

# Linear Progression Model for Predicting Price of Used Devices

## ReCell: Introduction to Supervised Learning

ABIONA ADEKUNLE ARIYO

Date: 16/06/2022

# Contents / Agenda

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

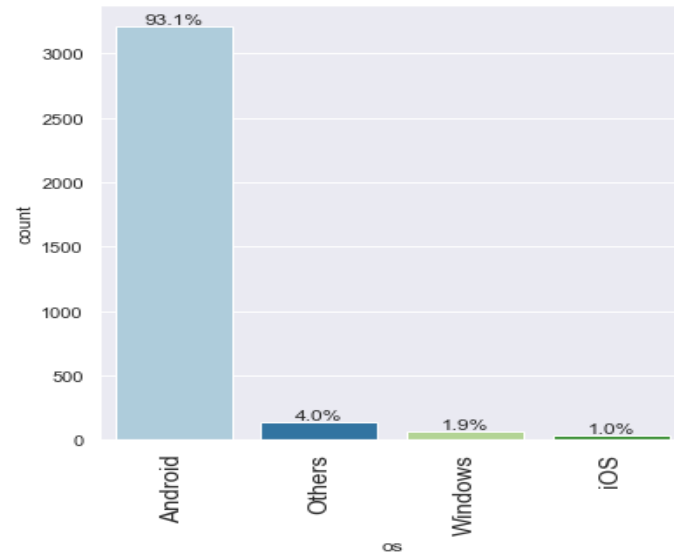
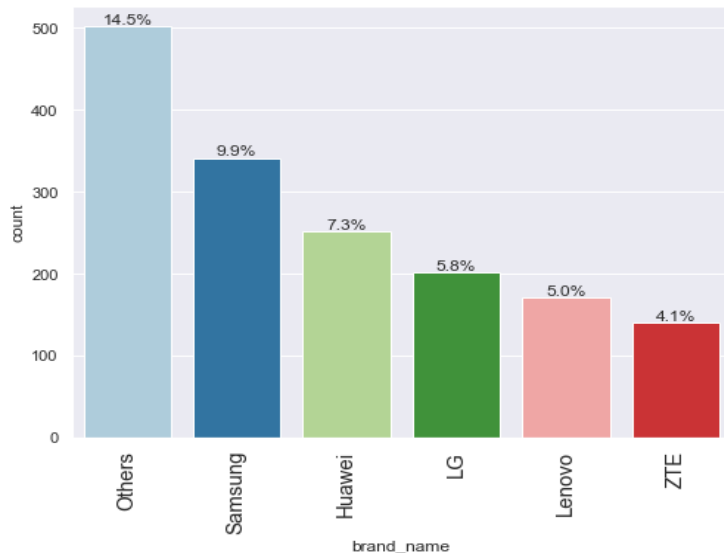
# Executive Summary

- Samsung, Huawei, LG, Lenovo and ZTE have a wider variety of devices and are the most popular in the used market.
- Features including main and selfie camera mp, ram, weight, normalized new price and year since release of a device are the best predictors of the normalized used price.
- Android operating system is very dominant in the used market ahead of iOS and others.
- Priority should be given to devices with main camera mp 13, 8 and 5 , selfie camera mp 5, 8 and 2, 4GB ram, weighing between 200g and 500g, having 4g/5g and with as little as possible the number of years since release to get the best normalized used price for the device at any given normalized new price.
- An Acer device with a similar specification will have used price advantage compared to Samsung and Sony brands, but will have a disadvantage compared to Karbonn and Xiaomi brands.

# Business Problem Overview and Solution Approach

- ReCell is a startup that is developing a dynamic pricing approach for used and refurbished devices in order to get into the rising market for used and refurbished devices.
- Over the last decade, the market for used and refurbished devices has grown significantly. The used phone industry is expected to be valued \$52.7 billion by 2023, with a compound annual growth rate (CAGR) of 13.6 percent between 2018 and 2023. The COVID 19 epidemic is predicted to drive up demand for used and refurbished devices even further.
- Used and refurbished devices are more cost-effective than new ones. They can be purchased with warranties and covered by evidence of purchase.
- The task is to analyze the data provided and build a linear regression model to predict the price of a used device and identify the factors that significantly influence it.

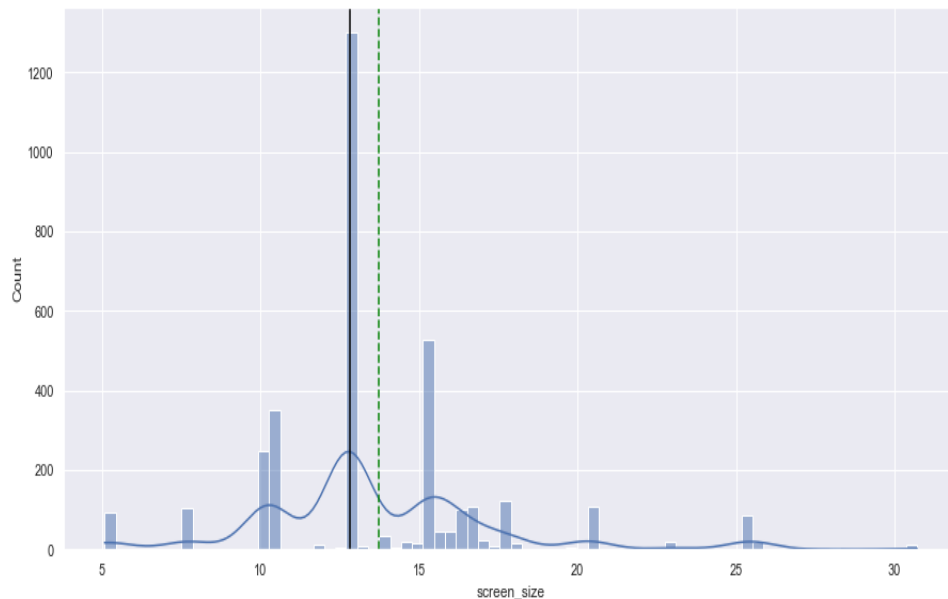
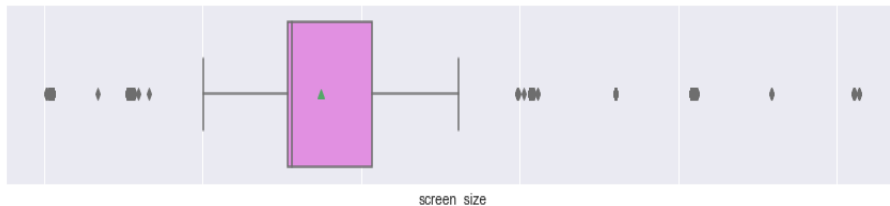
# EDA Results (Univariate Analysis) – Brand Name & OS



- Samsung is the most common brand among a total of 34. It is responsible for 9.9% all observations
- Completing the top five brands are Huawei, LG, Lenovo and ZTE
- The top five most popular brand names account for 32.1% of all the mobile devices
- Android operating systems is the most popular OS accounting for 93.1% of all devices
- Windows and iOS account for 1.9% and 1.0% respectively

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

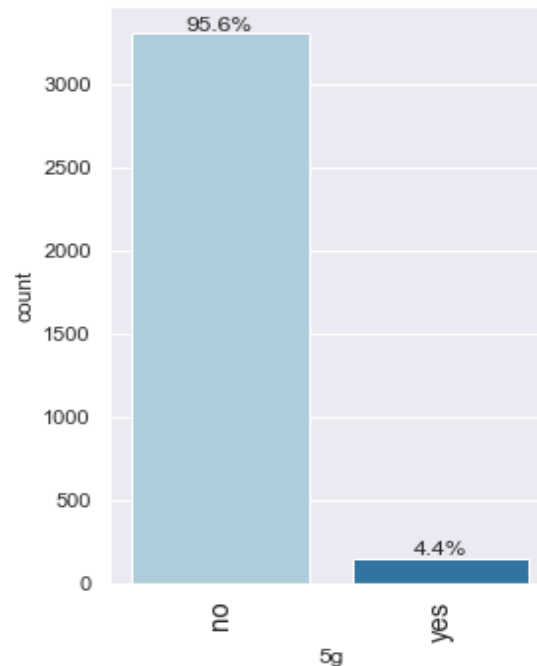
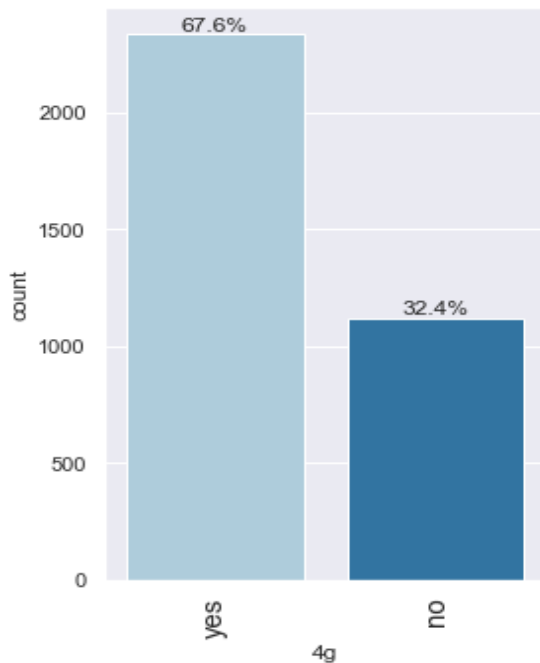
# EDA Results – Screen Size



- The screen size distributions is multimodal and slightly right-skewed
- The most common screen size on mobile devices is in the range of 12cm to 13cm
- The second most frequent screen size is in the range of 15cm to 16cm
- The third most frequent screen size is in the range of 10cm to 11cm
- There are outliers. However, they are not unusual as certain phones and devices have screen sizes represented here on the chart

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

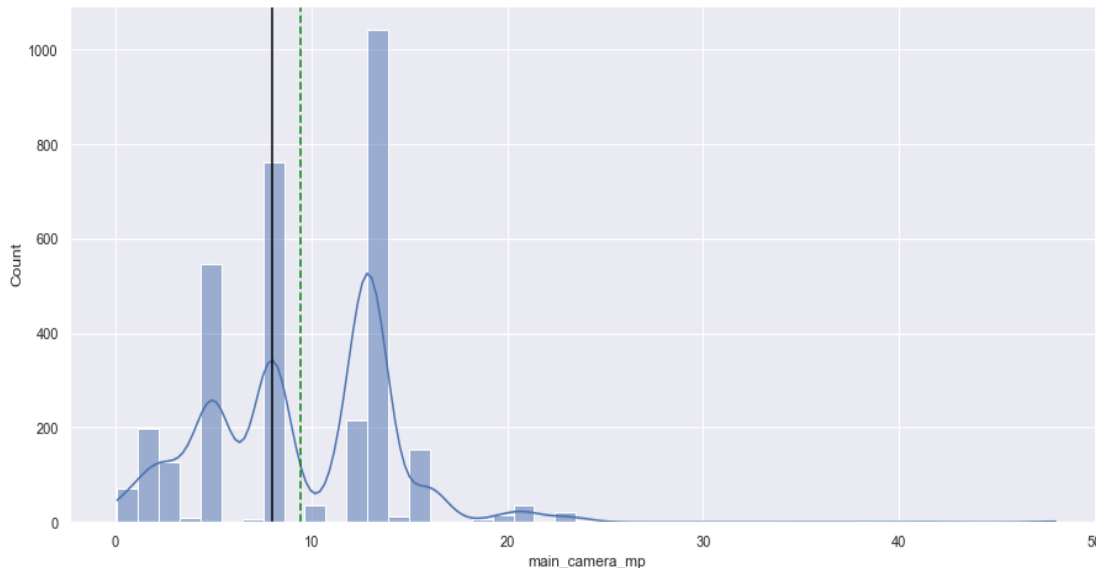
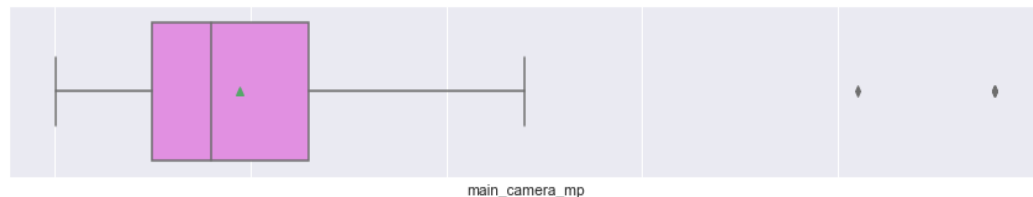
## EDA Results – 4g and 5g



- 67.6% of all the mobile devices have 4G capability
- 95.6% of all the mobile devices do not have 5G capability
- It appears that devices are more likely to have 4G capability than 5G

# EDA Results – Main Camera Megapixel

- The distribution is multimodal
- The most common main camera mps are 13, 8 and 5 megapixels respectively
- There are outliers with main camera mp of over 40. These are not improbable observations
- The median mp is 8 megapixel

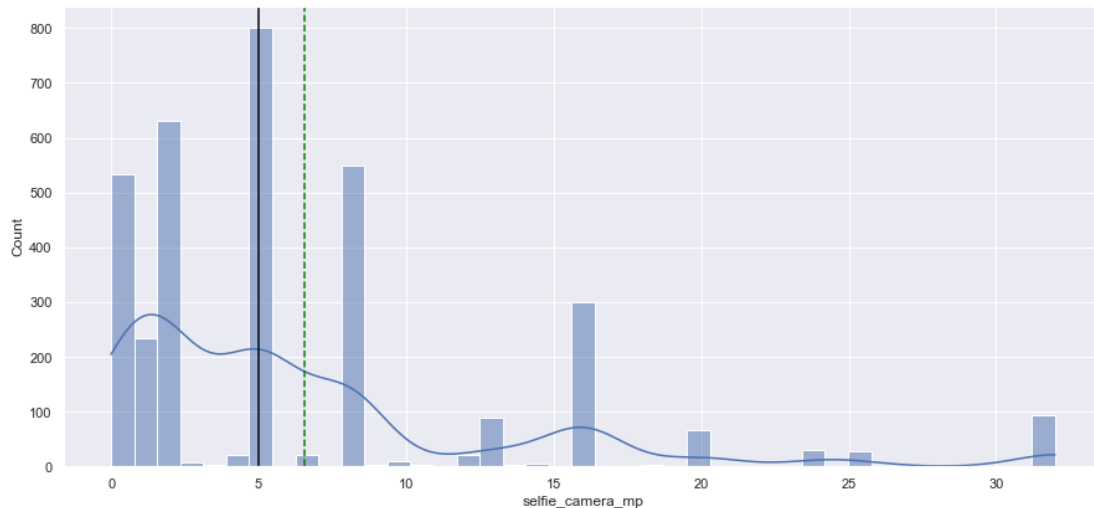
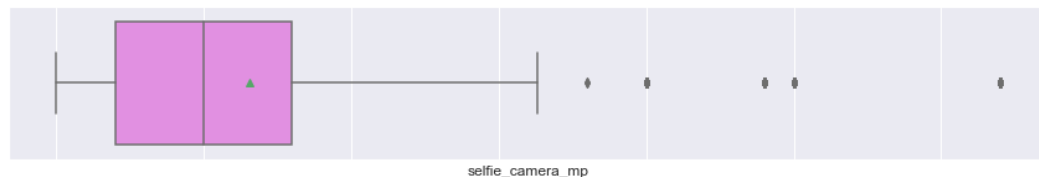


[Link to Appendix slide on data background check](#)



# EDA Results – Selfie Camera Megapixel

- The distribution is not normal. It is right skewed
- The most common selfie camera mps are 5, 8, 2 and 0.3 megapixels respectively
- There are outliers with selfie camera mp of over 16 megapixels. These are not improbable observations
- The median mp is 5 megapixel

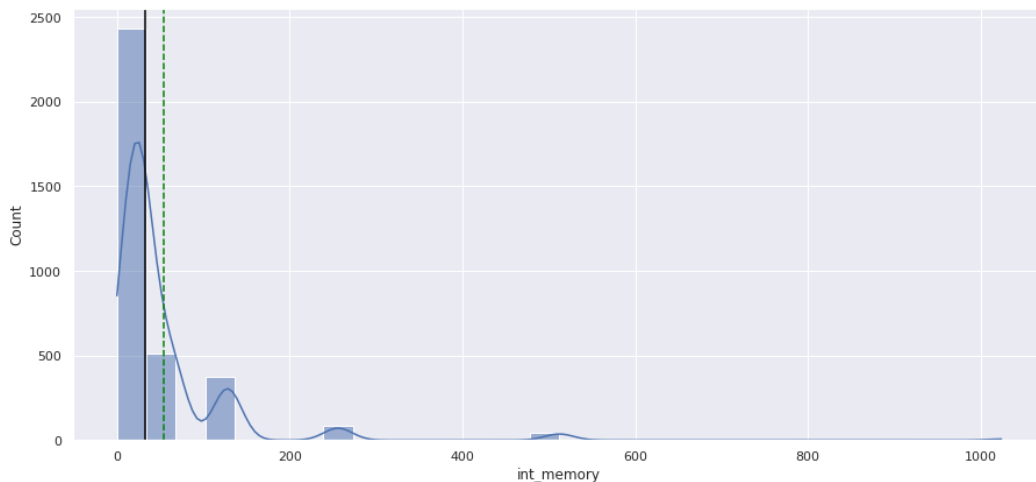


[Link to Appendix slide on data background check](#)

# EDA Results – Internal Memory

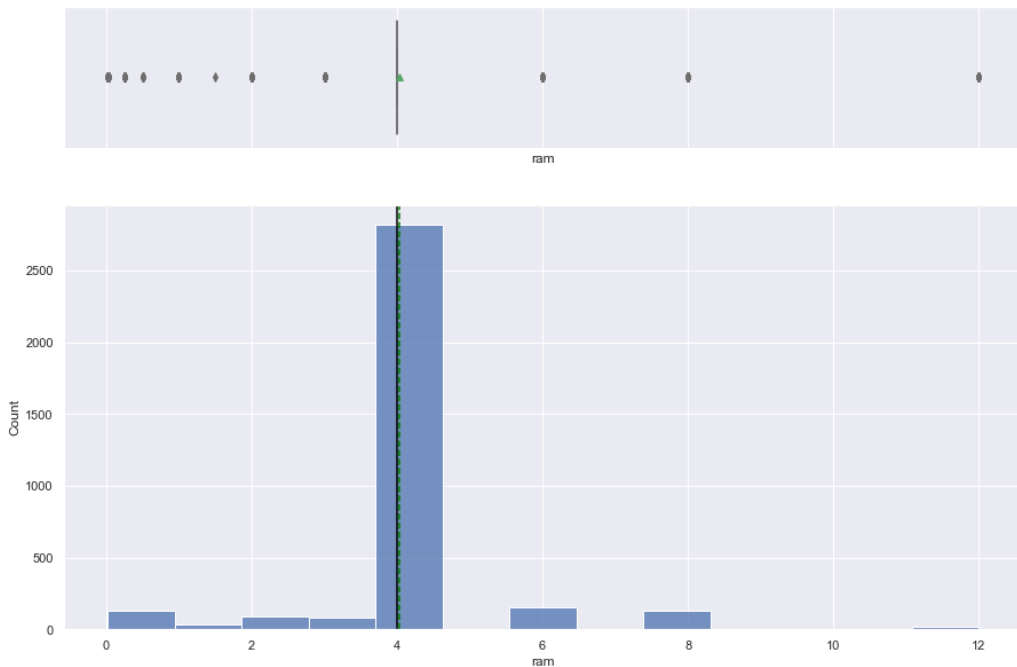


- The data distribution is not normal and right-skewed
- Most devices have internal memory between 0.01GB and 32GB
- There are outliers with approximately 250GB, 500 GB and 1TB. These are not unusual as certain mobile devices that have internal memory represented here on the chart



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

# EDA Results – Random Access Memory (ram)

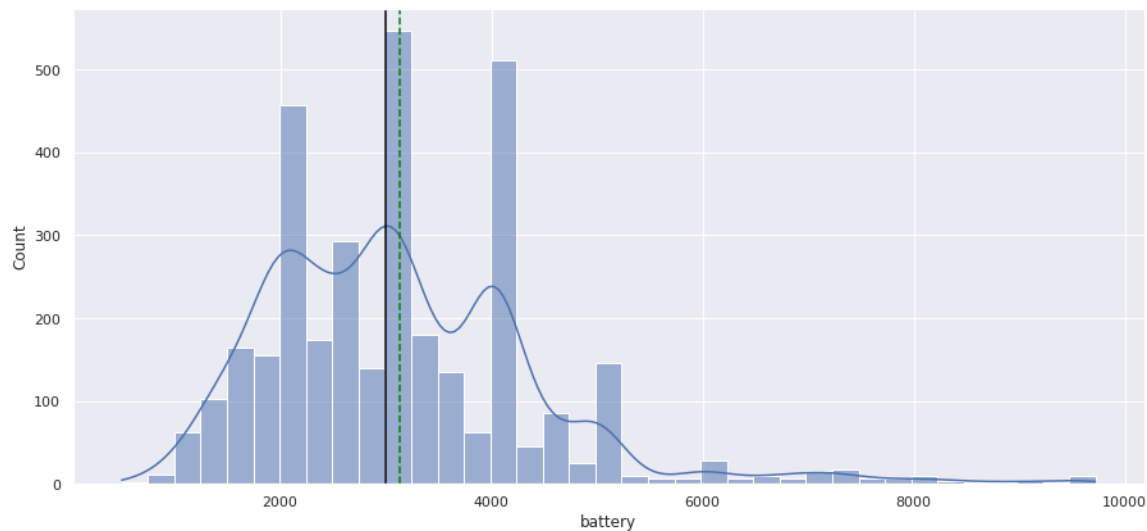
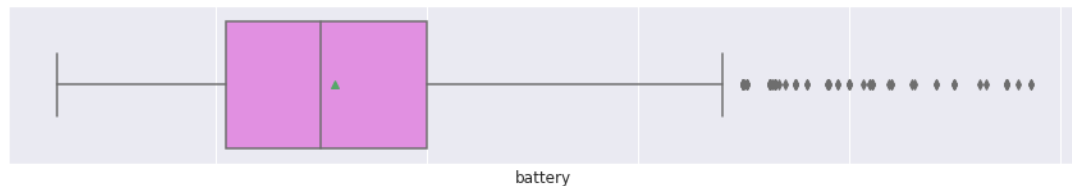


- The data distribution appear normal and not skewed
- The vast majority devices have random access memory around 4GB
- Most devices having a ram of 4GB has made other devices having different ram appear as outlier though they are known and not improbable

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

## EDA Results – Battery

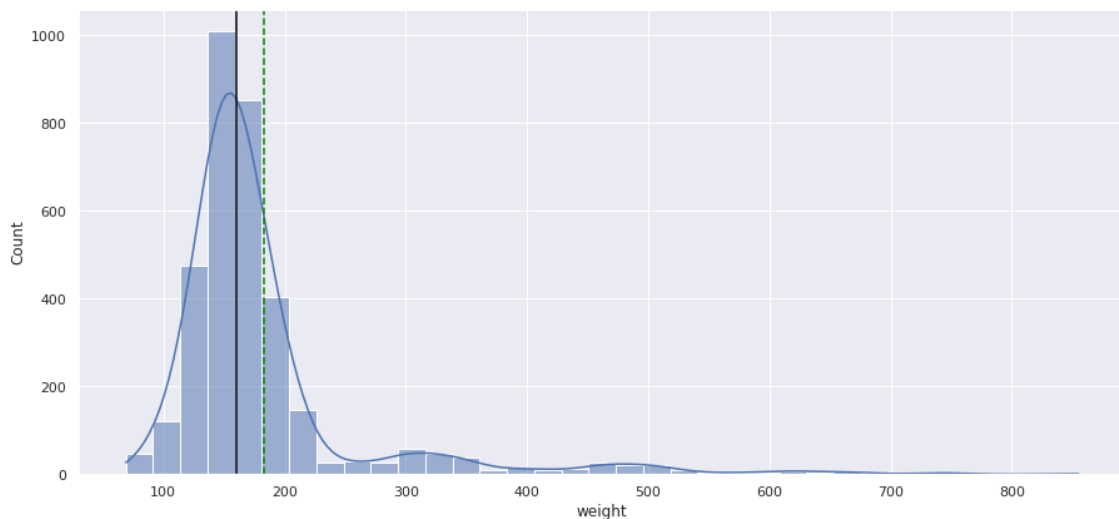
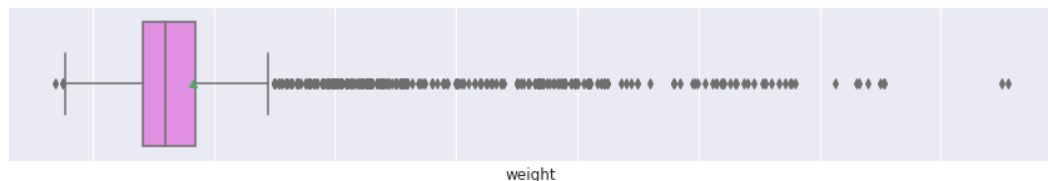
- The distribution is not normal. It is right skewed
- The distribution appears trimodal. More devices have battery capacities in the range 4000 – 4200mAh, 3000 – 3200mAh, and 2000 – 2200mAh
- The median battery capacity is 3000mAh
- The higher battery capacities that stand out as outliers likely reflect larger devices that need larger batteries



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

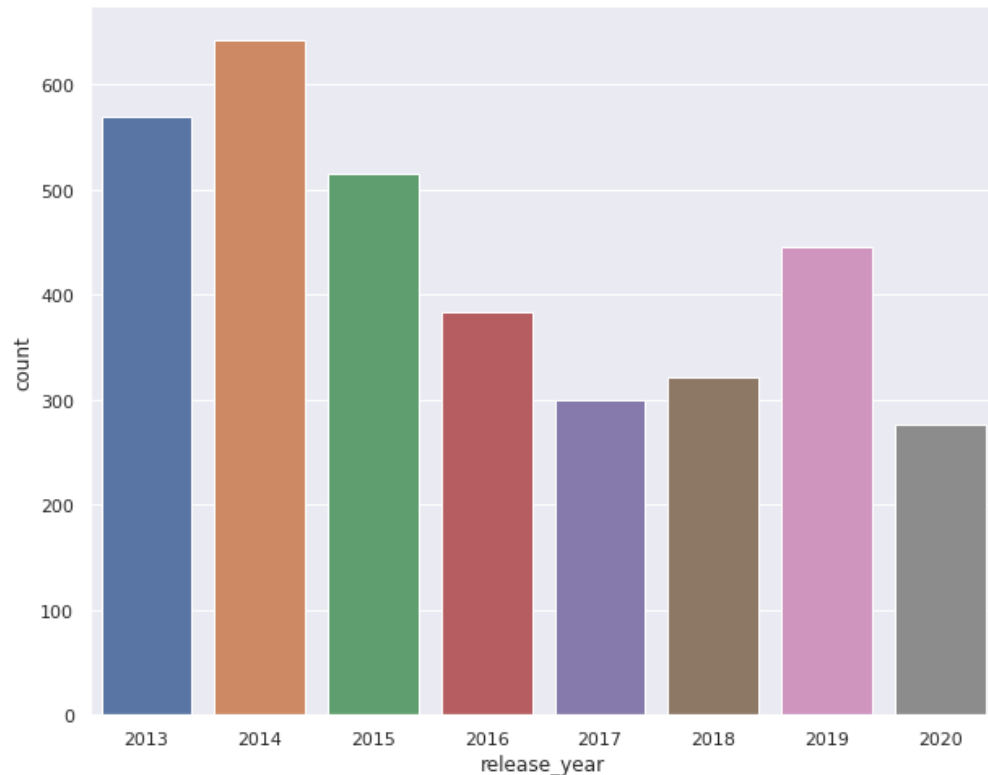
## EDA Results – Weight

- The distribution is normal and it is right skewed
- The median weight is 160g while the mean weight is 182.75g
- Most devices weight between 100g and 220g
- The higher device weights that stand out as outliers likely reflect larger devices that need larger batteries



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

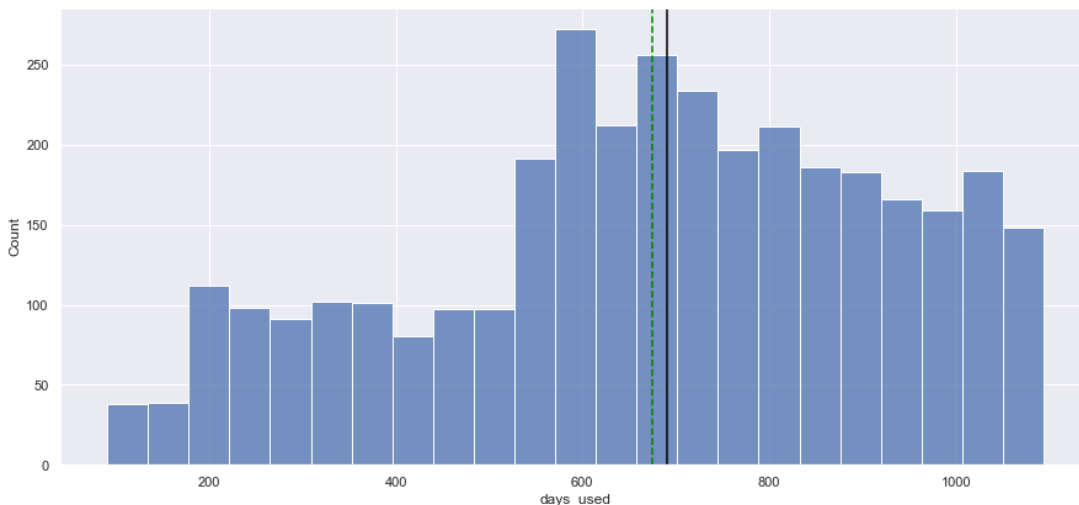
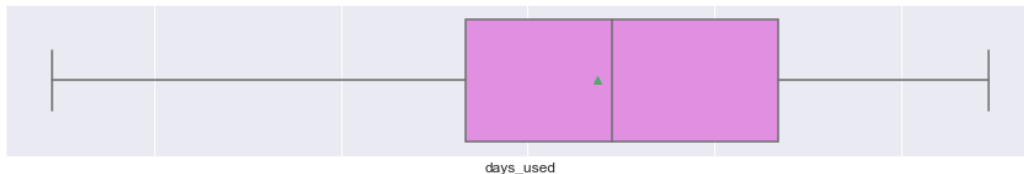
## EDA Results – Year of Release



- Generally, used devices have a higher probability of having early release year from the years under review
- There are more used devices released in 2014 than any other year reviewed
- 2020 produces the least amount of used devices from the years under review
- Used devices by year declined yearly between 2014 and 2017, after which it increased until 2020 which coincides with the beginning of Covid 19 and lockdowns

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

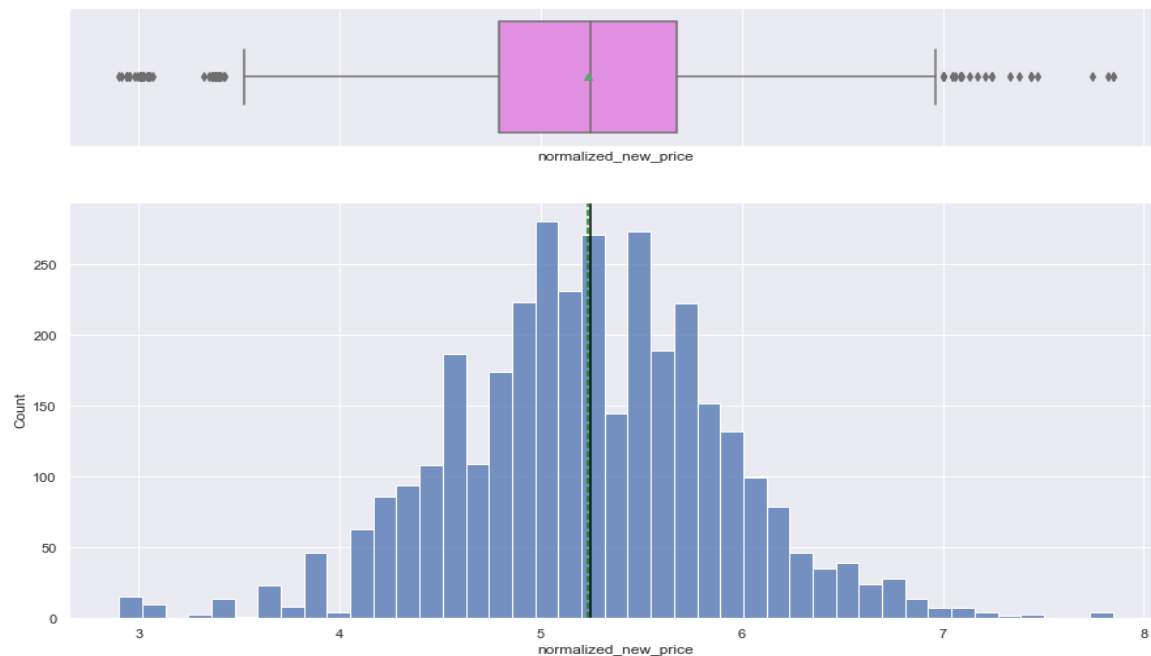
# EDA Results – Days Used



- The distribution is not normal and it is left skewed
- The median number of days is 690.5 while the mean number of days used is 674.9
- 75% of used devices are used for at least 533.5 days
- The longest a device was used is 1094 days while the least is 91 days

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

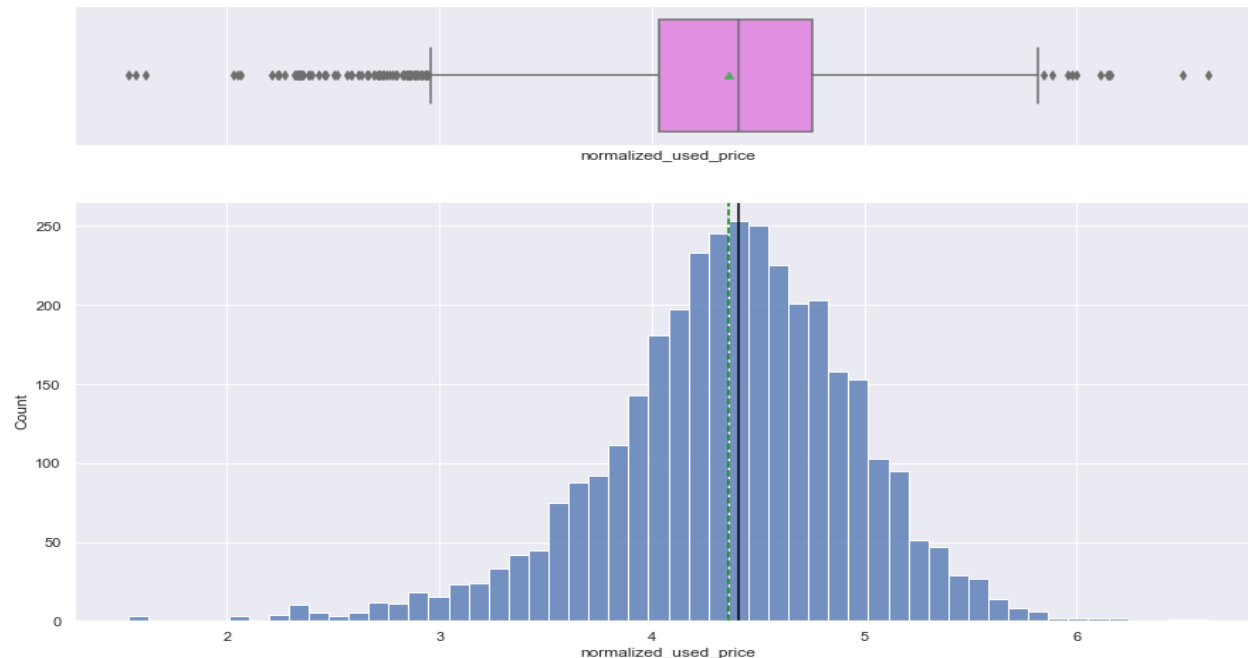
# EDA Results – Normalized New Price



- The distribution is normal
- The outliers represent the broad variety of devices available from feature phones to high-end phones and tablets
- The median normalized new price is 5.24
- The mean normalized new price is 5.23

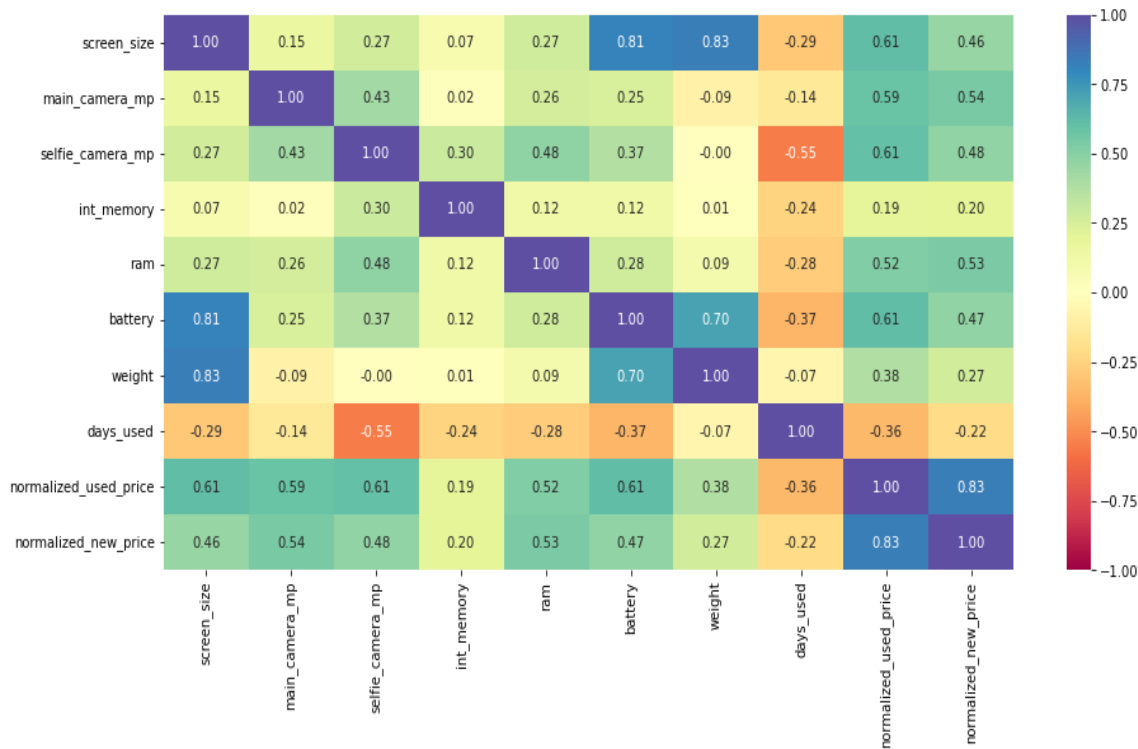


# EDA Results – Normalized Used Price



- The distribution is normal and slightly left skewed
- The median normalized new price is 4.41
- The mean normalized new price is 4.36
- The outliers represent the broad variety of devices available from feature phones to high-end phones and tablets

# EDA Results (Bivariate Analysis) – Heat Map



There is significant positive correlation between:

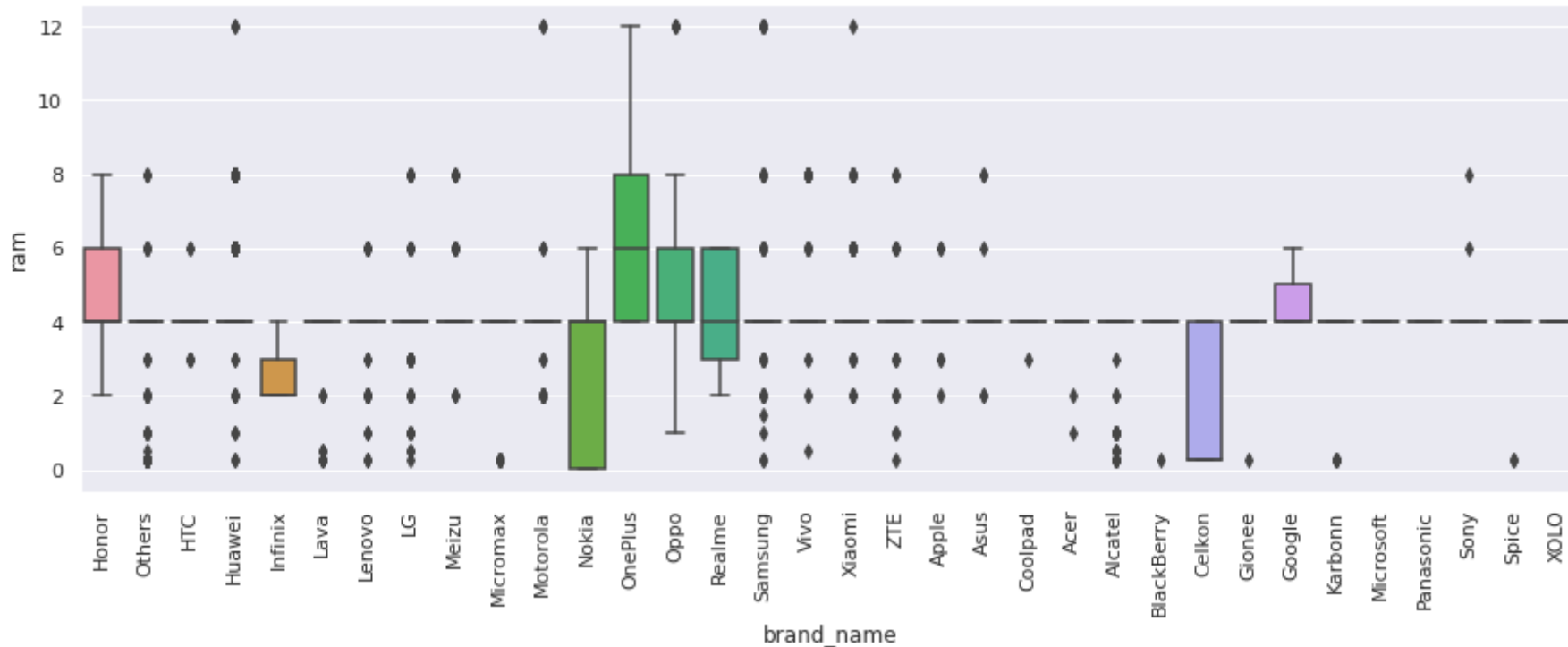
- screen size and battery
- screen size and weight
- screen size and normalized used price
- selfie camera mp and normalized used price
- normalized used price and battery
- normalized used price and normalized new price
- battery and weight

There is a slight negative relationship between:

- Days used and selfie camera mp

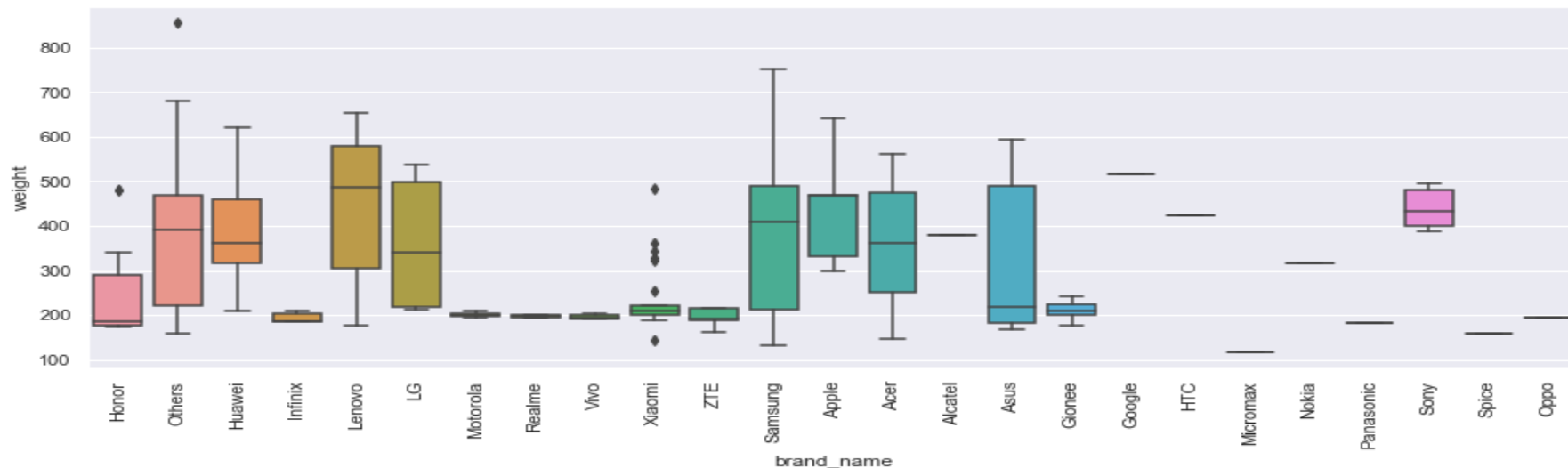
[Link to Appendix slide on data background check](#)

# Devices random access memory (ram) by brand name



- Most mobile devices have ram between the range of 0.02GB and 8GB
- The highest ram observed among the mobile devices is 12G

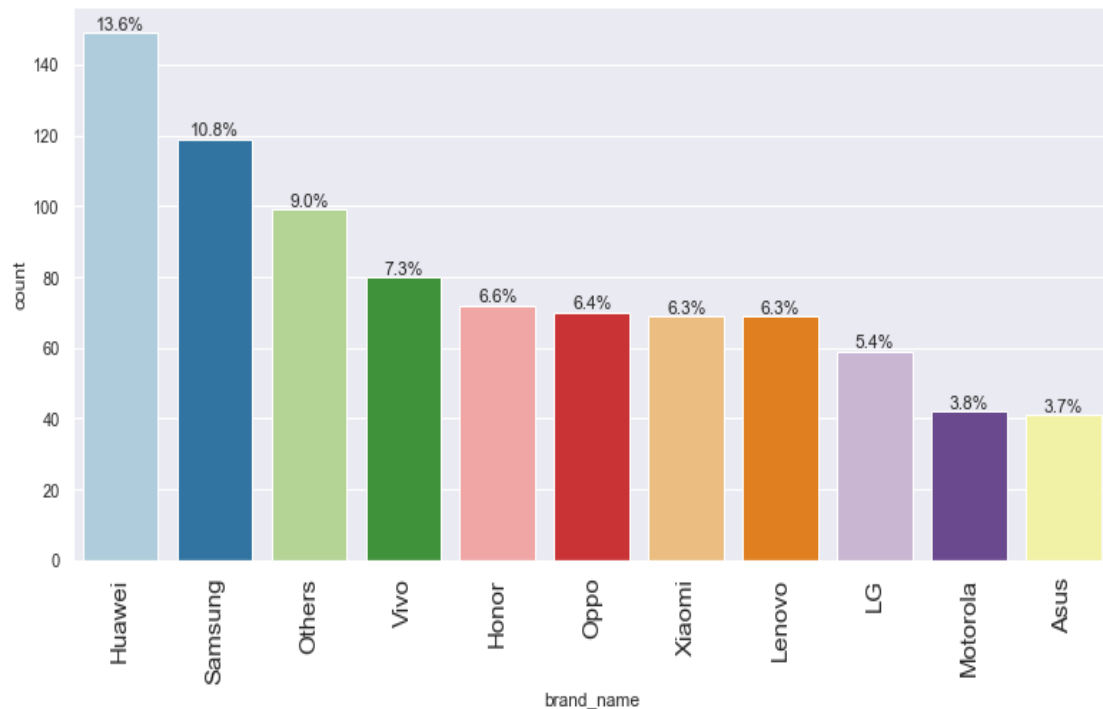
# Devices with large battery– Weight vs Brand Name



- Used devices with above 4500mAh battery capacity weight between 200 grams and 500 grams
- All Infinix, Motorola, Realme, Vivo, ZTE, Gionee Micromax, Panasonic, Spice and Oppo devices in this category weight below 250 grams
- Lenovo has the most devices weighting over 500 grams
- The highest brand in this category is Micromax
- The devices with the highest weight is among brands grouped as others then Samsung
- The median weight in this category is 300 grams while the mean weight is 332.27 grams

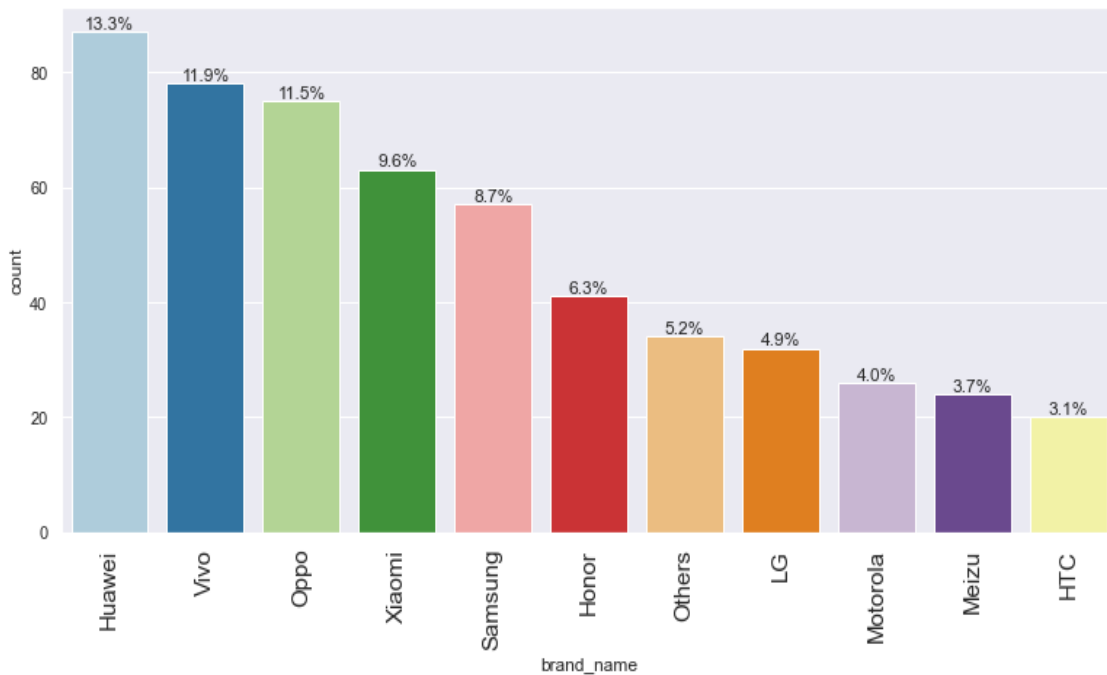
# Devices with large screen by Brand Name

- The top 10 brands of mobile devices with screen sizes more than 15.24cm are Huawei, Samsung, Vivo, Honor, Oppo, Xiaomi, Lenovo, LG, Motorola and Asus respectively.
- Huawei account for 13.6% of the large screen mobile devices
- The top 10 brands account for 72.9 of all mobile devices with screen sizes more than 15.24cm



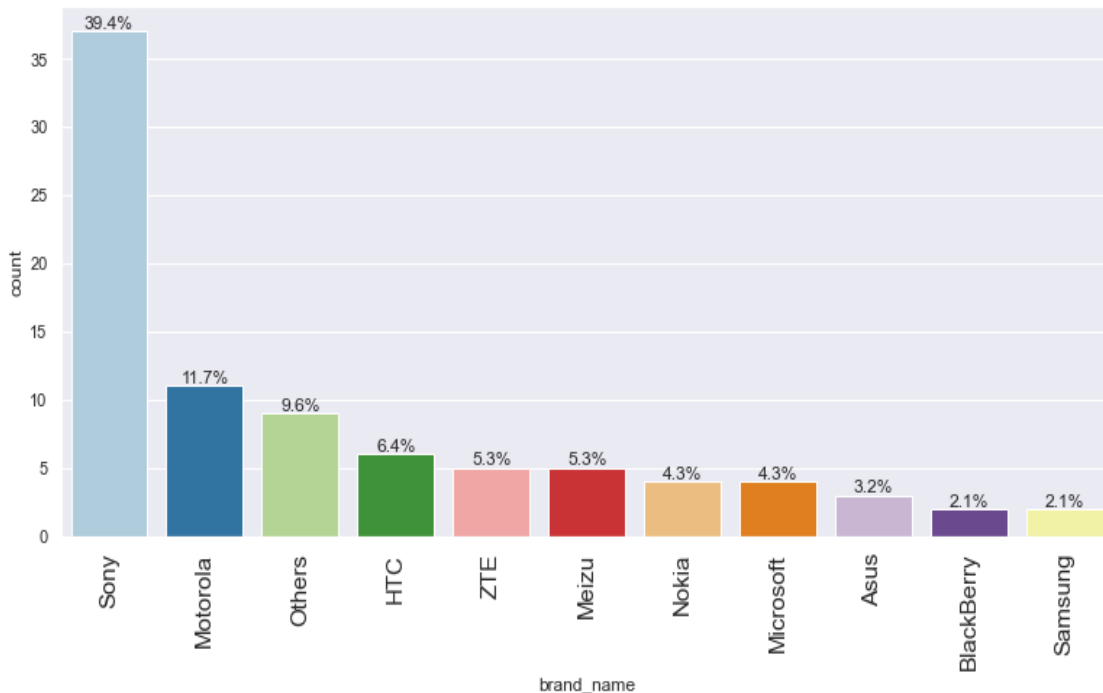
# Devices with high megapixel Selfie Camera by Brand Name

- The top 10 brands of mobile devices that have selfie camera mp more than 8 are Huawei, Vivo, Oppo, Xiaomi, Samsung, Honor, LG, Motorola, Meizu and HTC respectively.
- Huawei account for 13.3% of mobile devices that have selfie camera mp more than 8
- The top 10 brands account for 77% of all mobile devices that have selfie camera mp more than 8

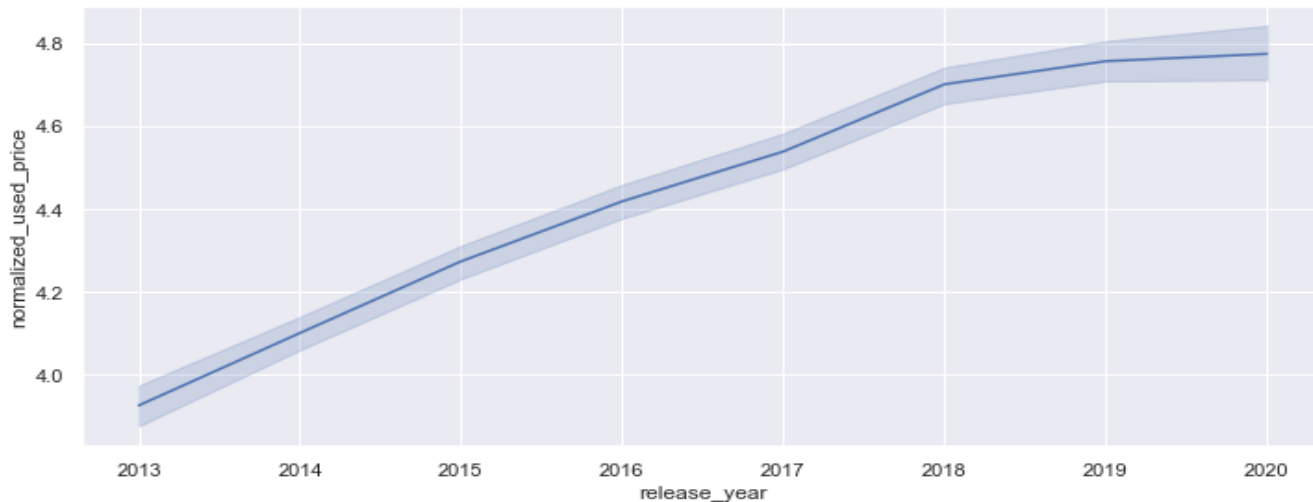


# Devices with high megapixel Main Camera by Brand Name

- The top 10 brands of mobile devices that have main camera mp more than 16 are Sony, Motorola, HTC, ZTE, Meizu, Nokia, Microsoft, Asus, Blackberry and Samsung respectively.
- Sony account for 39.4% of mobile devices that have main camera mp more than 16
- The top 10 brands account for 84.1% of all mobile devices that have main camera mp more than 16



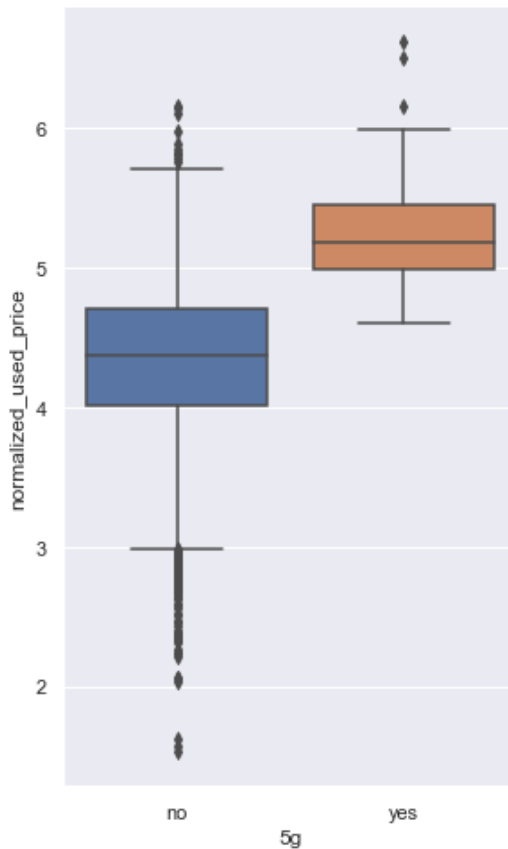
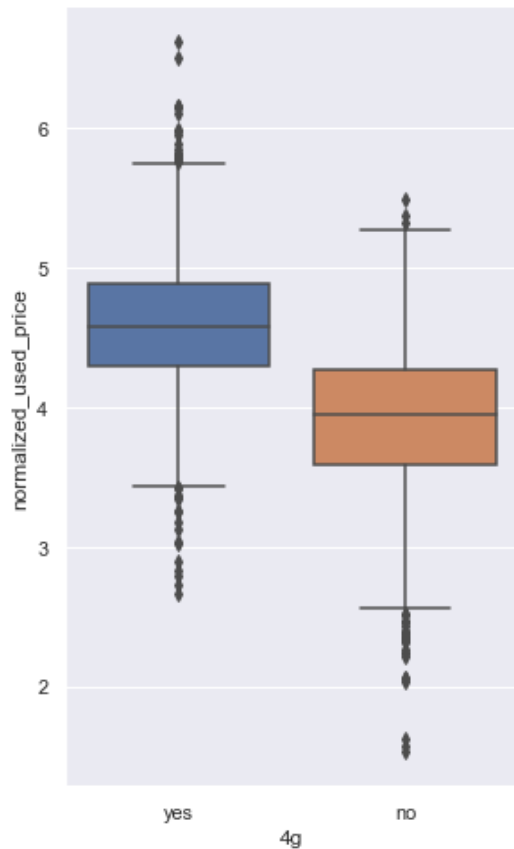
# Release Year vs Normalized Used Price



- The normalized used price has a positive relationship with the release year
- The relationship is generally linear but plateaus between 2019 and 2020
- The variation of normalized used price from year to year is fairly constant except between 2018 to 2022 where the variation fans out



# Normalized Prices vs 4g/5g



- Mobile devices having 4g are more likely to have a higher normalized used price than devices that do not have
- Mobile devices having 5g are more likely to have a higher normalized used price than devices that do not have
- Mobile devices having 5g command higher normalized used prices than mobile devices having 4g
- The presence of outliers may be due to brand and age of mobile device but not unusual

# Data Preprocessing

- Duplicate value check
  - There were no duplicate observations (checked with `data.duplicated().sum()`)
- Missing value treatment: Missing value were checked with `df1.isnull().sum()`. The results are

```
main_camera_mp    179
selfie_camera_mp    2
int_memory         4
ram                4
battery            6
weight             7
```

The missing values were treated with code `df1[col] = df1[col].fillna(value=df1.groupby(['release_year', 'brand_name'])[col].transform("median"))` on all the columns with missing values. The code imputes missing values in each column with the column median by grouping the data on release year and brand name. The result of the code is

```
main_camera_mp    179
selfie_camera_mp    2
int_memory         0
ram                0
battery            6
weight             7
```

## Data Preprocessing - Missing value treatment cont'd

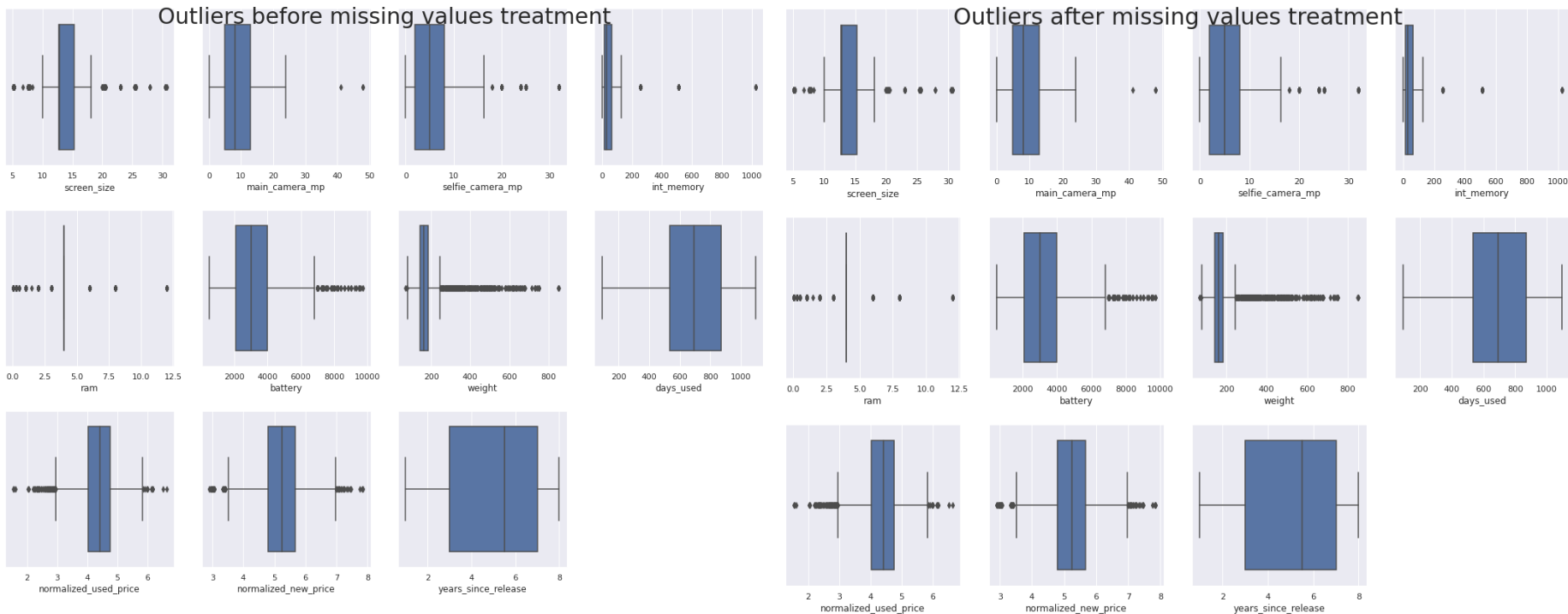
- Next, missing values in each column was treated with the column median by grouping the data on brand name only using the code

`df1[col].fillna(value=df1.groupby(['brand_name'])[col].transform("median"))`. The result is

```
main_camera_mp      10
selfie_camera_mp     0
int_memory           0
ram                  0
battery              0
weight               0
```

- The remaining missing values in the `main_camera_mp` column treated by imputing the column median using the code `df1["main_camera_mp"].fillna(df1["main_camera_mp"].median())`. All the missing values were thus treated.

# Data Preprocessing - Outlier check



- The outliers profile have not changed especially for columns that had missing values
- The distribution of observations for each column have been generally preserved

# Data Preprocessing - Feature engineering

- Feature engineering

A new column, *years\_since\_release*, was created to replace *release\_year* to allow proper interpretation of the age of the mobile device. 2021 was used as the baseline year. *release\_year* column was dropped from the data and the statistical summary of the new column, *years\_since\_release*, was imputed.

```
df1["years_since_release"] = 2021 - df1["release_year"]
df1.drop("release_year", axis=1, inplace=True)
df1["years_since_release"].describe()
```

```
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
```

- The average number of years since release is 5.0 years
- The median number of years is 5.5 years
- The minimum number of years is 1 year while the maximum is 8 years
- 75% of the devices have at least 3 years since year of release

# Data Preprocessing - Data preparation for modelling

- To predict the normalized price of used devices the column *normalized\_used\_price* was split from the data, and a constant was added

```
X = df1.drop(["normalized_used_price"], axis=1)
```

```
y = df1[["normalized_used_price"]]
```

```
X = sm.add_constant(X)
```

- Before the model was built categorical features were encoded by adding dummy variables to the data

```
X = pd.get_dummies(  
    X,  
    columns=X.select_dtypes(include=["object", "category"]).columns.tolist(),  
    drop_first=True,  
)
```

# Data Preprocessing - Data preparation for modelling

- The data was into train and test in 70:30 ratio to be able to evaluate the model that was build on the train data

```
x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

```
print("Number of rows in train data =", x_train.shape[0])  
print("Number of rows in test data =", x_test.shape[0])
```

```
Number of rows in train data = 2417  
Number of rows in test data = 1037
```

- The number of observations in the train data is 2,417 while the number of observations in the test data is 1,037

# Model Performance Summary

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:	Sun, 19 Jun 2022	Prob (F-statistic):	0.00			
Time:	23:25:47	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			

## Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 2.39e+03. This might indicate that there are strong multicollinearity or other numerical problems.

- R-squared of the model is 0.839 and adjusted R-squared is 0.838, which shows that the model is able to explain ~84% variance in the data.
- A unit increase in the main camera mp will result in a 0.0210 unit increase in the normalized used price, all other variables remaining constant.
- A unit increase in the selfie camera mp will result in a 0.0138 unit increase in the normalized used price, all other variables remaining constant.

[Link to Appendix slide on model assumptions](#)



# Model Performance Summary

- A unit increase in the ram will result in a 0.0207 unit increase in the normalized used price, all other variables remaining constant.
- A unit increase in the weight will result in a 0.0017 unit increase in the normalized used price, all other variables remaining constant.
- A unit increase in the normalized new price will result in a 0.4415 unit increase in the normalized used price, all other variables remaining constant.
- A unit increase in the year since release will result in a 0.0292 unit decrease in the normalized used price, all other variables remaining constant.
- The normalized used price of a Karbonn device will be 0.1156 units higher than the normalized used price of a Acer device, all other variables remaining constant.
- The normalized used price of a Samsung device will be 0.0374 units lower than the normalized used price of a Acer device, all other variables remaining constant.

[Link to Appendix slide on model assumptions](#)

# Model Performance Summary

- The normalized used price of a Sony device will be 0.0670 units lower than the normalized used price of a Acer device, all other variables remaining constant.
- The normalized used price of a Xiaomi device will be 0.0801 units higher than the normalized used price of a Acer device, all other variables remaining constant.
- The normalized used price of a device with others os will be 0.1276 units lower than the normalized used price of a device with Android os, all other variables remaining constant.
- The normalized used price of a device with iOS os will be 0.0900 units lower than the normalized used price of a device with Android os, all other variables remaining constant.
- The normalized used price of a device with 4g will be 0.0502 units higher than the normalized used price of a device without 4g, all other variables remaining constant.
- The normalized used price of a device with 5g will be 0.0673 units lower than the normalized used price of a device without 5g, all other variables remaining constant.


[Link to Appendix slide on model assumptions](#)

# Model Performance Summary

- The most important factors used by the model for prediction are the main camera megapixel, the selfie camera megapixel, the random access memory (ram), the device weight, the normalized price of the device when new and the number of years since its release.
- Key performance metrics for training and test data in tabular format for comparison


Training Performance

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



Test Performance

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



- Root Mean Square Error (RMSE) on the train and test sets are comparable. The model is not suffering from overfitting.
- Mean Absolute Error (MAE) indicates that the model is able to predict normalized used price within a comparable mean error of 0.186649 units on the test data.
- Hence, it can be concluded that the model is good for prediction as well as inference purposes.

# Actionable Insights and Recommendations

## Conclusions:

- There are 3,454 devices with 15 features recorded in the data
- The most important features to predict the normalized used price of a device are ram, weight, normalized new price and year since release
- The most popular brand of device is Samsung, then Huawei, LG, Lenovo and ZTE respectively
- Android os is by far the most popular operating system
- Devices with screen size 12cm to 13cm are the most popular
- Devices are more likely to have 4g than 5g or not having anything
- The most common main camera mp are 13, 8 and 5 megapixels respectively
- The most common selfie camera mp are 5, 8, 2 and 0.3 megapixels respectively
- Most devices have internal memory between 0.01GB and 32GB

# Actionable Insights and Recommendations

## Conclusions:

- The vast majority devices have random access memory around 4GB
- Most devices have battery capacities in the range 4000 – 4200mAh, 3000 – 3200mAh, and 2000 – 2200mAh
- Most devices weight between 100g and 220g
- Used devices have a higher probability of having earlier year of release from the years under review
- The median number of days device are used is 690.5 while the mean number of days used is 674.9
- The average normalized new price for devices is 5.23 while the average normalized used price is 4.36
- The top 10 brands of mobile devices with screen sizes more than 15.24cm are Huawei, Samsung, Vivo, Honor, Oppo, Xiaomi, Lenovo, LG, Motorola and Asus respectively.

# Actionable Insights and Recommendations

## Conclusions:

- Used devices with above 4500mAh battery capacity weight between 200 grams and 500 grams
- The top 10 brands of mobile devices that have selfie camera mp more than 8 are Huawei, Vivo, Oppo, Xiaomi, Samsung, Honor, LG, Motorola, Meizu and HTC respectively.
- The top 10 brands of mobile devices that have main camera mp more than 16 are Sony, Motorola, HTC, ZTE, Meizu, Nokia, Microsoft, Asus, Blackberry and Samsung respectively.
- The normalize used price has a positive relationship with the release year
- Mobile devices having 5g command higher normalized used prices than mobile devices having 4g

# Actionable Insights and Recommendations

## Recommendations:

- Priority should be given to the main and selfie camera mp, ram, weight, normalized new price and year since release of a device as these are the best predictors of the normalized used price.
- Priority should be given to brand such as Samsung, Huawei, LG, Lenovo and ZTE because they have a wider variety of devices and are very popular in the used market.
- Priority should be given to Android os as it is the most featured operating system on devices.
- Priority should be given to devices with main camera mp 13, 8 and 5 , selfie camera mp 5, 8 and 2, 4GB ram, weighing between 200g and 500g, having 4g/5g and with as little as possible the number of years since release to get the best normalized used price for the device at any given normalized new price.
- An Acer device with a similar specification will have used price advantage compared to Samsung and Sony brands, but will have a disadvantage compared to Karbonn and Xiaomi brands.

# APPENDIX



# Data Background and Contents

- The dataset contains 3,454 observations and 15 features
- 11 of the features (including brand\_name) are categorical in nature, while 4 are numerical
- There are 6 features with missing values.
- There are no duplicated observations

The statistical summary of the data shows us the following

- Normalized used prices for devices range between 1.54 to 6.61 with an average of 4.36. 75% of devices cost at least 4.03
- Variables '4g' and '5g' have only 2 categories while variable 'os' has 4 categories
- Variable 'release\_year' appears as a numerical variable but functions as a categorical variable
- The screen size range from 5.08 cm to 30.71 cm with an average of 13.71 cm  
75% of all devices have a screen size of all least 12.7 cm

## Data Background and Contents – cont'd

- The main camera mp range from 0.08 to 48. 75% of all devices have a main camera mp of all least 5.0
- The selfie camera mp range from 0.0 to 32. 75% of all devices have a selfie camera mp of all least 2.0
- The internal memory range from 0.01 GB cm to 1024 GB with an average of 54.6 GB 75% of all devices have a screen size of all least 16.0 GB
- The ram range from 0.02 GB cm to 12 GB with an average of 4.0 GB. 75% of all devices have a screen size of all least 4.0 GB
- The battery capacity range from 500 mAh to 9270 mAh with an average of 3133.40 mAh. 75% of all devices have a screen size of all least 2100 mAh
- The device weight range from 69 g to 855 g with an average of 182.75 g. 75% of all devices have a screen size of all least 142 g
- The release years for the devices was between 2013 and 2020 with 75% of the devices released by 2018

# Data Background and Contents

- The number of days the devices were used range between 91 days to 1094 days. 75% of devices were used for at least 533.5 days. The number of days used is 675 days.
- Normalized new prices for devices range between 2.90 to 7.85 with an average of 5.23. 75% of devices cost at least 4.79 when new

# Model Assumptions

The tests conducted for checking model assumptions and the results obtained are:

- No Multicollinearity
- Linearity of variables and independence of error terms
- Normality of error terms
- Homooscedasticity or equality of variance

# Model Assumptions – Test for Multicollinearity

- The test is performed using variance inflation factor (VIF)
- If VIF is 1 then there is no correlation between the  $k^{th}$  predictor and the remaining predictor variables.
- If VIF exceeds 5 or is close to exceeding 5, we say there is moderate multicollinearity.
- If VIF is 10 or exceeding 10, it shows signs of high multicollinearity. Such predictors are removed as their effect is explained or contained in another predictor.
- The VIF values values are:

# Test for Multicollinearity

```
[ ] checking_vif(x_train)
```

	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
11	brand_name_Alcatel	3.405693
12	brand_name_Apple	13.057668
13	brand_name_Asus	3.332038



14	brand_name_BlackBerry	1.632378
15	brand_name_Celkon	1.774721
16	brand_name_Coolpad	1.468006
17	brand_name_Gionee	1.951272
18	brand_name_Google	1.321778
19	brand_name_HTC	3.410361
20	brand_name_Honor	3.340687
21	brand_name_Huawei	5.983852
22	brand_name_Infinix	1.283955
23	brand_name_Karbonn	1.573702
24	brand_name_LG	4.849832
25	brand_name_Lava	1.711360
26	brand_name_Lenovo	4.558941
27	brand_name_Meizu	2.179607
28	brand_name_Micromax	3.363521
29	brand_name_Microsoft	1.869751

30	brand_name_Motorola	3.274558
31	brand_name_Nokia	3.479849
32	brand_name_OnePlus	1.437034
33	brand_name_Oppo	3.971194
34	brand_name_Others	9.711034
35	brand_name_Panasonic	2.105703
36	brand_name_Realme	1.946812
37	brand_name_Samsung	7.539866
38	brand_name_Sony	2.943161
39	brand_name_Spice	1.688863
40	brand_name_Vivo	3.651437
41	brand_name_XOLO	2.138070
42	brand_name_Xiaomi	3.719689
43	brand_name_ZTE	3.797581
44	os_Others	1.859863
45	os_Windows	1.596034
46	os_iOS	11.784684
47	4g_yes	2.467681
48	5g_yes	1.813900

# Model Assumptions – Test for Multicollinearity

- Columns with VIF values 5 and above and made the least change to adjusted R-squared and RMSE were dropped – screen\_size
- Next, predictor variables with p-value higher than 0.05 significant level were also dropped
- The remaining predictor variables used in the model are:  
*'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'*
- The model performance after using these predictor variables are

Training Performance

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



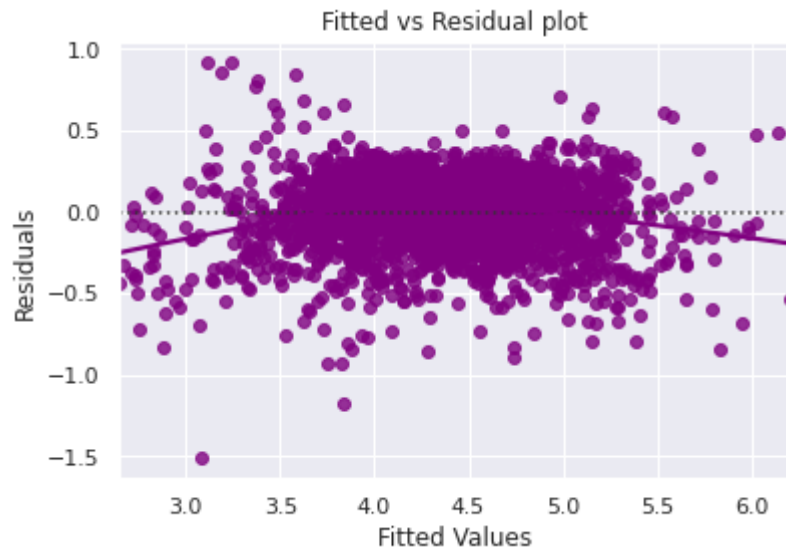
Test Performance

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



# Model Assumptions – Linearity and Independence of Errors

- A plot of residuals against fitted values was made to test for linearity and independence of variables
- The residuals are the difference in actual and fitted values. There should be no pattern determinable from the plot ideally.



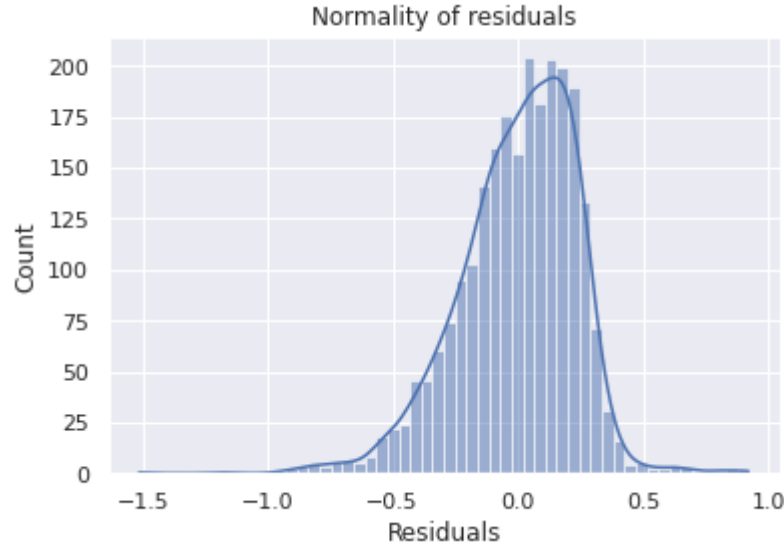
- The scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values).
- There is no pattern in the plot. Hence, the assumptions of linearity and independence are satisfied.



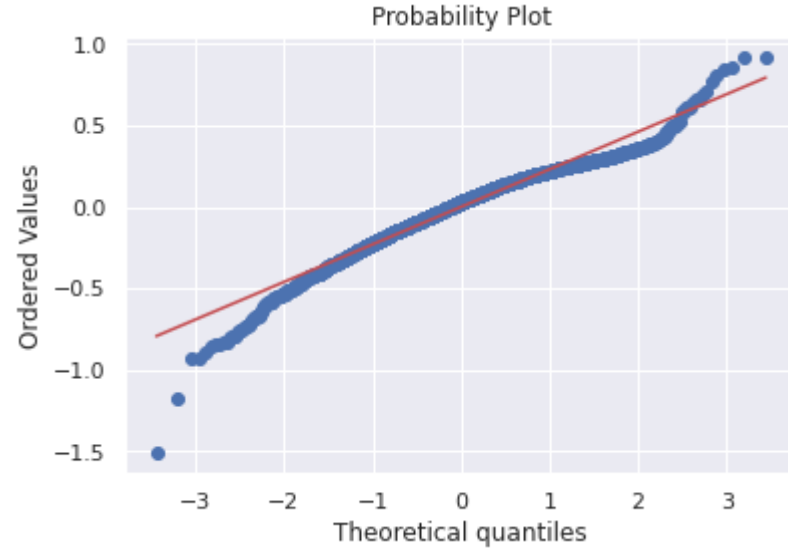
# Model Assumptions – Normality of error terms

- Test for normality performed 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.
- If the p-value of the Shapiro-Wilk test is greater than 0.05, we can say the residuals are normally distributed.

# Model Assumptions – Normality of error terms



Distribution of residuals



Q-Q plot of residuals

- The histogram of residuals does have a bell shape though not perfect
- The residuals more or less follow a straight line except for the tails.

## Model Assumptions – Normality of error terms

- The null hypothesis for the Shapiro-Wilk test

$H_0$ : Data is normally distributed.

was tested against the alternative hypothesis

$H_a$ : Data is not normally distributed.

- A p-value of 6.995328206686811e-23 was obtained, which is lower than the significance level of 0.05  
Therefore, we should reject the null hypothesis
- However, as an approximation, we can accept this distribution as close to being normal.  
So, the assumption is satisfied

# Model Assumptions – Test for Homooscedasticity

- 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.
- The null hypothesis for the Goldfeldquandt test

$H_0$ : The residuals are homoscedastic.

was tested against the alternative hypothesis

$H_a$ : The residuals are heteroscedastic.

- A p-value of 0.4402 was obtained, which is higher than the significance level of 0.05. Therefore, we fail to reject the null hypothesis.
- The residuals are homoscedastic. So, this assumption is satisfied.



**Happy Learning !**

