ReCell – Linear Regression Project

UT- Data Science and Business Analytics

April 2023

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Objective

- 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 the data to be analyzed and a linear regression model built to predict the price of a used phone/tablet and identify factors that significantly influence it.

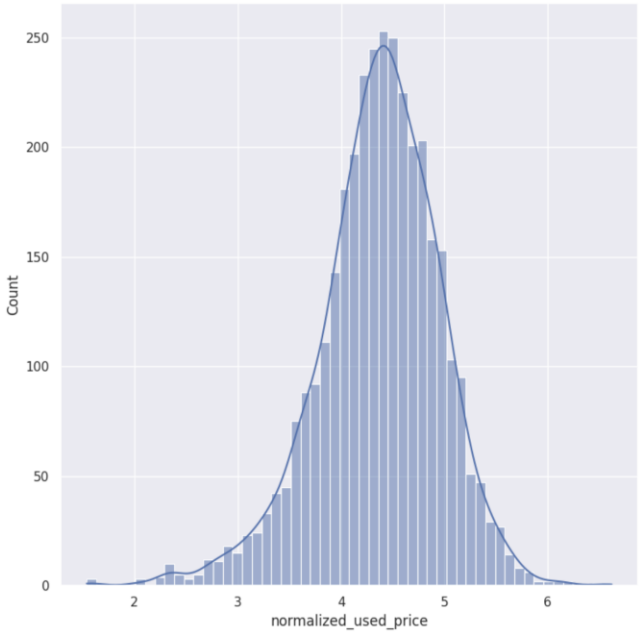
Data Overview

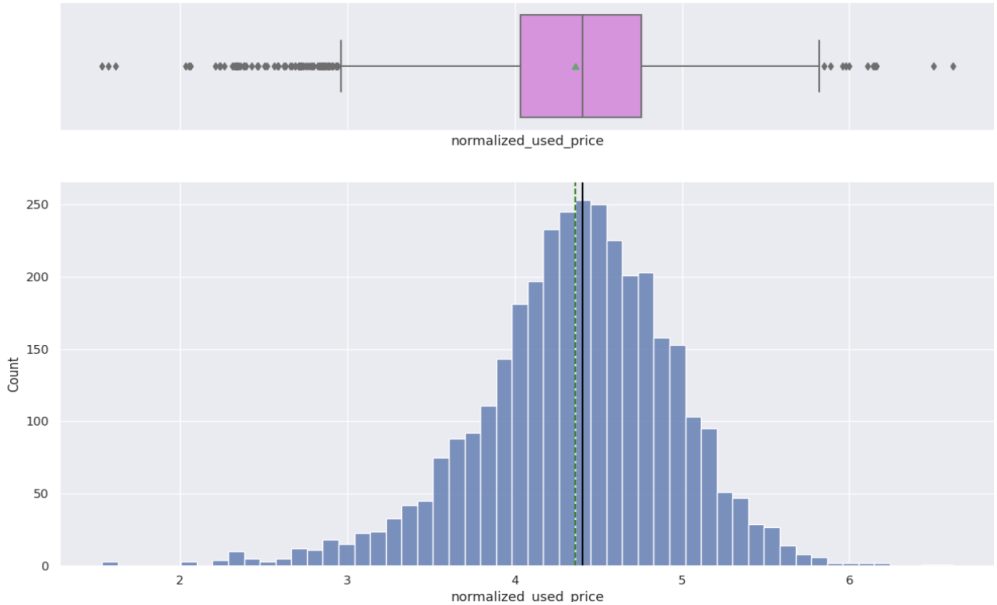
- There are 3,454 rows and 15 columns
- There are 9 float types, 2 integers, and 4 objects
- 34 brands of phones
- 4 different operating systems
- Mean screen size is 13.7
- There are no duplicate values.

Data Overview

- Missing Values:
  - Main_camera_mp is missing 179 values
  - Selfie_camera_mp is missing 2 values
  - Int_memory is missing 4 values
  - Ram is missing 4 values
  - Battery is missing 6 values
  - Weight is missing 7 values

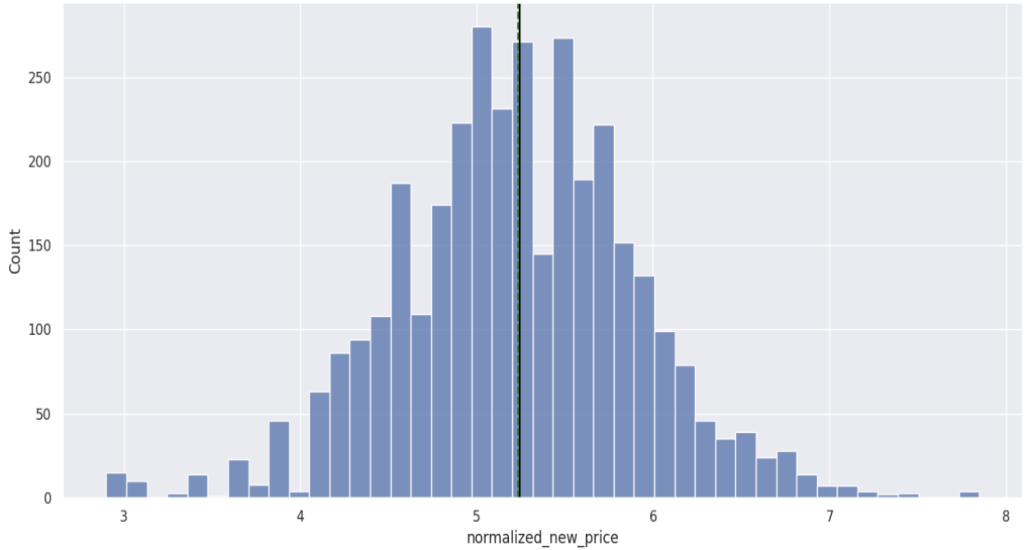
EDA- Normalized Used Price





- Normalized used prices follow a normal distribution curve with a mean of 4.36.
- There are outliers present.

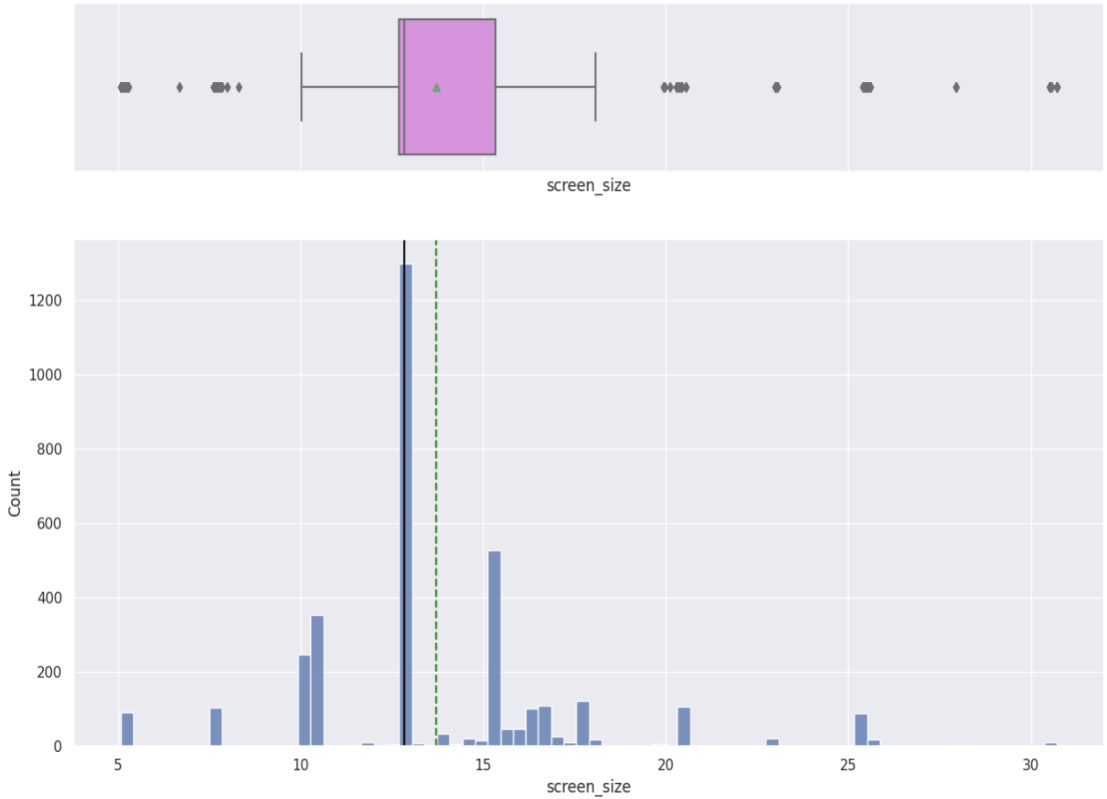
EDA- Normalized New Price



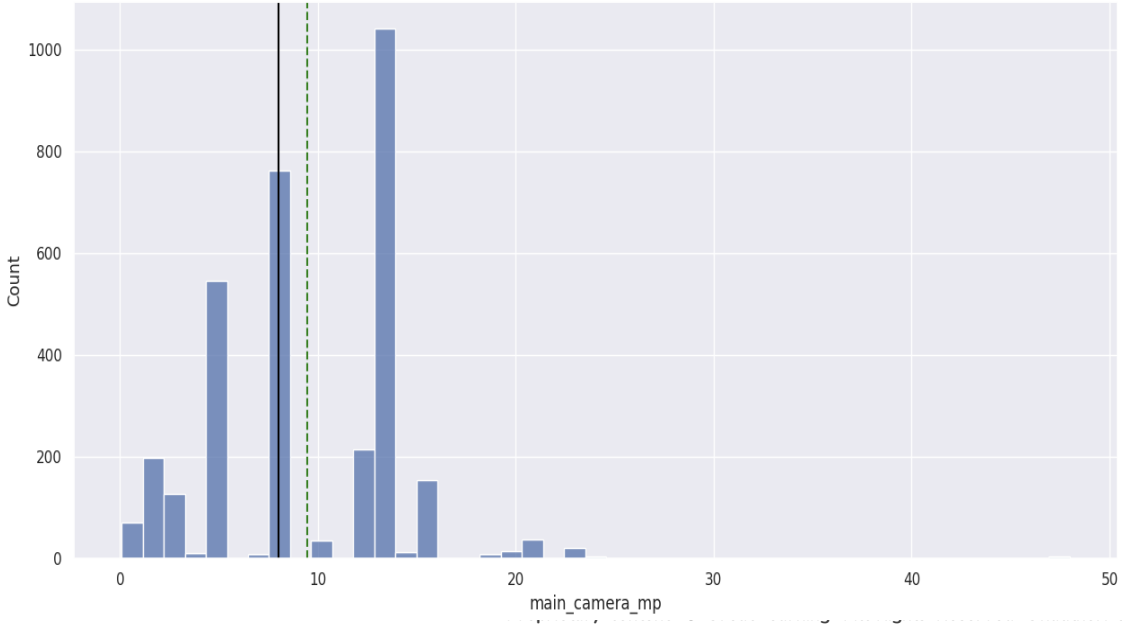
- The prices of new devices are approaching a normal distribution.
- The mean normalized new price is 5.23, 25% of devices are priced below 4.79 euros, 50% are 5.24 euros and below, and 75% are 5.67 euros and below.

EDA- Screen Size

- The mean screen size is 13.7
- There are many outliers in the data.
- The screen size is not evenly distributed. There are many outliers with screen size.
- The screen size is 13.7 mp, 25% are below 12.7 mp, and 75% are below 15.34 mp.



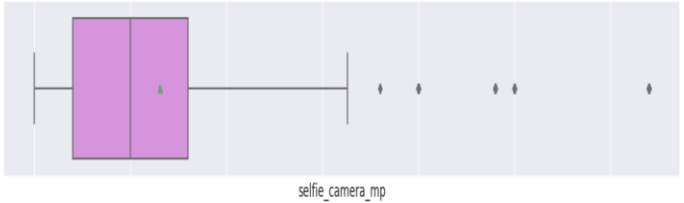
EDA- Main Camera MP

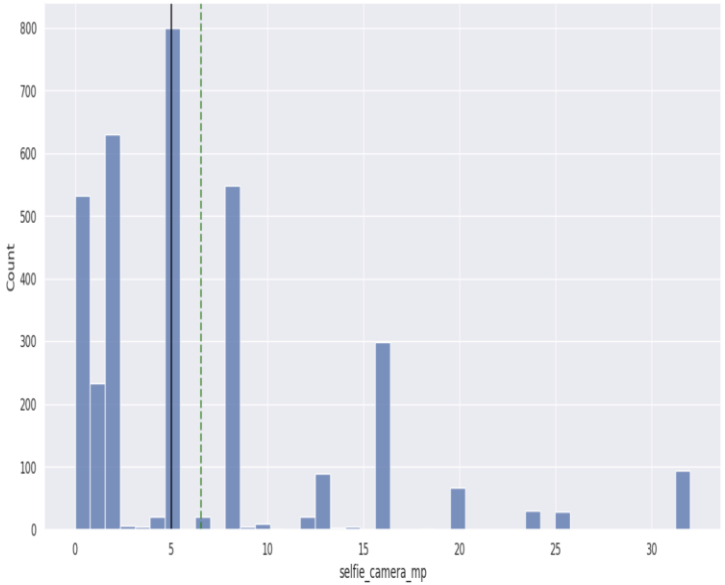


- Main camera mp is not a normal distribution
- Most buyers selected cameras with 5 megapixels.
- Very few buyers selected cameras over 15 mp, probably due to cost.
- The main camera mp data is right skewed. The mean main camera mp is 9.46mp, 25% fall below 5mp, 50% are below 8mp, and 75% are below 13mp.

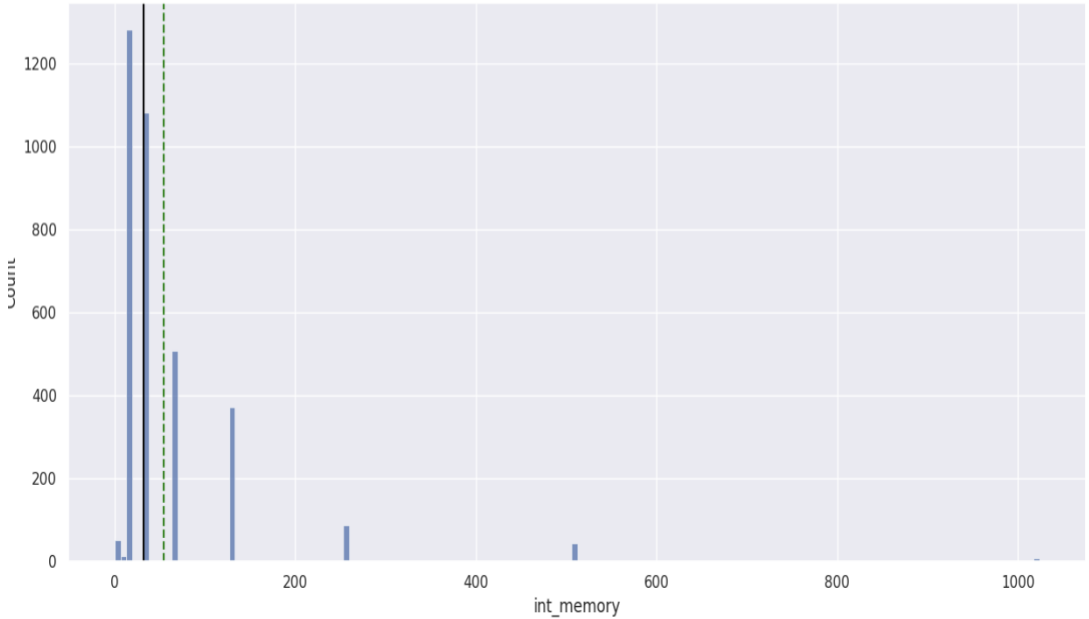
EDA- Selfie Camera MP

- The distribution for the selfie camera mp is not normal.
- The more popular choices included selfie cameras that had 3, 5, or 8 megapixels.
- There are outliers present, with some buyers willing to take on the expense of more expensive phones. The selfie camera megapixel data is heavily right skewed. There are outliers present on the high end of price. Obviously the less expensive phones are the more popular choice for people as they are more accessible to all and they will have fewer megapixels while the more expensive models will have more megapixels.
- The mean selfie camera megapixel is 6.54, 25% is below 2, 50% below 5, and 75% below 8.



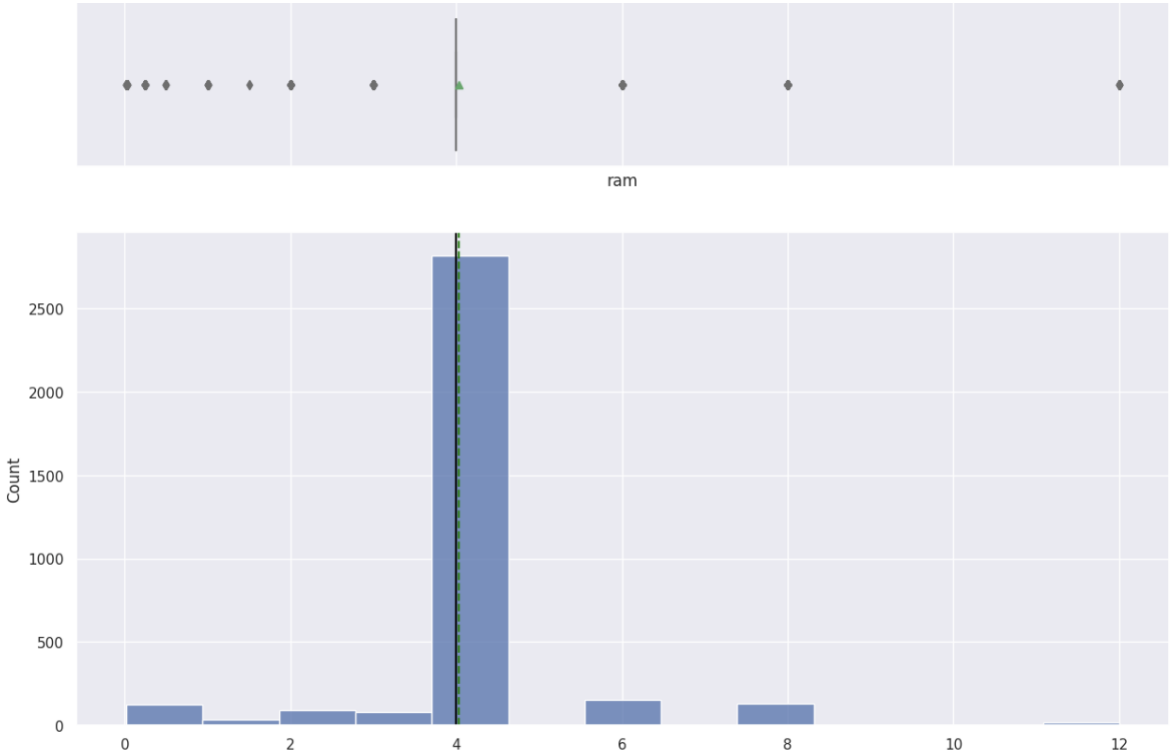


EDA- Internal Memory



- The internal memory is not normally distributed.
- Most of the memory of devices is under 150 GB
- The mean is 54.5GB, 25 % of devices have less than 16 GB, 50% have less than 32GB, and 75% have less than 64GB.
- The max amount was 1,024GB.

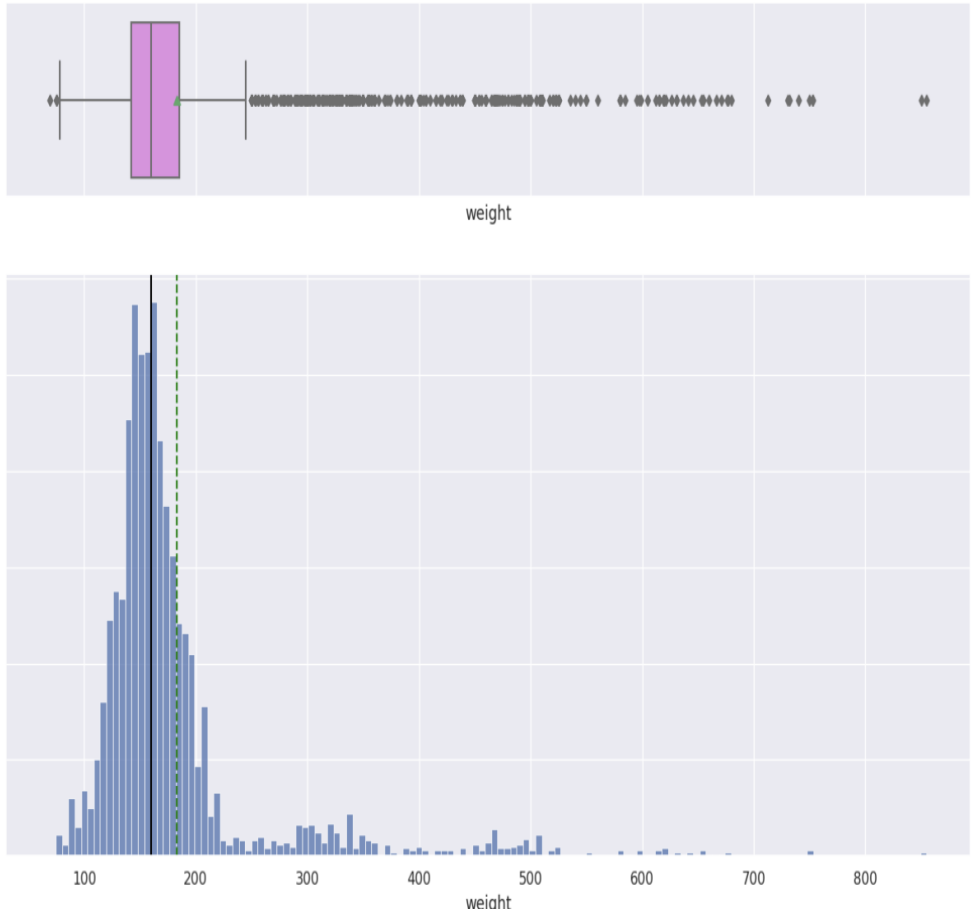
EDA - RAM



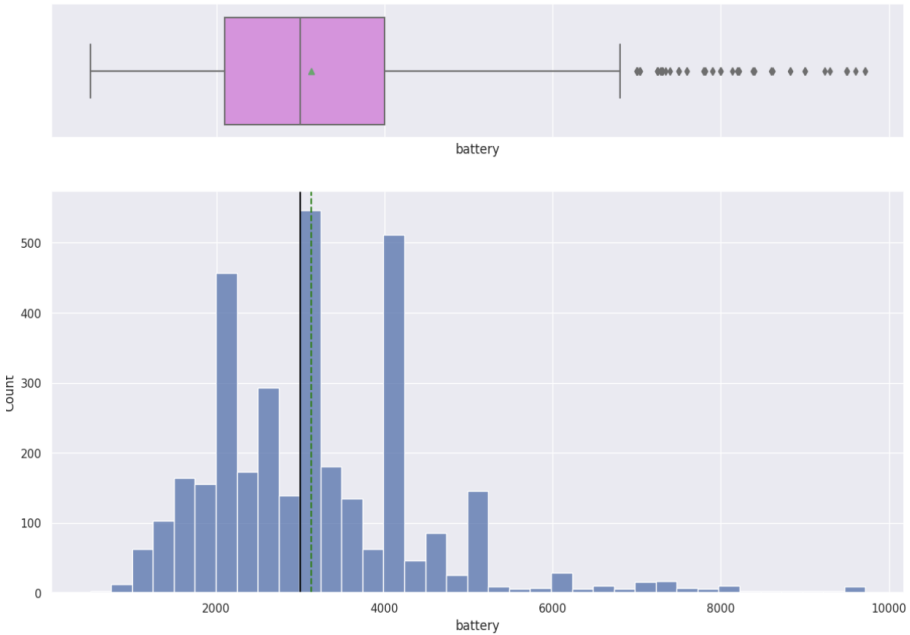
- The mean for RAM was 4.0G, and 75 % of the devices had 4G or less.
- The distribution is normal.

EDA - Weight

- The distribution for weight is heavily right skewed with many outliers.
- The mean weight is 182.75g, with 25% falling below 142g, 50% below 160g, and 75% falling below 185g.
- The heavier devices are not as popular as the lighter ones probably because they are more cumbersome to carry.



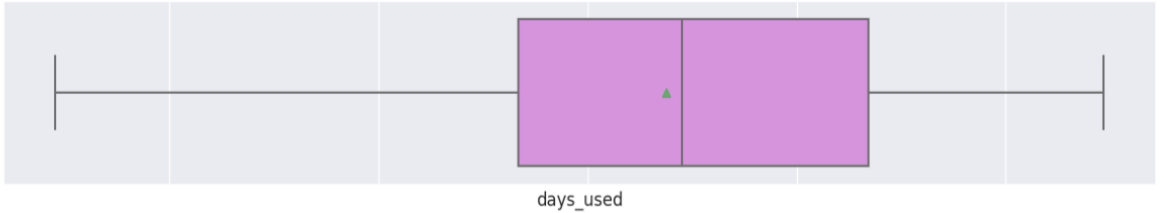
EDA – Battery

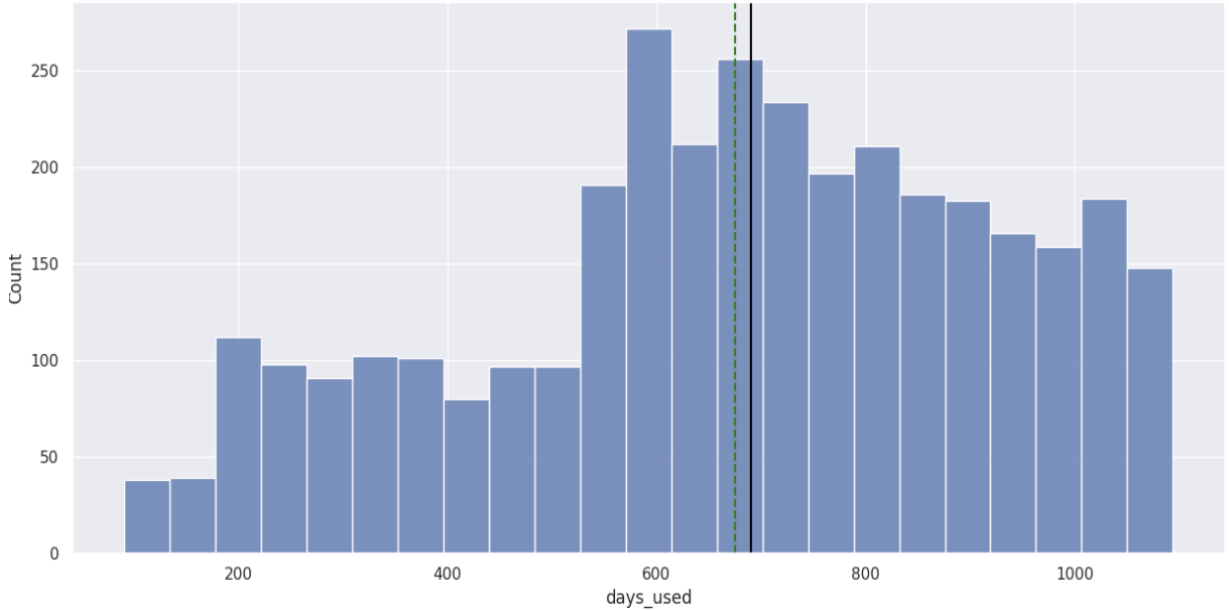


- The distribution is approaching normal, with outliers present.
- The mean battery is 1299mAh,
  with 25% 500mAh and below, 50% 2100mAh and below, and 75% are at 3000mAh and below.

EDA – Days Used

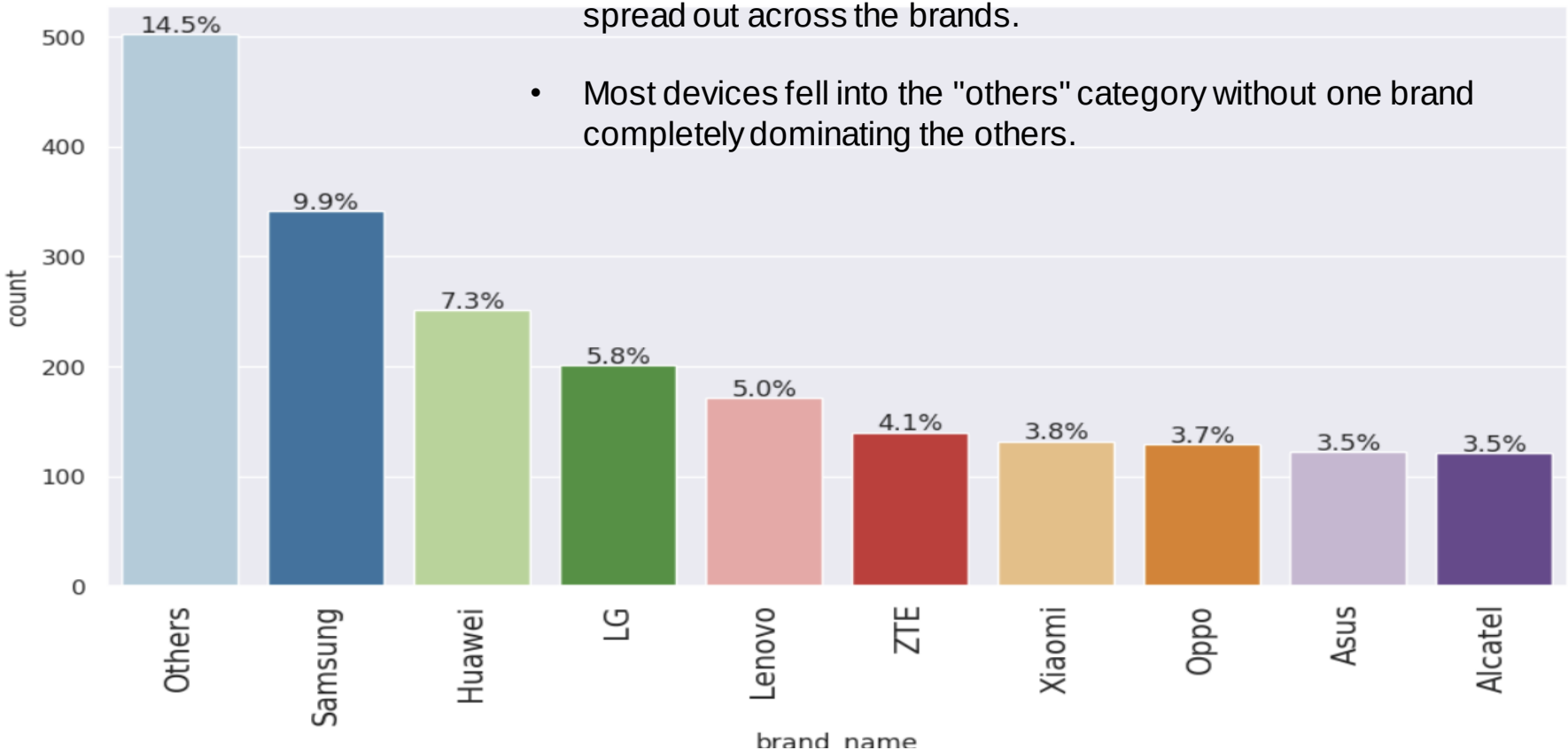
- The distribution is not normal and there are no outliers present.
- The mean days used is 674 days.
- 25% are used 533 days and less, 50% 690 days or less, and 75% are used 868 days and below.





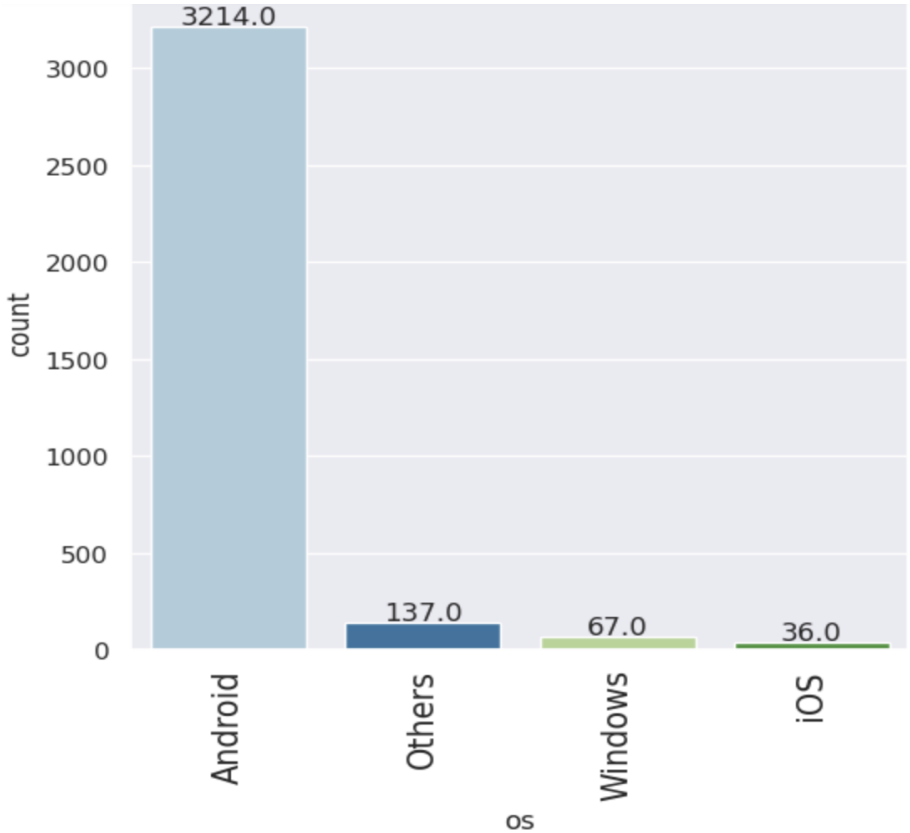
EDA – Brand Name

- There are 34 brand names in the data sets with the devices spread out across the brands.
- Most devices fell into the "others" category without one brand completely dominating the others.

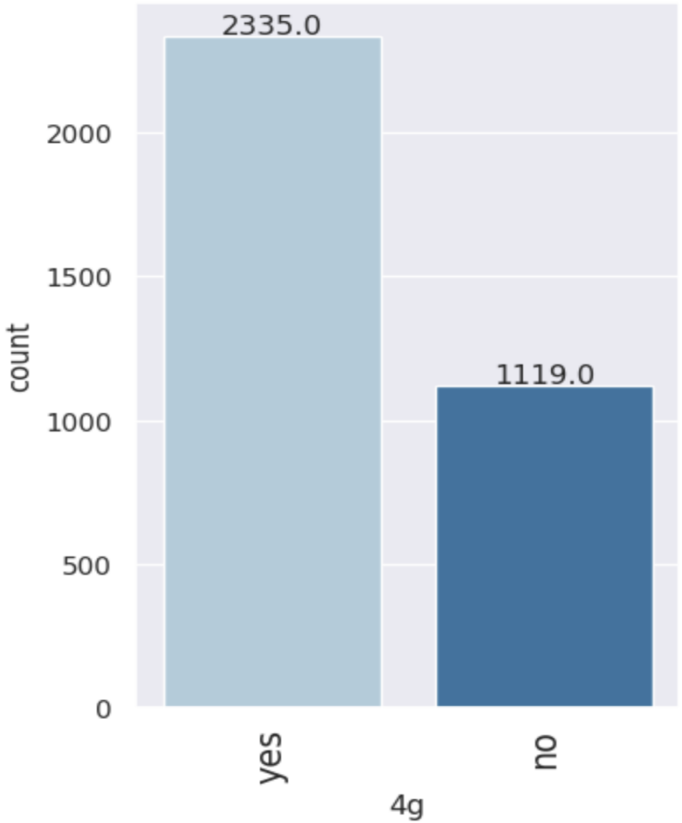


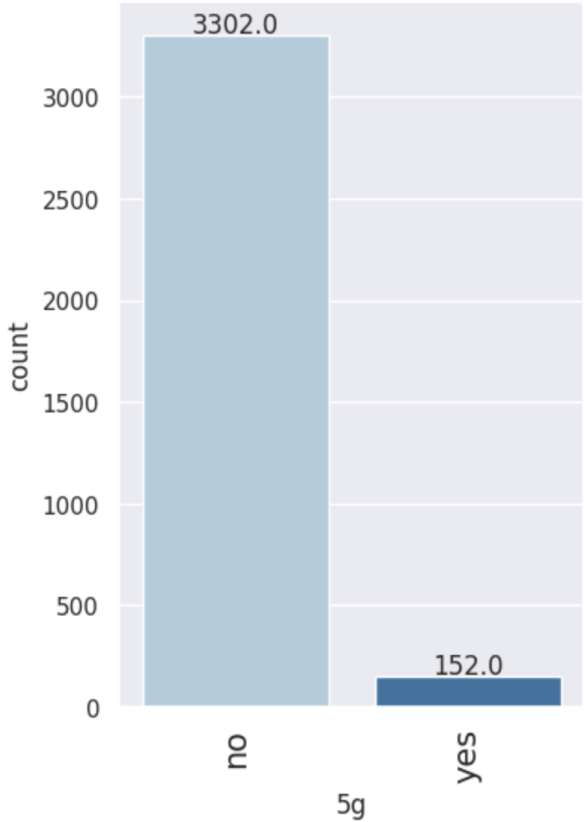
EDA - Operating Systems

Android devices are in the overwhelming majority with 3, 214 devices. No other devices even came close.



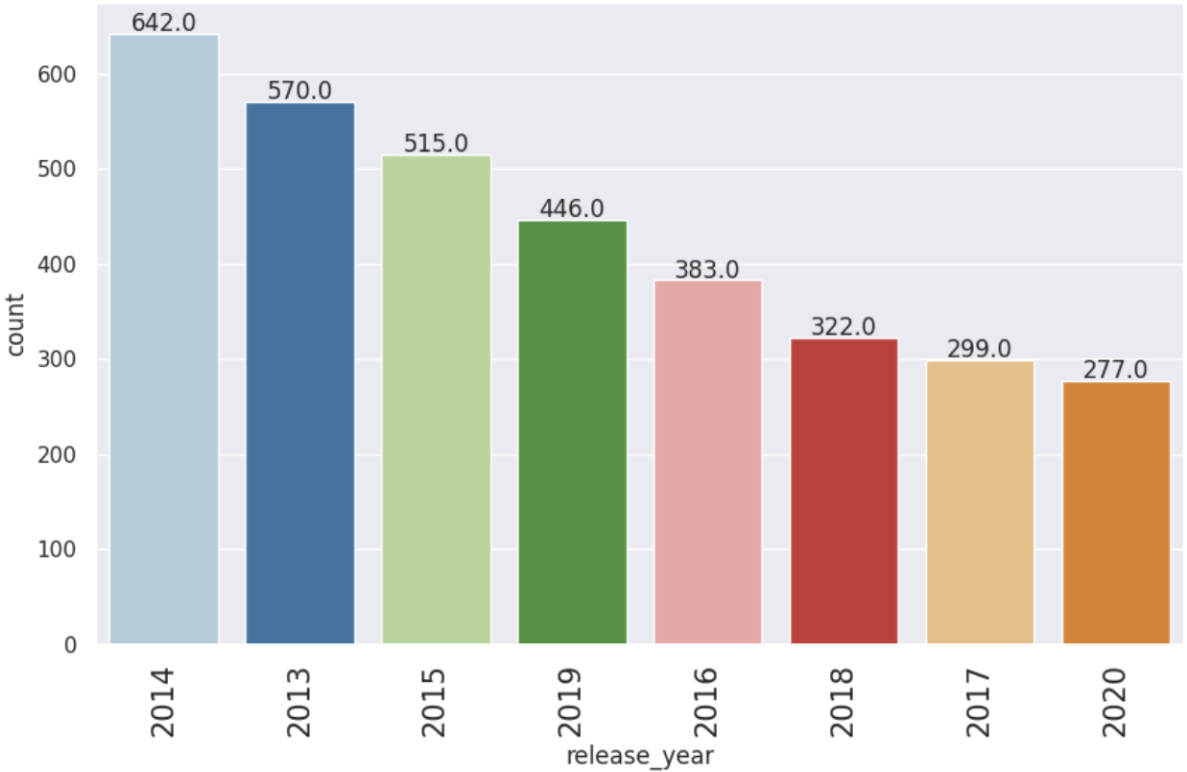
EDA – 4G & 5G





- Most have availability of 4G while many do not have 5G available.

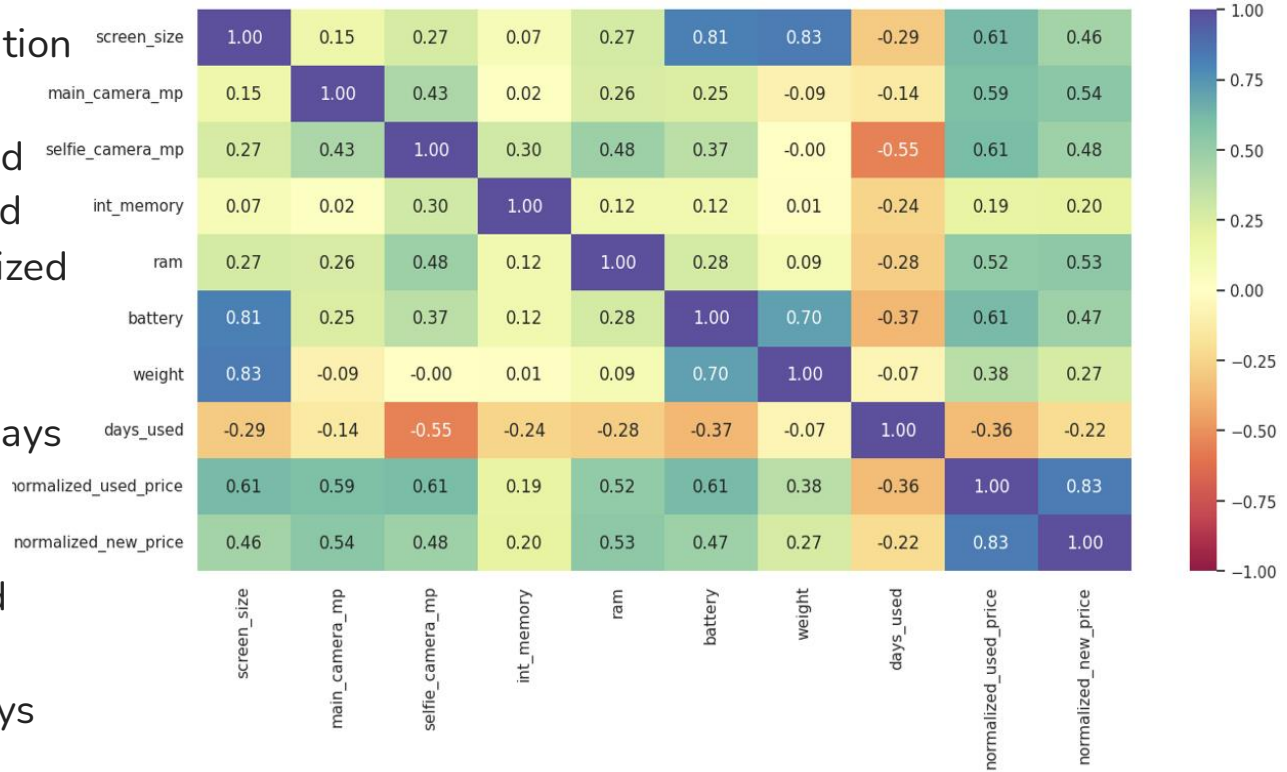
Release Year



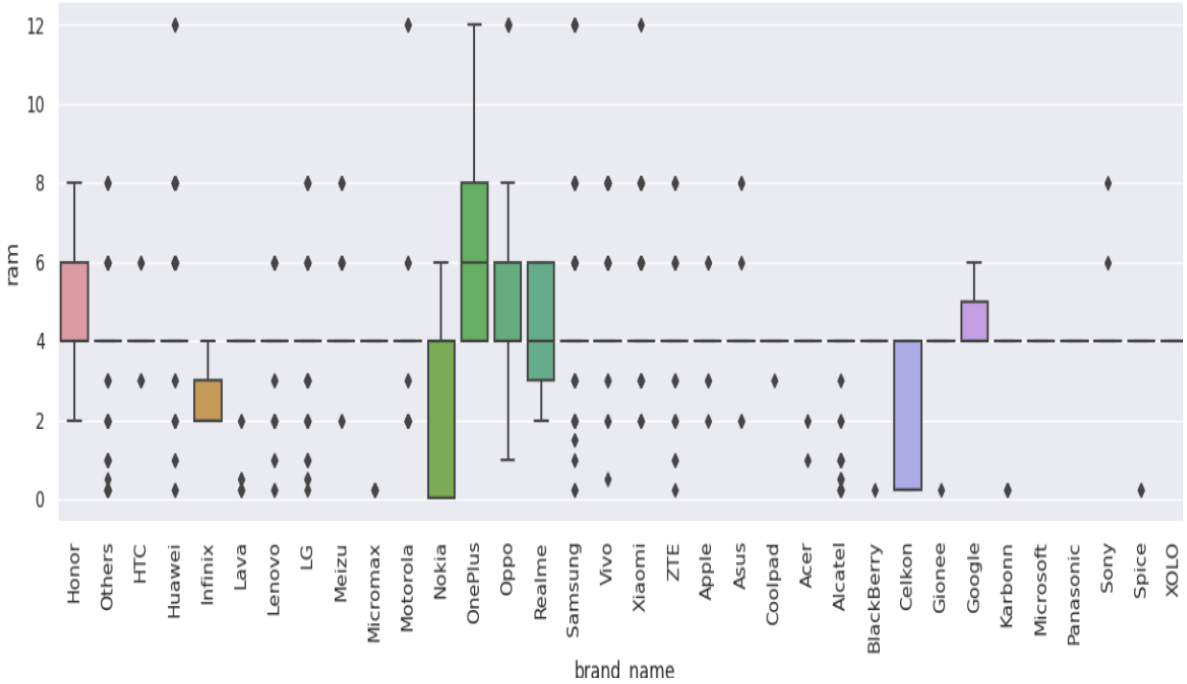
- Most of the devices refurbished were from 2014.

Correlation

- There is a high correlation between: battery and screen size, weight and screen size, normalized new price and normalized used price.
- There is a negative correlation between days used and screen size, days used and selfie camera mp, days used and internal memory, RAM and days used.



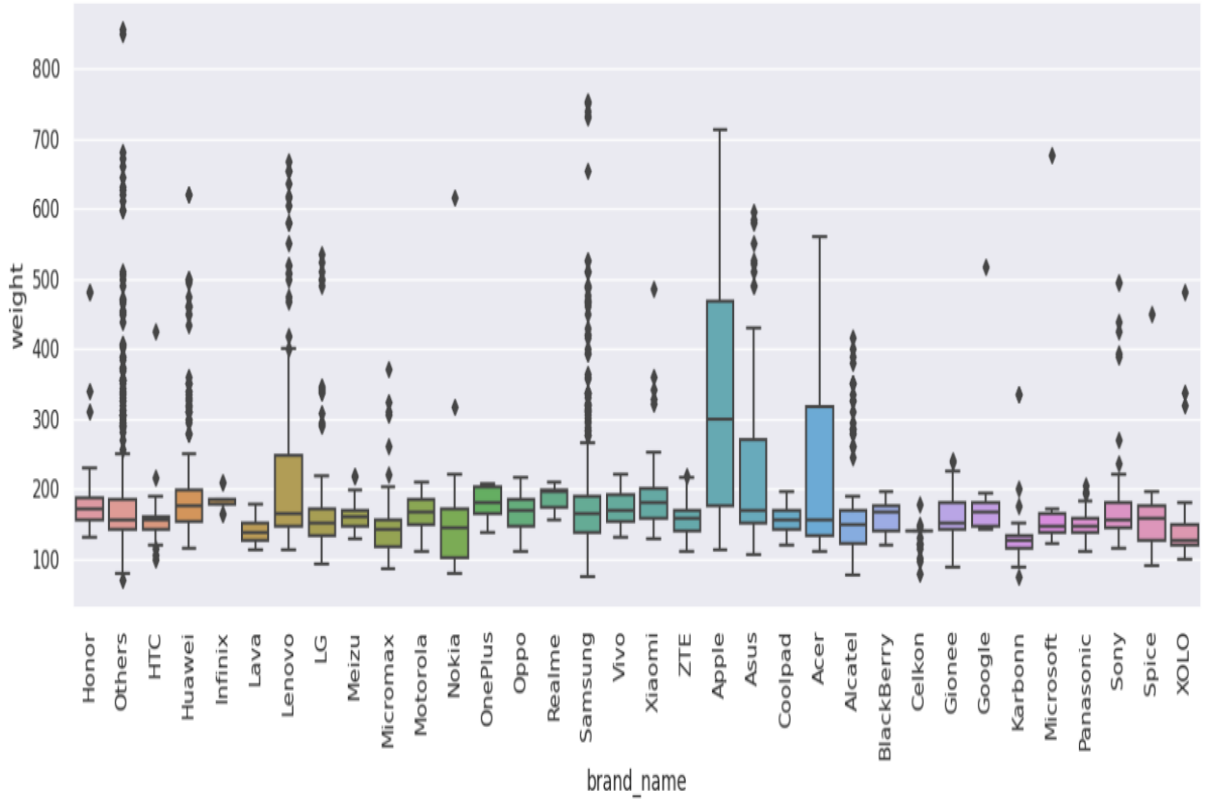
EDA – Brand Name



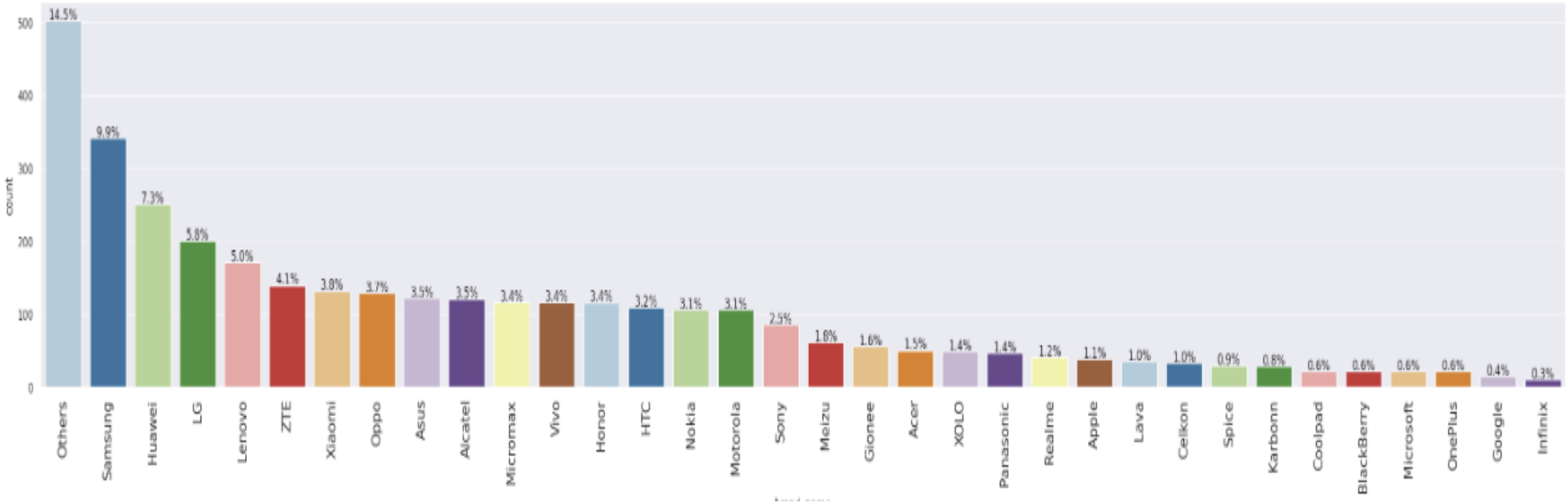
- One Plus phones had the most RAM, followed by Honor, Oppo and Realme.

Brand Name vs. Weight

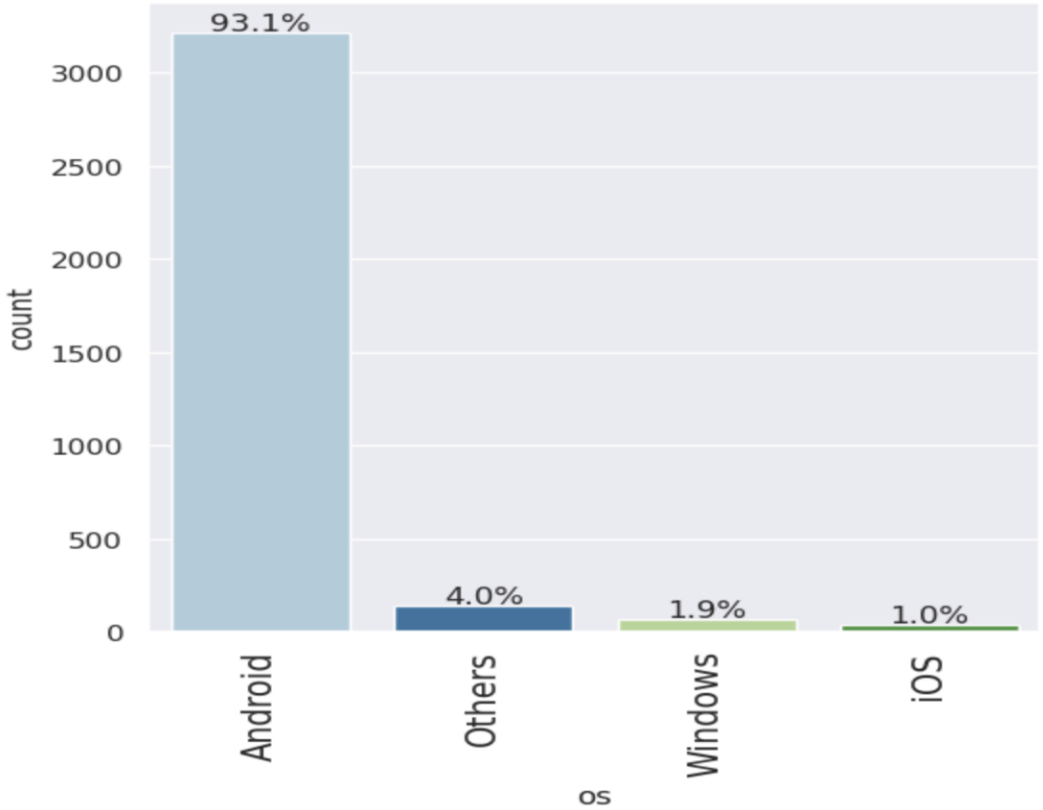
- Apple phones are by far the heaviest of the brands.



Brand Name Count



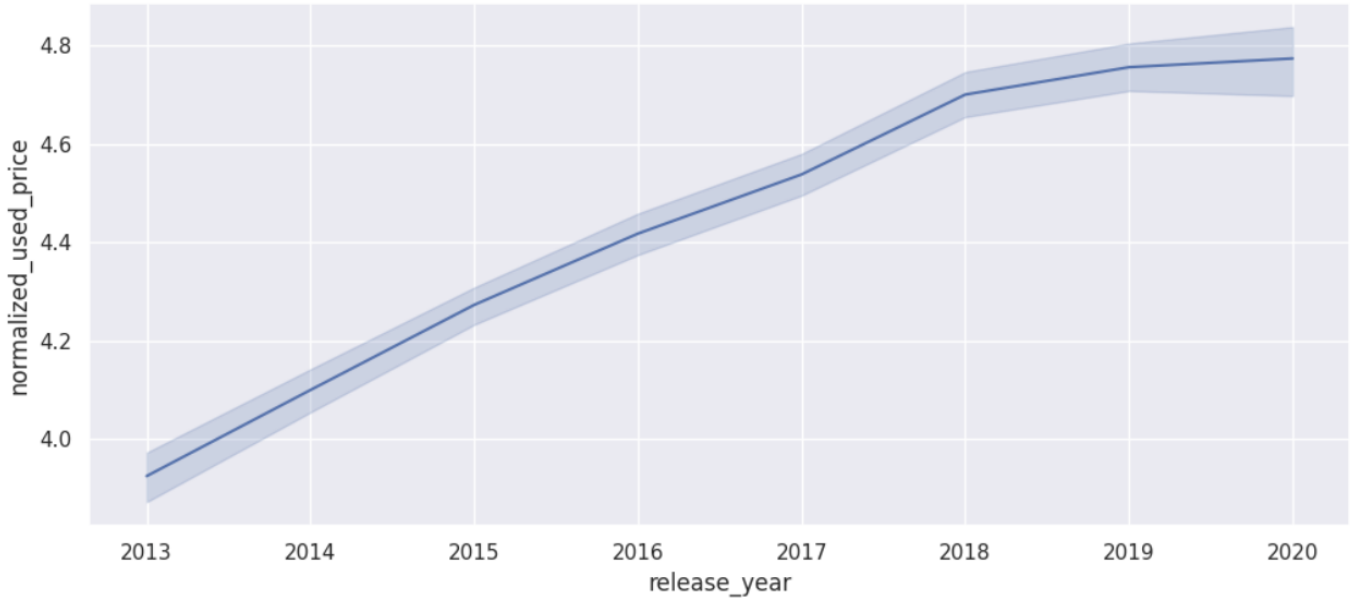
Operating System



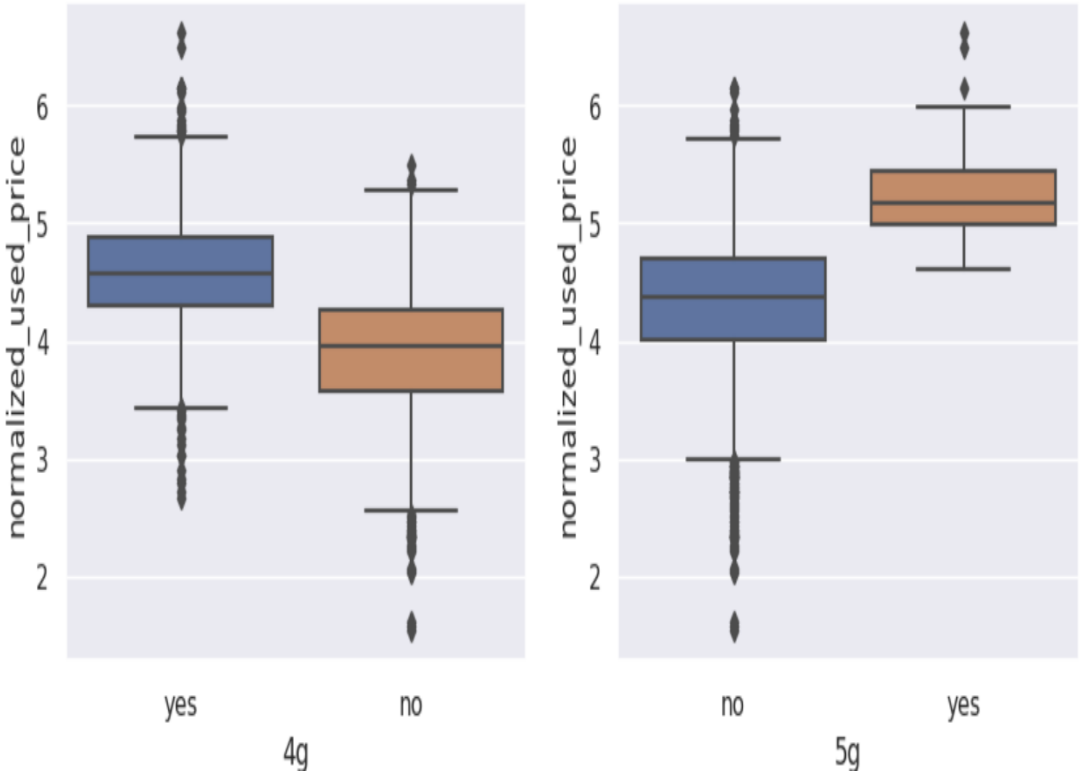
- 93.1% of the devices operate on the Android operating system.
- Only 1% were operating on iOS.

Release year vs. Normalized used price

- The new the phone is, the more expensive the device.



Normalized Used Price for 4G and 5G



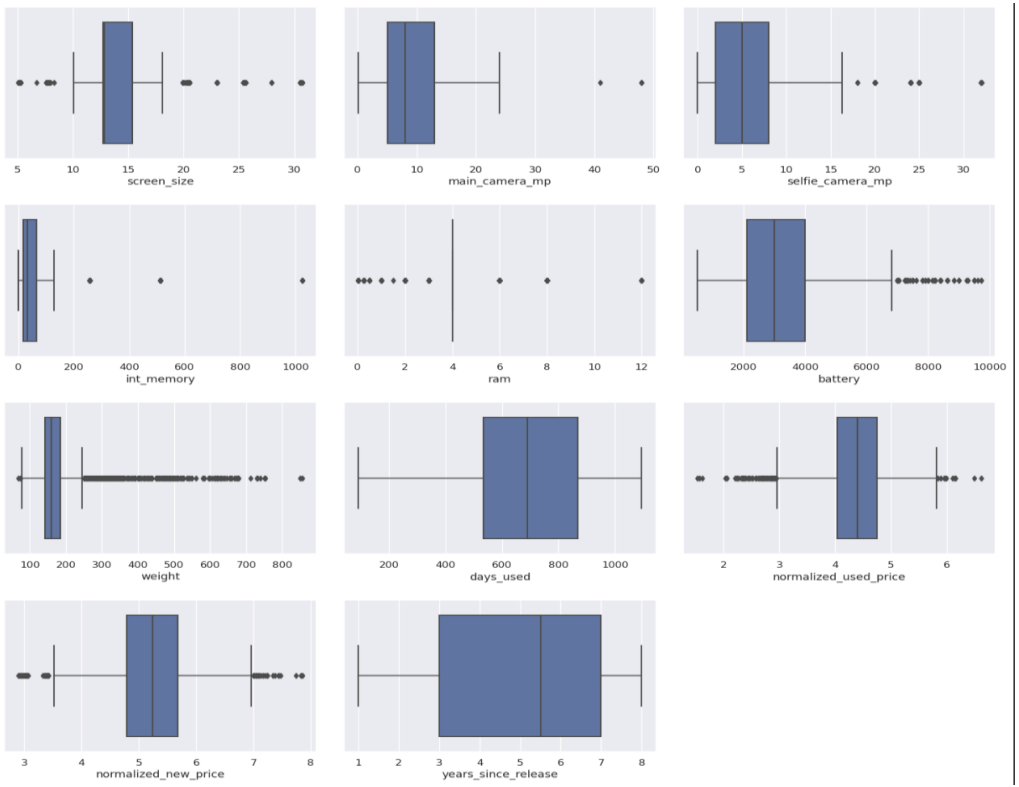
- Prices were higher for used phones that offered 5G.

Feature Engineering

- A new column was created years since release from the release year column.
- Since the year of data collections was 2021, it was used as the baseline and the release
  year column was dropped.

Outlier Check

- There are several outliers in the data.



Train and Test Data

- The number of rows in the train data = 2,417

- The number of rows in the test data = 1037

Linear Regression Model Results

- The R-squared value is at 0.845
- The Adjusted R-squared value is 0.842.
- The constant coefficient is 1.3286
- The coefficient value of main camera is mp is 0.02
- The coefficient value of selfie camera mp is 0.0136
- The coefficient value of int memory is 0.0001
- The coefficient of RAM is 0.02
- The coefficient of battery is –1.585e-05
- The coefficient of weight is 0.0010
- The coefficient for days used is 3.485e-05
- The coefficient of normalized new price is 0.43.
- The coefficient of years since release is –0.02

Model Performance Check on train set V. Test Set

- **Train Performance**

- RMSE = 0.229849
- MAE = 0.180336
- R-squared = 0.844933
- Adj. R-squared = 0.841723
- MAPE = 4.326958

- The train and test RMSE and MAE are both low so the model is not overfitting.

Test Performance

- RMSE = 0.238306
- MAE = 0.184064
- R-squared = 0.842547
- Adjusted R-squared = 0.834731
- MAPE = 4.488006

Model Performance Summary

- Overview of ML model and its parameters
- Summary of most important factors used by the ML model for prediction
- Summary of key performance metrics for training and test data in tabular format for comparison

***Note**: You can use more than one slide if needed*

Link to Appendix slide on model assumptions

APPENDIX