.)



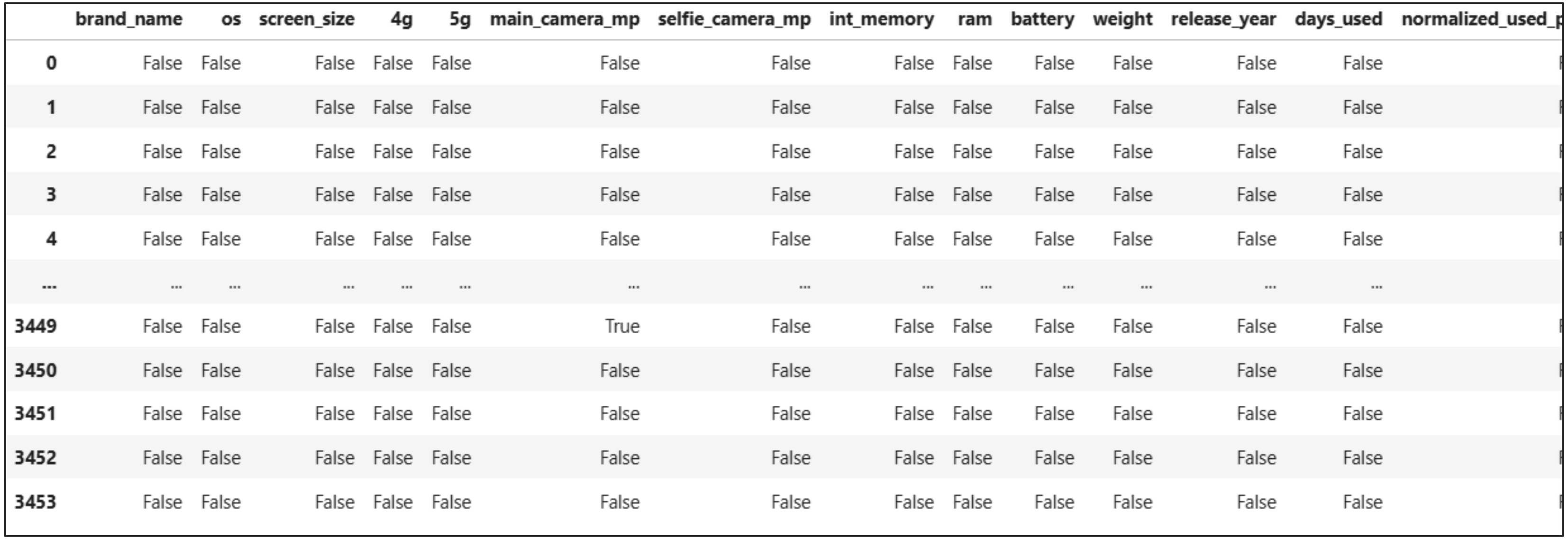
	screen_size	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	release_year	days_used	normalized_used_price	normali
count	3454.000000	3275.000000	3452.000000	3450.000000	3450.000000	3448.000000	3447.000000	3454.000000	3454.000000	3454.000000	
mean	13.713115	9.460208	6.554229	54.573099	4.036122	3133.402697	182.751871	2015.965258	674.869716	4.364712	
std	3.805280	4.815461	6.970372	84.972371	1.365105	1299.682844	88.413228	2.298455	248.580166	0.588914	
min	5.080000	0.080000	0.000000	0.010000	0.020000	500.000000	69.000000	2013.000000	91.000000	1.536867	
25%	12.700000	5.000000	2.000000	16.000000	4.000000	2100.000000	142.000000	2014.000000	533.500000	4.033931	
50%	12.830000	8.000000	5.000000	32.000000	4.000000	3000.000000	160.000000	2015.500000	690.500000	4.405133	
75%	15.340000	13.000000	8.000000	64.000000	4.000000	4000.000000	185.000000	2018.000000	868.750000	4.755700	
max	30.710000	48.000000	32.000000	1024.000000	12.000000	9720.000000	855.000000	2020.000000	1094.000000	6.619433	

Table 3: Statistical Summary of the Dataset

Data Background and Contents (cont.)

- **Observations**
  - The normalized_used_price ranges between 1.54 and 6.62, indicating a large variation probably due to age, specification, brand reputation and battery life. This suggests opportunity to model or segment phones by value tiers
  - Since normalized_new_price average is around 5.23 and the average of the normalized_used_price is 4.36, it suggest that the average used price is about 16.5% lower than the new price
  - The median of 4.41 with a mean of 4.36 suggest that the distribution is roughly symmetric though outliers exist
  - The minimum value of 1.54 indicates some very cheap used phones, possibly with poor specs or heavy usage.

Data Background and Contents (cont.)

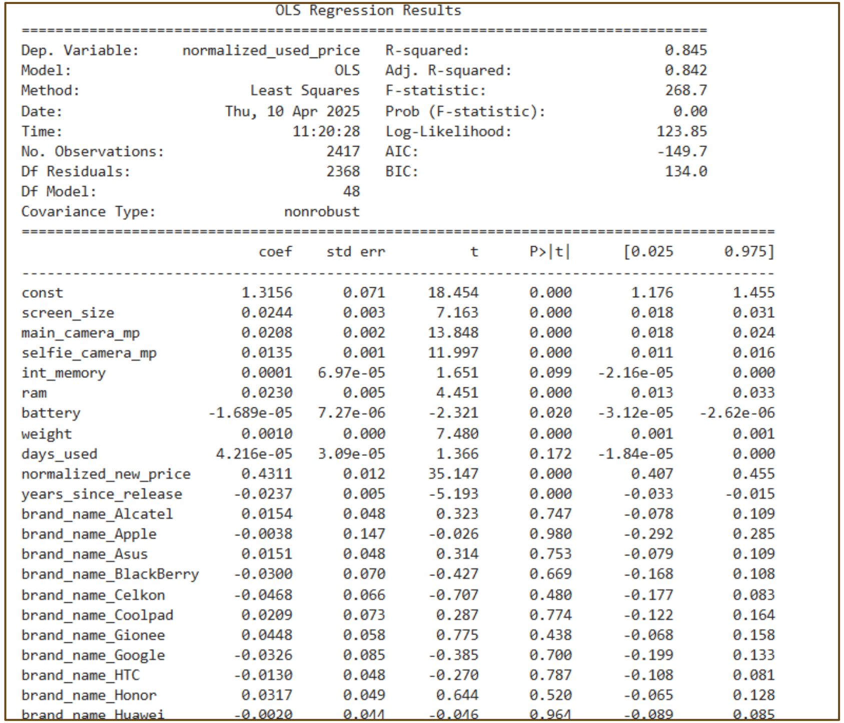


	brand_name	os	screen_size	4g	5g	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	release_year	days_used	normalized_used_p
0	False	False	False	False	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3449	False	False	False	False	False	True	False	False	False	False	False	False	False	
3450	False	False	False	False	False	False	False	False	False	False	False	False	False	
3451	False	False	False	False	False	False	False	False	False	False	False	False	False	
3452	False	False	False	False	False	False	False	False	False	False	False	False	False	
3453	False	False	False	False	False	False	False	False	False	False	False	False	False	

- There are missing values in some columns in the dataset
- There are no duplicates in the dataset

Model Building - Linear Regression

- **First Model (olsmodel1)**



```
                              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:                   Thu, 10 Apr 2025   Prob (F-statistic):               0.00
Time:                           11:20:28   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
```

- R-squared: 0.845 - The model explains about 84.5% of the variance in used device prices.

- Adjusted. R-squared: 0.842 - It reflects a strong fit of the model even after adjusting for number of predictors.

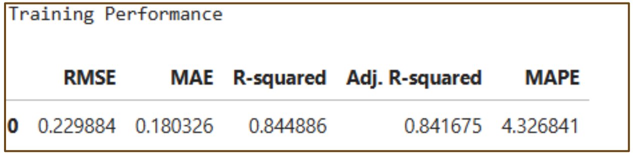
- const coefficient: 1.3156

- Coefficient of a predictor variable:

- selfie_camera_mp has a coefficient of 0.0135. This means for every additional megapixel in the selfie camera, the normalized used price is expected to increase by 0.0135 units, assuming all other factors remain the same.

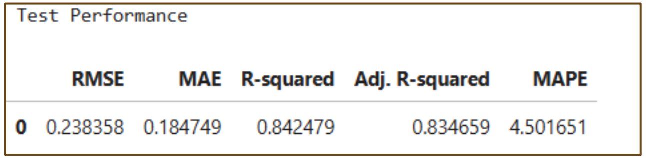
- 5g_yes has a coefficient of -0.0714. This suggests that 5G-capable devices tend to have a slightly lower normalized used price, when controlling for all other variables

- ## Model Performance Check



Training Performance

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



Test Performance

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

The training R-squared is 0.845, so the model is not underfitting

The train and test RMSE and MAE are comparable, so the model is not overfitting either

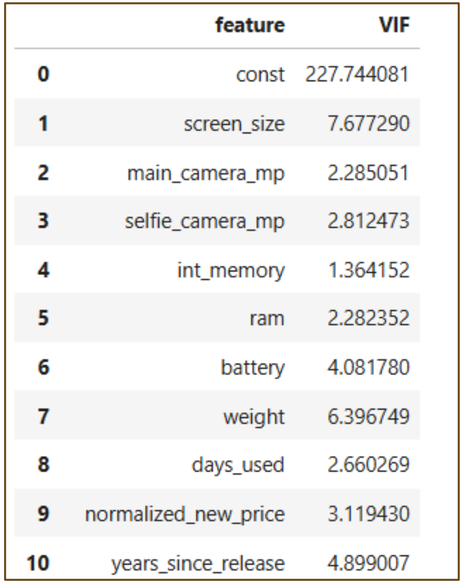
MAE of 0.1847 implies that the model's average absolute prediction error is about 18.5% of the used price range on the test data

MAPE of 4.50 on the test data means that we are able to predict within 4.50% of the normalized used price

Model Assumptions

- **Checking Linear Regression Assumptions**

  - Test for Multicollinearity

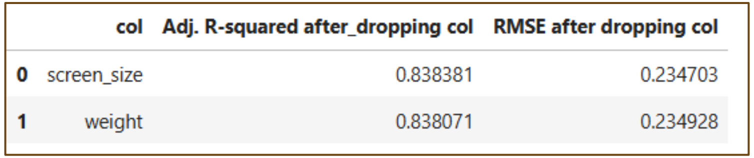


	feature	VIF
0	const	227.744081
1	screen_size	7.677290
2	main_camera_mp	2.285051
3	selfie_camera_mp	2.812473
4	int_memory	1.364152
5	ram	2.282352
6	battery	4.081780
7	weight	6.396749
8	days_used	2.660269
9	normalized_new_price	3.119430
10	years_since_release	4.899007

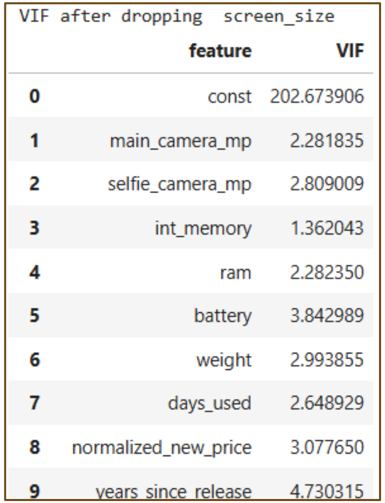
There are multiple columns with very high VIF values, indicating presence of strong multicollinearity.

- Removing Multicollinearity

Specifying columns with high VIF



	col	Adj. R-squared after_dropping col	RMSE after dropping col
0	screen_size	0.838381	0.234703
1	weight	0.838071	0.234928



VIF after dropping screen_size

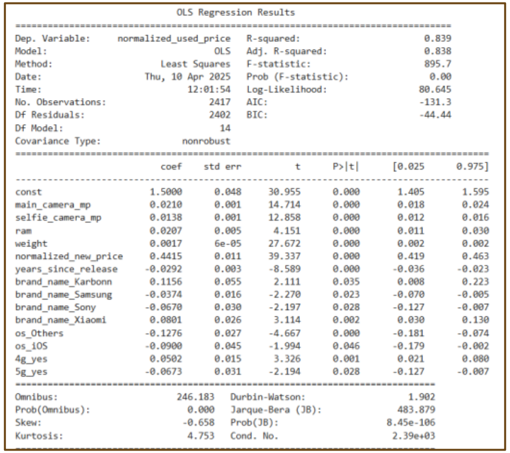
	feature	VIF
0	const	202.673906
1	main_camera_mp	2.281835
2	selfie_camera_mp	2.809009
3	int_memory	1.362043
4	ram	2.282350
5	battery	3.842989
6	weight	2.993855
7	days_used	2.648929
8	normalized_new_price	3.077650
9	years_since_release	4.730315

- Dropping screen_size slightly improves model fit (higher adjusted R²) and reduces RMSE compared to dropping weight.
- This suggests that screen_size may be more responsible for multicollinearity without sacrificing predictive power.

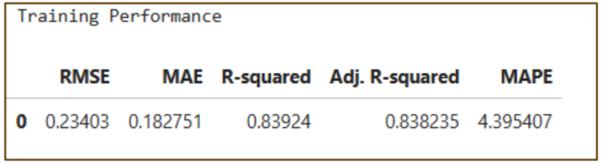
Removing p values >0.05
Below are the selected_features after removing p>0.05:

['const', 'main_camera_mp', 'selfie_camera_mp', 'ram', 'weight', 'normalized_new_price', 'years_since_release', 'brand_name_Karbonn', 'brand_name_Samsung', 'brand_name_Sony', 'brand_name_Xiaomi', 'os_Others', 'os_iOS', '4g_yes', '5g_yes']

- ## Second Model (olsmodel2)

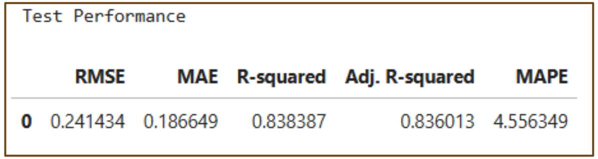


```
                            OLS Regression Results
==============================================================================
Dep. Variable:     normalized_used_price   R-squared:                       0.839
Model:                               OLS   Adj. R-squared:                  0.838
Method:                    Least Squares   F-statistic:                     895.7
Date:                   Thu, 10 Apr 2025   Prob (F-statistic):               0.00
Time:                           12:01:54   Log-Likelihood:                 80.645
No. Observations:                   2417   AIC:                            -131.3
Df Residuals:                       2402   BIC:                            -44.44
Df Model:                             14
Covariance Type:               nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                     1.5000      0.048     30.955      0.000       1.405       1.595
main_camera_mp            0.0210      0.001     14.714      0.000       0.018       0.024
selfie_camera_mp          0.0138      0.001     12.858      0.000       0.012       0.016
ram                       0.0207      0.005      4.151      0.000       0.011       0.030
weight                    0.0017      6e-05     27.672      0.000       0.002       0.002
normalized_new_price      0.4415      0.011     39.337      0.000       0.419       0.463
years_since_release      -0.0292      0.003     -8.589      0.000      -0.036      -0.023
brand_name_Karbonn        0.1156      0.055      2.111      0.035       0.008       0.223
brand_name_Samsung       -0.0374      0.016     -2.270      0.023      -0.070      -0.005
brand_name_Sony          -0.0670      0.030     -2.197      0.028      -0.127      -0.007
brand_name_Xiaomi         0.0801      0.026      3.114      0.002       0.030       0.130
os_Others                -0.1276      0.027     -4.667      0.000      -0.181      -0.074
os_iOS                   -0.0900      0.045     -1.994      0.046      -0.179      -0.002
4g_yes                    0.0502      0.015      3.326      0.001       0.021       0.080
5g_yes                   -0.0673      0.031     -2.194      0.028      -0.127      -0.007
==============================================================================
Omnibus:                      246.183   Durbin-Watson:                   1.902
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              483.879
Skew:                          -0.658   Prob(JB):                    8.45e-106
Kurtosis:                       4.753   Cond. No.                     2.39e+03
==============================================================================
```



Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.23403	0.182751	0.83924	0.838235	4.395407

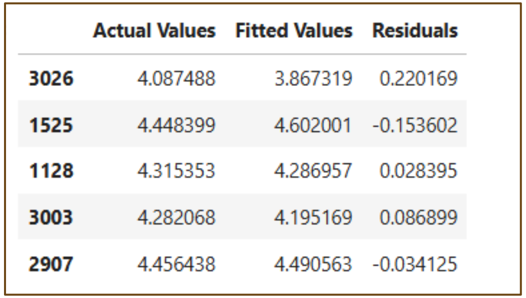


Test Performance

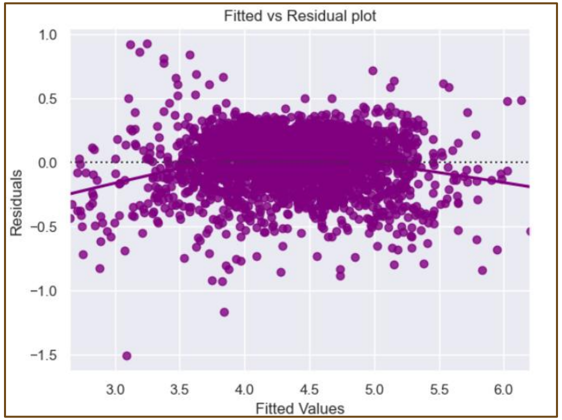
	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.241434	0.186649	0.838387	0.836013	4.556349

- The similarity between training and test metrics confirms that the model generalizes well. There's no sign of overfitting.
- An R² of ~0.84 means the model explains about 84% of the variation in the normalized used price — that's quite strong for real-world pricing models.
- Both RMSE and MAE are low and close across datasets, which confirms model stability and high accuracy.
- With an MAPE < 5% in suggest the model is very precise, with predictions deviating on average by only 4.5% from actual values.

Test for Linearity and Independence

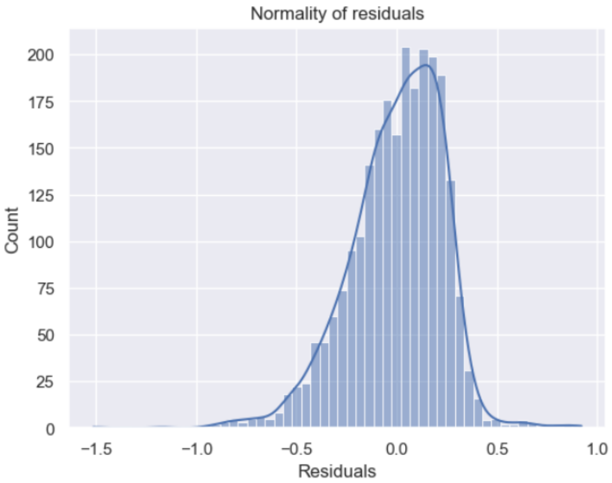


	Actual Values	Fitted Values	Residuals
3026	4.087488	3.867319	0.220169
1525	4.448399	4.602001	-0.153602
1128	4.315353	4.286957	0.028395
3003	4.282068	4.195169	0.086899
2907	4.456438	4.490563	-0.034125



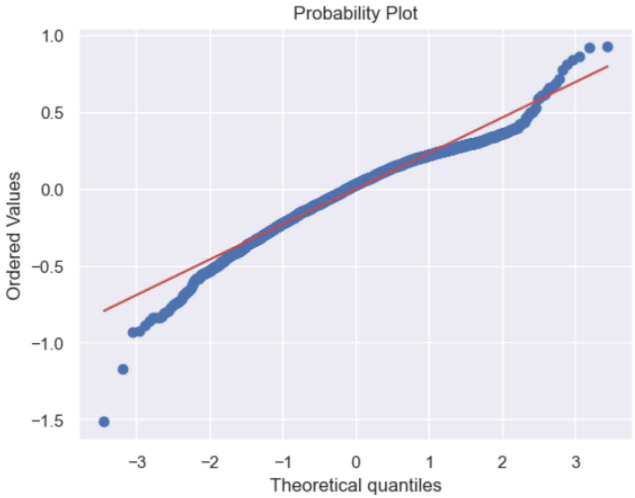
- Residuals are mostly scattered randomly around the horizontal line at 0. This supports linearity of the relationship between the independent and dependent variables.
- There's a slight curve at both ends (left and right), which might suggest minor non-linearity at very low or very high predicted values.
- There is slight funneling on the sides, where residual spread seems a bit wider. This could suggest mild heteroscedasticity, though it's not severe.
- No obvious clustering or trends in residuals, so residuals appear mostly independent of the fitted values.

- Test for Normality



- The residuals form a rough bell-shaped curve, centered around zero.
- This suggests that residuals are approximately normally distributed, which supports one of the key assumptions of linear regression.
- There is a slight left skew (longer tail on the negative side), but it's not extreme.
- Residuals are close enough to normal to meet the assumption.
- The model seems statistically sound, with residuals that: Are approximately normal

- Probability Plot (Q-Q Plot)



ShapiroResult(statistic=0.9676950829900569, **pvalue**=6.983856712612207e-23)

- Since the p-value is much less than 0.05, we reject the null hypothesis.
- This means the residuals are not normally distributed at the 5% significance level.
- So, the assumption is satisfied

- Slight Non-Normality in Tails: There is some deviation in the extreme residuals, indicating heavier tails than a normal distribution (potential leptokurtosis).
- Overall Fit Still Acceptable: The majority of residuals lie close to the line, suggesting the normality assumption holds reasonably well, especially for predictive modeling purposes.
- Residuals are approximately normal with some mild deviations in the tails.
- For regression modeling, this level of deviation is generally acceptable.

- Test for Homoscedasticity

Test Result:

- The test for Homoscedasticity produce the following result:

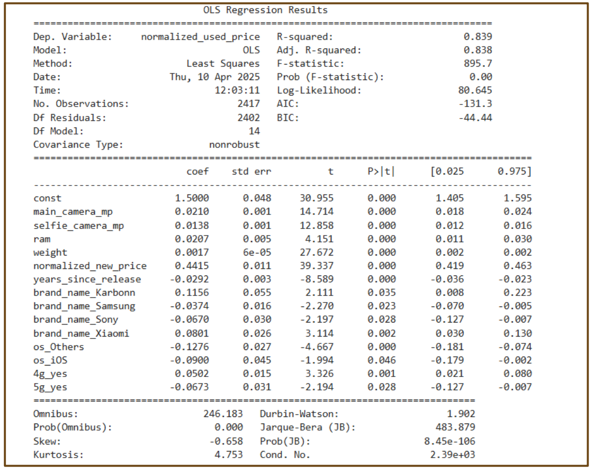
[('F statistic', 1.0087504199106676), ('p-value', 0.4401970650667301)]

Interpretation:

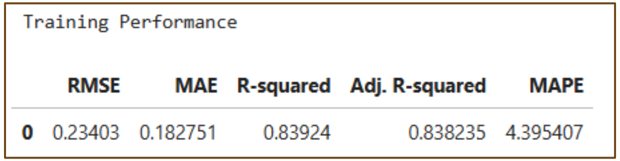
- Since the p-value > 0.05, we fail to reject the null hypothesis.
- This means there's no strong evidence of heteroscedasticity (non-constant variance).
- So, the assumption of constant variance (homoscedasticity) is satisfied.

Final Model Summary

- **Final Model (olsmodel_final)**

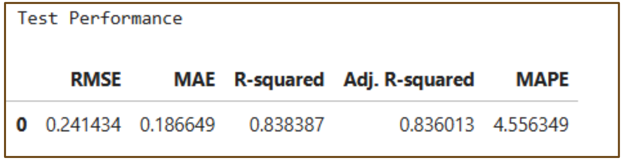


```
                            OLS Regression Results
==============================================================================
Dep. Variable:     normalized_used_price   R-squared:                       0.839
Model:                               OLS   Adj. R-squared:                  0.838
Method:                    Least Squares   F-statistic:                     895.7
Date:                   Thu, 10 Apr 2025   Prob (F-statistic):               0.00
Time:                           12:03:11   Log-Likelihood:                 80.645
No. Observations:                   2417   AIC:                            -131.3
Df Residuals:                       2402   BIC:                            -44.44
Df Model:                             14
Covariance Type:               nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                    1.5000      0.048     30.955      0.000       1.405       1.595
main_camera_mp           0.0210      0.001     14.714      0.000       0.018       0.024
selfie_camera_mp         0.0138      0.001     12.858      0.000       0.012       0.016
ram                      0.0207      0.005      4.151      0.000       0.011       0.030
weight                   0.0017      6e-05     27.672      0.000       0.002       0.002
normalized_new_price     0.4415      0.011     39.337      0.000       0.419       0.463
years_since_release     -0.0292      0.003     -8.589      0.000      -0.036      -0.023
brand_name_Karbonn       0.1156      0.055      2.111      0.035       0.008       0.223
brand_name_Samsung      -0.0374      0.016     -2.270      0.023      -0.070      -0.005
brand_name_Sony         -0.0670      0.030     -2.197      0.028      -0.127      -0.007
brand_name_Xiaomi        0.0801      0.026      3.114      0.002       0.030       0.130
os_Others               -0.1276      0.027     -4.667      0.000      -0.181      -0.074
os_iOS                  -0.0900      0.045     -1.994      0.046      -0.179      -0.002
4g_yes                   0.0502      0.015      3.326      0.001       0.021       0.080
5g_yes                  -0.0673      0.031     -2.194      0.028      -0.127      -0.007
==============================================================================
Omnibus:                      246.183   Durbin-Watson:                   1.902
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              483.879
Skew:                          -0.658   Prob(JB):                    8.45e-106
Kurtosis:                       4.753   Cond. No.                     2.39e+03
```



Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.23403	0.182751	0.83924	0.838235	4.395407



Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.241434	0.186649	0.838387	0.836013	4.556349

- The train and test RMSE and MAE are low and comparable. So, our model is not suffering from overfitting
- The MAPE on the test set suggests we can predict within 4.55% of the normalized_used_price
- The MAE for training and test data barely changed indicating a stable performance
- There is minimal difference between the MAPE for training and test data which supports model robustness
- The model is robust and generalizes well to unseen data and thus can confidently be used for predicting normalized used prices

References

Great Learning. (n.d.) *Supervised Learning - Foundations*. **Great Learning.** https://olympus.mygreatlearning.com/courses/124965/modules/items/6397704?pb_id=18483

Happy Learning !