ReCell - a project on used cell phone buying and selling

ReCell - used phones Data Analysis
The University of Texas at Austin
McCombs School of Business

Itu Mukherjee
Date : 11.11.2022

Contents / Agenda

Executive Summary

- The used and refurbished device market has grown considerably over the past decade. Refurbished and used devices continue to provide cost-effective alternatives to both consumers and businesses that are looking to save money when purchasing one. There are plenty of other benefits associated with the used device market.
- Exploratory analysis of the data show that expensive brands have the maximum number of refurbished phones with large screen size and better selfie camera compared to cheaper brands.
- After analyzing the data, we can see that main camera, selfie camera, screen size, weight and normalized new price are significant parameters in predicting used price. An increment of any of the above is expected to increase the used price for the device, as indicated by the positive coefficients for these parameters in a multivariate regression model.

Business Problem Overview and Solution Approach

- We want to 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:

- EDA (univariate and multivariate analysis), duplicate value check, missing value treatment, outlier check (treatment if needed), feature engineering
- Linear Regression model building
- Train, test data split and model performance check
- Checking linear regression assumptions such as multicollinearity, linearity of variables, independence of error terms, normality of error terms, test for heteroscedasticity

Data Overview

	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	normalized_new_price
0	Honor	Android	14.50	yes	no	13.0	5.0	64.0	3.0	3020.0	146.0	2020	127	4.307572	4.715100
1	Honor	Android	17.30	yes	yes	13.0	16.0	128.0	8.0	4300.0	213.0	2020	325	5.162097	5.519018
2	Honor	Android	16.69	yes	yes	13.0	8.0	128.0	8.0	4200.0	213.0	2020	162	5.111084	5.884631
3	Honor	Android	25.50	yes	yes	13.0	8.0	64.0	6.0	7250.0	480.0	2020	345	5.135387	5.630961
4	Honor	Android	15.32	yes	no	13.0	8.0	64.0	3.0	5000.0	185.0	2020	293	4.389995	4.947837

- We have 3454 rows and 15 columns in the Data Frame
- Brand name, operating system, 4g and 5g are categorical while all others are numerical data types
- Normalized used price is the dependent variable
- There are 34 different manufacturing brands, 4 different operating systems, 4gs or 5gs have either values yes or no for different phones

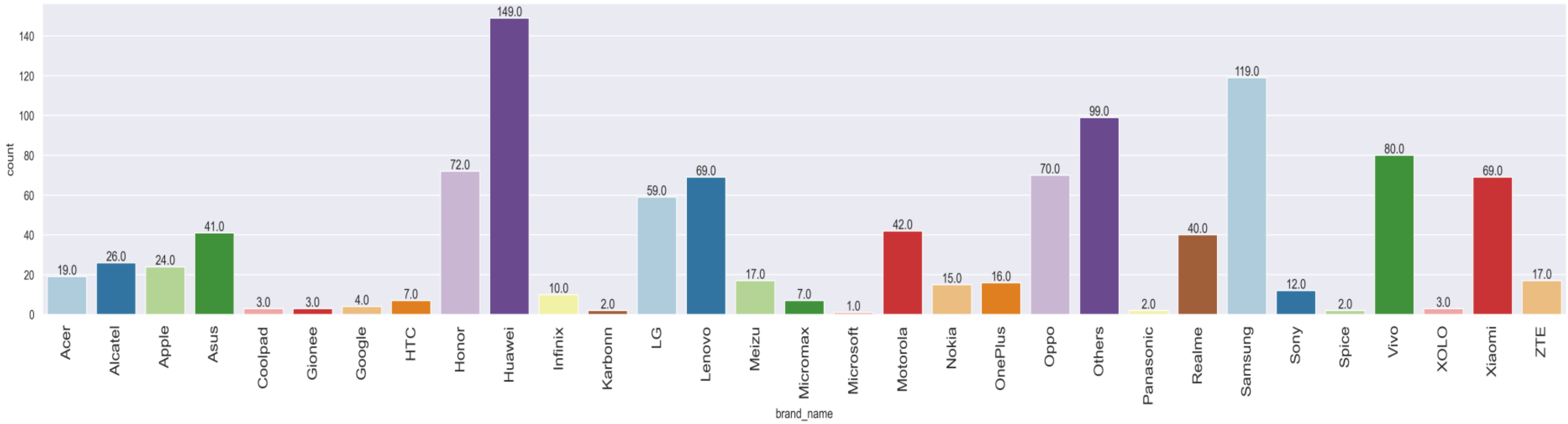
Link to Appendix slide on data background check

Exploratory Data Analysis

- Android is the most popular operating system, with 3214 phones running on the same
- 2335 phones have 4g available
- The average values for most numerical data types like screen size, main camera, selfie camera, internal memory, battery, weight, normalized new price and normalized used price are larger than median values, indicating that data may be skewed right.
- The average values and median values are almost similar for amount of RAM in GB, indicating very little skewness, if any
- The average values for number of days the used/ refurbished phone has been used is less than median value, indicating data maybe skewed left

Exploratory Data Analysis

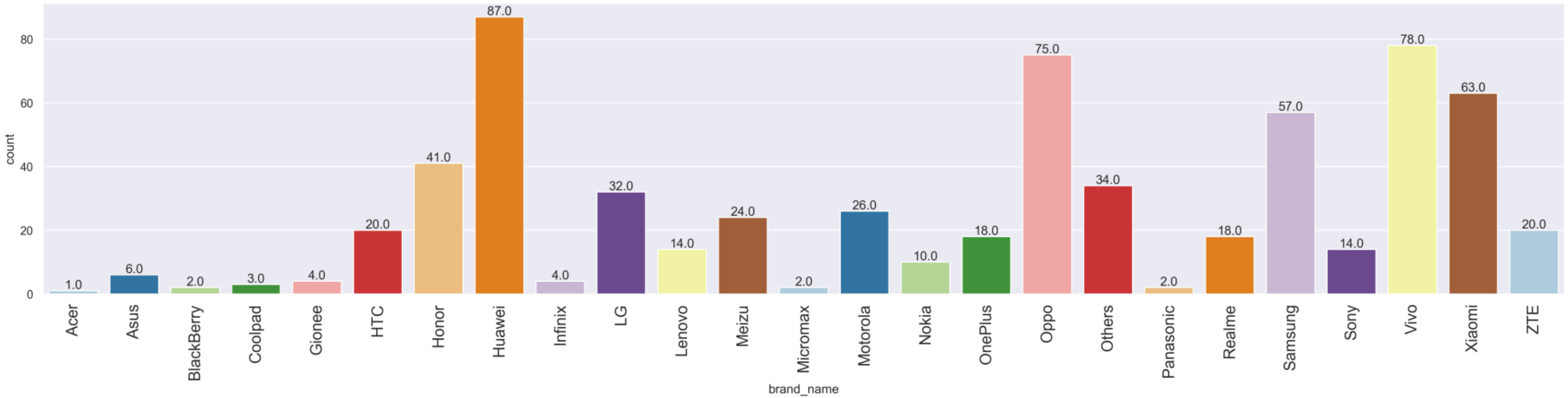
Phones with large screen size



- Huawei has the highest number of refurbished phones with large screen (i.e. 149 phones), followed by Samsung (119 phones), Honor (72 phones), Vivo (80 phones), Xiaomi (69 phones) and Oppo (70 phones) among known manufacturing brands
- Microsoft (1 phone), Karbonn/Panasonic/Spice (2 phones) have the lowest number of refurbished phones with large screen size

Exploratory Data Analysis

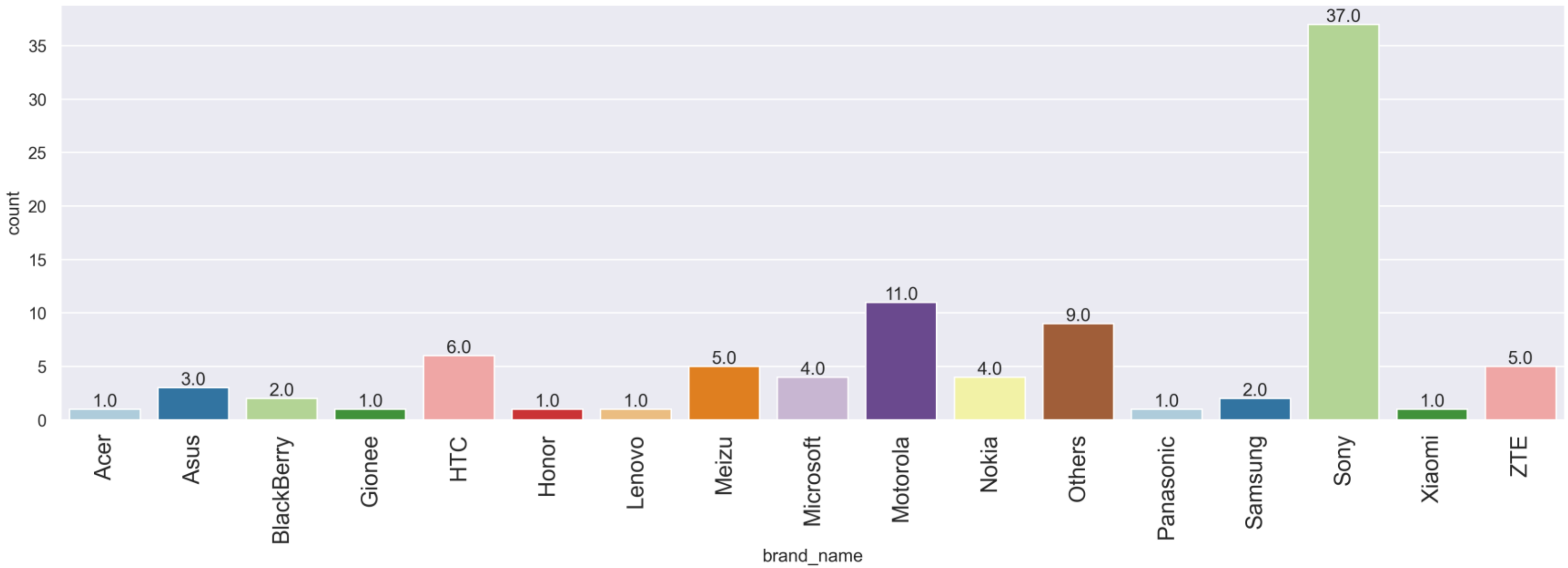
Phones with greater selfie camera



- Huawei (87 phones), Oppo (75 phones), Vivo (78 phones), Xiaomi (63 phones) and Samsung (57 phones) have some of the highest number of refurbished phones with a great selfie camera (>8MP) -similar brand names observed as for phones with large screen size
- Acer (1 phone), Blackberry/Microsoft/Panasonic (2 phones) have some of the lowest number of refurbished phones with a great selfie camera (>8MP)

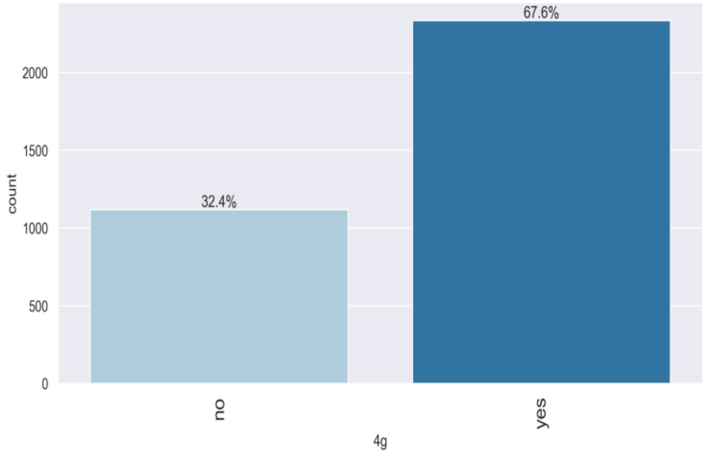
Exploratory Data Analysis

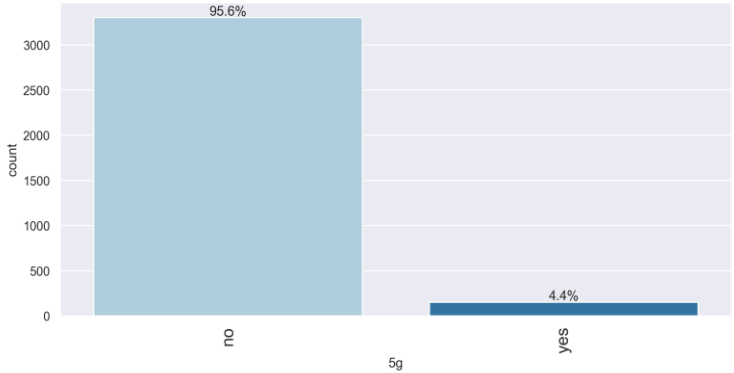
Phones with greater main camera



- Only Sony (57 phones) have the highest number of refurbished phones with a great main camera (>8MP) - similar brand names observed as for phones with large screen size

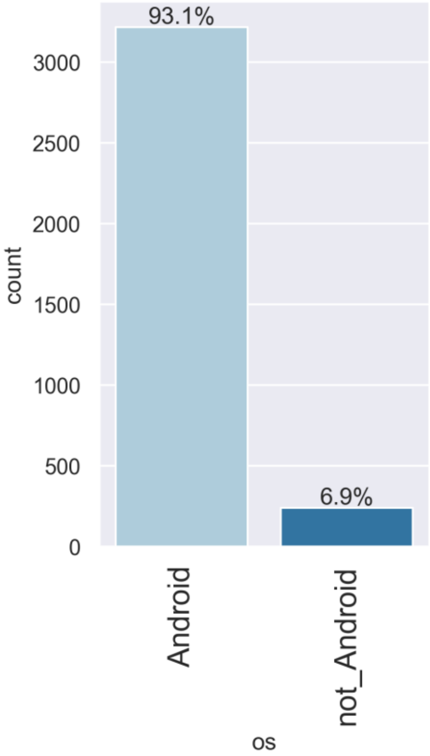
Exploratory Data Analysis



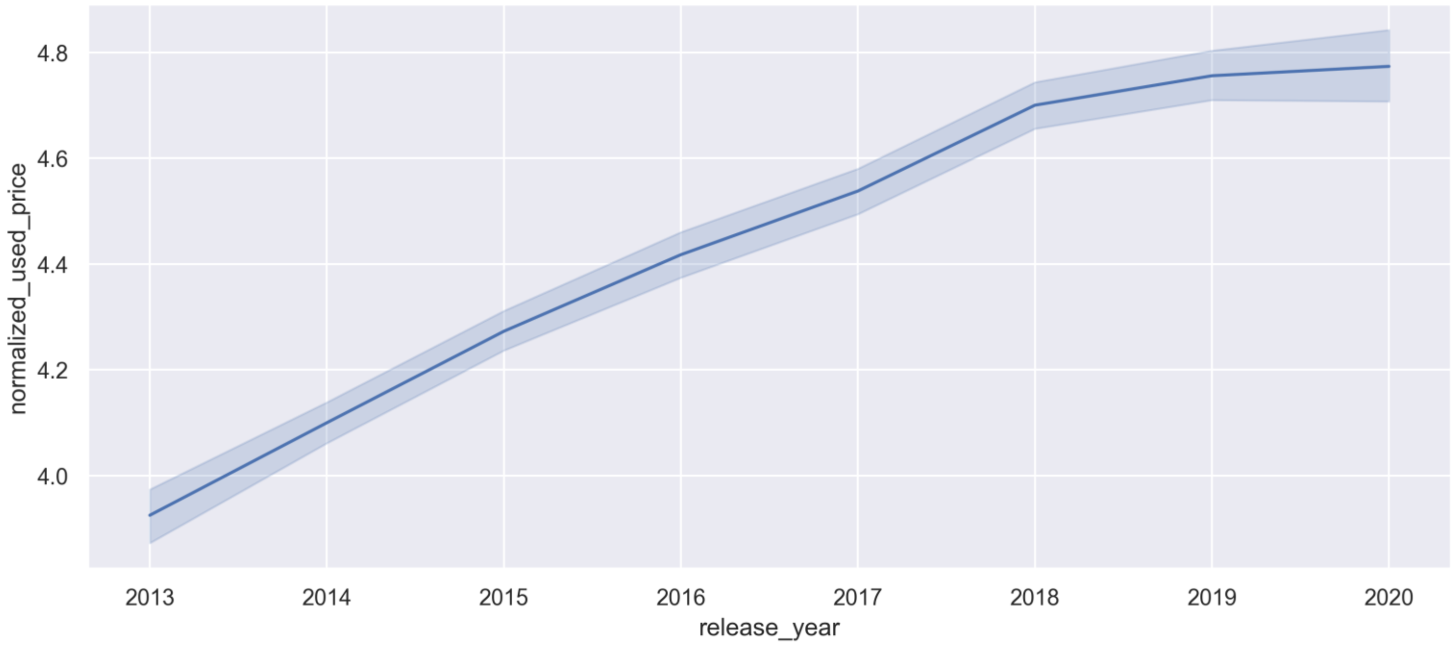


- More than 90% of the used phone market is dominated by Andorid devices

- Number of 4g phones are higher to 5g phones.

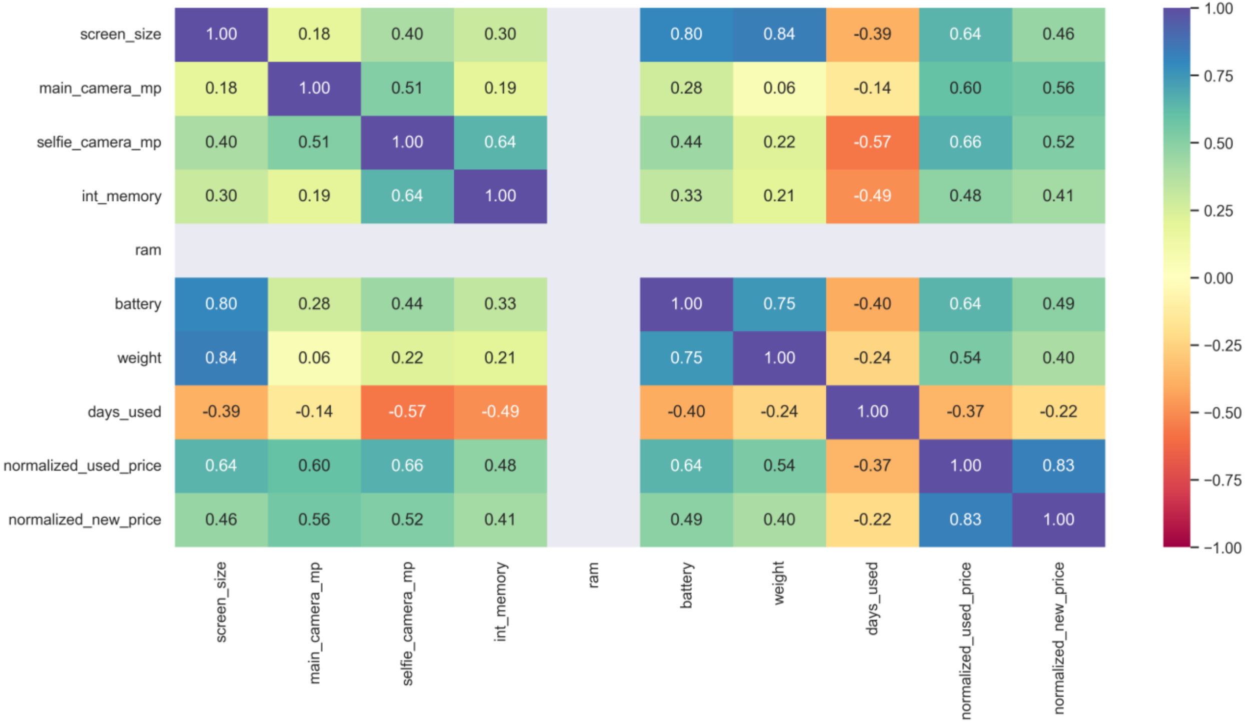


Exploratory Data Analysis



- Used prices increase linearly from 2013 till 2018.
- Since then, there have not been any noticeable increase in used phone prices.

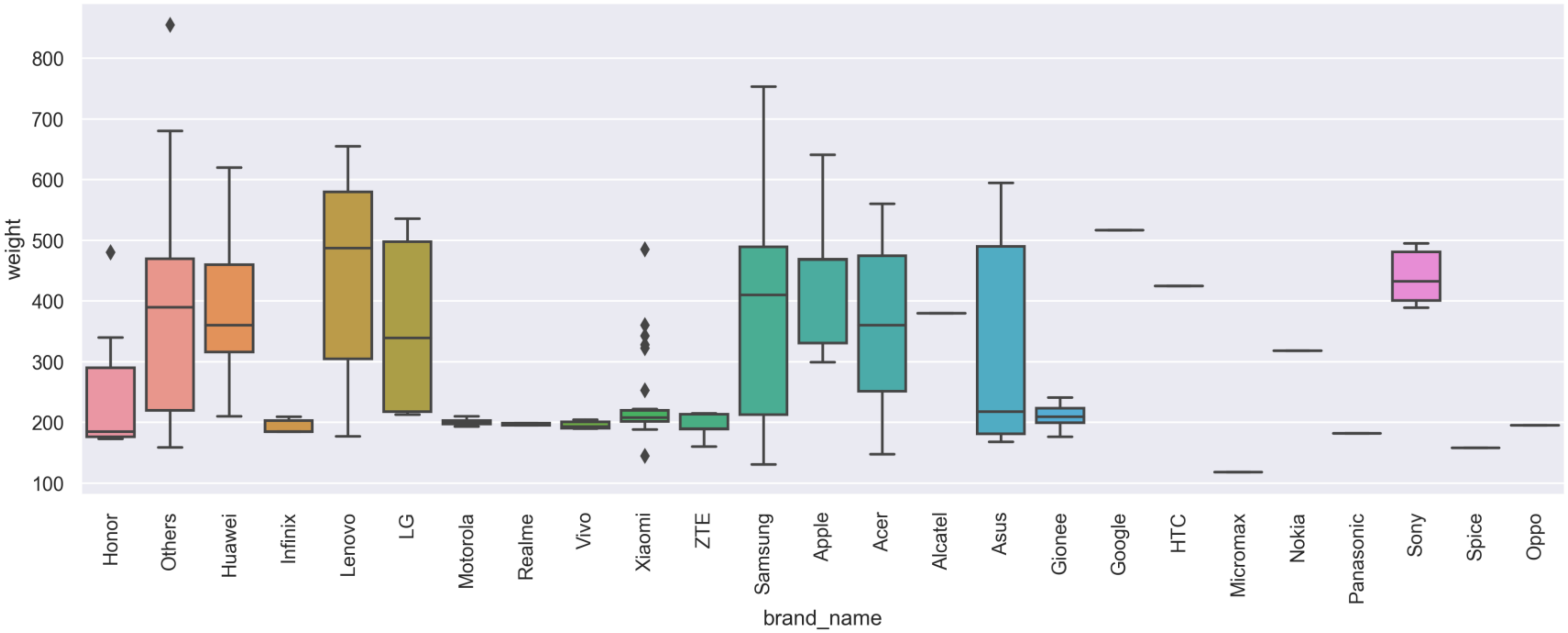
Exploratory Data Analysis



- days_used and selfie_camera_mp are negatively correlated (-0.57 respectively)
- Weight and screen_size and normalised_used_price and normalised_new_price are strongly-positively correlated (with 0.84 and 0.83 values respectively)

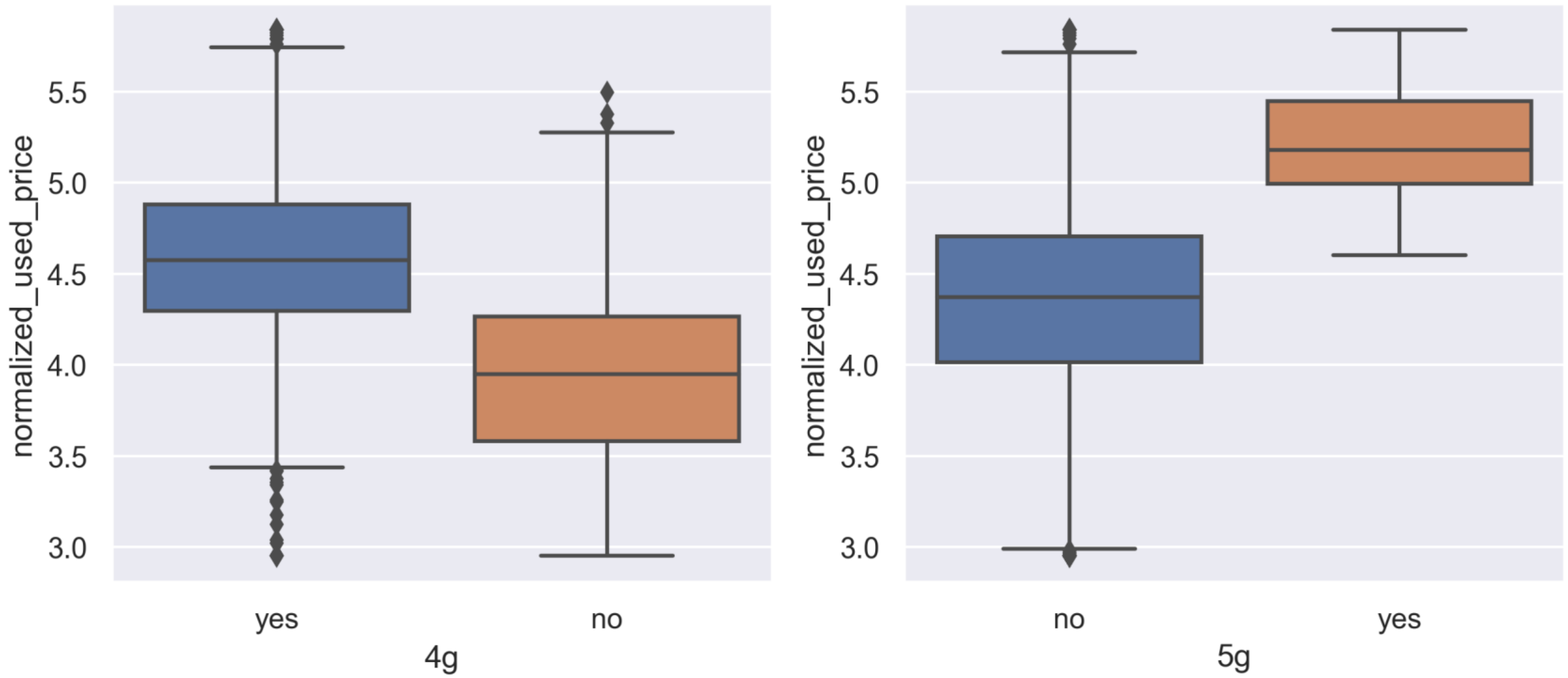
Exploratory Data Analysis

Phones with large batteries > 4500 mAh



- Among all phones with large batteries Motorola, Realme, Vivo, Xiaomi, Infinity, Zte, Gionee have relatively lighter weights compare to the other brands.

Exploratory Data Analysis



- The price for 5g is slightly higher than 4g

Data Preprocessing

- **Duplicate value check** - There are no duplicate values in the Data Frame
- **Missing value treatment** - Main camera, selfie camera, internal memory, ram, battery and weight have missing values
- **Outlier check (treatment if needed)** – Found some insignificant outliers.
- **Feature engineering**
  - we created a new column ”years_since_release” from the ”release_year column”.
  - We considered the year of data collection, 2021, as the baseline.
  - We dropped the ”release_year column”

Data Processing

Data preparation for modeling

- We predicted the normalized price of used devices.
- To evaluate the model, we split the data into train and test in the ratio of 70-30.
- We built a Linear Regression model using the train data and then check its performance. Before model building, we encoded the categorical features.
- Normalized used price is the dependent variable. Number of rows in train data is 2417 Number of rows in test data is 1037

Data Processing

❖ Normalized used price is the dependent variable (y) and all other variables are independent variables (x)

```
  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
```

Train Test Split Data

brand_name_Samsung	brand_name_Sony	brand_name_Spice	brand_name_Vivo	brand_name_XOLO
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

brand_name_Xiaomi	brand_name_ZTE	os_not_Android	4g_yes	5g_yes
0	0	0	1	0
0	0	0	1	1
0	0	0	1	1
0	0	0	1	1
0	0	0	1	0

Encoded Categorical Data

Model Performance Summary

- We used OLS model from Linear Regression
- We used metric functions defined in sklearn for RMSE, MAE, and $R2$ and defined a function to calculate MAPE and adjusted $R2$.
- We created a function which will print out all the above metrics in one go.

Model performance on train set (seen 70% data)

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.2299	0.180312	0.844864	0.841786	4.326774

Model performance on train set (seen 30% data)

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.238391	0.184803	0.842435	0.834947	4.503298

- Factors used by our machine learning model for prediction are "RMSE" "MAE“, "R-squared", “Adj. R-squared", "MAPE"

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     normalized_used_price   R-squared:                   0.845
Model:                               OLS   Adj. R-squared:              0.842
Method:                    Least Squares   F-statistic:                 280.6
Date:                   Wed, 09 Nov 2022   Prob (F-statistic):           0.00
Time:                           09:27:16   Log-Likelihood:             123.68
No. Observations:                   2417   AIC:                        -153.4
Df Residuals:                       2370   BIC:                         118.8
Df Model:                             46
Covariance Type:               nonrobust
```

Model Assumptions

- **Linear Regression Assumptions**

  - Multicollinearity check
  - Linearity of variables
  - Independence of error terms
  - Normality of error terms
  - Heteroscedasticity

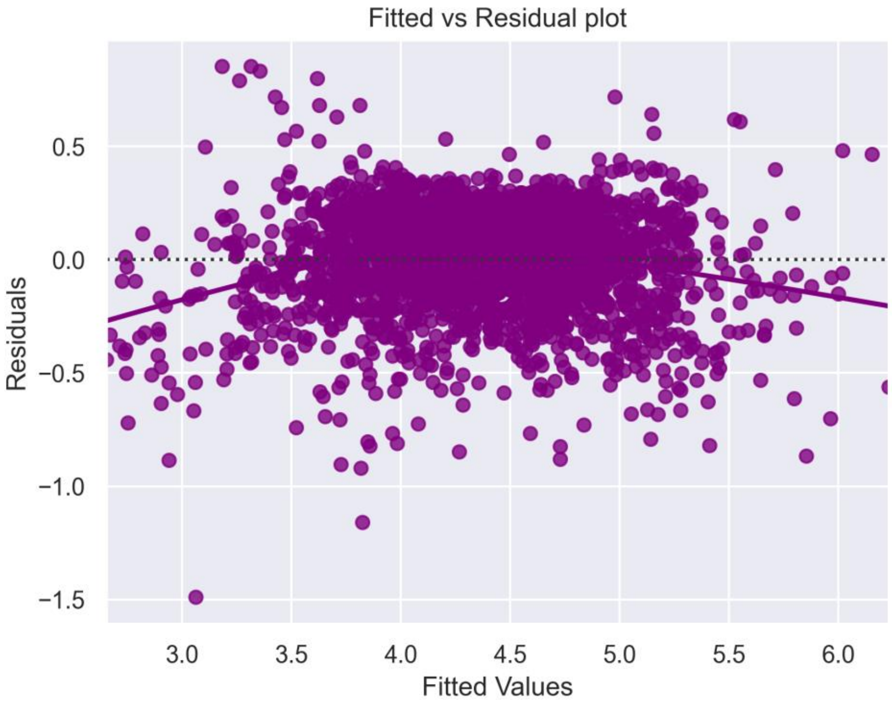
Multicollinearity check

- We tested for Multicollinearity using VIF (Variance Inflation Factor)

	col	Adj. R-squared after_dropping col	RMSE after dropping col
0	brand_name_Huawei	0.841919	0.232120
1	brand_name_Others	0.841917	0.232121
2	brand_name_Samsung	0.841884	0.232145
3	weight	0.838183	0.234847
4	screen_size	0.838172	0.234855

- It was observed that brand_name_Huawei, brand_name_Others, brand_name_Samsung, weight and screen_size are more than the value 5 which shows moderate to high multicollinearity.
- After dropping brand_name_Huawei, brand_name_Others, brand_name_Samsung and screen_size columns one by one we got VIF values less than 5.

Test for Linearity and Independence

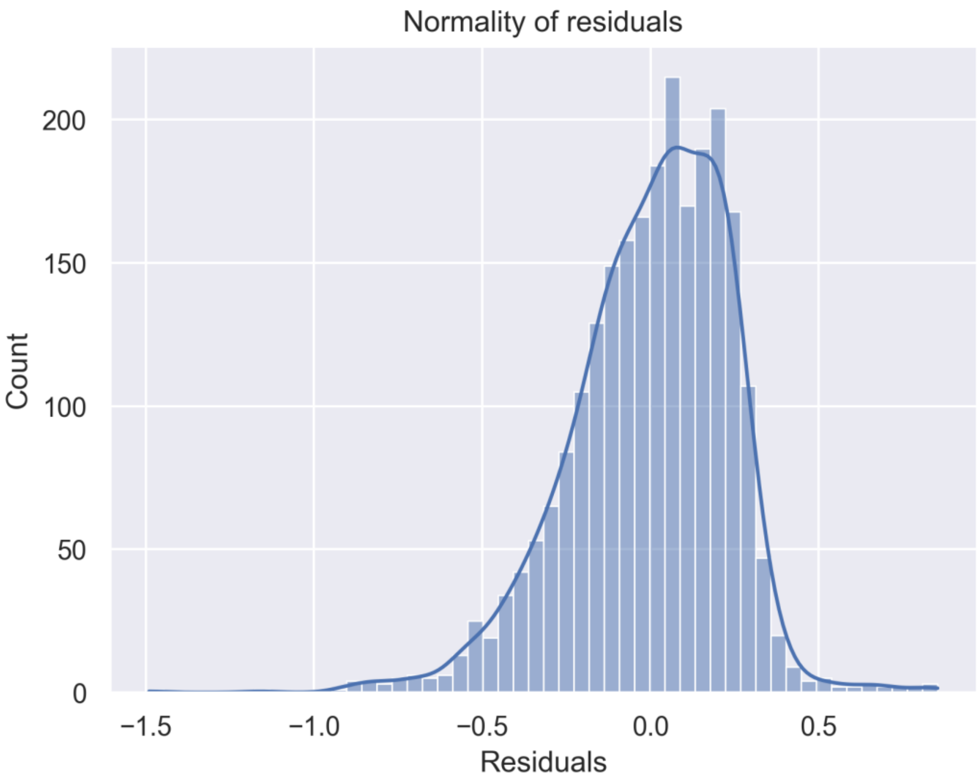


	Actual Values	Fitted Values	Residuals
3026	4.087488	3.860511	0.226977
1525	4.448399	4.645695	-0.197295
1128	4.315353	4.282477	0.032875
3003	4.282068	4.185717	0.096351
2907	4.456438	4.482911	-0.026473

- There is no pattern in the plot of fitted values vs residuals as below.
- So, we can say the model is linear and residuals are independent.

Test for Normality

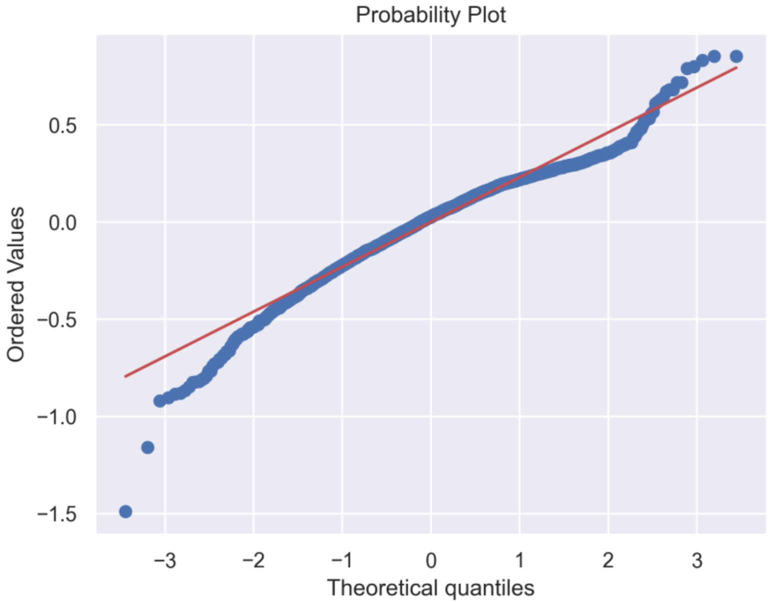
- ➢ Tested for normality by checking the Q-Q plot of residuals



- ❖ Null hypothesis: Residuals are normally distributed

- ❖ Alternate hypothesis: Residuals are not normally distributed

Test for Normality



Shapiro Wilk's Result :- (statistic=0.968555748462677, pvalue=1.3796865474485753e-22)

- The residuals shows approximately a straight line except for the tails
- p-value < 0.05, as per the Shapiro-Wilk's test the residuals are not normal.
- The distribution of the residuals is close to being normal

Test for Homoscedasticity

- Null hypothesis: Residuals are homoscedastic
- Alternate hypothesis: Residuals have heteroscedasticity

- We have done goldfeldquandt test to check for homoscedasticity. When the variance of the residuals is symmetrically distributed across the regression line, then the data is said to be homoscedastic and whereas it is said to be heteroscedastic when the same is unequal.

[('F statistic', 1.01667630253455), ('p-value', 0.38762657199753797)]

- Since p-value > 0.05, we can say that the residuals are homoscedastic. So, this assumption is satisfied

Final Model 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:                     835.3
Date:                   Wed, 09 Nov 2022   Prob (F-statistic):               0.00
Time:                           09:32:02   Log-Likelihood:                 80.236
No. Observations:                   2417   AIC:                            -128.5
Df Residuals:                       2401   BIC:                            -35.83
Df Model:                             15
Covariance Type:               nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                     1.4708      0.046     32.138      0.000       1.381       1.561
main_camera_mp            0.0216      0.001     15.186      0.000       0.019       0.024
selfie_camera_mp          0.0138      0.001     12.984      0.000       0.012       0.016
ram                       0.0240      0.005      4.820      0.000       0.014       0.034
weight                    0.0017   6.05e-05     27.329      0.000       0.002       0.002
normalized_new_price      0.4418      0.011     40.953      0.000       0.421       0.463
years_since_release      -0.0287      0.003     -8.402      0.000      -0.035      -0.022
brand_name_Karbonn        0.1248      0.055      2.279      0.023       0.017       0.232
brand_name_Lenovo         0.0514      0.022      2.370      0.018       0.009       0.094
brand_name_Microsoft      0.1842      0.069      2.685      0.007       0.050       0.319
brand_name_Nokia          0.0669      0.031      2.128      0.033       0.005       0.129
brand_name_Sony          -0.0613      0.030     -2.014      0.044      -0.121      -0.002
brand_name_Xiaomi         0.0871      0.026      3.385      0.001       0.037       0.137
os_not_Android           -0.1027      0.022     -4.688      0.000      -0.146      -0.060
4g_yes                    0.0479      0.015      3.182      0.001       0.018       0.077
5g_yes                   -0.0746      0.030     -2.453      0.014      -0.134      -0.015
==============================================================================
Omnibus:                      245.508   Durbin-Watson:                   1.910
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              459.932
Skew:                          -0.672   Prob(JB):                    1.34e-100
Kurtosis:                       4.661   Cond. No.                     2.97e+03
```

Training Results

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.2299	0.180312	0.844864	0.841786	4.326774

Test Results

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.238391	0.184803	0.842435	0.834947	4.503298

Final Model Summary

- The train and test RMSE and MAE (~0.22 and ~0.23) are low and comparable. So, we can say our model is not having data overfitting
- The model is able to explain ~84% of the variation in the data
- The MAPE on the test set suggests we can predict within 4.5% of normalized used price)
- So, we can conclude that our model ***olsmodel_final*** is good for prediction as well as inference purposes

Actionable insights & recommendations

- Linear correlation between screen size and weight ; through EDA we found that there is a strong positive (0.84) correlation, thereby reaffirming model validity. The company should consider the screen size and weight of the mobile for deciding the price.

- We can see that main camera, selfie camera, screen size, weight and normalized new price are significant parameters. As these increases, by default normalized used price is expected to increase. This is indicated by positive coefficients for these parameters predicted by the model. As per the exploratory data analysis we can find these features are high in brands like Huawei, Samsung, Oppo and Vivo. So, the company should focus on these brands.

Actionable insights & recommendations

- Post exploratory data processing also indicated high_brand (i.e., expensive brands) have the maximum number of refurbished phones with large screen_size and better selfie_camera and low_brand (i.e, cheaper brands) have the lowest number of such refurbished phones, thereby reaffirming the validity of the model. Definitely it is a good idea for the company to invest on high_brand phones more.

- Almost 93% of phones were found to be operating on Android operating system, also an significant factor for prediction.

- The linear predictive model is able to predict ~84% of the variance in the data, within a mean absolute percentage error of ~4.5%. Considering these we can say the model is good. We should strongly consider the parameters suggest by the model for deciding the price.

Actionable insights & recommendations

- All of the assumptions for linear regression were met for the model - multicollinearity or predictor VIFs<5, normality of error terms and homoscedasticity. While the independence and linearity assumption can be assumed met after suitable transformation/data preprocessing, the data gave the impression that non-linear models may be more suited for prediction

Happy Learning !