ReCell start up

Supervised Learning - Foundations Project

Contents / Agenda

Executive Summary

- **Actionable Insights and Recommendations**
  - The most important feature is the new price.
    - A unit increase in a new price increases the used price by 0.4 unit with other features are constant.
    - If the device was 4G, the used price will increase by 0.04 with other features are constant.
  - Lenovo, Nokia and Xiamo brand name are in very high demand and have and affect the price.
    - if the brand name one of these brands the used price with increase by 0.04,0.09,0.07 unite with other features are constant. Nokia has the higher price effect.
  - Days of used, weight, and intel memoires are the least important features, and we probably can drop them from the model.
  - Years since release and 5G features decrease the price of the used devices.
    - One unit increase in years since release will decease the used price by 0.015 unit. While if it is 5G will decrease the price by 0.06 unit.
  - To increase the profit in this company We recommend to not include 5G and old devices because that will reduce the price. New Price and brand are more important than other physical features. Weight and int memory are less import features which will not affect the price.

Business Problem Overview and Solution Approach

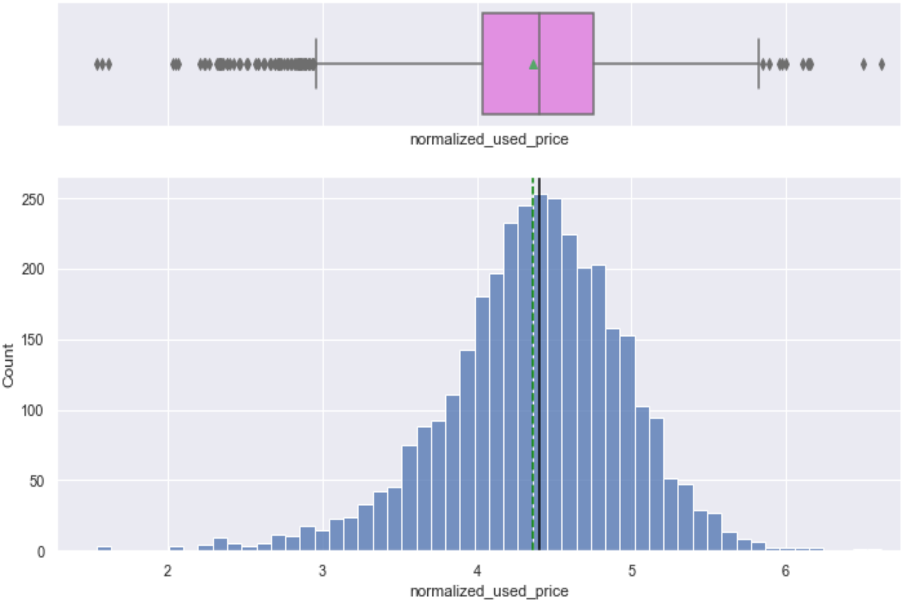
- ReCell, a startup aiming to tap the potential in used and refurbished devices market. They want to predict the price of a used phone/tablet and identify factors that significantly influence it.

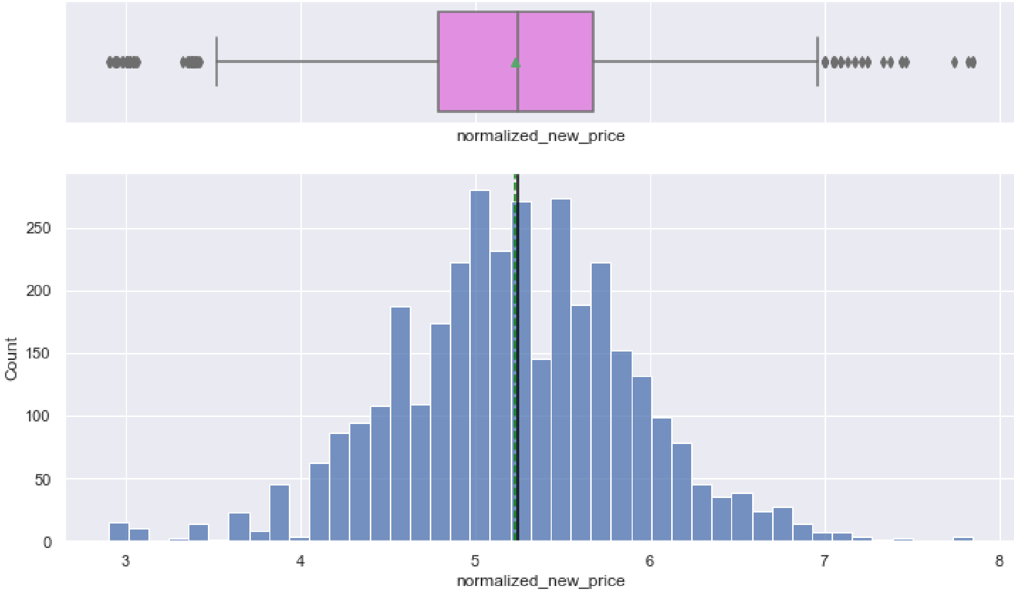
- To solve this problem, I will apply multiple linear regression model to predict the the used phone/table price and find the main features that mostly affect price prediction.

Link to Appendix slide on data background check

Univariate Analysis

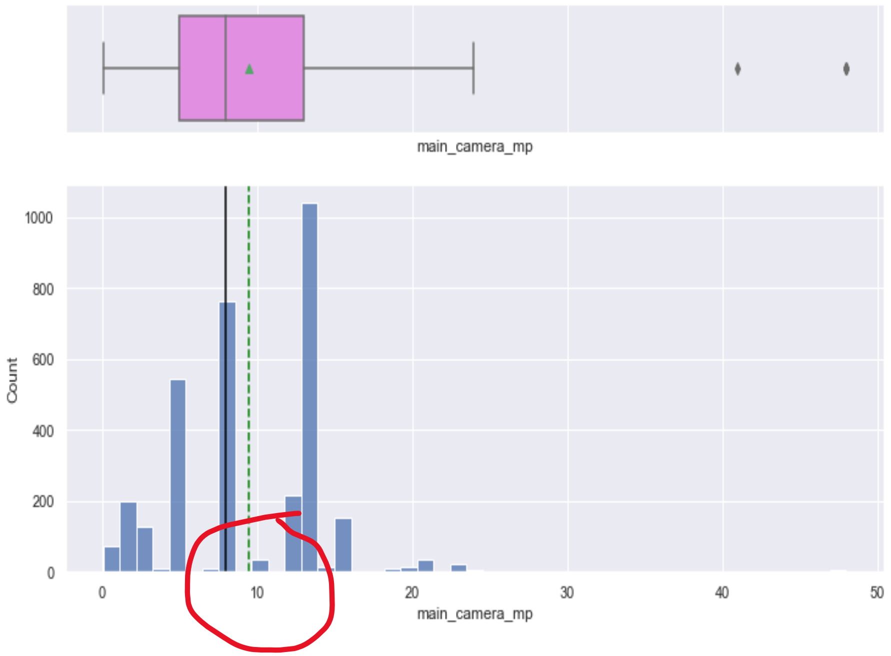
- Normalized Prices
    - Normalized Prices look close to the normal distribution.



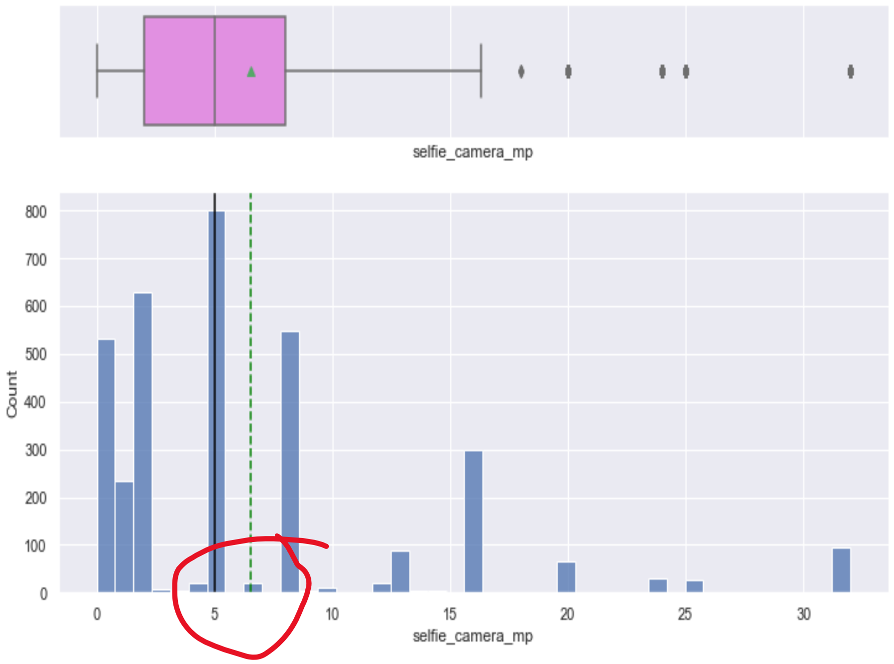


Univariant Analysis

- Resolution of the rear camera in megapixels

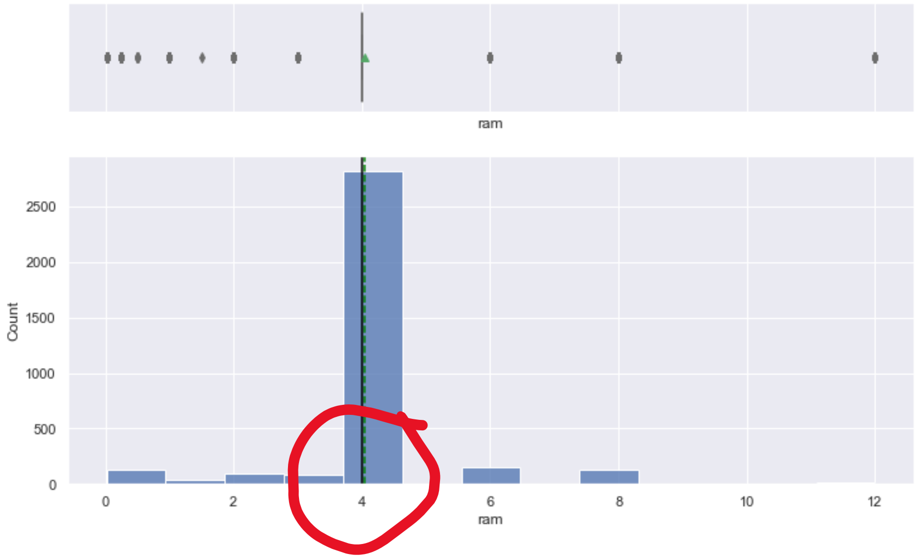


- Resolution of the front camera in megapixels

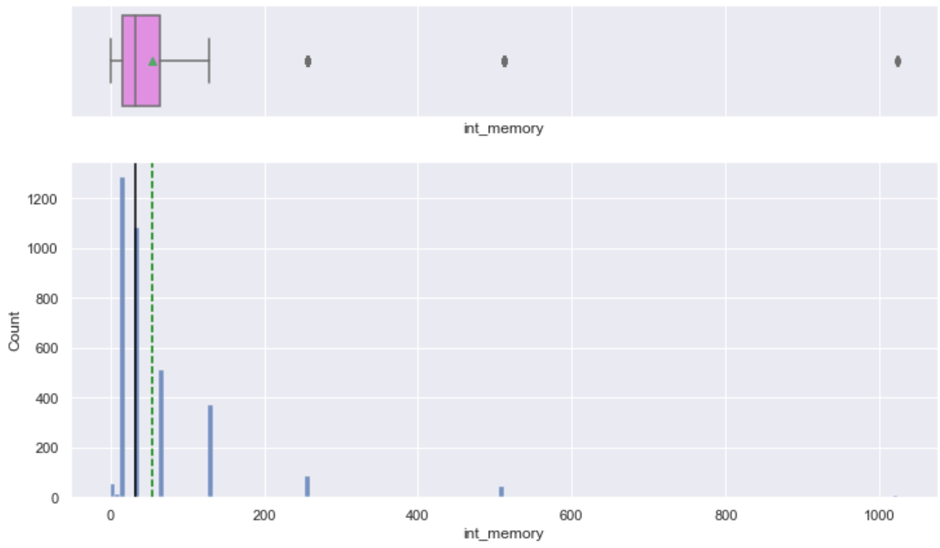


Univariant Analysis

- Amount of RAM in GB

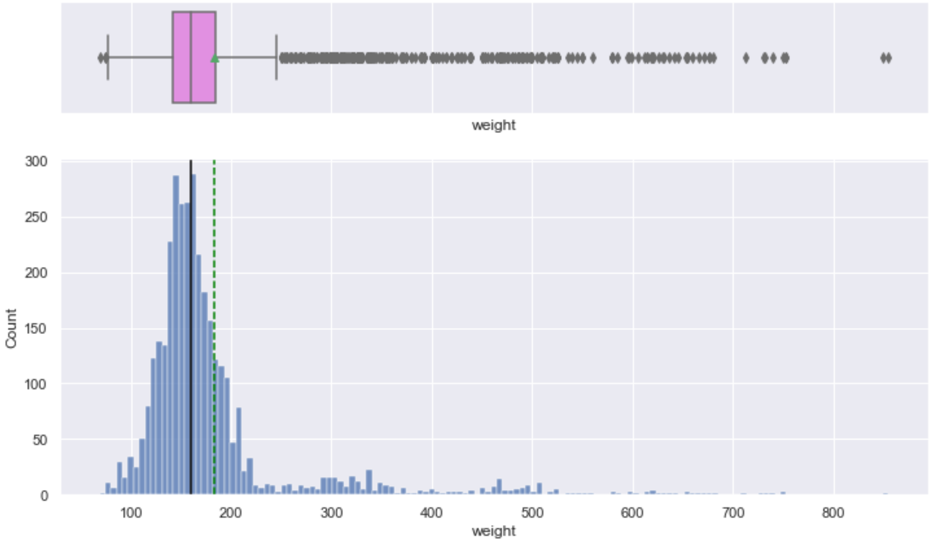


- Amount of internal memory (ROM) in GB

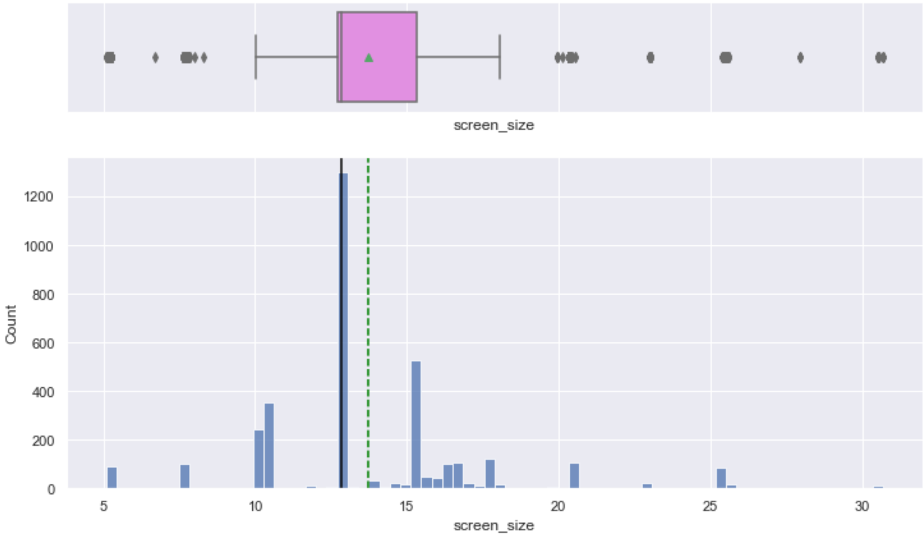


Univariant Analysis

- Weight of the device in grams

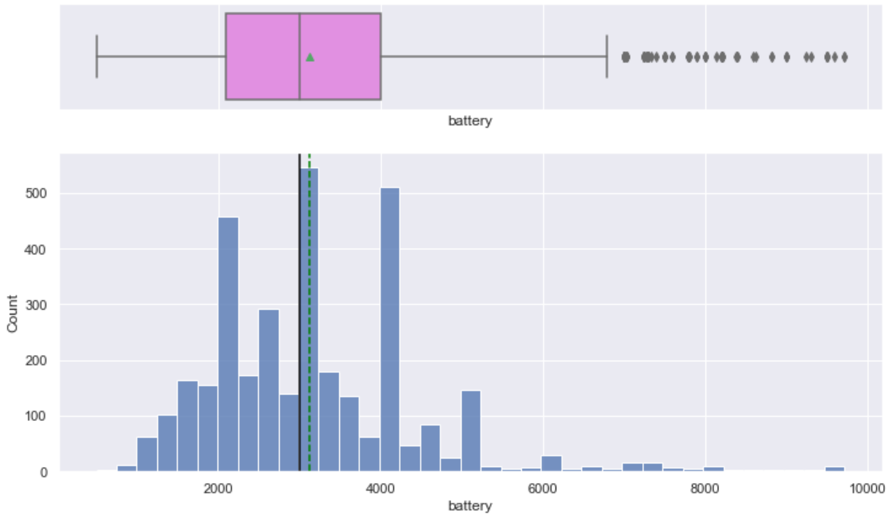


- Screen size

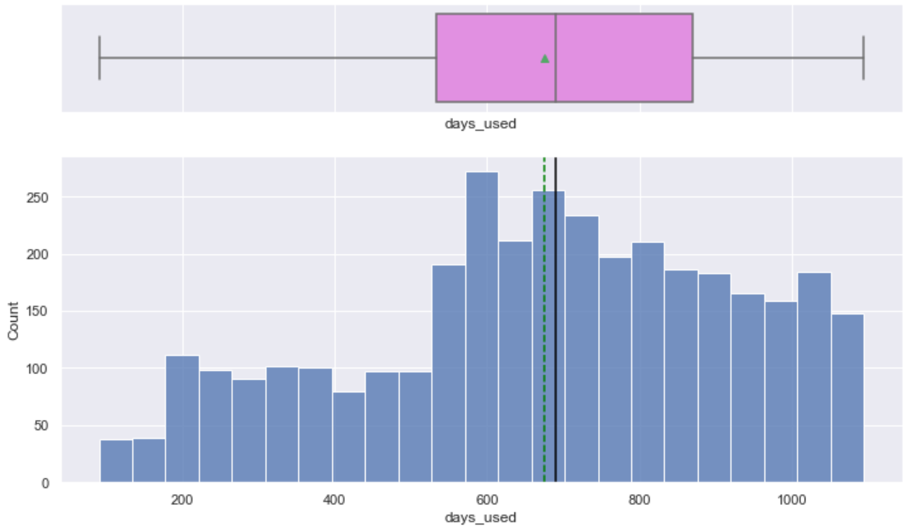


Univariant Analysis

- Energy capacity of the device battery in mAh

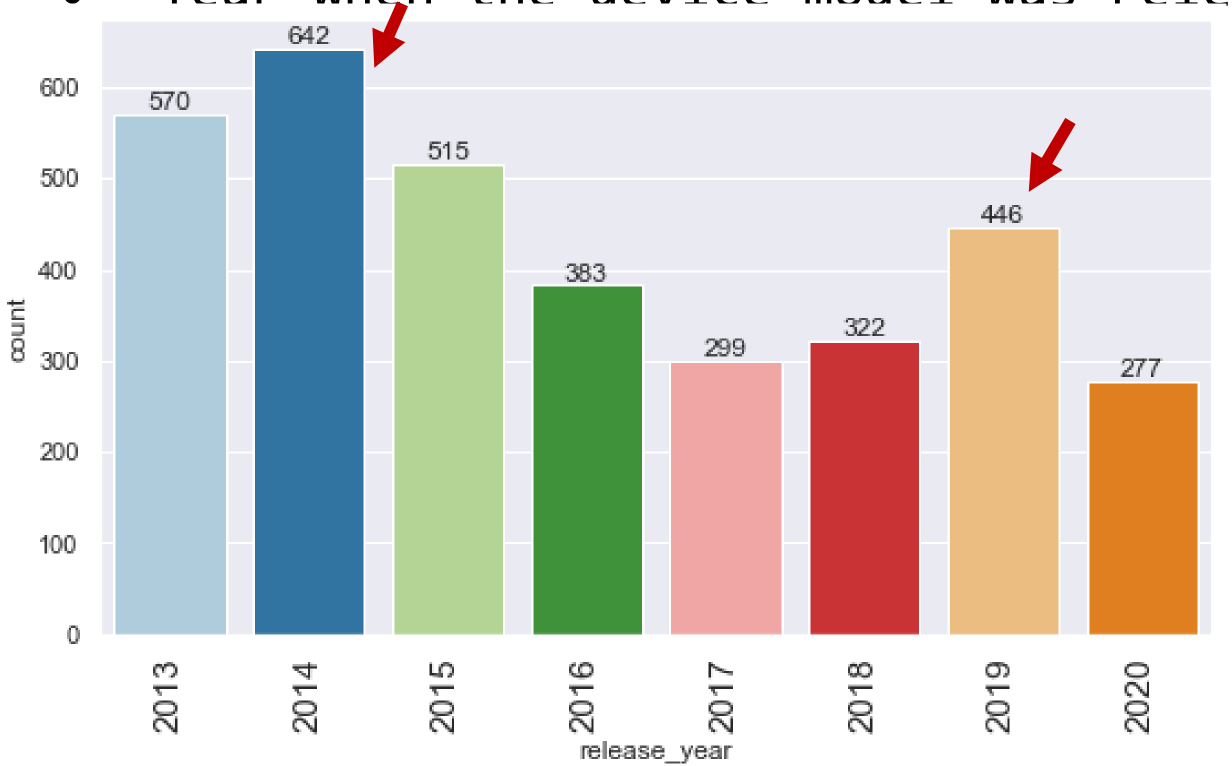


- Number of days the used/refurbished device has been used



Univariant Analysis

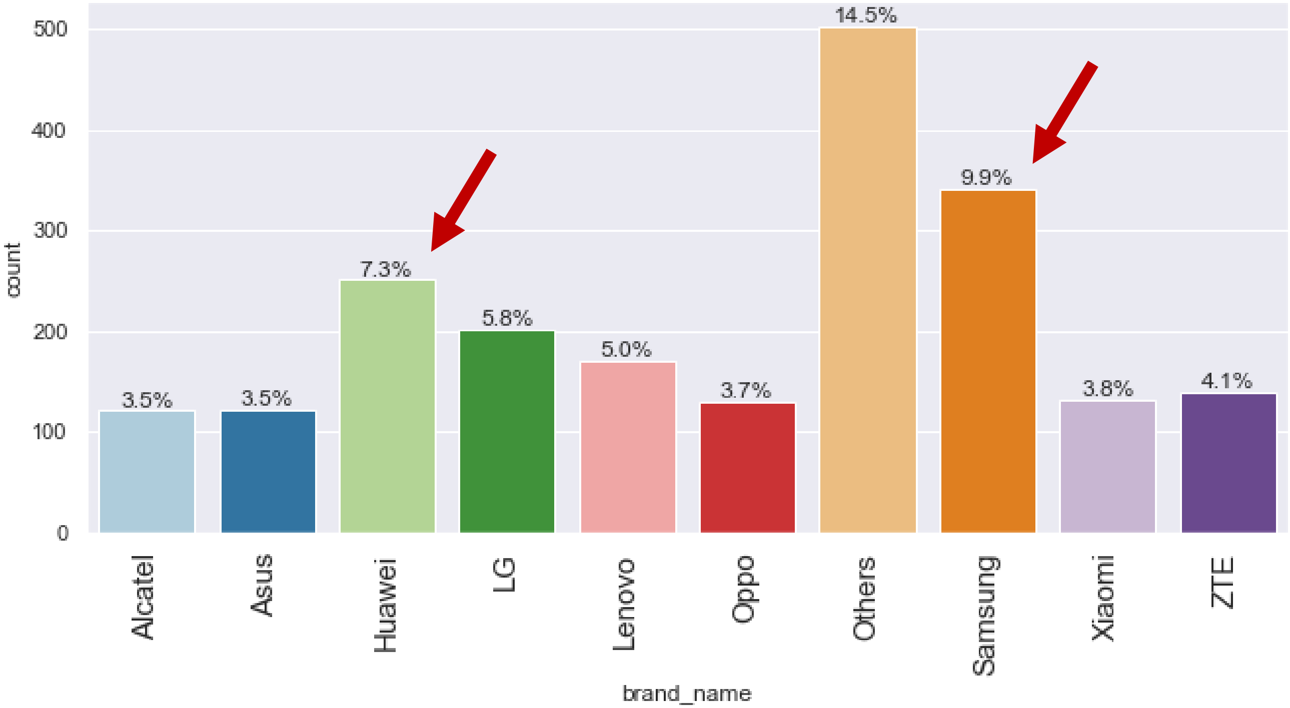
- Year when the device model was released



2014 and 2019 show a jump in the number of the devices were released.

Univariant Analysis

- `Name of manufacturing brand`

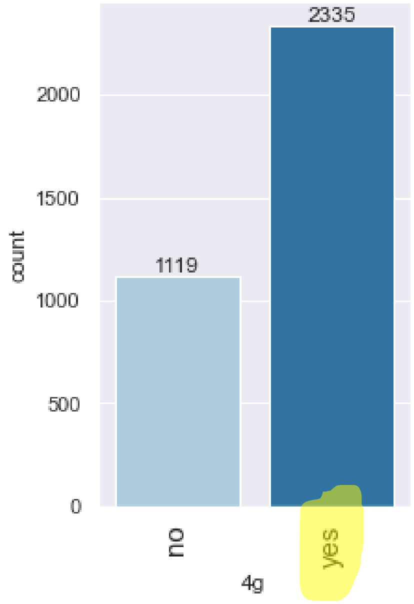


Samsung and Huawei are the most salable device brands

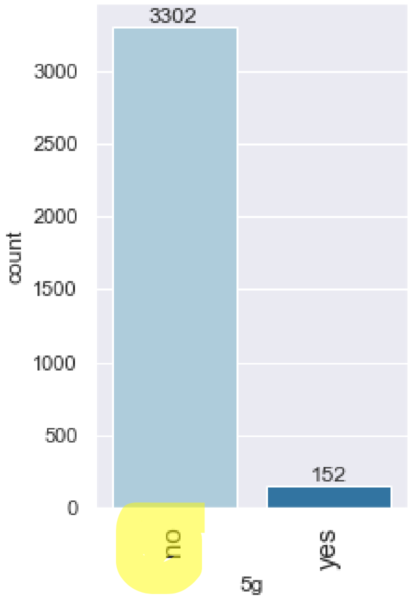
[Link to Appendix slide on data background check]()

Univariant Analysis

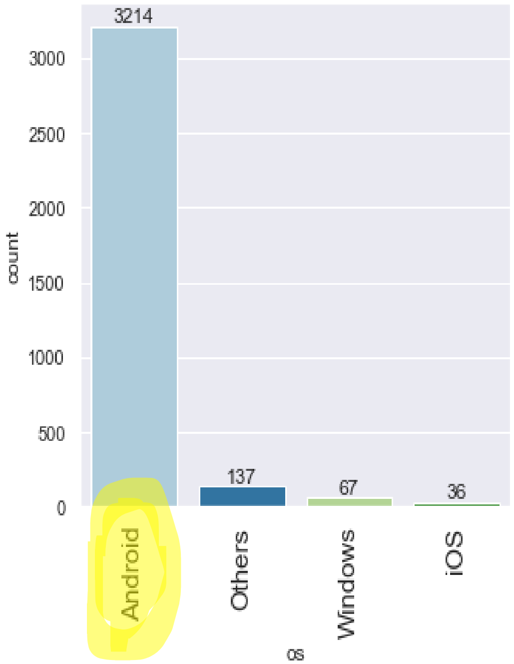
- Whether 4G is available or not



- Whether 5G is available or not



- OS on which the device runs

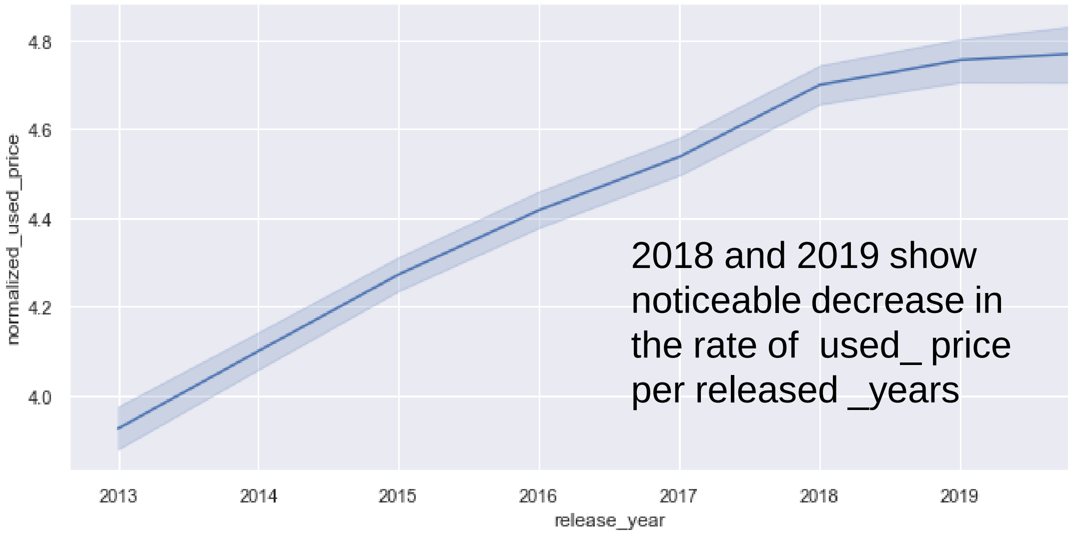


Bivariate Analysis

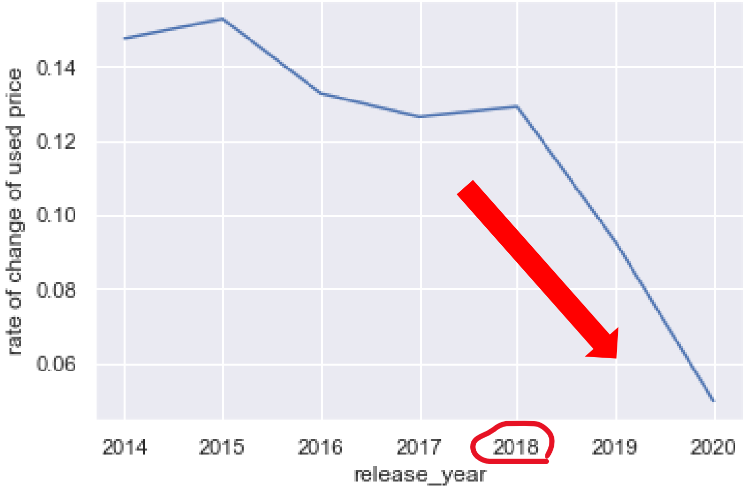
- Feature Correlation

Bivariant Analysis

- Variation normalized used price with the release year

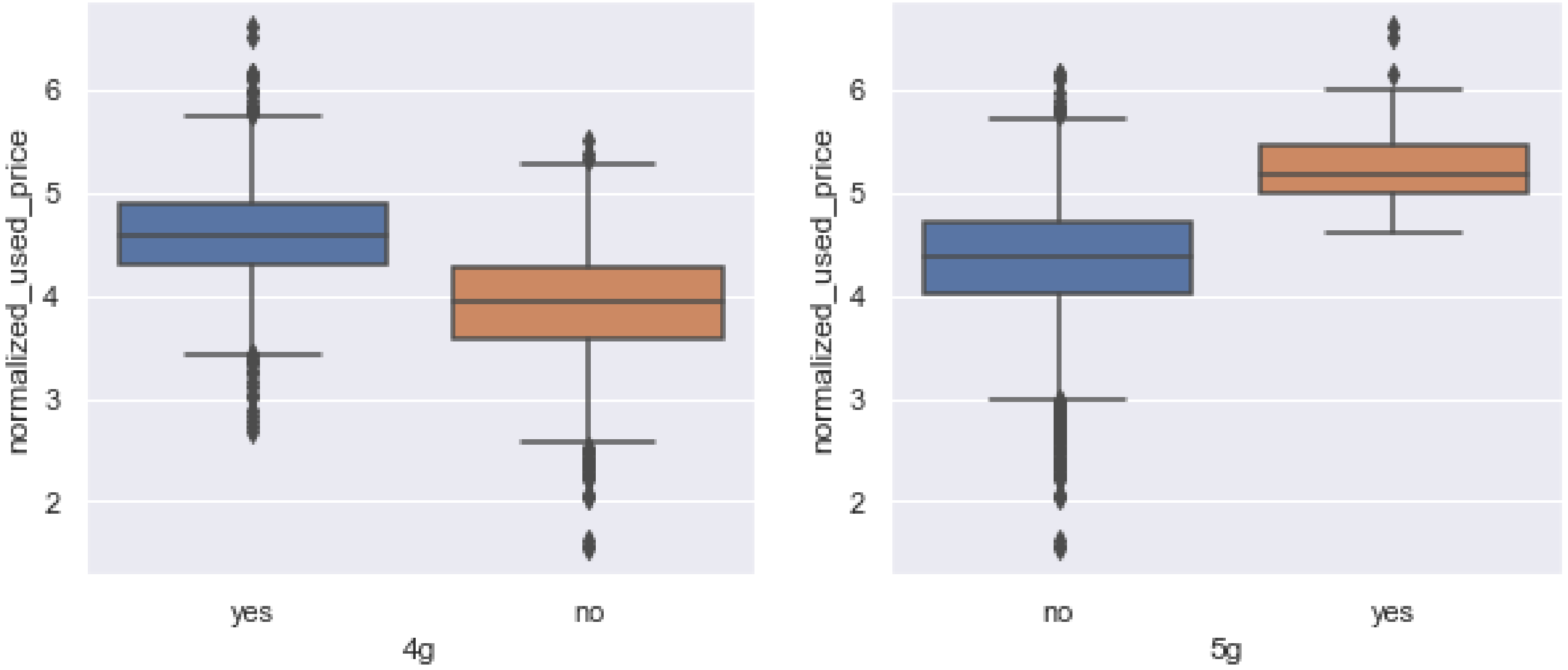


2018 and 2019 show noticeable decrease in the rate of used_ price per released _years



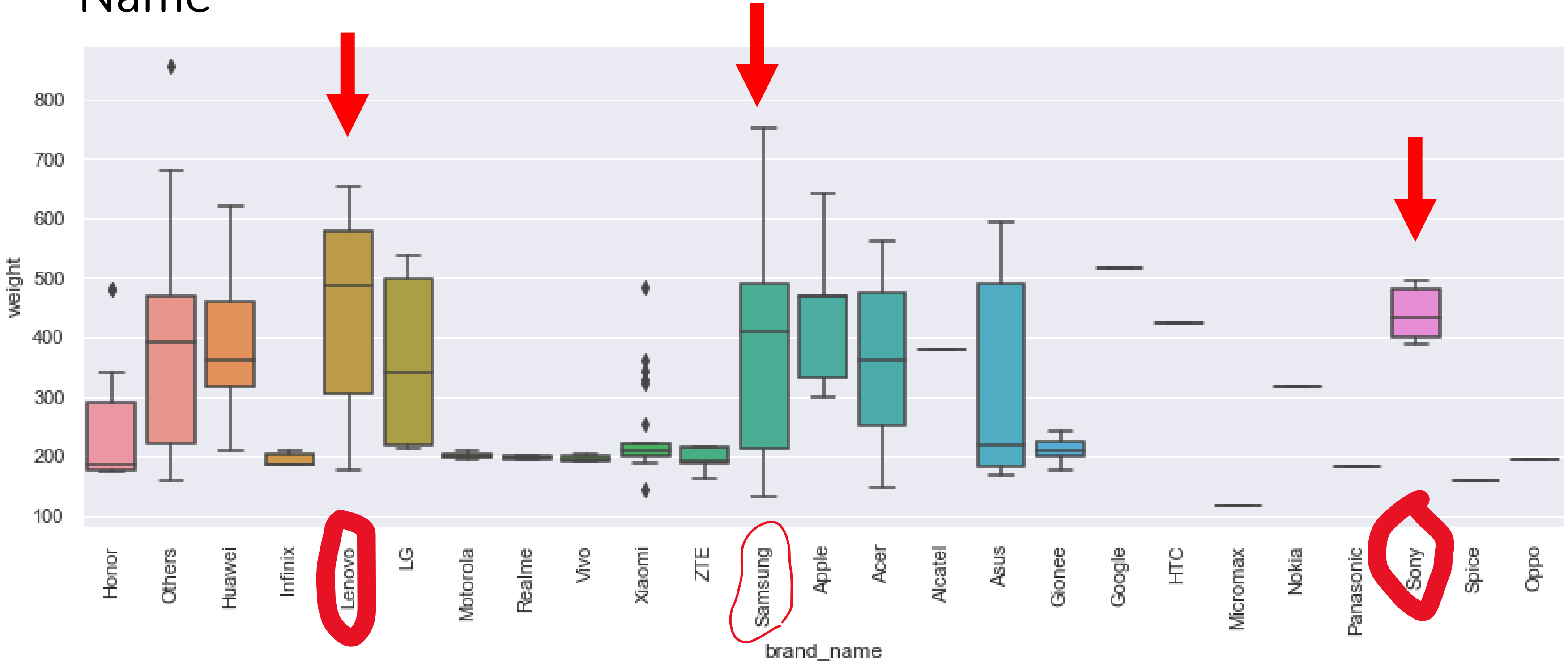
Bivariant Analysis

- 4|5 G versus normalized used price



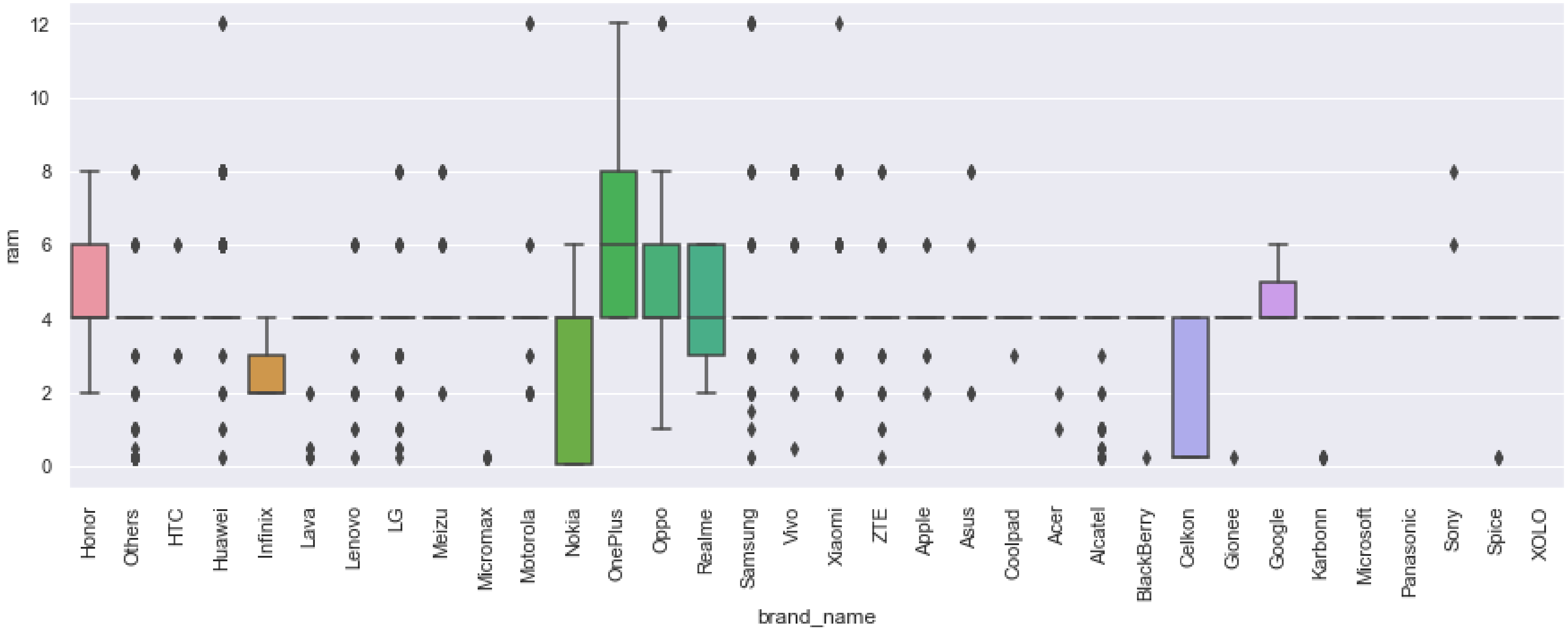
Bivariant Analysis

- For battery > 4500 How device weight varies with Brand Name



Bivariant Analysis

- Ram Versus Brand Name



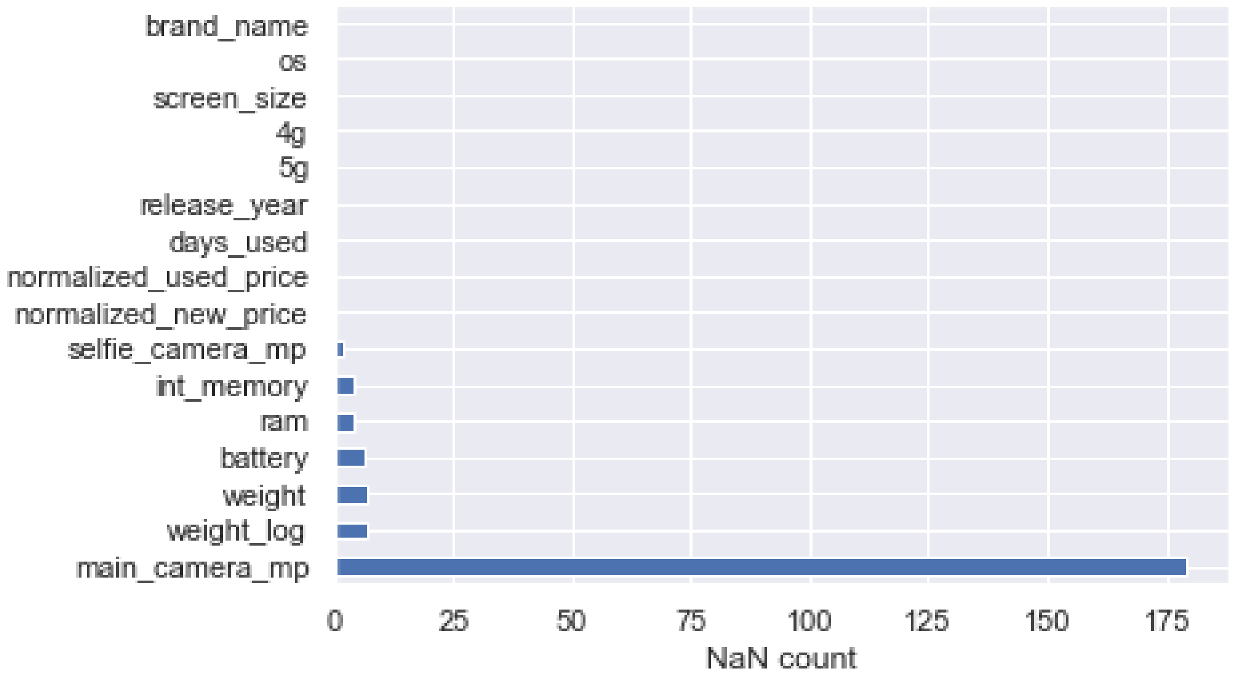
Data Preprocessing

- Duplicate value check
- Missing value treatment
- Outlier check (treatment if needed)
- Feature engineering
- Data preparation for modeling

Data Preprocessing

- **There is no duplicate value**

- **Missing value treatment**

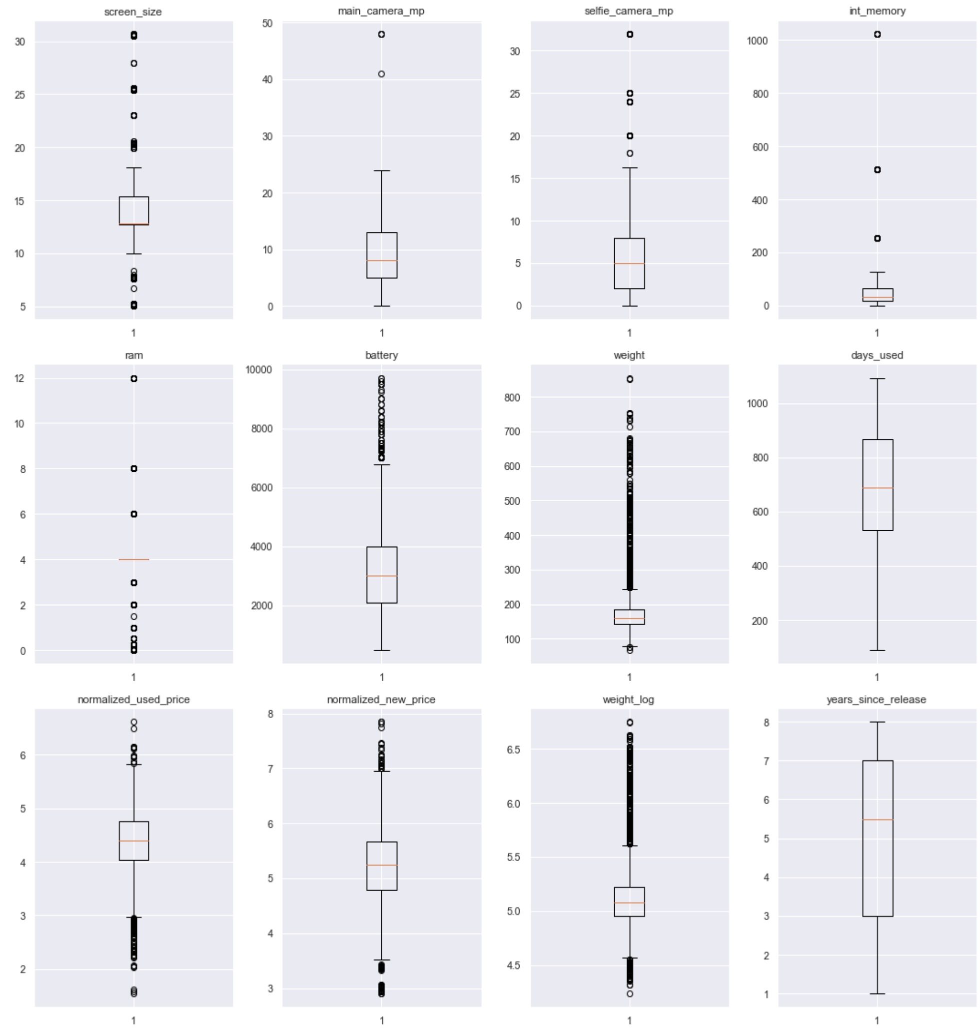


We imputed the missing values in the data by the column medians grouped by e.g."brand_name", and /or `release year` or one of them. We choose median to avoid the outlier effects

Data Preprocessing

- **Outlier check (treatment if needed)**

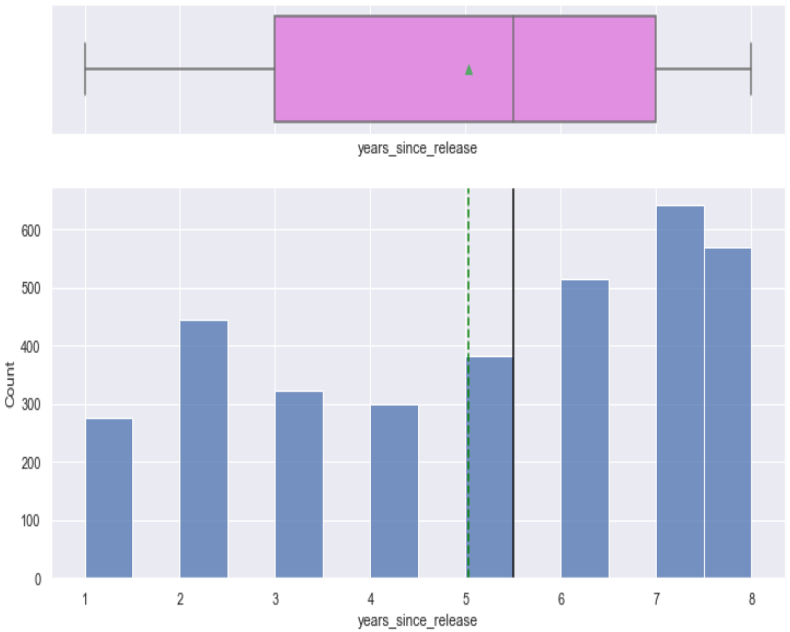
We used skew() to choose the feature with most dramatic outliers. The result shows the int_memory and weight_log are the highest outliers. However, I try both with outliers and without outliers and the result shows that outliers improved the accuracy.



Data Preprocessing

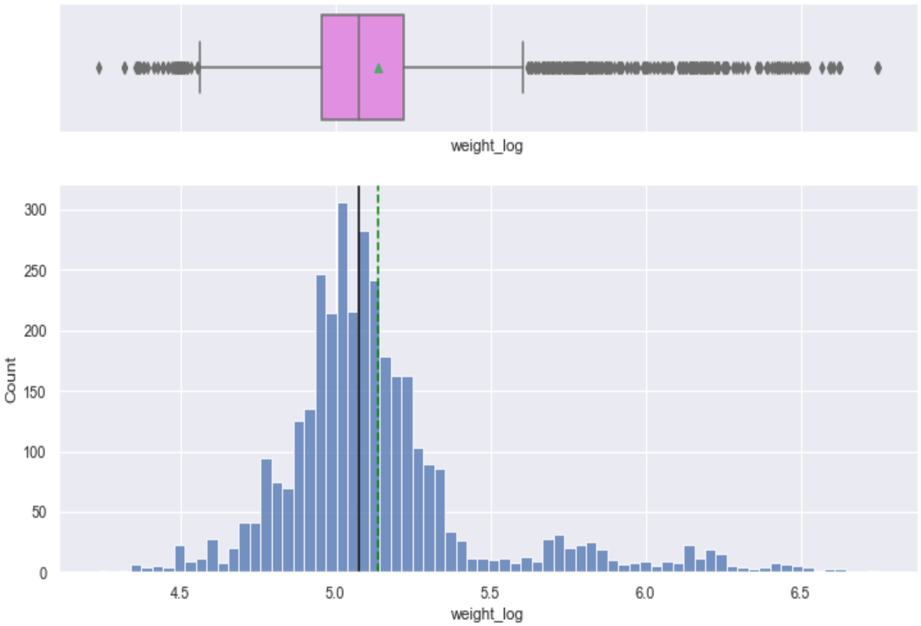
- **Feature engineering**

I create a new column `years_since_release` from the `release_year` column. Considering the year of data collection, 2021, as the baseline then droping the `release_year` column.



- **Feature engineering**

We transform the feature Weight to logarithmic Weight in order to reduce skewness.



Model Performance Summary

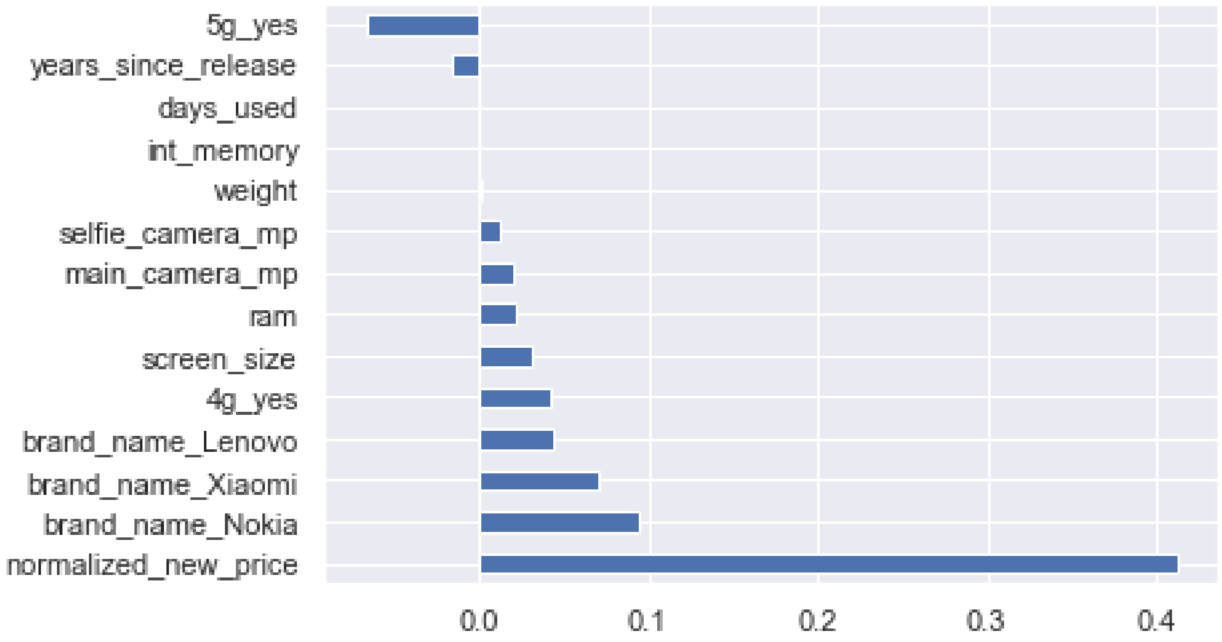
	coef	std err	t	P>|t|	[0.025	0.975]
const	1.137800	0.04700	24.428	0.000	1.046000	1.229
normalized_new_price	0.412500	0.01100	36.676	0.000	0.390000	0.435
brand_name_Nokia	0.093500	0.03000	3.128	0.002	0.035000	0.152
brand_name_Xiaomi	0.071300	0.02500	2.802	0.005	0.021000	0.121
brand_name_Lenovo	0.043100	0.02100	2.016	0.044	0.001000	0.085
4g_yes	0.042700	0.01500	2.832	0.005	0.013000	0.072
screen_size	0.030700	0.00200	13.420	0.000	0.026000	0.035
ram	0.021600	0.00500	4.434	0.000	0.012000	0.031
main_camera_mp	0.020500	0.00100	14.918	0.000	0.018000	0.023
selfie_camera_mp	0.012800	0.00100	11.441	0.000	0.011000	0.015
weight	0.001600	0.00000	7.468	0.000	0.001000	0.002
int_memory	0.000400	0.00000	2.180	0.029	0.000039	0.001
days_used	0.000064	0.00003	2.108	0.035	0.000004	0.000
years_since_release	-0.015800	0.00400	-3.601	0.000	-0.024000	-0.007
5g_yes	-0.067300	0.03100	-2.206	0.027	-0.127000	-0.007

- Model: OLS
- Method: Least Squares
- Dep. Variable: normalized_used_price
- R-squared: 0.843
- Adj. R-squared: 0.843
- F-statistic: 924.1

We test the multicollinearity by Variance inflation facto, VIF features >5 were dropped feature with high P values were dropped as well. P>0.05 shows that there is no relationship between the feature and the target.

Model Performance Summary

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



Model Performance Summary

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

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
Training	0.23097	0.179765	0.843416	0.842438	4.317809
Test	0.236781	0.183961	0.844557	0.842273	4.477063

- The train and test R^2 are 0.843 and 0.844, indicating that the model explains 84% and 84% of the total variation in the train and test sets respectively. Also, both scores are comparable.

- RMSE values on the train and test sets are also comparable.

- This shows that the model is not overfitting.

- MAE indicates that our current model is able to predict used price within a mean error of 0.18 on the test set.

- MAPE of 4.4 on the test data means that we are able to predict within 4% of used price.

Model Assumptions

- The linear model assumptions are valid, the residuals are normally distributed, independent, and homoscedastic.
- We conclude that model is linear with RMSE =0.23.
- We can predict [normalized_used_price +_ 0,23] using main features: normalized new price, brand name, and 4G,screen, ram, camera, 5G, and years since release .

Link to Appendix slide on data background check

APPENDIX

Data Background and Contents

Data Description

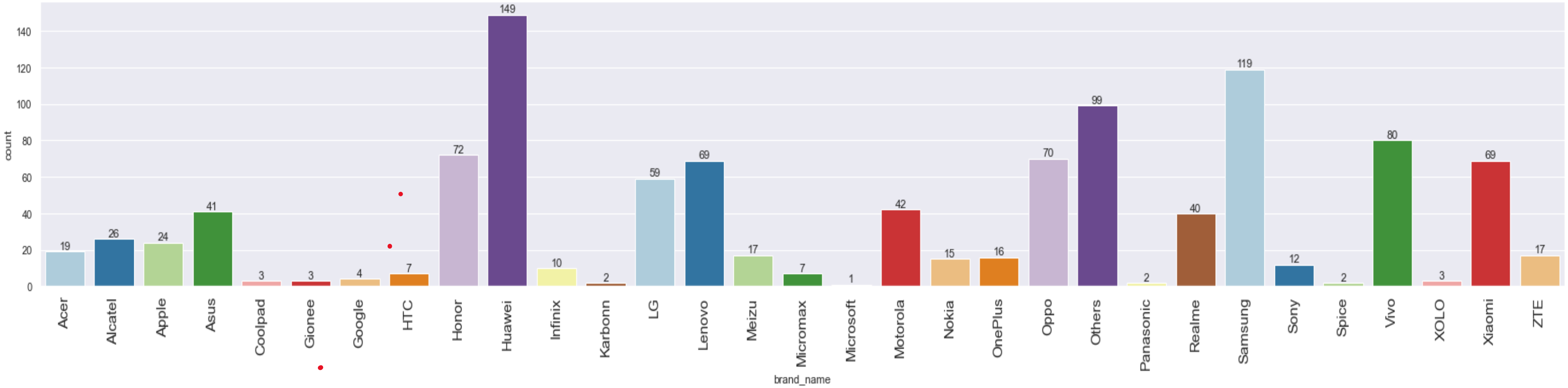
The data contains the different attributes of used/refurbished phones and tablets. The data was collected in the year 2021. The detailed data dictionary is given below.

Data

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

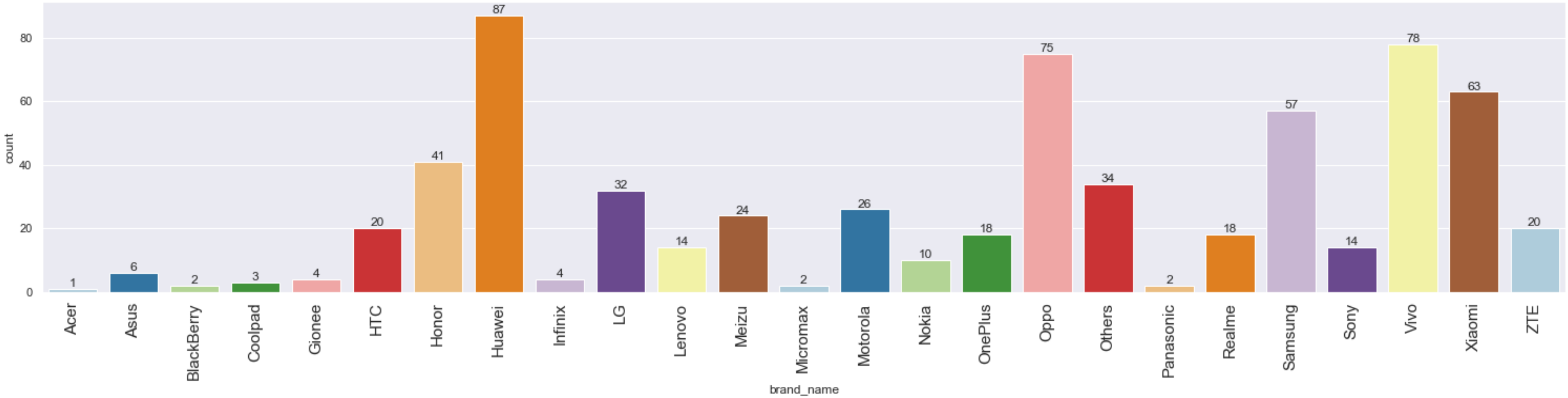
Data Background and Contents

- large_screen best brands

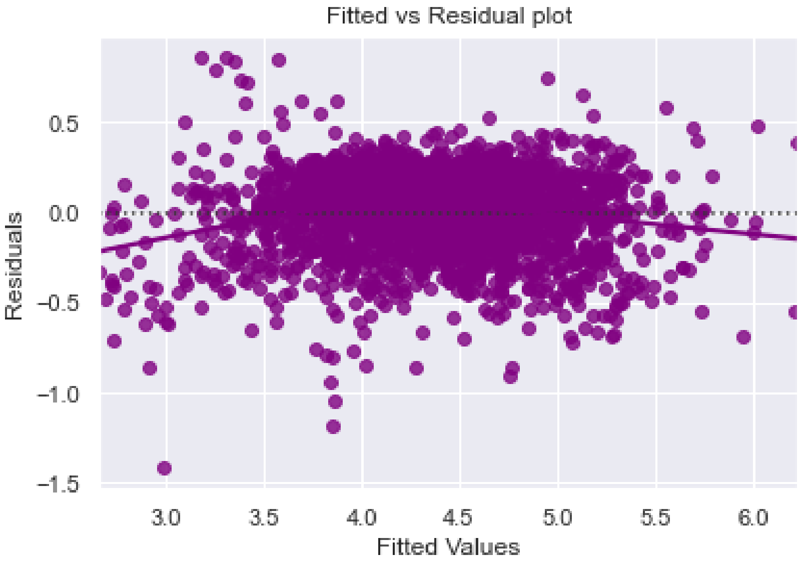


Data Background and Contents

- selfie_camera best brands



Model Assumptions

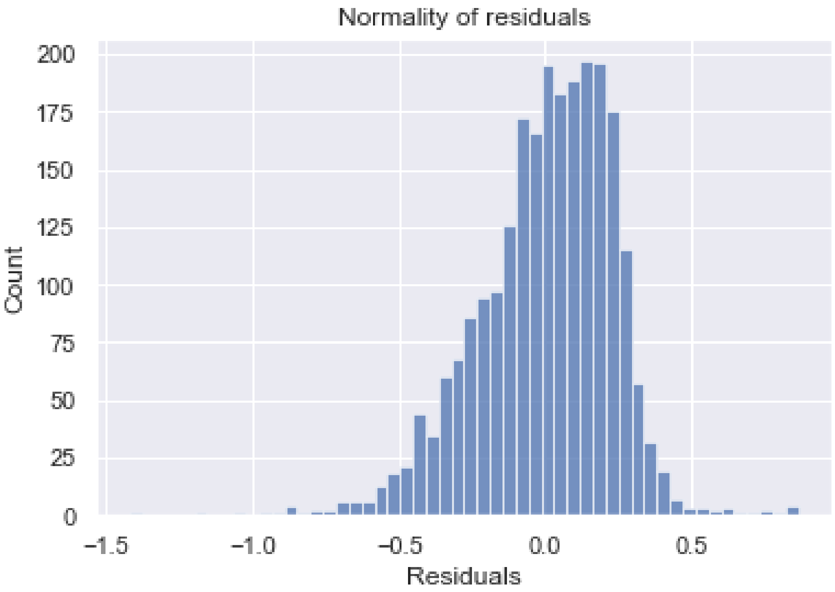


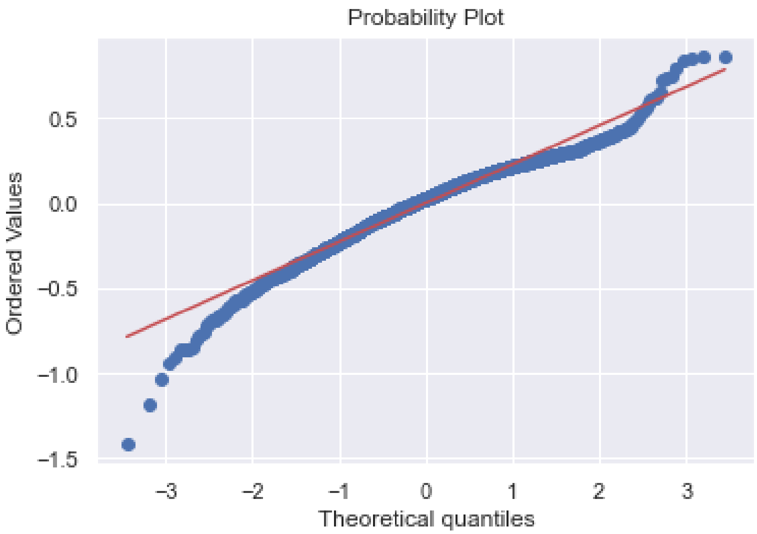
- **TEST FOR LINEARITY AND INDEPENDENCE**
- We will test for linearity and independence by making a plot of fitted values vs residuals and checking for patterns.
- If there is no pattern, then we say the model is linear and residuals are independent. There is no clear pattern

Model Assumptions

- **TEST FOR NORMALITY**

We will test for normality by checking the distribution of residuals, by checking the Q-Q plot of residuals, and by using the Shapiro-Wilk test.





ShapiroResult(statistic=0.9696815013885498, pvalue=3.4333233457361316e-22)

We assume the residuals are normally distributed although P<0.05 since the plot show decent shape.

Model Assumptions

- **TEST FOR HOMOSCEDASTICITY**
- We will test for homoscedasticity by using the goldfeldquandt test.
- If we get a p-value greater than 0.05, we can say that the residuals are homoscedastic. Otherwise, they are heteroscedastic.
- ('F statistic', 1.0705808661951246), ('p-value', 0.11945184546372178)
- 0.11>0.05 so the residuals are homoscedastic.