Business Presentation
ReCell

Contents

- Exploratory Data Analysis
- Data Preprocessing
- Model Building- Linear regression
- Testing the assumptions of Linear regression model
- Model Performance Evaluation
- Actionable Insights & Communication

Business Problem Overview and Solution Approach

- Core business idea
  - Maximizing the longevity of used phones and tablets through second-hand trade thus reducing their environmental impact and help in recycling and reducing waste
- Problem to tackle
  - Develop a machine learning based solution to develop a dynamic pricing strategy for used and refurbished devices
- Financial implications
  - Increase in revenue through utilizing the potential in the used phone and tablets market

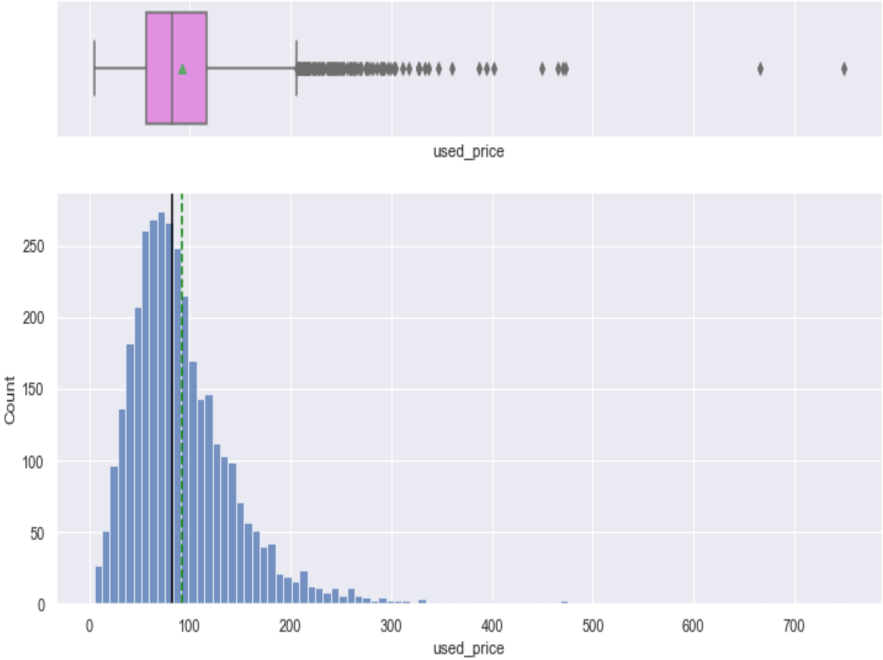
Data Overview

- Brief description of data provided
  - Brand_name: Name of manufacturing brand
  - OS: Operating System 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
  - New_price: Price of a new device of the same model in euros
  - Used_price: Price of the used/refurbished device in euros
- There are 3454 rows and 15 columns

Data Overview – Univariate Analysis

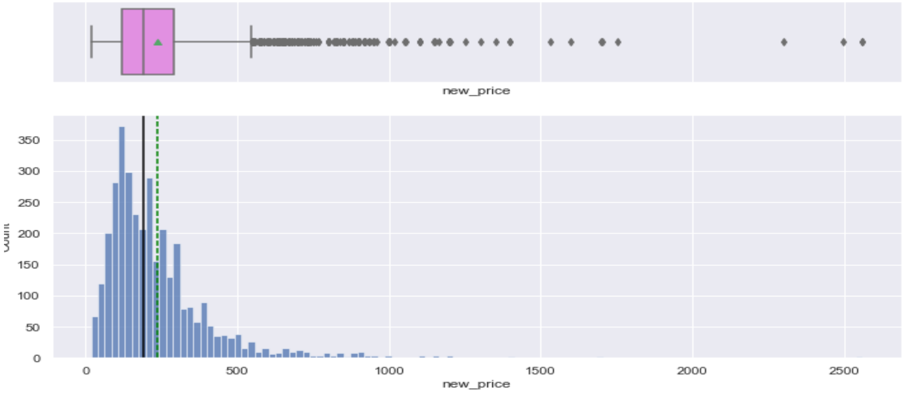
- There are 34 unique brand names within the dataset
- The tablets and phones run on 4 unique operating systems
- The tablets and phones run on the 4g and 5g network
- The release years run from 2013 to year 2020 with a pre-owned days of usage of between 91 and 1094 days

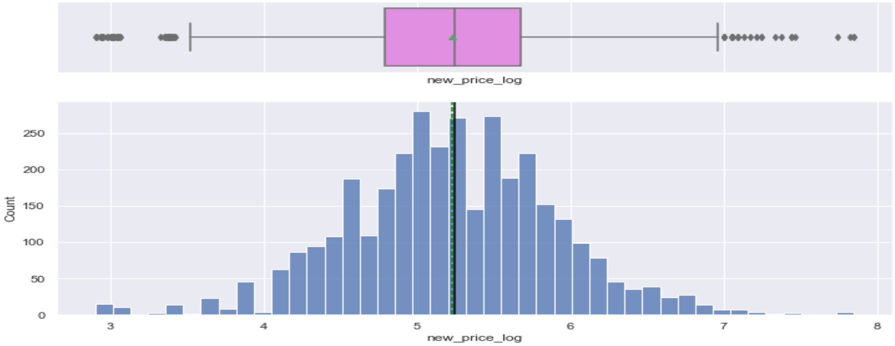
Exploratory Data Analysis: Univariate Analysis



- The used price are kind of normally distributed with a mean of about 98
- The prices appear to be close to each other with few outliers
- The average cost of the order is near the median cost indicating that the distributions is nearly symmetrical

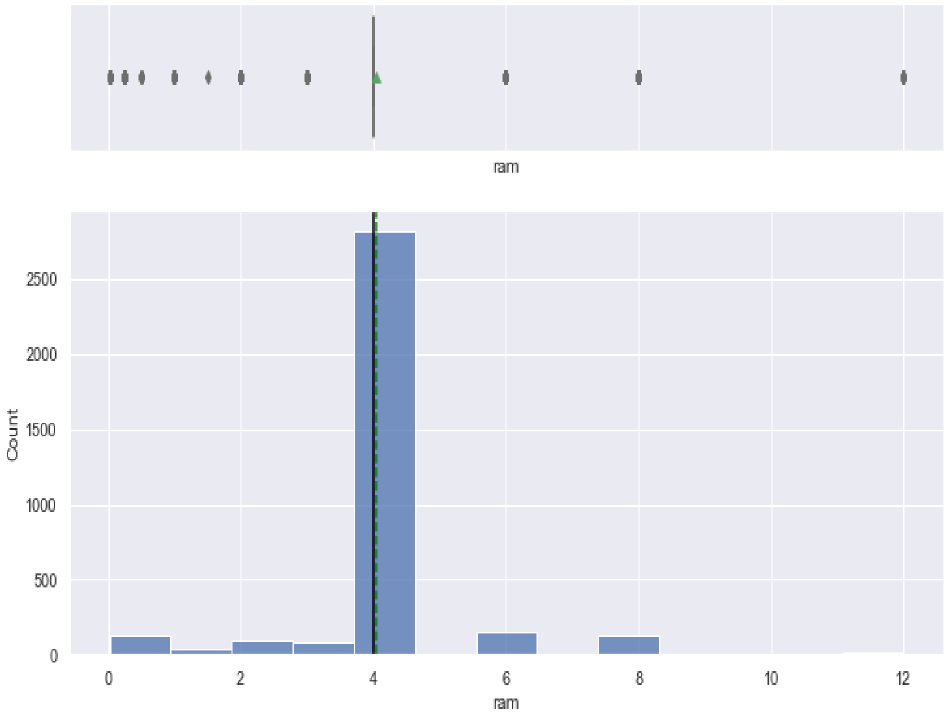
Exploratory Data Analysis: Univarate Analysis-New price





- The data appears fairly normally distributed with a bit of a right skewed
- The average new_price is almost equal to the median new_price indicating that the distribution is nearly symmetrical
- A lot of the prices centered around each other

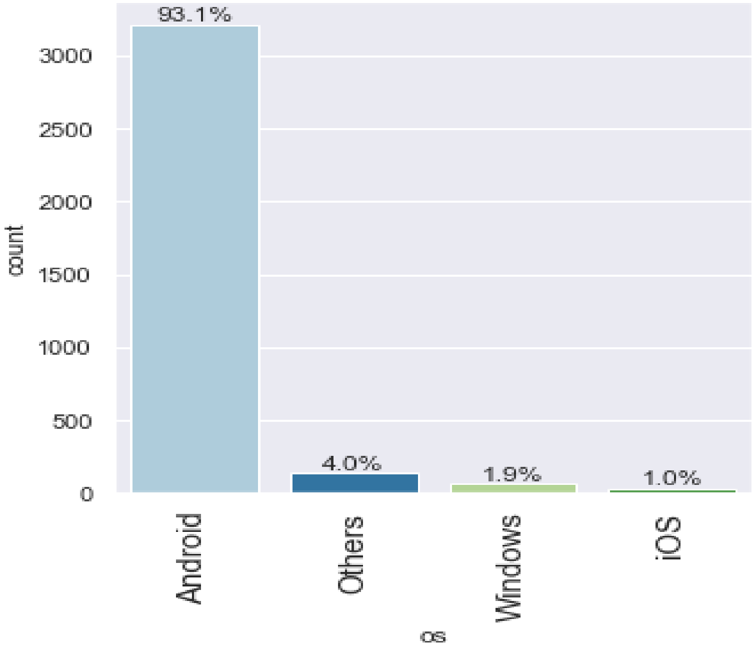
Exploratory Data Analysis: Univarate Analysis-RAM



- The data appears mostly left skewed with near symmetry

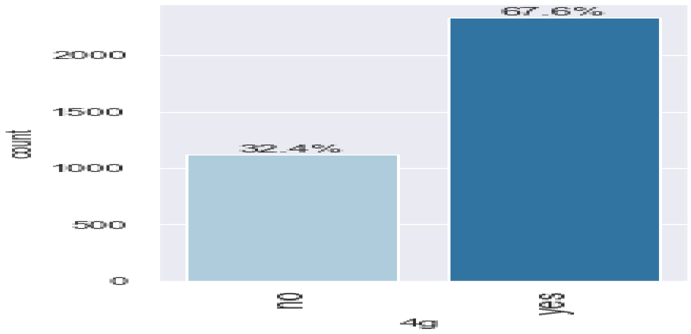
- The used phones and tablets are majority 4 ram

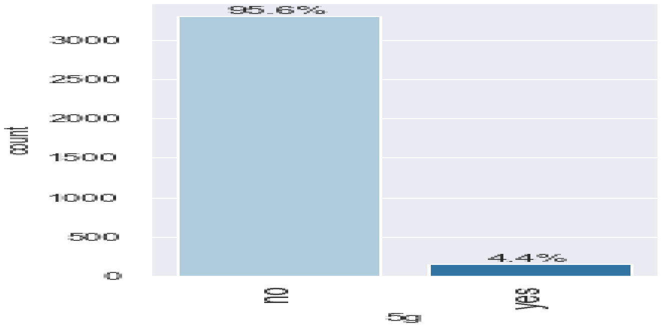
Exploratory Data Analysis (EDA): Univarate Analysis- Operating System



- The 93.1% of phones and tablets mostly use the Android system with the ios operating system being used the least at 1.0%

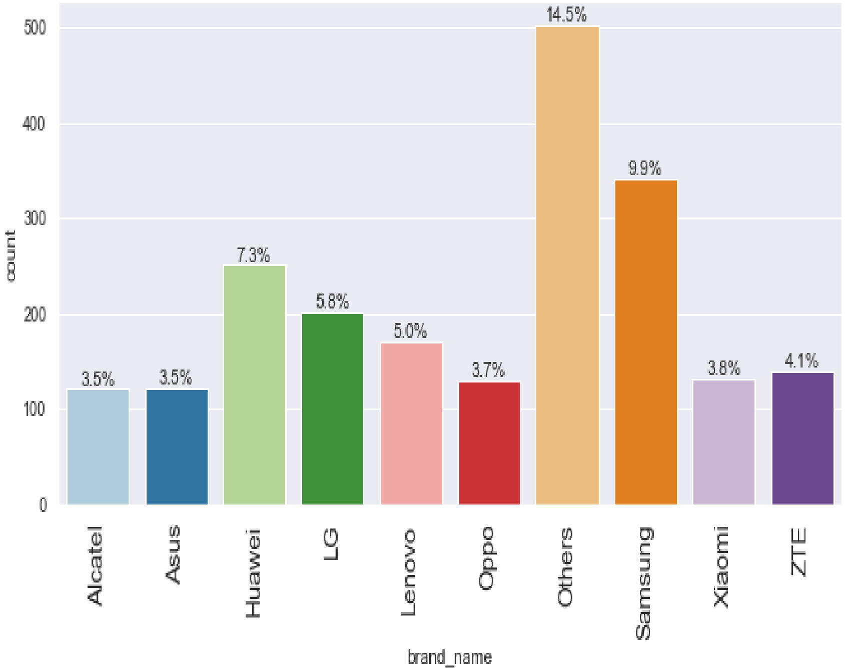
Exploratory Data Analysis (EDA): Univarate Analysis- Network





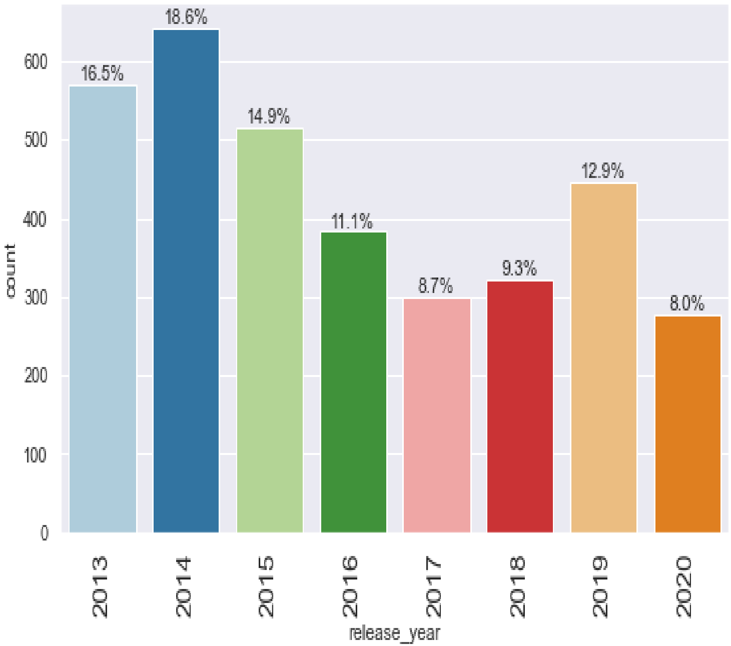
- With respect to the used phones and tablets the distribution shows that:
- 4g accounts for 67.6% usage
- 5g account for only 4.4% of phones using it

Exploratory Data Analysis (EDA): Univarate Analysis- Brand Name



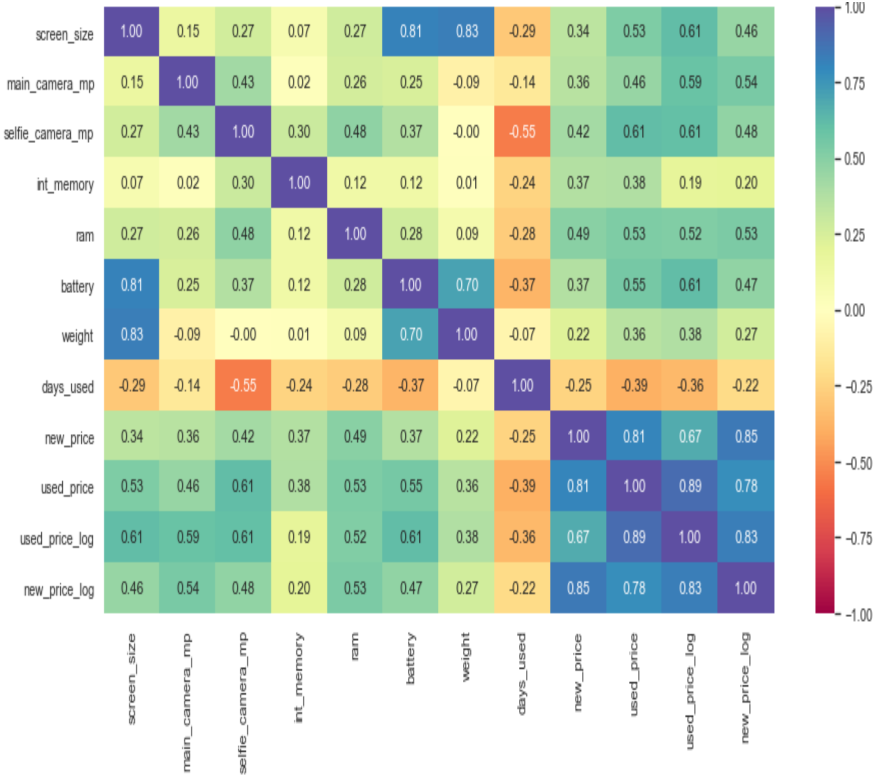
- It appears others accounts for 14.5% of the market share
- While brand name Samsung with 9.9% market share is second, followed by Huawei with 7.3% market share
- The brand Alcatel and Asus have the same market share of 3.5%

Exploratory Data Analysis (EDA): Univarate Analysis- Release Year



- Year 2014 accounts for 18.6% of the used phones and tablets
- Followed by year 2015 14.9% and 2019 12.9%
- The year 2020 has the least number of used phone and tablets at 8%

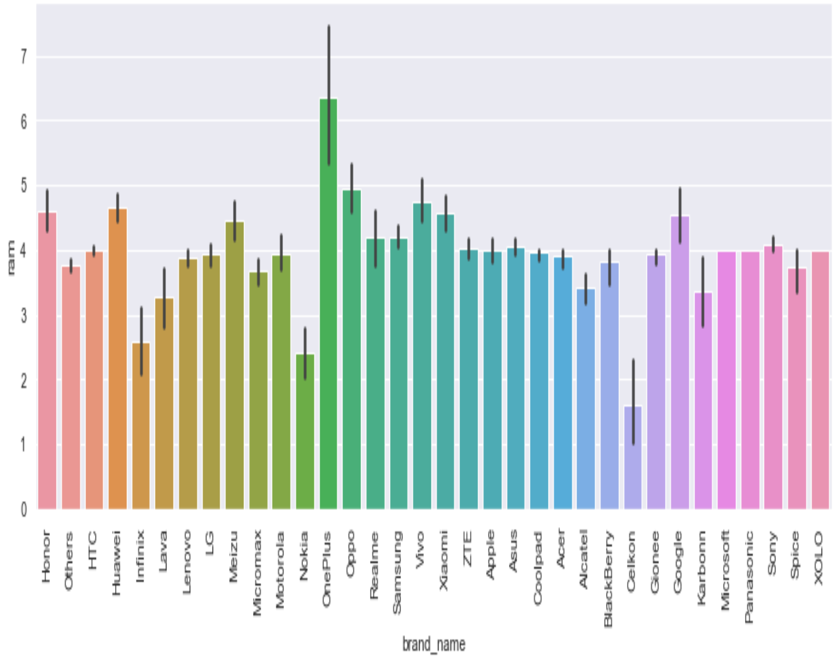
Exploratory Data Analysis (EDA): Bivarate Analysis



	screen_size	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	days_used	new_price	used_price	used_price_log	new_price_log
screen_size	1.00	0.15	0.27	0.07	0.27	0.81	0.83	-0.29	0.34	0.53	0.61	0.46
main_camera_mp	0.15	1.00	0.43	0.02	0.26	0.25	-0.09	-0.14	0.36	0.46	0.59	0.54
selfie_camera_mp	0.27	0.43	1.00	0.30	0.48	0.37	-0.00	-0.55	0.42	0.61	0.61	0.48
int_memory	0.07	0.02	0.30	1.00	0.12	0.12	0.01	-0.24	0.37	0.38	0.19	0.20
ram	0.27	0.26	0.48	0.12	1.00	0.28	0.09	-0.28	0.49	0.53	0.52	0.53
battery	0.81	0.25	0.37	0.12	0.28	1.00	0.70	-0.37	0.37	0.55	0.61	0.47
weight	0.83	-0.09	-0.00	0.01	0.09	0.70	1.00	-0.07	0.22	0.36	0.38	0.27
days_used	-0.29	-0.14	-0.55	-0.24	-0.28	-0.37	-0.07	1.00	-0.25	-0.39	-0.36	-0.22
new_price	0.34	0.36	0.42	0.37	0.49	0.37	0.22	-0.25	1.00	0.81	0.67	0.85
used_price	0.53	0.46	0.61	0.38	0.53	0.55	0.36	-0.39	0.81	1.00	0.89	0.78
used_price_log	0.61	0.59	0.61	0.19	0.52	0.61	0.38	-0.36	0.67	0.89	1.00	0.83
new_price_log	0.46	0.54	0.48	0.20	0.53	0.47	0.27	-0.22	0.85	0.78	0.83	1.00

- There is a strong correlation between battery, weight and screen size
- There is also a correlation between the new_price and the used_price which corresponds with the new_price_log and used_price_log
- There is negative correlation between selfie_camera and days_used

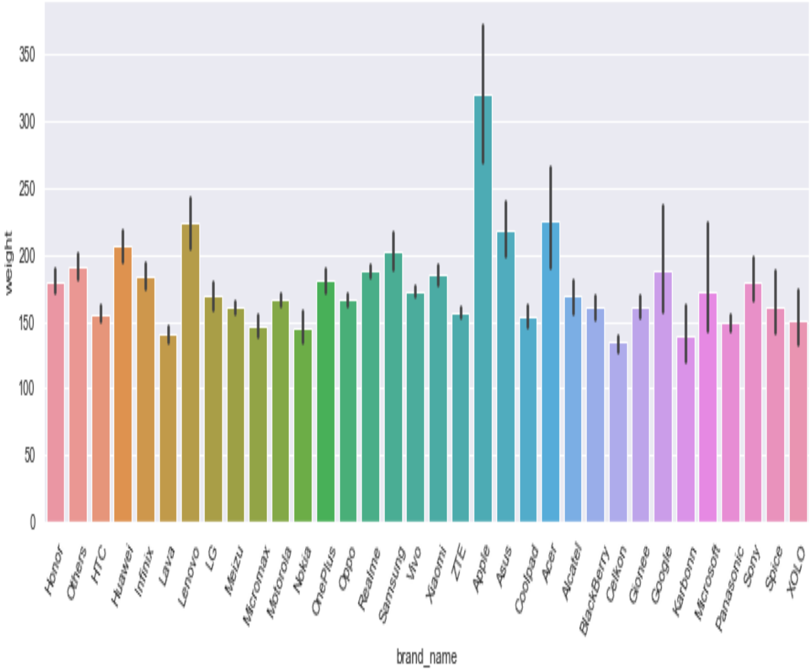
Exploratory Data Analysis (EDA): Bivarate Analysis



- OnePlus has the most RAM followed by brand OPPO and Google

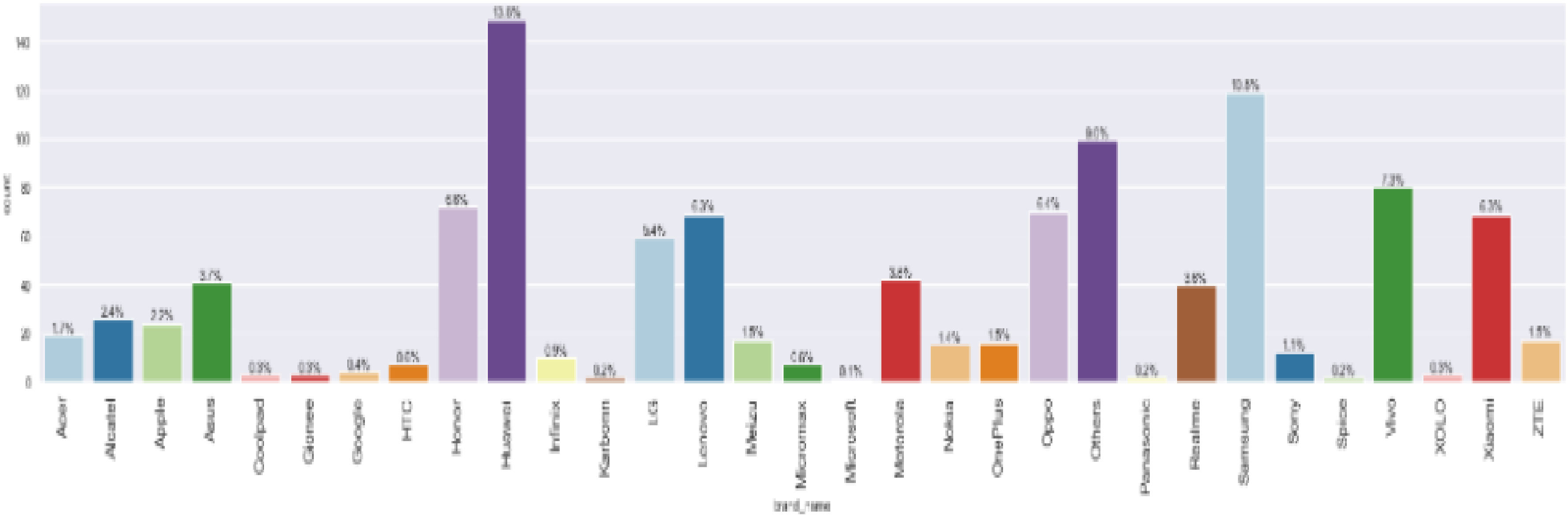
- Brand Celkon has the least RAM followed by Nokia

Exploratory Data Analysis (EDA): Bivarate Analysis



- Brand Apple has the most weight followed by Acer and Asus
- Brand Celkon has the least weight followed by Karbonn

Exploratory Data Analysis (EDA): Bivarate Analysis



- Huawei and Samsung have the largest screen followed by Vivo

Data Preprocessing: Duplicate Value Checks

- The data consists of 4 object types, 9 float 64 type and 2 int 64 type

- There are no duplicated data noted

- The missing values tend to center around main_camera_mp which on further analysis has NaN values in it

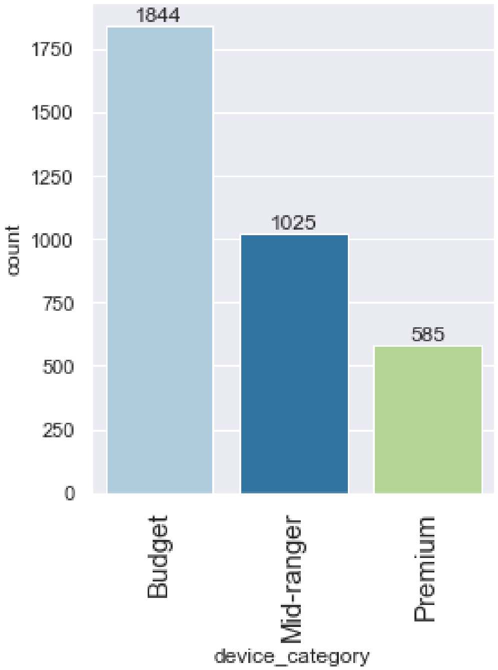
Data Preprocessing: Missing Values

- The column main_camera_mp has the most missing values at 179
- Followed by weight (7), battery (6), int_mermory(4), ram(4), selfie_camera_mp(2)
- The data is then grouped by release year and brand name and inputed with the median value

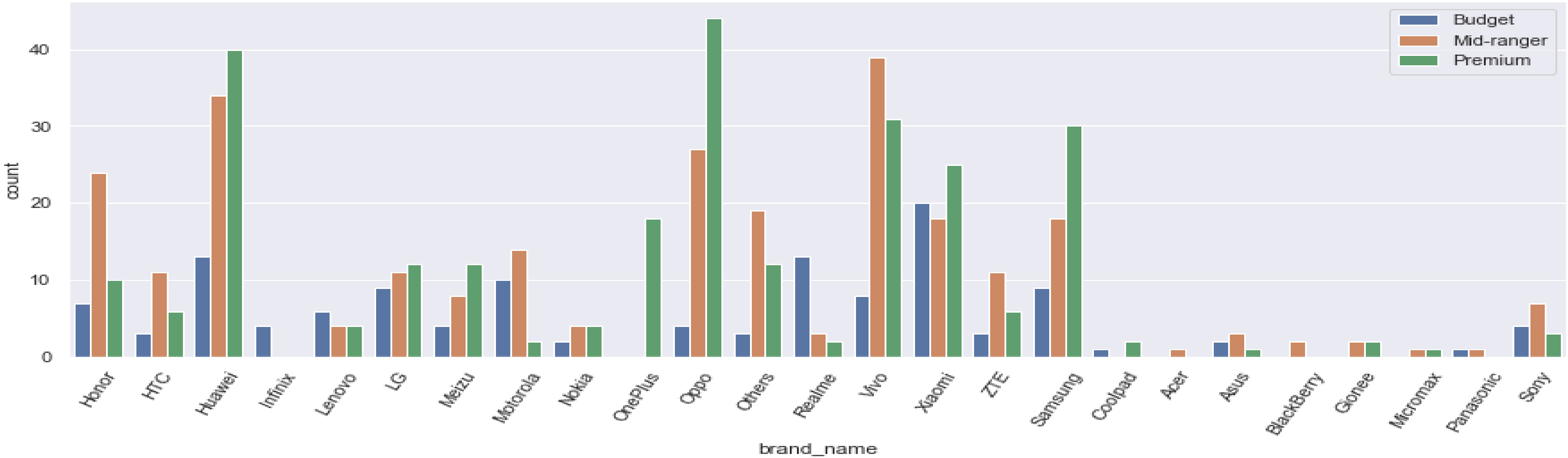
Data Preprocessing: Feature Engineering

- The device category is broken down using new_price into three main headers:

• Budget	1844
• Mid_ranger	1025
• Premium	585

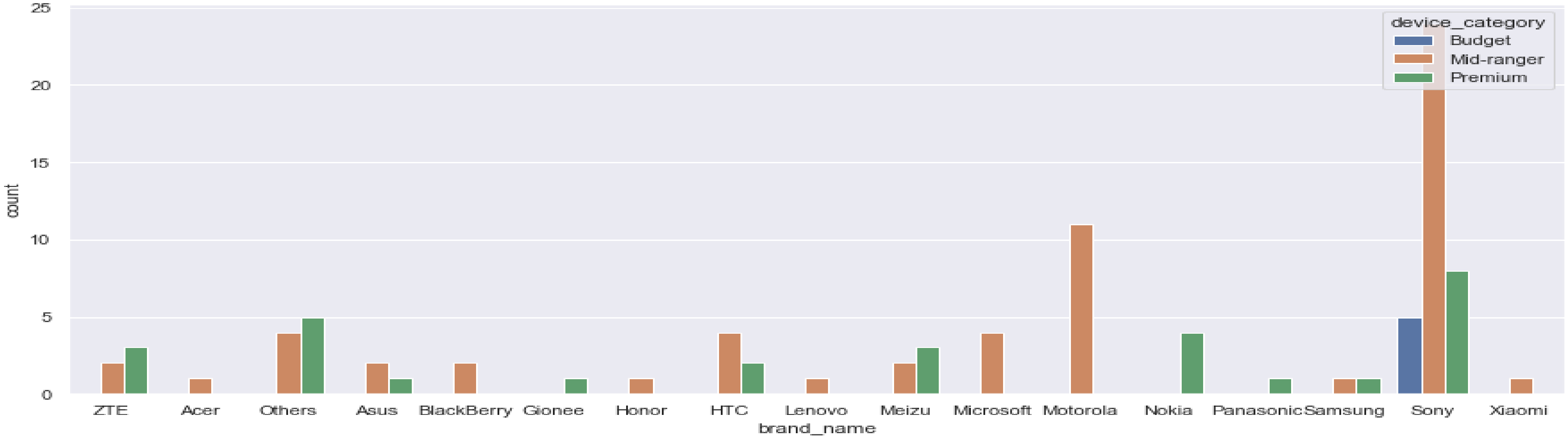


Data Preprocessing



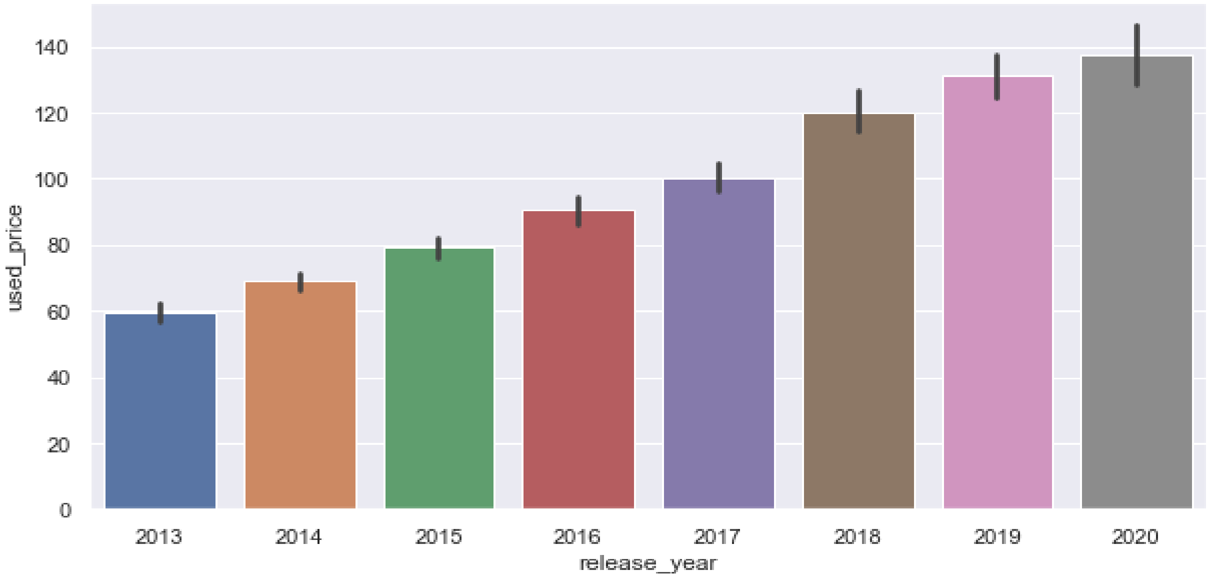
Break down of the data by brand and pricing range with Oppo, Vivo, Samsung and Huwaie having the largest numbers by brand

Data Preprocessing



Break down of the brands by rear cameras with Samsung having the most number of users with rear_camera_mp greater than 16

Data Preprocessing: Used price based on year



The year 2021 has the highest used price followed by 2019 but 2019 has prices very close to 2020 and 2018

Testing the assumptions of Linear Regression Model

- Test for Multicollinearity
  - Screen size (102.63), release year(221.57), new_price_log(174.81) have the highest VIF and show signs of high multicollinearity
  - Therefore they have high correlation with other independent variables and can be predicted from other independent variables
  - This can lead to undermining of the statistical significance of the independent variable

Testing the assumptions of Linear Regression Model

THE OLS REGRESSION RESULTS INTERPRETATIONS

- The R-squared value tells us that our model can explain 99.7% of the variance in the training set.

- The coefficients tell us how one unit change in X can affect y.

- The sign of the coefficient indicates if the relationship is positive or negative.

- In this data set, for example, an increase of 1 in screen size with a 0.031 increase in used_price_log, and a unit increase in new_price_log occurs with a 0.4115 increase in the used_price_log.

Testing the assumptions of Linear Regression Model

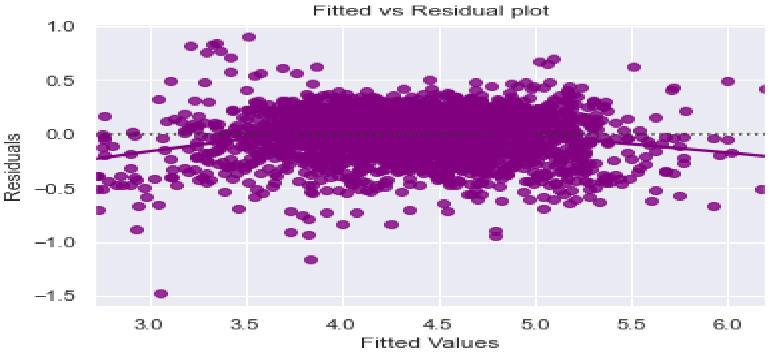
- Earlier we saw that the relationship of used_price_log with screen_size, selfie_camera_mp and is almost the same (as used_price_log increases, the variable increase and vice versa). This suggests that all the 3 factors have similar effect on used_price_log, i.e., the increase in either of the 3 increases used_price_log. Therefore, the signs of the coefficents should be the same. But we observe that it is not so. This indicates the presence of multicollinearity in our data.

- Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, we might not be able to trust the p-values to identify independent variables that are statistically significant.

- When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

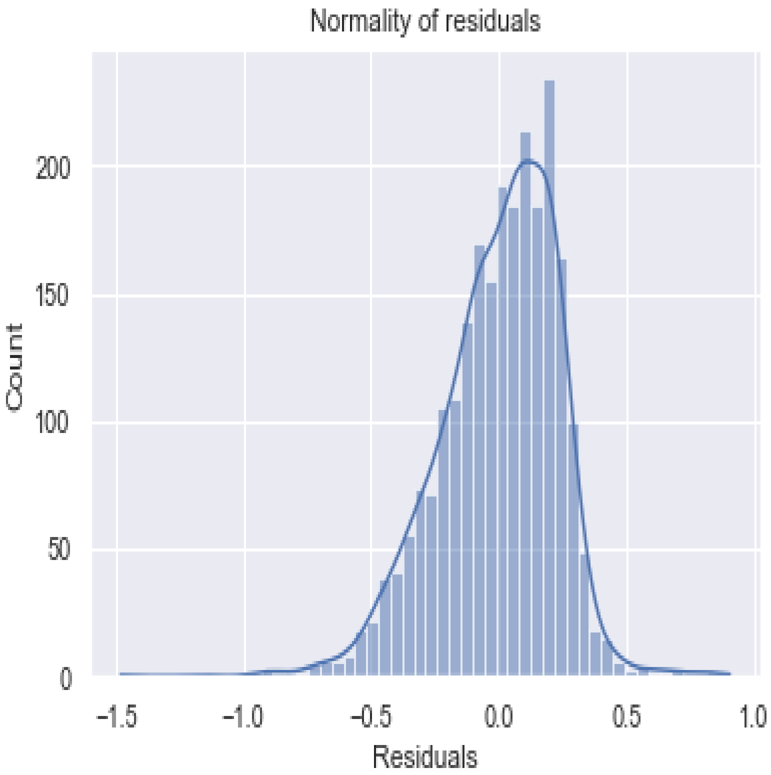
Testing the assumptions of Linear Regression Model

Test for Linearity and Independence



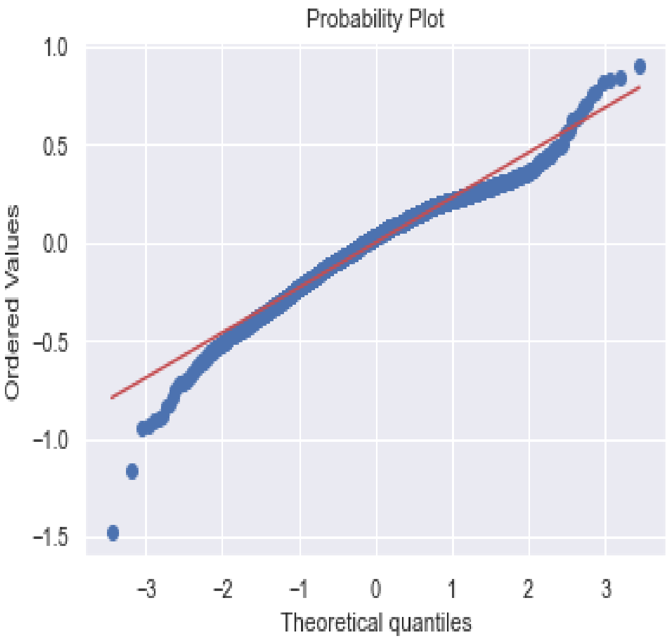
- Linearity describes a straight-line relationship between two variables, predictor variables must have a linear relation with the dependent variable.
- The plot of fitted values vs residuals, we observe a pattern in the values, hence we can say the model is linear
- follow any pattern (the curve is a straight line), then we say the model is linear otherwise model is showing signs of non-linearity.

Testing the assumptions of Linear Regression Model



- Error terms/residuals should be normally distributed
- The plot mirrors a normal distribution with residual terms normally distributed

Testing the assumptions of Linear Regression Model



Test for Normality

The normal probability plot of residuals follows an approximate straight line

Shapiro-Wilk Test

- Null hypothesis- Data is normally distributed
- Alternate Hypothesis- Data is not normally distributed

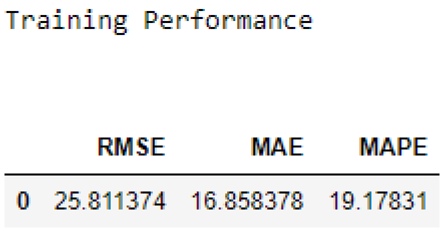
The p-value is 1.088-21 is less than 0.05 the residuals are not normal as per the shapiro test

Testing the assumptions of Linear Regression Model

Test for Homoscedasticity

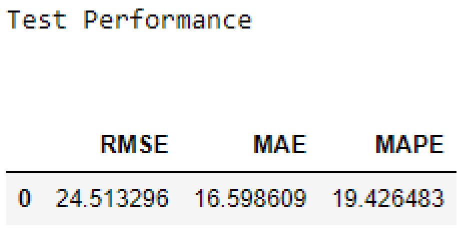
- Test was done using the goldfeldquandt test
  - Null Hypothesis: Residuals are homoscedastic
  - Alternate Hypothesis: Residuals are hetrosedasticity
  - Using the statsmodels we have a p-value of 0.205398
  - Since P-value is less than 0.5 the residuals hetrosdeasticity

Testing the assumptions of Linear Regression Model



Training Performance

	RMSE	MAE	MAPE
0	25.811374	16.858378	19.17831



Test Performance

	RMSE	MAE	MAPE
0	24.513296	16.598609	19.426483

- RMSE values on the train and test sets are also comparable.
- The train and test RMSE and MAE are comparable, this shows that the model is not overfitting.
- MAE indicates that our current model is able to predict used_car_log ratings within a mean error of 0.16 on the test set.
- MAPE of 19.42 on the test data means that we are able to predict within 19% of the rating.

Model Performance Evaluation

Summary of Final Model

- R-squared of the model is 0.997 and adjusted R-squared is 0.997, which shows that the model is able to explain ~99.7% variance in the data. This is quite good.
- A unit increase in the screen size will result in a 0.0311 unit increase in the used_price_log, all other variables remaining constant.
- The brand_name_nokia will have a higher impact on used_price_log than the brand_name_Samsung and brand_name_Sony, all other variables remaining constant.
- We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.
- MAE indicates that our current model is able to predict used price log within a mean error of 16.85 units on the test data.
- Hence, we can conclude the model "ols" is good for prediction as well as inference purposes.

Business Insights and Recommendations

- Actionable insights based on the results of the analysis
  - The used_price_log average price $92 with a maximum price of $749.50 and minimum of $4.65
  - 75% of the new price are below 291.15 while only 25% of the new price are below $20
  - Based on the data Samsung and Huwaei are the most popular brands with both also having the largest screens at 149 and 119
- Factors that appear to influence the used tablet or phone are:
  - Brand: The popular brands are Samusung and Huwaie and a mix of brands which on their own account for 14.5%
  - Operating System: The preferred Operating System is the Android system which accounts for 93.1%
  - The 4g network right now is the most popular accounting for 67.6%
  - The most used release year on the used products is 2014 about 18% followed by 2013 16.5% and 2019 12.9%

Business Insights and Recommendations

- ReCell should tap into the potential of used and and refurbished devices especially of the brand Huawei and Samsung and ensure their integration on the 4g network.

- ReCell should provide promotional offers on the refurbished items given that the new prices on those items are still very high and thus they still have a market for the items.

- .Screen size is a very important factor to gauge customer satisfaction. The company should investigate the reason behind the importance of screen size.

- The release year of the phones is also very important with the year 2014 being the most sought after release year, however given the ever changing climate with respect to technology they should focus more on 2019 which is closer to the year 2021 thus ensuring some form of recent technology in the phones sold

THANK YOU