ReCell - Supervised Machine Learning

Michail Mersinias

03/15/2023

Contents / Agenda

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

Executive Summary

- The executive summary (insights and recommendations) of the ReCell Supervised Machine Learning presentation is as follows:
    - **Higher normalized new prices significantly boost used normalized prices:** Aim for higher normalized new prices for comparable devices to boost the prices of equivalent used devices.
    - **Brand name matters a lot:** Focus on devices with good brand name recognition, as the normalized used price may be increased or decreased significantly, based on the chosen brand.
    - **4g capability is of high importance:** Ensure the devices provide at least 4g capability as it provides a strong increase of the normalized used price.
    - **RAM is of medium importance:** High RAM moderately increases the normalized used price.
    - **High number of megapixels in both the main and the selfie camera is appreciated:** Prefer devices with a high number of megapixels in both the main and the selfie cameras, as they both lead to a moderate increase of the normalized used price.

Business Problem Overview and Solution Approach

- Problem Definition:
  - ReCell, is aiming to tap the potential in the market by formulating an ML-driven dynamic pricing strategy for used and refurbished phone/tablet devices.
  - The goal is to analyze the data provided and build a linear regression model to predict the price of a used phone/tablet and identify factors that significantly influence it.
- Solution Approach and Methodology:
  - For EDA, both univariate and multivariate analysis will be performed to find insights.
  - For Data Preprocessing, missing value imputation and feature engineering will be performed.
  - For Model Building, we will use a Linear Regression model after performing data manipulation to ensure that the following five assumptions are satisfied: no multicollinearity, linearity of variables, independence of error terms, normality of error terms, no heteroscedasticity.
  - Metrics used include: RMSE, MAE, R-squared, Adjusted R-squared and MAPE.

EDA Results: Data Overview

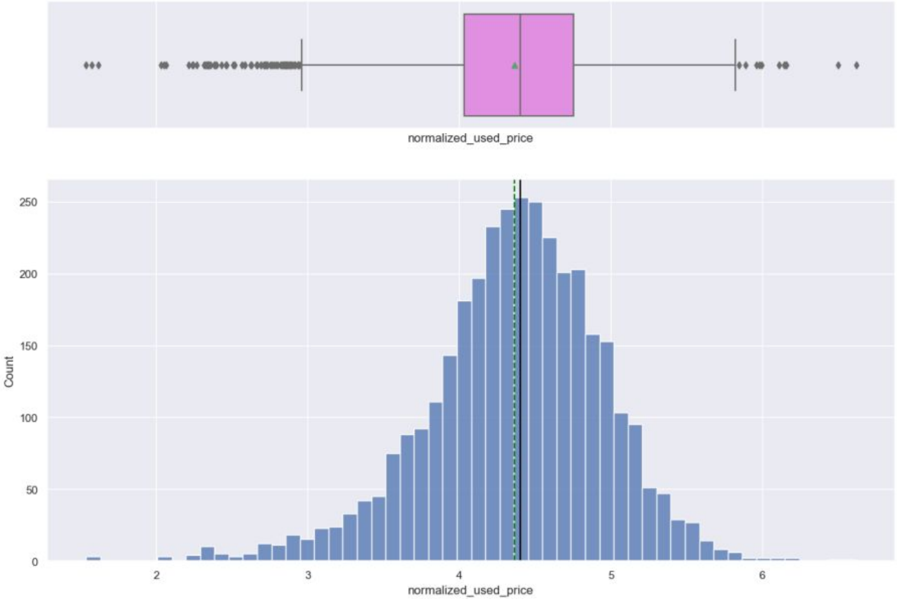
- The dataset is comprised of 3454 rows and 15 columns.
- The columns are as follows: brand_name, os, screen_size, 4g, 5g, main_camera_mp, selfie_camera_mp, int_memory, ram, battery, weight, release_year, days_used, normalized_used_price and normalized_new_price.
- There are no duplicate values in the dataset.
- There are missing values in the dataset. They exist in the following columns: main_camera_mp (179), selfie_camera_mp (2), int_memory (4), ram (4), battery (6) and weight (7).

EDA Results: Univariate Analysis

- In this section, we perform univariate analysis on the data.
  - For each attribute, we perform a descriptive statistical analysis where only that attribute is involved as a variable.
  - We also analyze the corresponding data and present the distribution of the attribute, with confidence intervals to signify variance and statistical significance.
  - Finally, we write the conclusion based on both quantitative and qualitative observations.

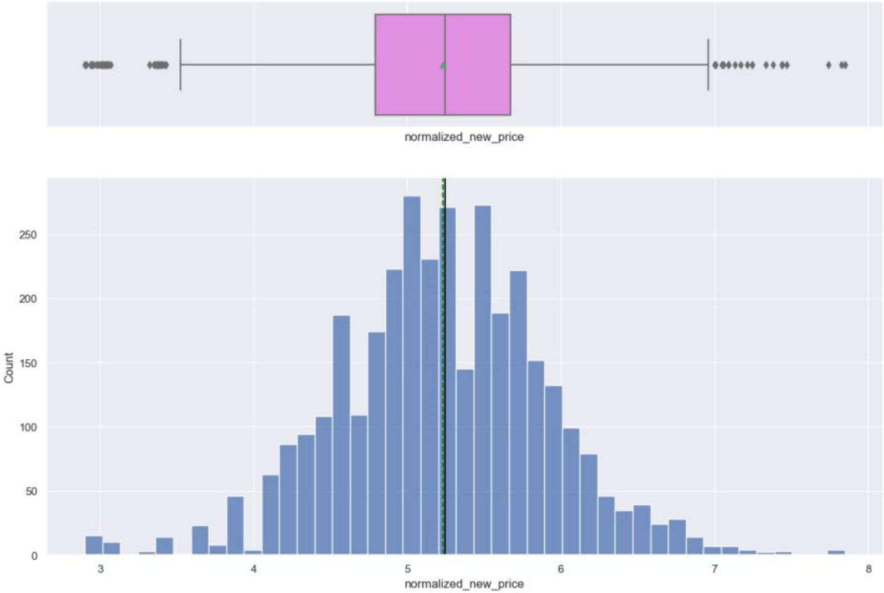
EDA Results: Univariate Analysis

- Analyzing the normalized_used_price attribute, we report a mean value of 4.36, with a standard deviation of 0.59. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



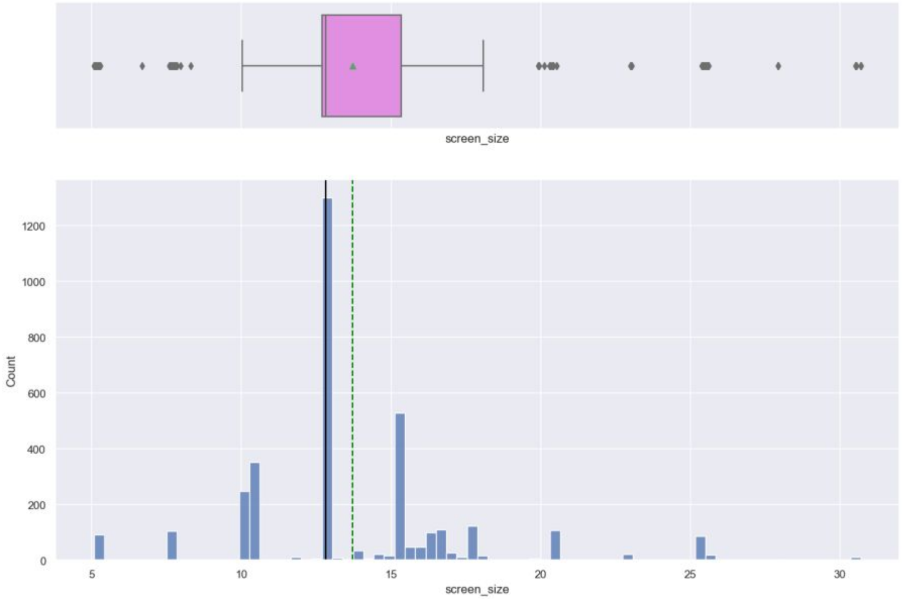
EDA Results: Univariate Analysis

- Analyzing the normalized_new_price attribute, we report a mean value of 5.23, with a standard deviation of 0.68. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



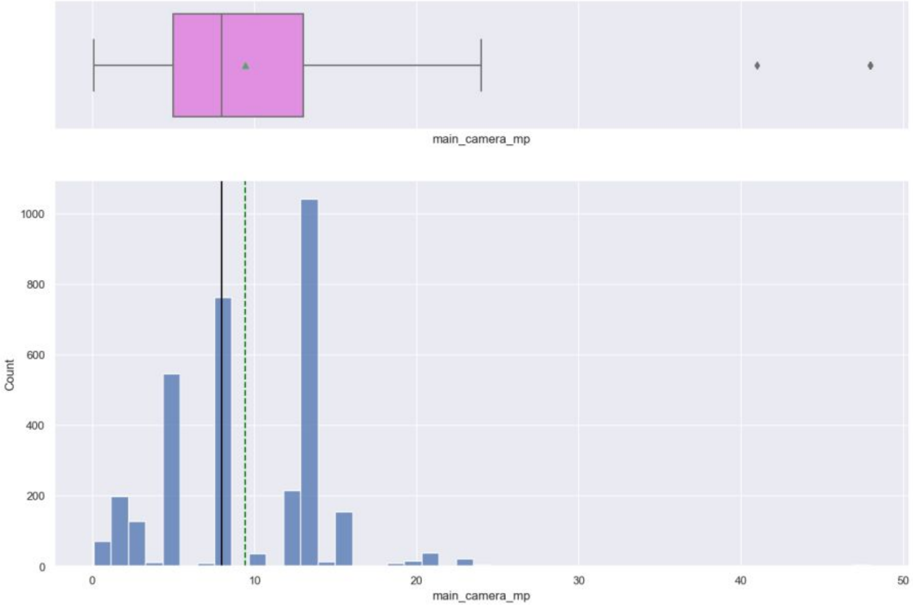
EDA Results: Univariate Analysis

- Analyzing the screen_size attribute, we report a mean value of 13.71, with a standard deviation of 3.81. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



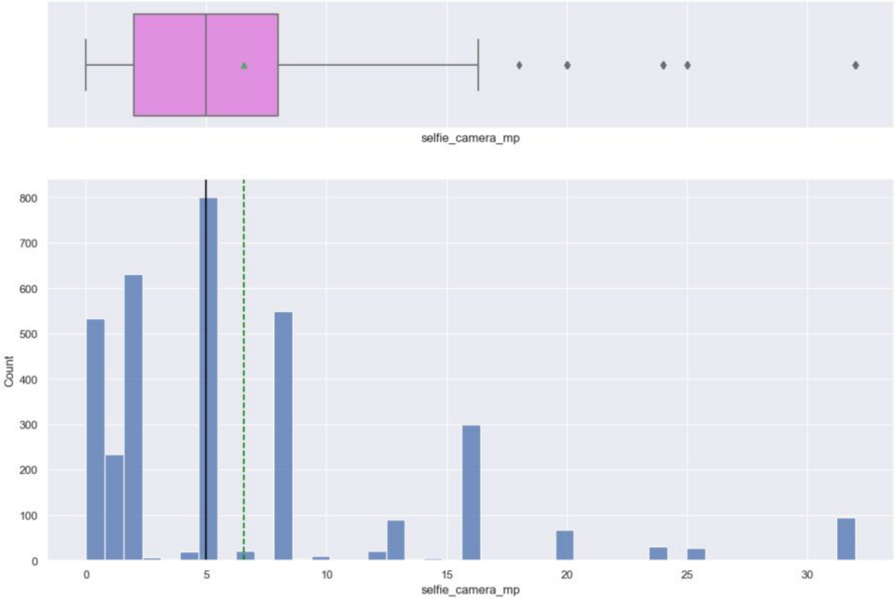
EDA Results: Univariate Analysis

- Analyzing the main_camera_mp attribute, we report a mean value of 9.46, with a standard deviation of 4.82. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



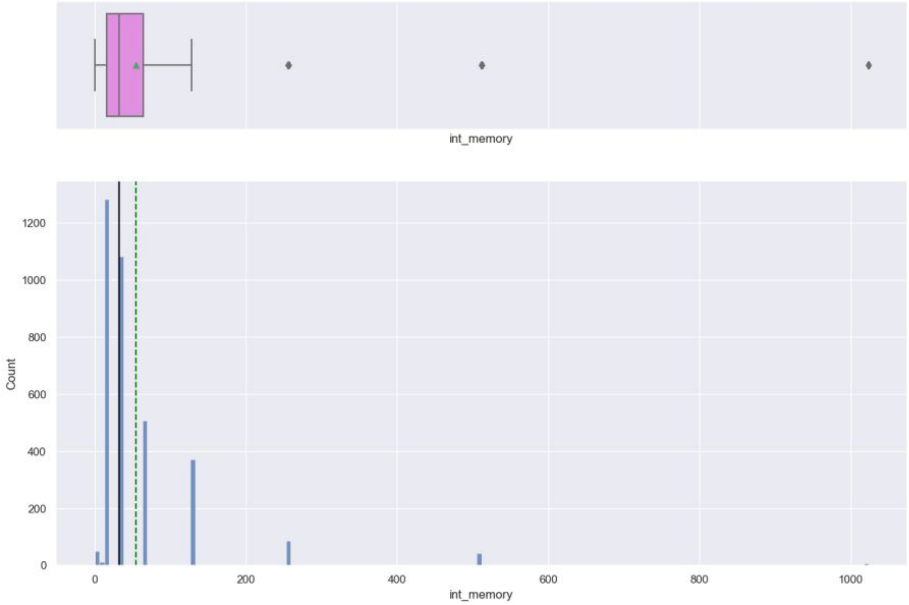
EDA Results: Univariate Analysis

- Analyzing the selfie_camera_mp attribute, we report a mean value of 6.55, with a standard deviation of 6.97. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



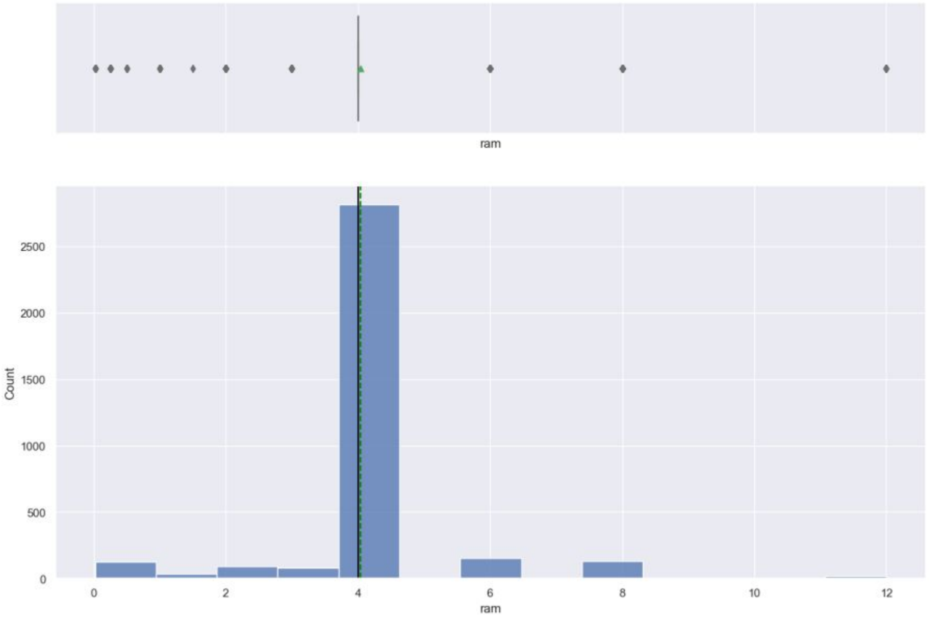
EDA Results: Univariate Analysis

- Analyzing the int_memory attribute, we report a mean value of 54.57, with a standard deviation of 84.97. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



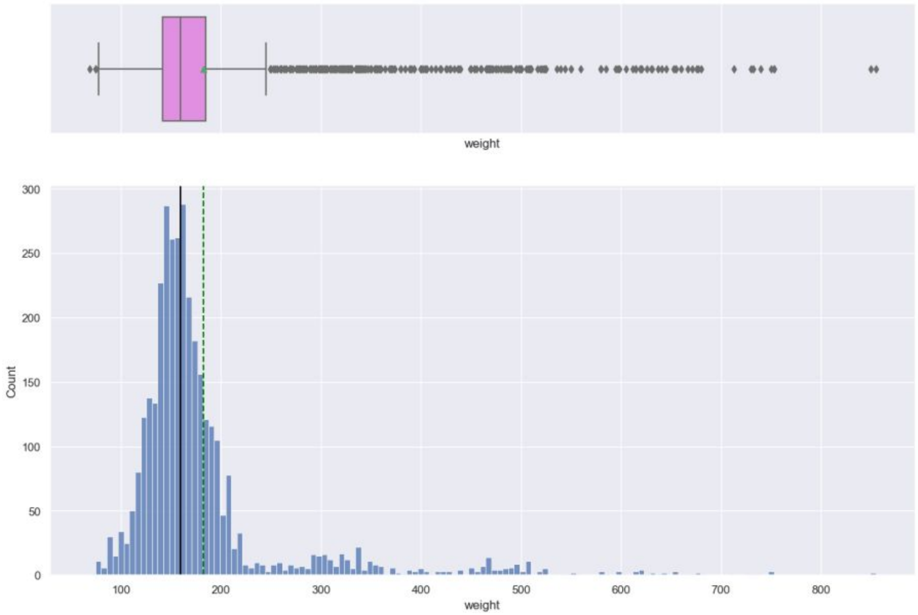
EDA Results: Univariate Analysis

- Analyzing the ram attribute, we report a mean value of 4.04, with a standard deviation of 1.37. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



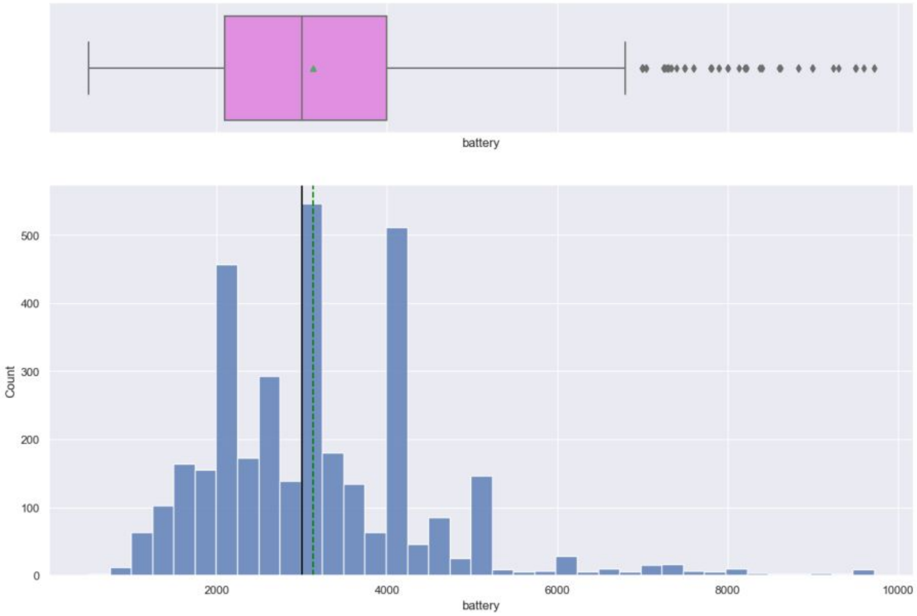
EDA Results: Univariate Analysis

- Analyzing the weight attribute, we report a mean value of 182.75, with a standard deviation of 88.41. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



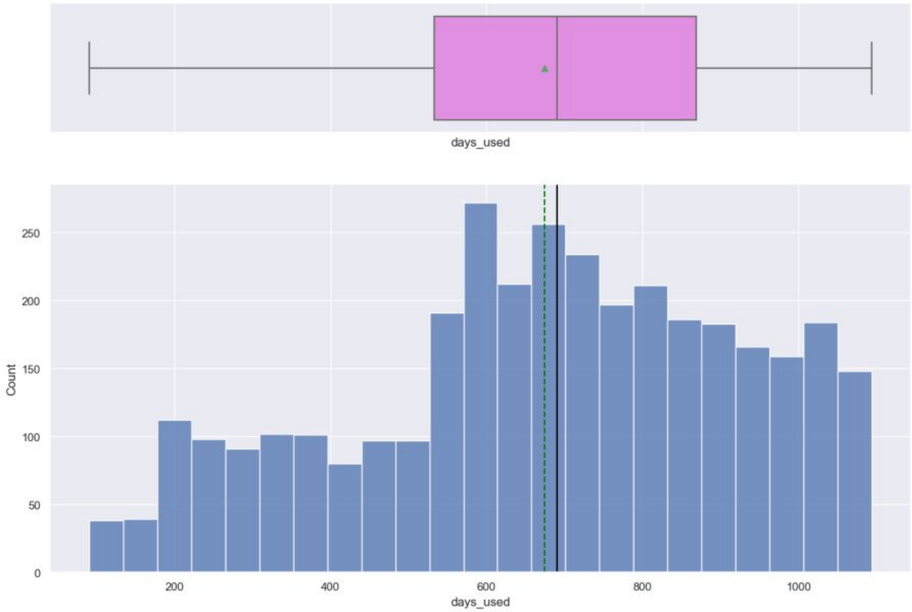
EDA Results: Univariate Analysis

- Analyzing the battery attribute, we report a mean value of 3133.40, with a standard deviation of 1299.68. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



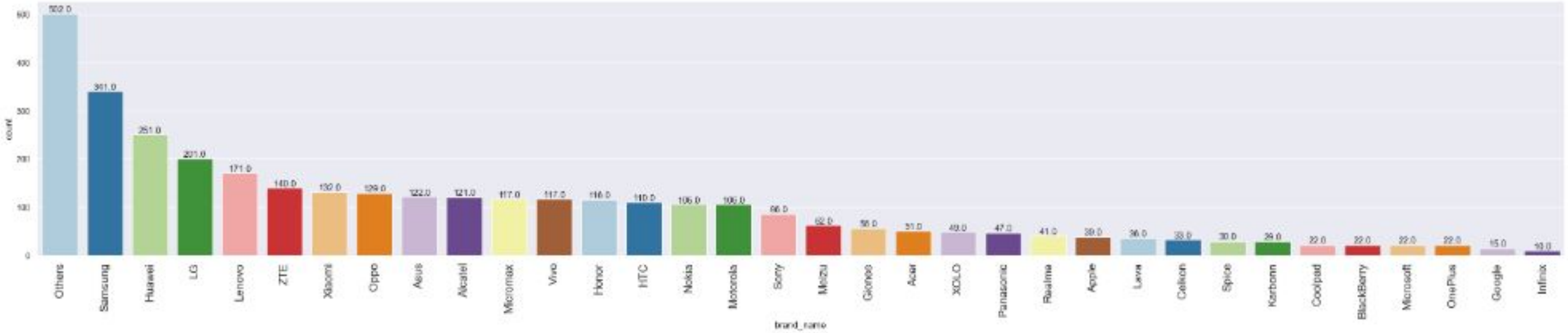
EDA Results: Univariate Analysis

- Analyzing the days_used attribute, we report a mean value of 674.87, with a standard deviation of 248.58. Two graphs, a histplot and a boxplot, are presented to depict this more clearly.



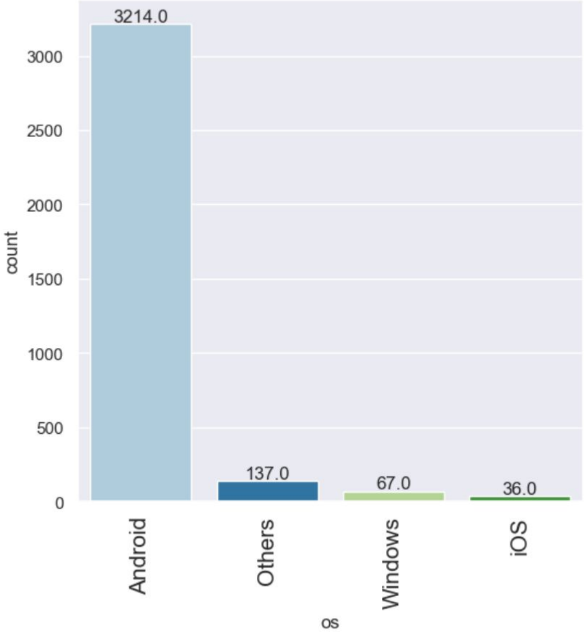
EDA Results: Univariate Analysis

- Analyzing the brand_name attribute, we report that most of the devices belong to Samsung (9.9%), Huawei (7.3%), LG (5.8%), Lenovo (5%) and ZTE (4.1%). The remaining brands follow with less than 4% each. A barplot is presented to depict this more clearly.



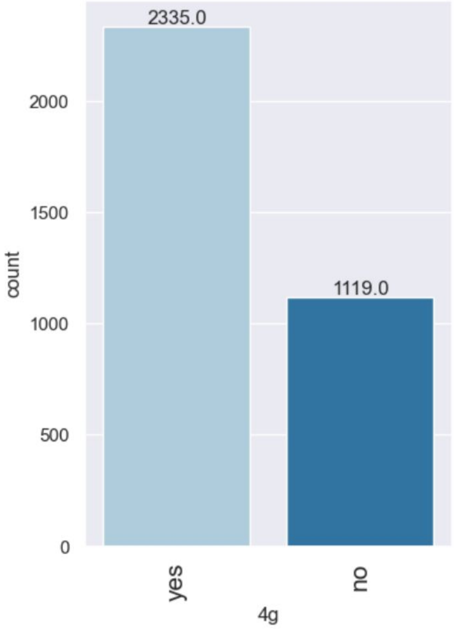
EDA Results: Univariate Analysis

- Analyzing the os attribute, we report that the vast majority of devices (3214) has an Android operating system. A barplot is presented to depict this more clearly.



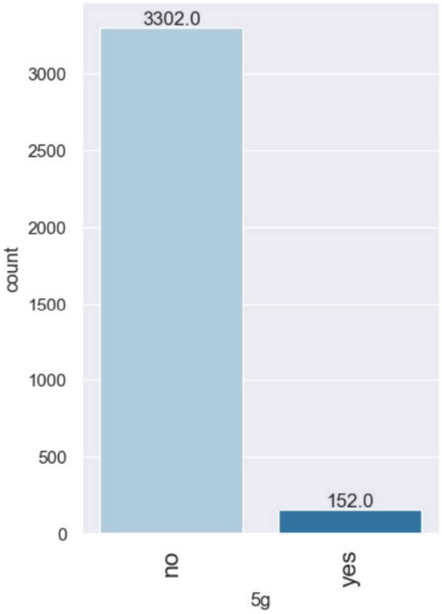
EDA Results: Univariate Analysis

- Analyzing the 4g attribute, we report that the majority of devices (2235) has 4g capability. A barplot is presented to depict this more clearly.



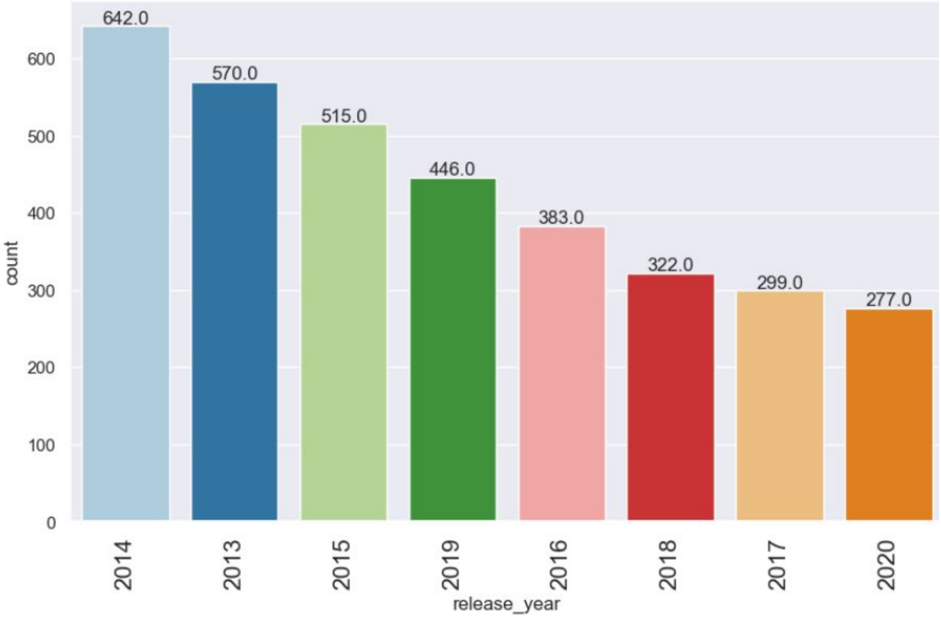
EDA Results: Univariate Analysis

- Analyzing the 5g attribute, we report that the vast majority of devices (3302) does not have 5g capability. A barplot is presented to depict this more clearly.



EDA Results: Univariate Analysis

- Analyzing the release_year attribute, we report that most of the phones were released in the years 2014 (642), 2013 (570) and 2015 (515). A barplot is presented to depict this more clearly.

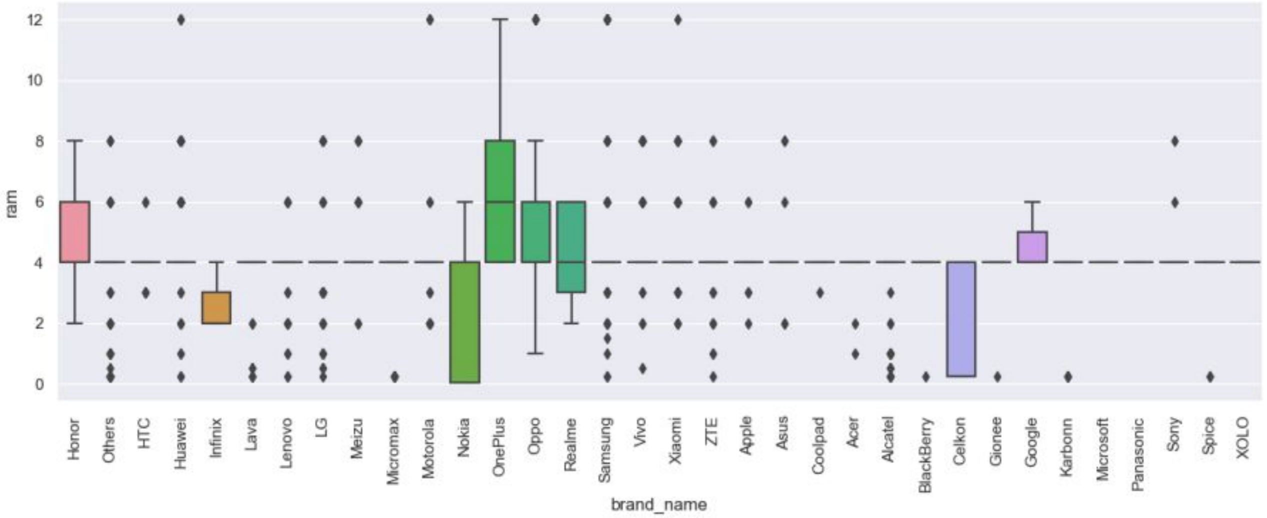


EDA Results: Multivariate Analysis

- In this section, we perform multivariate analysis on the data.
    - For selected attributes, we perform a descriptive statistical analysis where each attribute is jointly examined along with other attributes in order to evaluate their relationship and correlation.
    - We also analyze the corresponding data and present the joint distribution of the attributes, with confidence intervals to signify variance and statistical significance.
    - Finally, we write the conclusion based on both quantitative and qualitative observations.

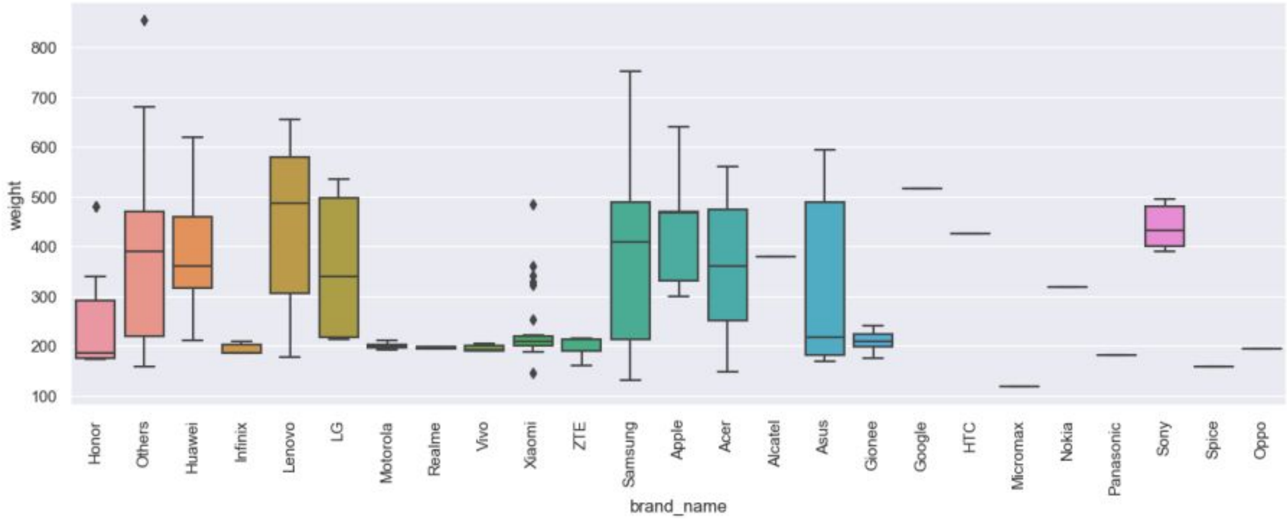
EDA Results: Multivariate Analysis

- Analyzing the relationship between the ram and brand_name attributes, we observe that there is a lot of variance across devices. Specifically, OnePlus seems to have consistently high levels of RAM, while Celkon has low levels. It is worth noting that devices such as Huawei, Micromax, Oppo, Samsung and Xiaomi have outliers that display very high levels of RAM.



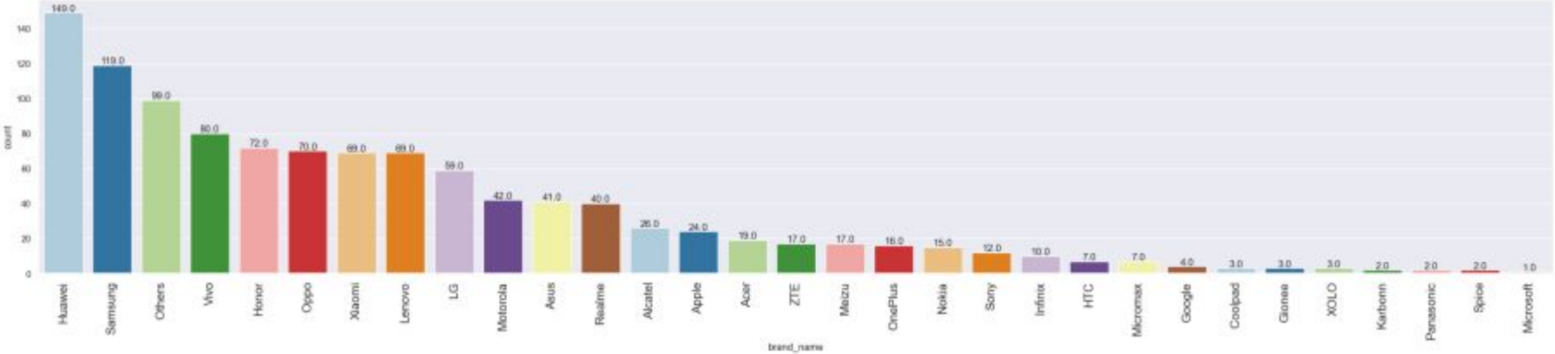
EDA Results: Multivariate Analysis

- Analyzing the relationship between the weight and brand_name attributes, we observe that once again there is a lot of variance across devices. Specifically, the heaviest devices appear in the outliers of the Others category, followed by Samsung, while Lenovo has the highest mean weight. Infinix, Motorola, Realme, Vivo, Xiaomi, ZTE, Gionee, Micromax, Panasonic, Spice and Oppo are the lightest.



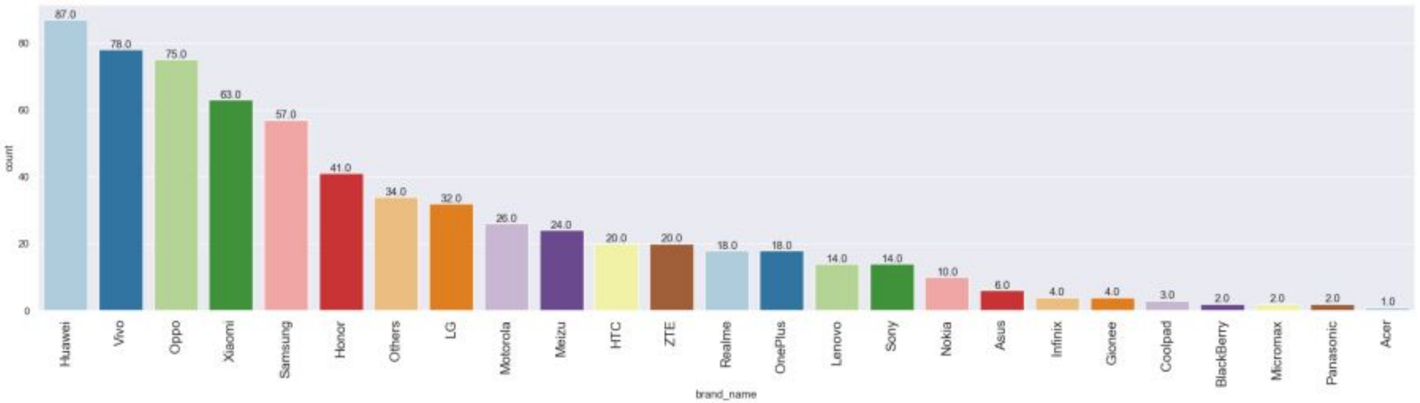
EDA Results: Multivariate Analysis

- Analyzing the relationship between the screen_size and brand_name attributes, we observe that the vast majority of large screens appear in devices from Huawei (149), Samsung (119), Others (99) and Vivo (80). They are followed by Honor (72), Oppo (70), Xiaomi (69), Lenovo (69) and LG (59). The remaining brands follow with smaller values.



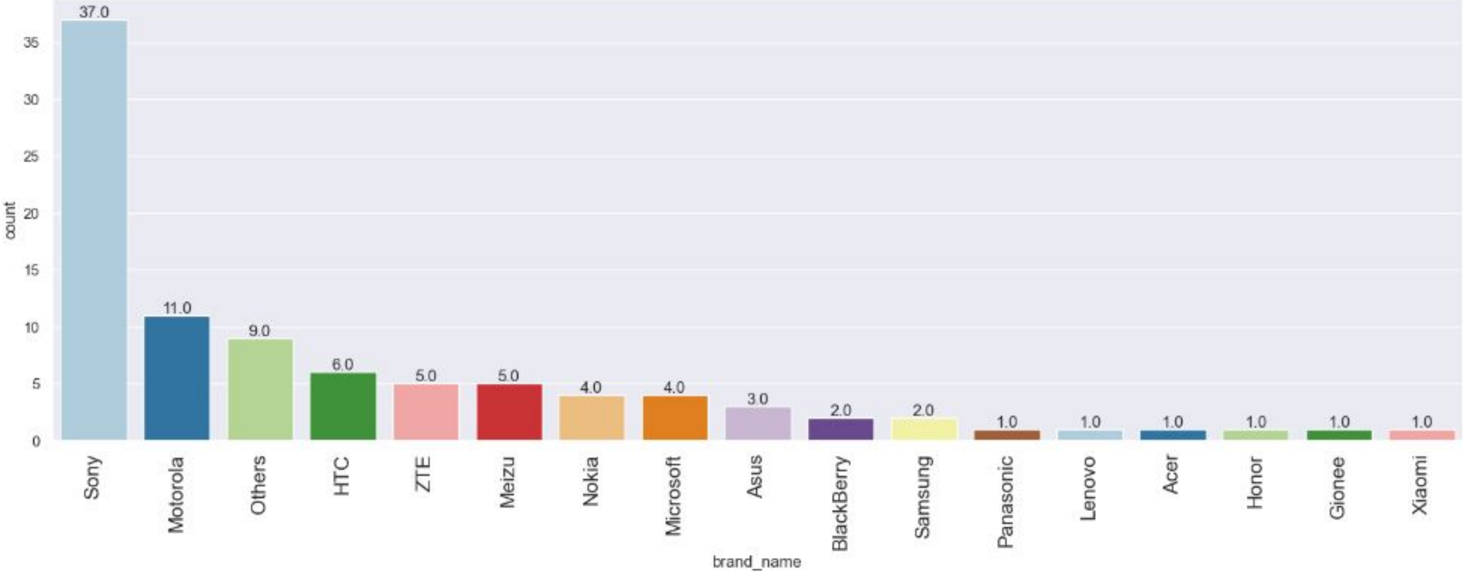
EDA Results: Multivariate Analysis

- Analyzing the relationship between the selfie_camera_mp and brand_name attributes, we observe that the vast majority of devices with over 8 megapixels selfie cameras are from Huawei (87), Vivo (78), Oppo (75), Xiaomi (63) and Samsung (57). The remaining brands follow with smaller values.



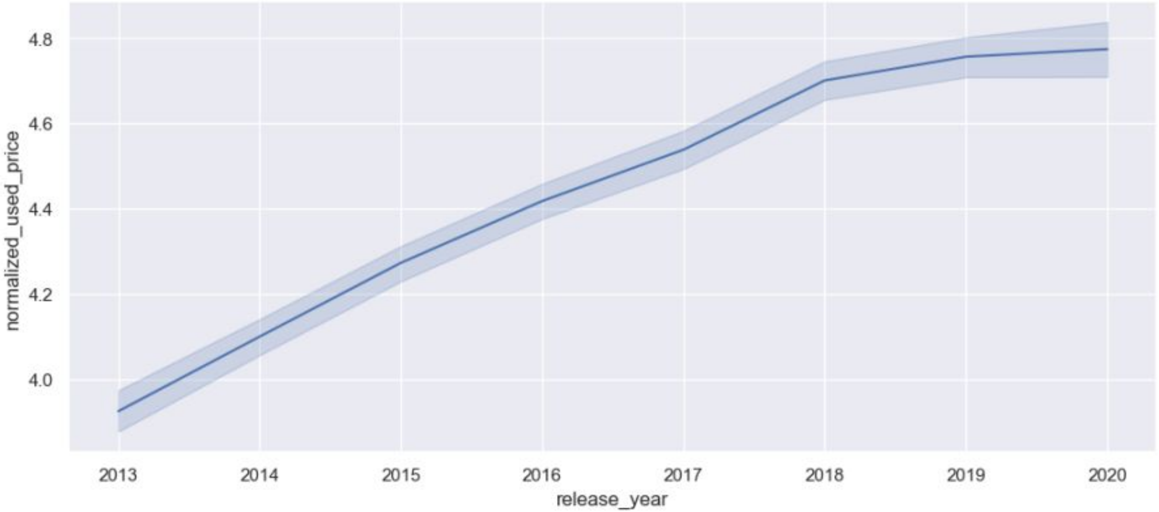
EDA Results: Multivariate Analysis

- Analyzing the relationship between the main_camera_mp and brand_name attributes, we observe that the vast majority of devices with over 16 megapixels main rear cameras come from Sony, a renown brand for its camera quality, followed by Motorola (11%), while the rest are under 6%.



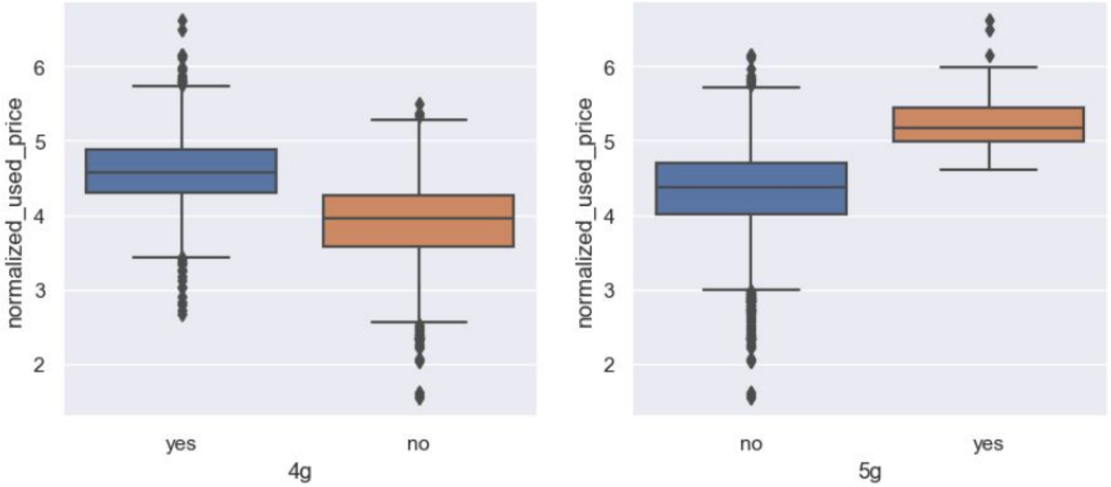
EDA Results: Multivariate Analysis

- Analyzing the relationship between the normalized_used_price and release_year attributes, we observe that the newer the device is, the higher the normalized used price it has. Specifically, if the device was released within the past three years, it has retained most of its value as it is after the three years point that the drop in price becomes steep.



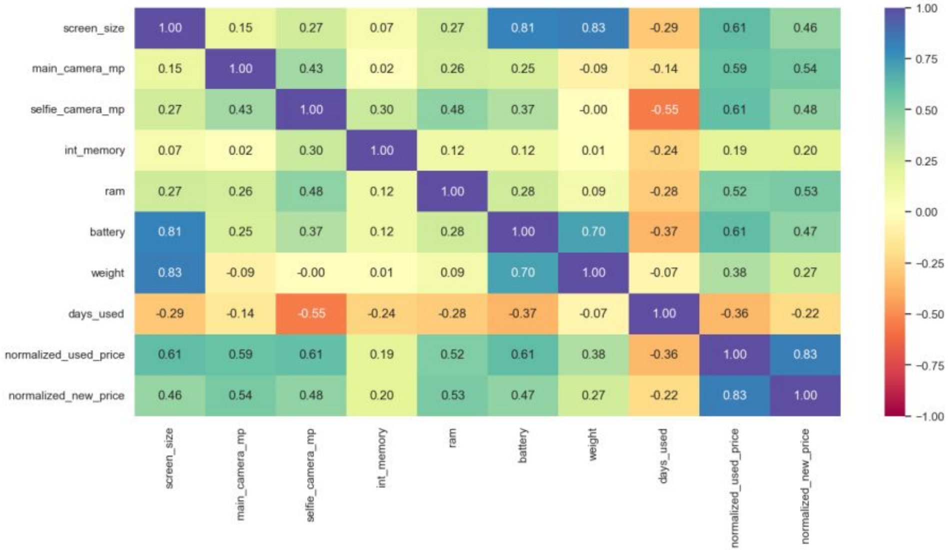
EDA Results: Multivariate Analysis

- Analyzing the relationship between the normalized_used_price and 4g/5g attributes, we observe that it is statistically significant to state that if the device offers 4g or 5g capability, then the normalized used price will also be higher.



EDA Results: Multivariate Analysis

- Analyzing the correlation matrix, we can make certain observations. Normalized used price is mostly correlated with normalized new price (0.83), screen size (0.61), battery (0.61) as well as the main and the selfie camera megapixels (0.61 and 0.59) respectively. On the other hand, normalized new price is mostly correlated with normalized used price (0.83), main camera megapixels (0.54) and ram (0.53) while screen size, selfie camera megapixels and battery show relatively weaker correlations.



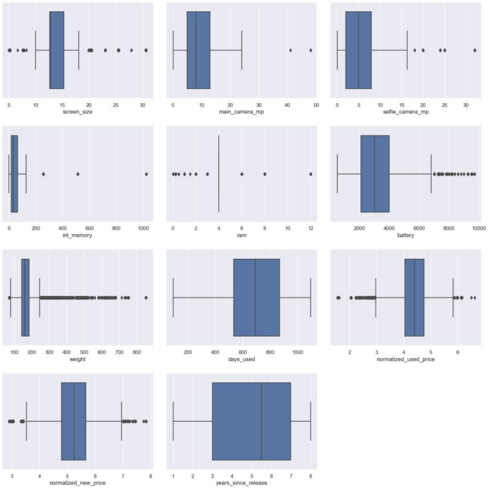
	screen_size	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	days_used	normalized_used_price	normalized_new_price
screen_size	1.00	0.15	0.27	0.07	0.27	0.81	0.83	-0.29	0.61	0.46
main_camera_mp	0.15	1.00	0.43	0.02	0.26	0.25	-0.09	-0.14	0.59	0.54
selfie_camera_mp	0.27	0.43	1.00	0.30	0.48	0.37	-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

Data Preprocessing

- Missing Value Imputation:
    - First, we impute the missing values in the data by the column medians grouped by release_year and brand_name.
    - Then, we impute the remaining missing values in the data by the column medians grouped by brand_name.
    - Finally, we fill the remaining missing values in the main_camera_mp column by the column median.
- Feature Engineering:
    - New Column: years_since_release, defined as the difference between 2021 and release_year.
    - Delete Column: release_year, as it is now redundant.

Data Preprocessing

- Outlier Check: There are numerous outliers in this most columns of this dataset, however no action will be taken as they are accurate values that contain important information about each feature.

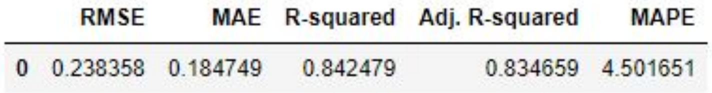


- Preparing Data for Modeling: We use normalized_used_price as our target variable which we seek to predict accurately, and all the other columns as features to train our model.

- Train/Test Dataset Split: We split the data in 70:30 ratio for train to test data.

Model Performance Summary

- Modeling Setup:
  - First, we choose our model, which is Logistic Regression.
  - Then, we define functions for the following metrics of performance: RMSE, MAE, R-squared, Adjusted R-squared and MAPE.
- First Model Evaluation:
  - The results on the test data are the following:



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

  - However, we still need to check the five assumptions of Linear Regression and also attempt to improve our the performance of our model in regard to these metrics.

Model Performance Summary

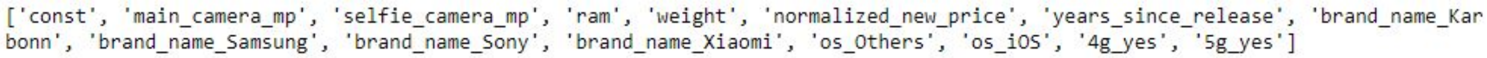
- Five Assumptions for Linear Regression:
  - No Multicollinearity.
  - Linearity of variables.
  - Independence of error terms.
  - Normality of error terms.
  - No Heteroscedasticity.

Model Performance Summary

- Test for Multicollinearity:
  - We test for multicollinearity with a function that calculates the VIF score for each feature.
  - If VIF is greater than 5, there is moderate multicollinearity and if it is greater than 10, there is high multicollinearity. The ideal VIF of 1 means there is no correlation between the kth predictor and the remaining predictor variables.
  - Thus, we perform the following algorithm to tackle multicollinearity in the training data:
    - We examine the adjusted R-squared and RMSE of the model after dropping every variable one-by-one that has a VIF score greater than 5. Dummy variables are ignored.
    - We drop the variable that makes the least change in the adjusted R-squared metric.
    - We repeat this process until there is no column with a VIF score greater than 5.
  - Column dropped to remove multicollinearity: screen_size.

Model Performance Summary

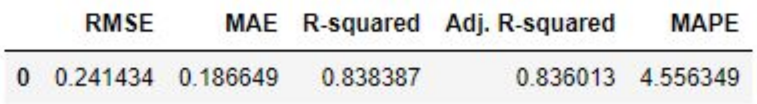
- Dropping high p-value variables:
  - We drop the predictor variables having a p-value greater than 0.05 as they do not significantly impact the target variable.
  - We will drop the predictor variables one-by-one and iteratively build a model and check the new p-values, as sometimes p-values change after dropping a variable.
  - We repeat the process as described above until there are no columns with p-value > 0.05.
  - Final selected features:



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

Model Performance Summary

- Modeling Setup:
  - Our chosen model is Logistic Regression, with the selected training features.
  - Our metrics of performance are: RMSE, MAE, R-squared, Adjusted R-squared and MAPE.
- Second Model Evaluation:
  - The results on the test data are the following:

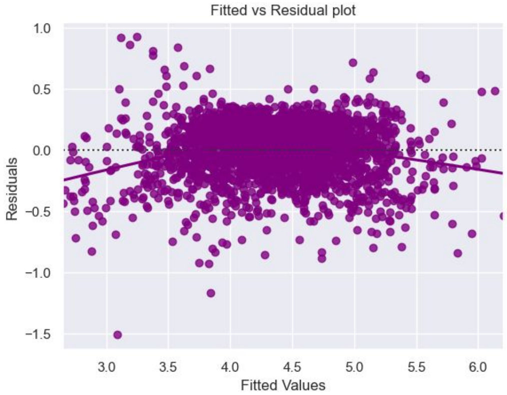


	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.241434	0.186649	0.838387	0.836013	4.556349

  - However, we still need to check the remaining assumptions of Linear Regression and also further attempt to improve our the performance of our model in regard to these metrics.

Model Performance Summary

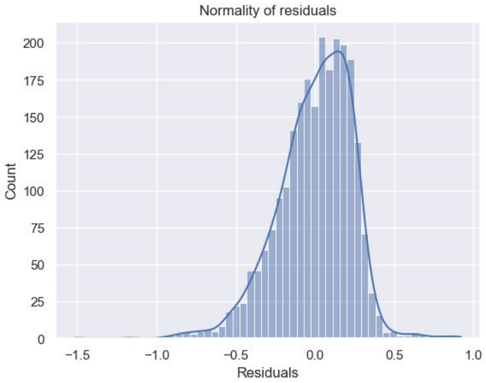
- Test for Linearity and Independence:
  - We test for linearity and independence by making a plot of fitted values vs residuals and checking for patterns.
  - If there is no pattern, then we say the model is linear and residuals are independent. Otherwise, the model is showing signs of non-linearity and residuals are not independent.

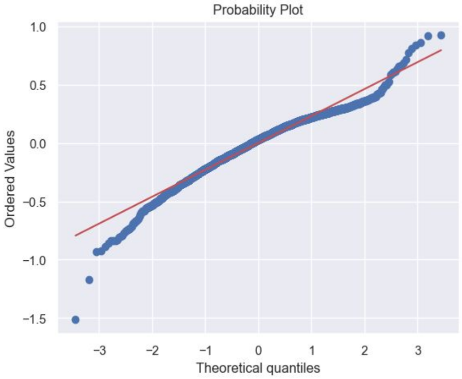


  - No pattern can be observed, thus the model is linear and the residuals are independent.

Model Performance Summary

- Test for Normality:
  - We test for normality by checking the distribution of residuals, by checking the Q-Q plot of residuals, and by using the Shapiro-Wilk test.
  - If the residuals follow a normal distribution, they will make a straight line plot, otherwise not. Alternatively, if the p-value of the Shapiro-Wilk test is greater than 0.05, we can say the residuals are normally distributed.





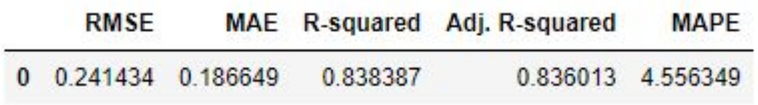
  - From the plots above, we can conclude that the residuals follow a normal distribution.

Model Performance Summary

- Test for Homoscedasticity:
  - We test for homoscedasticity by using the Goldfeld-Quandt test.
  - If we get a p-value greater than 0.05, we can say that the residuals are homoscedastic. Otherwise, they are heteroscedastic.
  - Reported p-value: 0.44
  - As the p-value of the Goldfeld-Quandt test is greater than 0.05, the residuals are homoscedastic.
  - Therefore, all five assumptions of Logistic Regression are now satisfied.

Model Performance Summary

- Modeling Setup:
  - Our chosen model is Logistic Regression, with the selected features. All assumptions satisfied.
  - Our metrics of performance are: RMSE, MAE, R-squared, Adjusted R-squared and MAPE.
- Third and Final Model Evaluation:
  - The results on the test data are the following:



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

  - The metrics are generally improved compared to the initial model: RMSE is 0.2414 (up from 0.2383), MAE is 0.1866 (up from 0.1847), R-Squared is 0.8383 (down from 0.8424), Adjusted R-Squared is 0.8360 (up from 0.8346) and MAPE is 4.55 (up from 4.50).
  - Thus, we achieve the optimal performance for the task of predicting the price of a used phone/tablet. We also view the coefficients to identify factors that significantly influence it.

Executive Summary

- The executive summary (insights and recommendations) of the ReCell Supervised Machine Learning presentation is as follows:
    - **Higher normalized new prices significantly boost used normalized prices:** Aim for higher normalized new prices for comparable devices to boost the prices of equivalent used devices.
    - **Brand name matters a lot:** Focus on devices with good brand name recognition, as the normalized used price may be increased or decreased significantly, based on the chosen brand.
    - **4g capability is of high importance:** Ensure the devices provide at least 4g capability as it provides a strong increase of the normalized used price.
    - **RAM is of medium importance:** High RAM moderately increases the normalized used price.
    - **High number of megapixels in both the main and the selfie camera is appreciated:** Prefer devices with a high number of megapixels in both the main and the selfie cameras, as they both lead to a moderate increase of the normalized used price.

Thank you for your time!

Michail Mersinias

03/15/2023