ReCell: Used Phone Pricing

CONTENT

- Business Overview
- Problem Statement
- Data Overview
- Exploratory Data Analysis
- Model Performance Summary
- Business Insights & Recommendations

Business Overview

The used and refurbished phone market has grown considerably over the past decade, and a new International Data Corporation forecast predicts that the used phone market would be worth $52.7bn by 2023.

ReCell is an online store that sells used and refurbished smartphones.

Objective

The rising potential of this comparatively under-the-radar market fuels the need for an ML-based solution to develop a dynamic pricing strategy for used and refurbished smartphones. ReCell, a startup aiming to tap the potential in this market.

Analyze the data provided and build a linear regression model to predict the price of a used phone and identify factors that significantly influence it.

Data Overview: Data Fields

brand_name: Name of manufacturing brand	os: OS on which the phone 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 phone battery in mAh	weight: Weight of the phone in grams	release_year: Year when the phone model was released
days_used: Number of days the used/refurbished phone has been used	new_price: Price of a new phone of the same model in euros	used_price: Price of the used/refurbished phone in euros

Data Overview: Sypnosis

There are 3,751 observations and 15 features in the original data set

The column types includes 2 integer, 9 float and 4 object/string

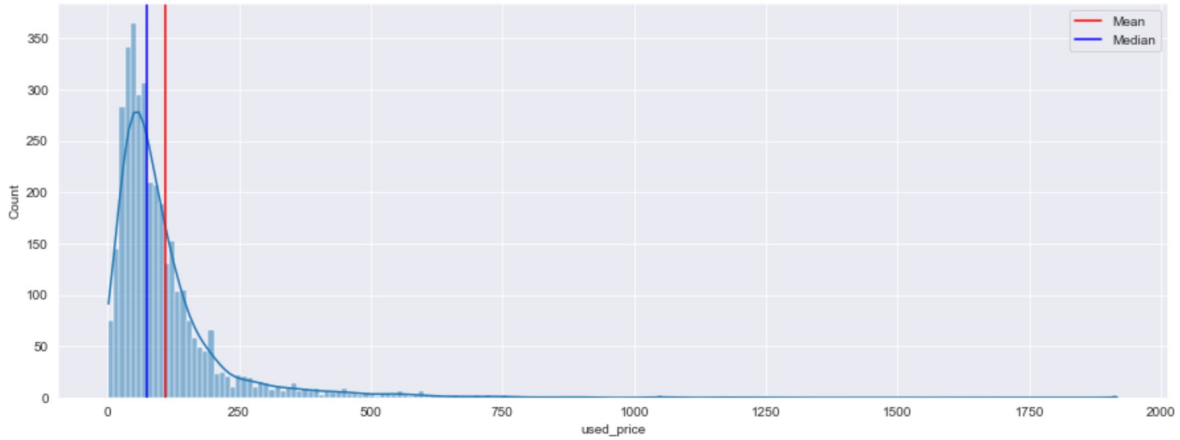
There are 53 unique brand names with an additional others type

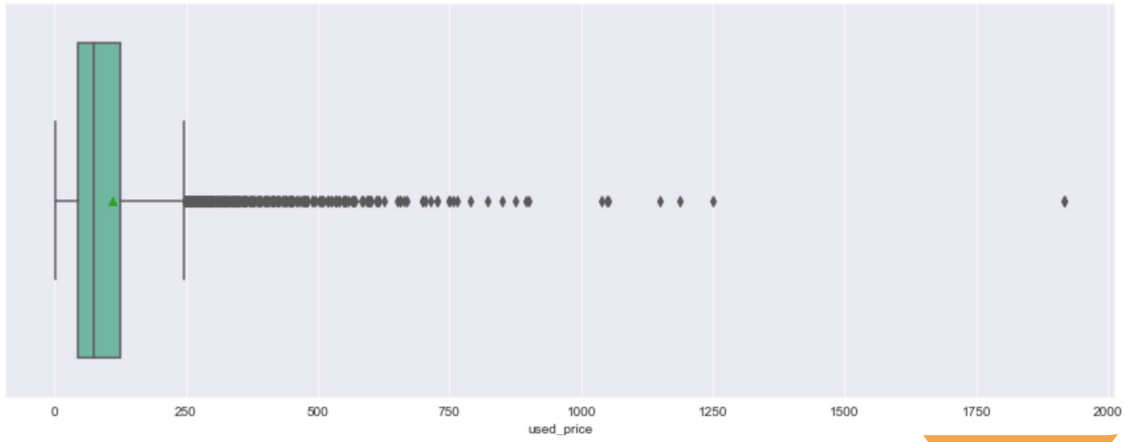
OS has 4 unique values

4g and 5g are yes or no responses

EDA – Used price

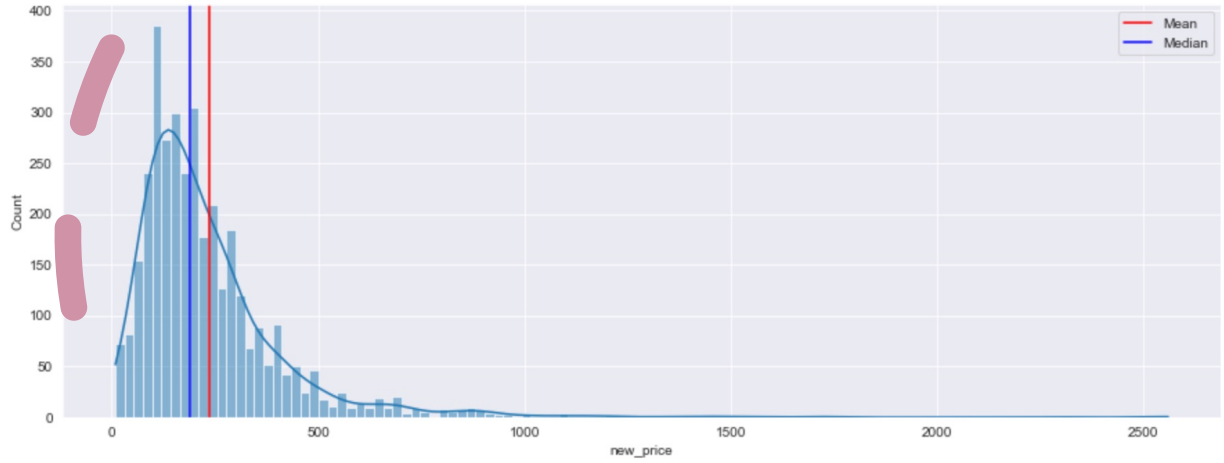
- The used_price is highly right skewed
- Very long tail on the right
- Used price goes up to 2,000, a lot of observations lie between 250-1,250



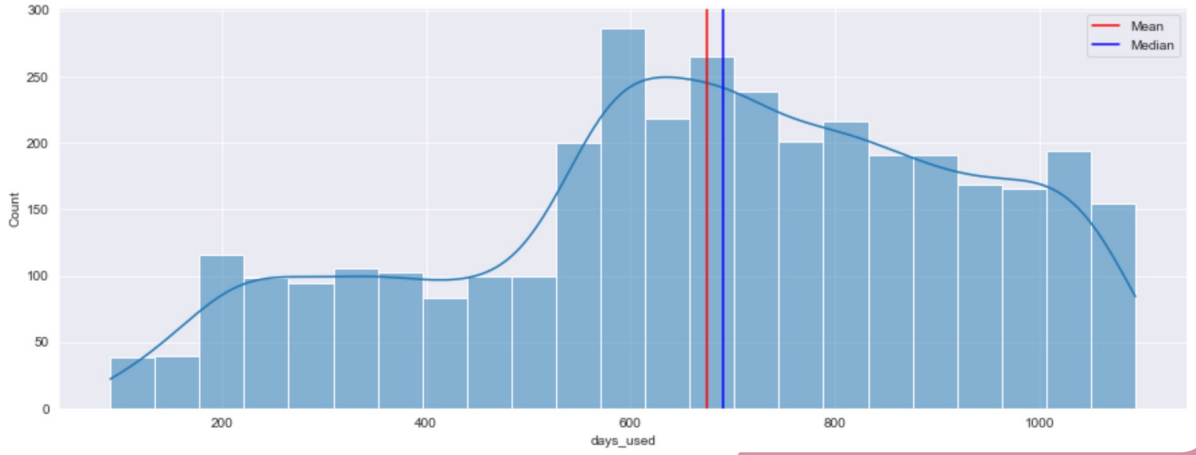


EDA – New Price & Days Used

- New price is right skewed
- New price of the phone range between $0-$2600

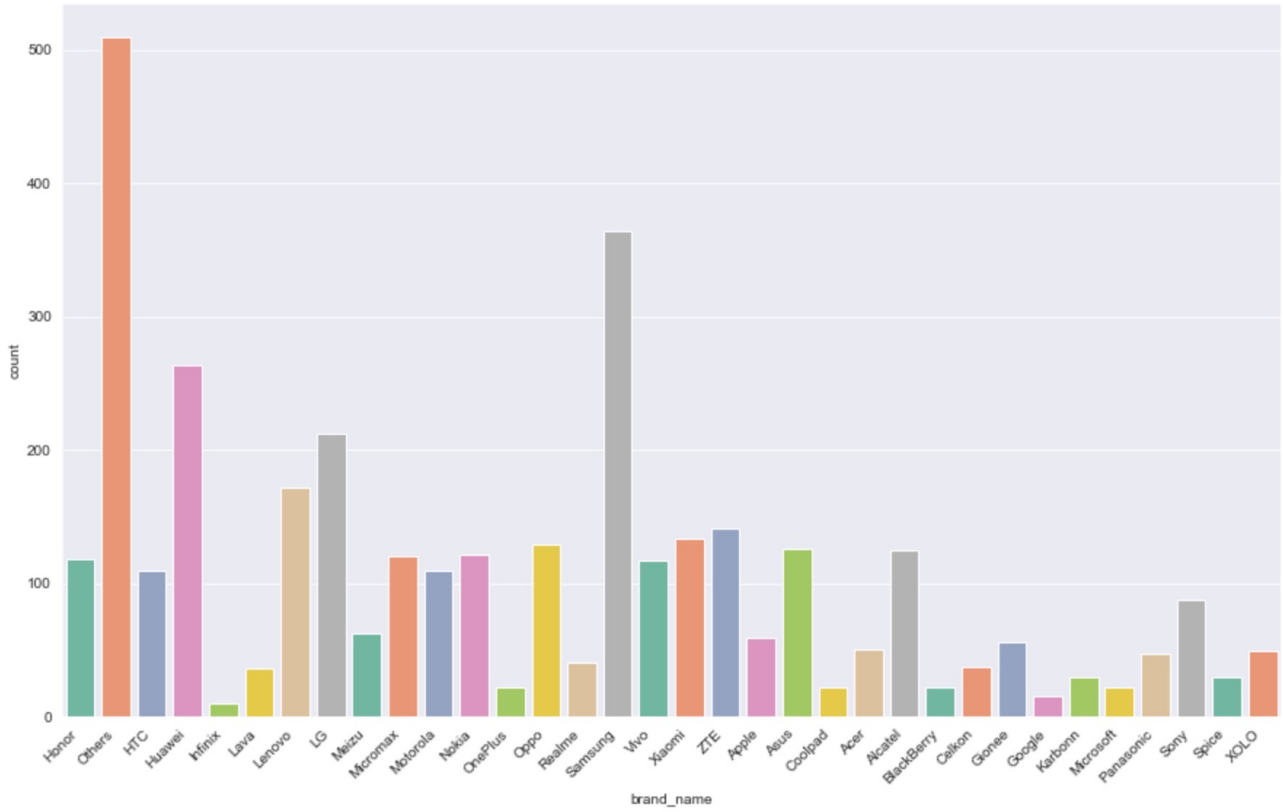


- Days used are more uniformly distributed
- Days used range between 0-1000+ days



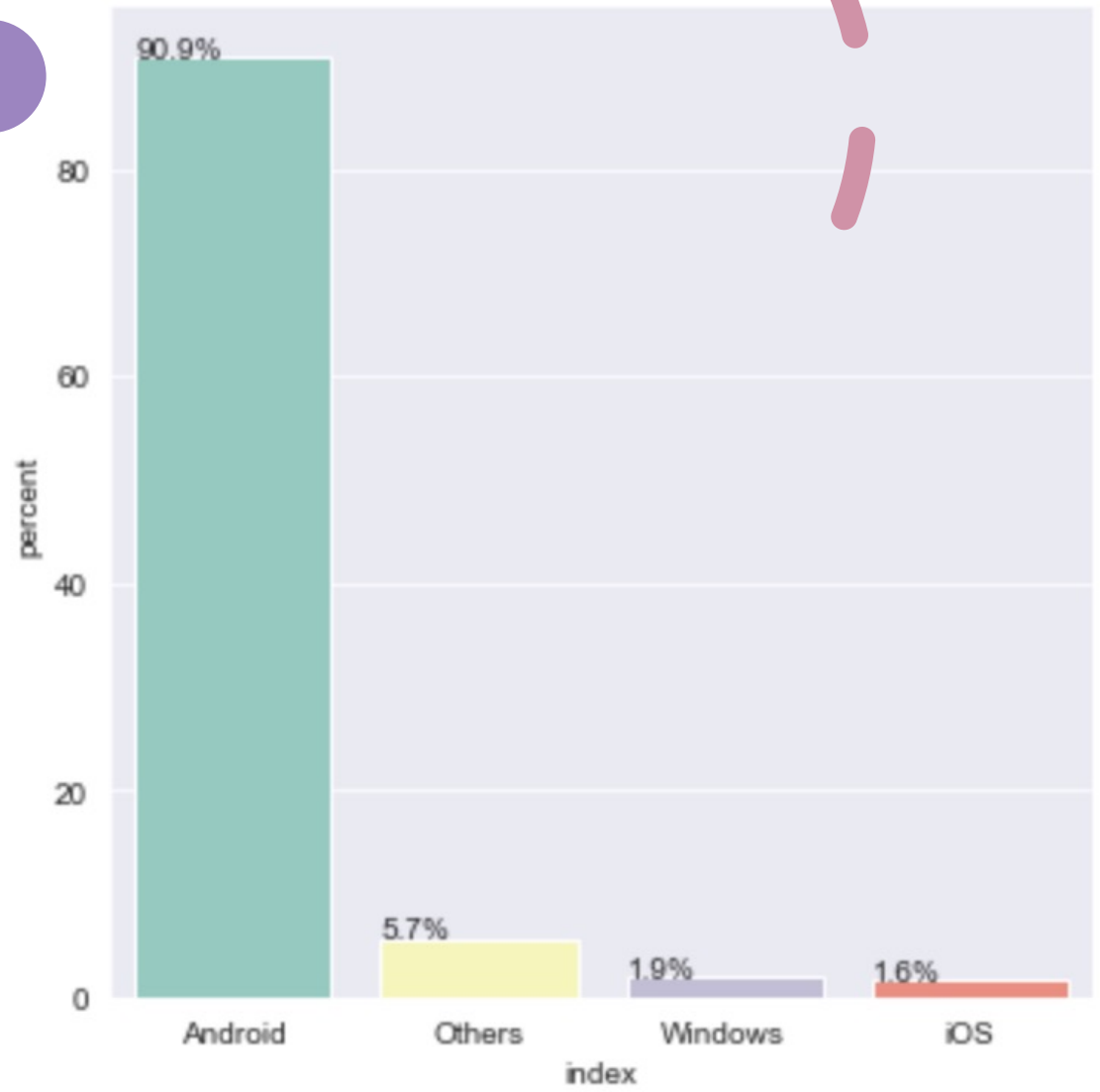
EDA – Phones by Brand

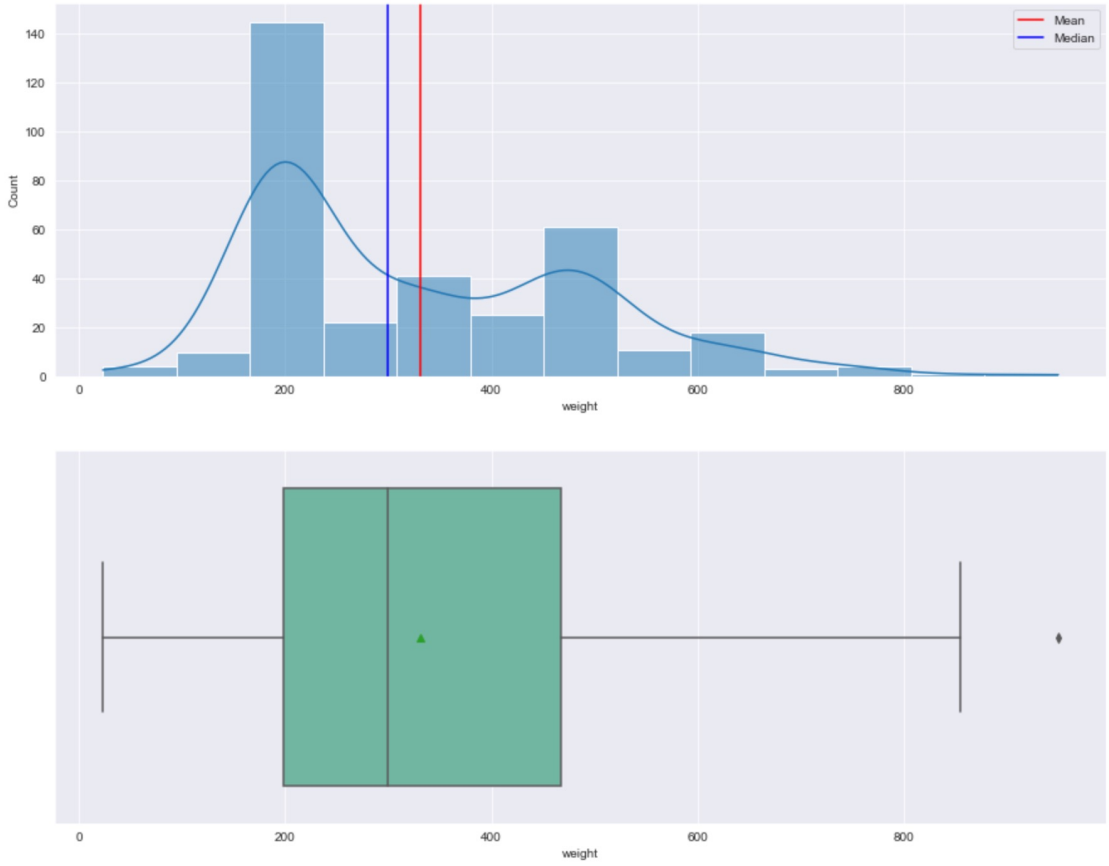
- Most of the phones are made by Samsung
- There are also lot of unidentified(Other) brand phones
- High number of Huawei, LG, ZTE, Xiaomi, Lenovo, Asus and Alcatel phones



EDA – OS Distribution

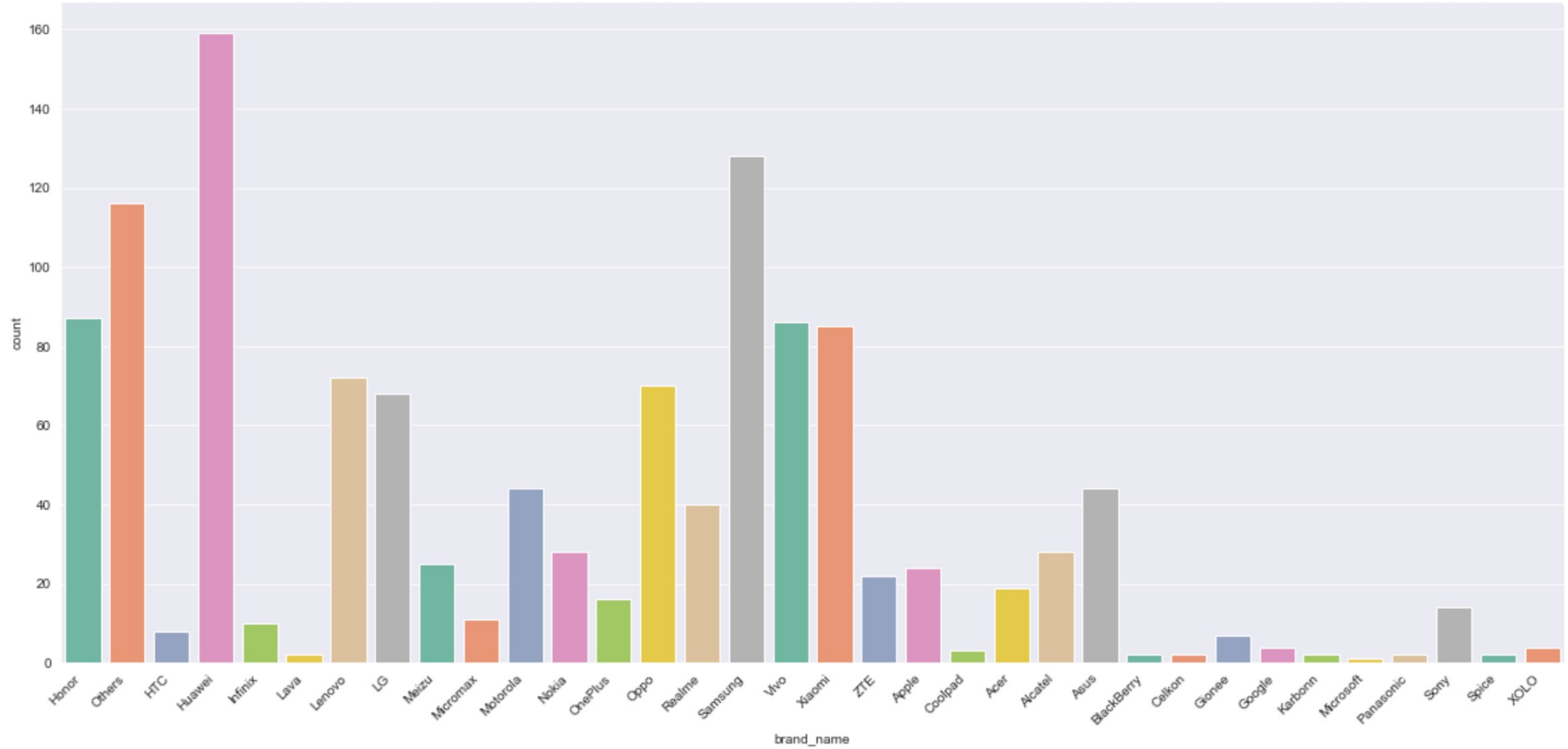
- 90% Android phones for re-sale
- 1.6% of iOS available





EDA – Weight of Large Battery Phones

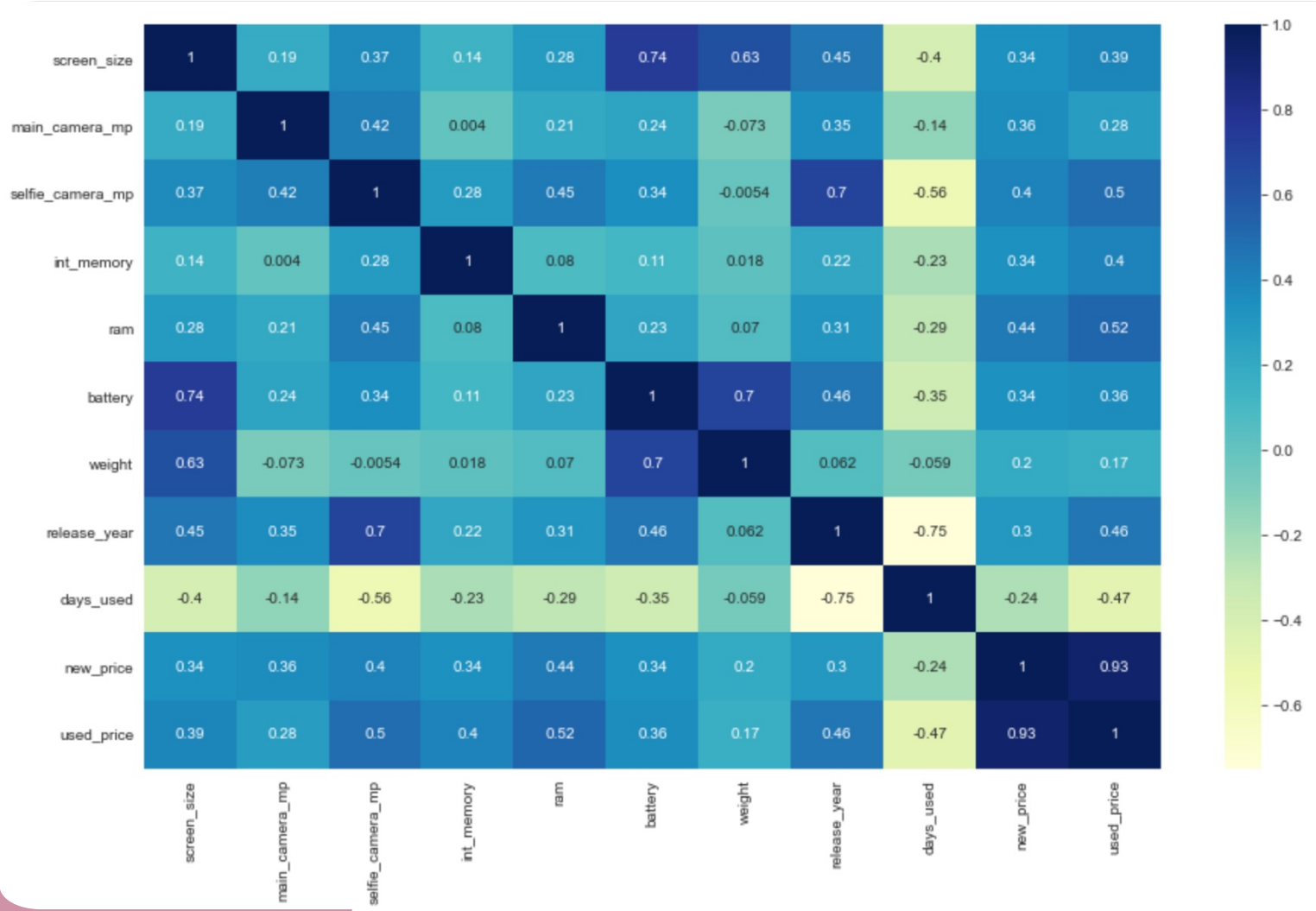
- Most of the phone weigh 200 and above
- There aren't outliers
- The large battery phones tend to fall between normal range of 200-850 weight



EDA – Phone with Screen size greater than 6”

- Huawei has the largest screen phones
- Samsung has large screen phones as well

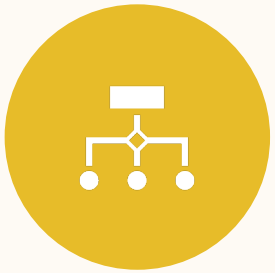
EDA – Correlation Matrix



- Used price is highly positively correlated with the new price
- Used price is negatively correlated with the number of days used
- Screen size and battery have positive relation
- Recent release year's phone have higher selfie camera MP
- Larger screen size have higher weight

Data Processing

MISSING VALUE TREATMENT



VARIABLE TRANSFORMATION (LOG, STANDARDIZING, SCALING)

OUTLIER DETECTION & TREATMENT

FEATURE TRANSFORMATION

Missing Values

Missing values by columns

brand_name	0
os	0
screen_size	0
4g	0
5g	0
main_camera_mp	180
selfie_camera_mp	2
int_memory	10
ram	10
battery	6
weight	7
release_year	0
days_used	0
new_price	0
used_price	0

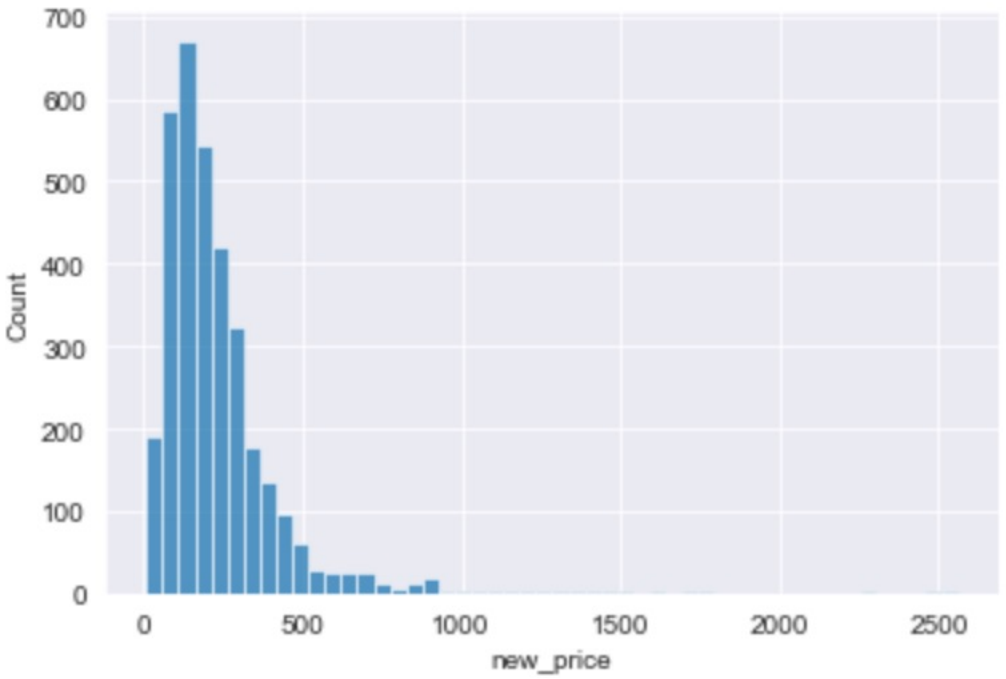
dtype: int64

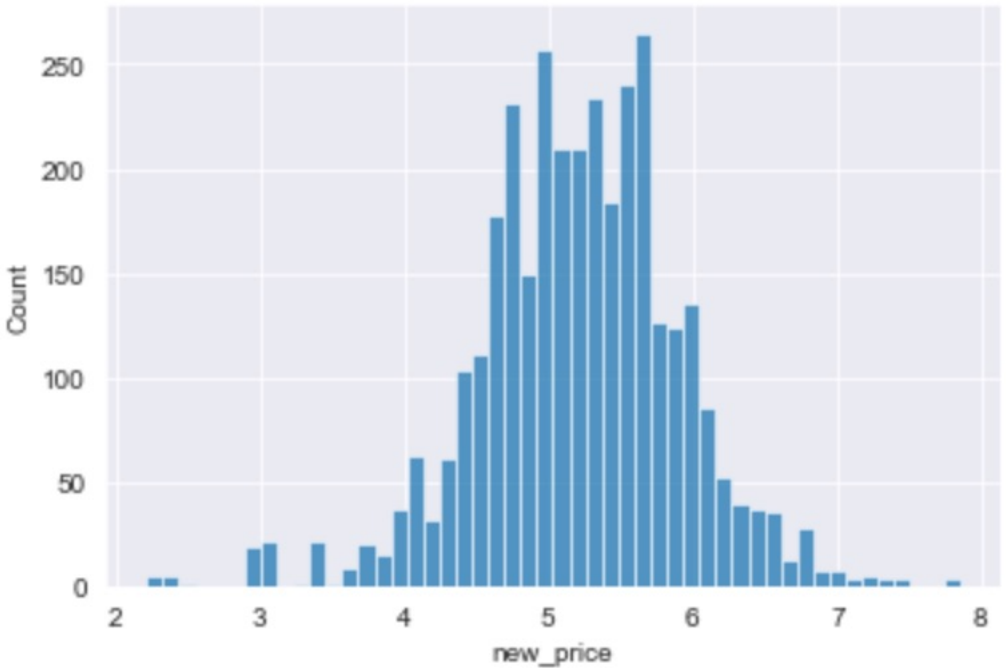
Missing values by observations

0	3368
1	193
2	8
3	2

dtype: int64

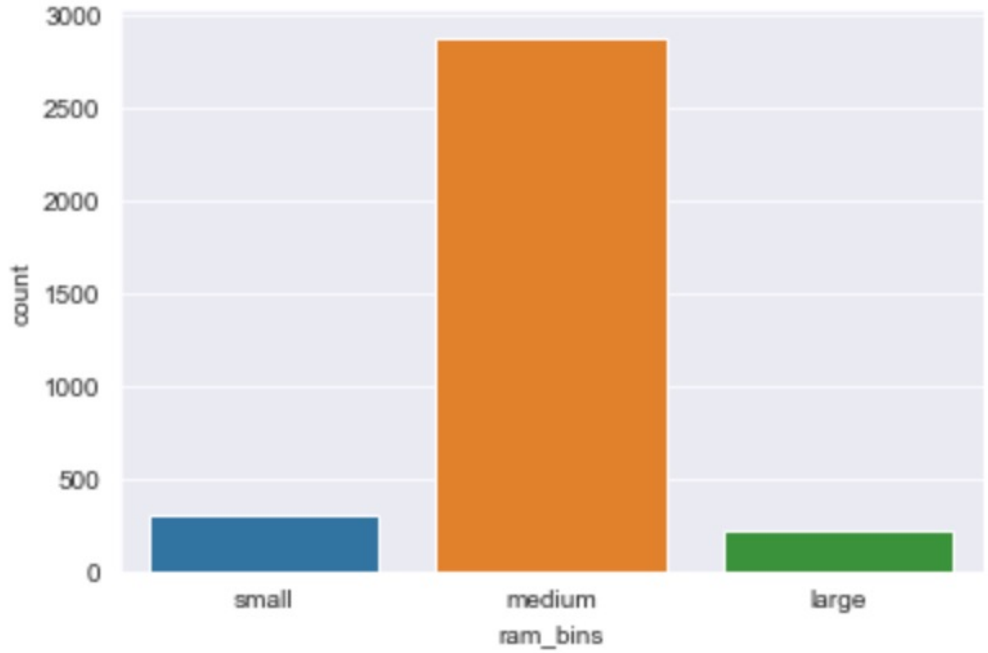
New Price Log Transformation





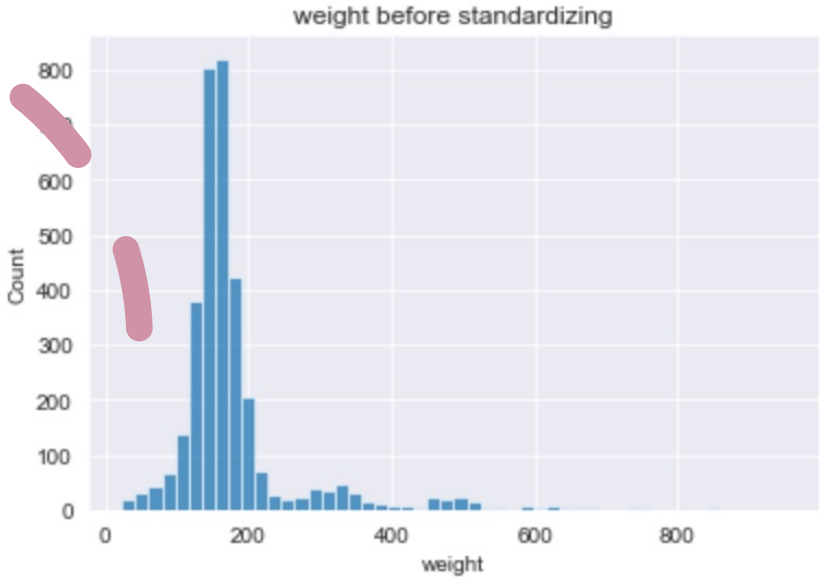
RAM - Binning

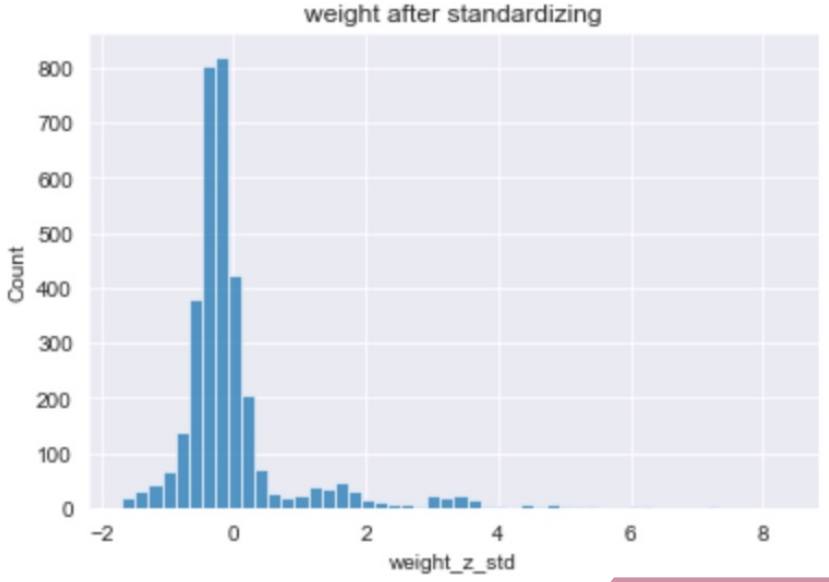
Grouped RAM into “small”, “medium” and “large”



Standardizing Variables

- Screen size
- Main Camera
- Selfie Camera
- Init Memory
- Battery
- Weight





Linear Model Performance Summary

OLS Regression Results

Dep. Variable:	used_price_log	R-squared:	0.987
Model:	OLS	Adj. R-squared:	0.987
Method:	Least Squares	F-statistic:	3.032e+04
Date:	Thu, 19 Aug 2021	Prob (F-statistic):	0.00
Time:	13:10:41	Log-Likelihood:	2380.9
No. Observations:	2373	AIC:	-4748.
Df Residuals:	2366	BIC:	-4707.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.5180	0.017	144.182	0.000	2.484	2.552
new_price_log	0.6781	0.002	324.970	0.000	0.674	0.682
days_used_z_std	1.7815	0.014	124.808	0.000	1.753	1.809
brand_name_BlackBerry	0.0555	0.026	2.113	0.035	0.004	0.107
os_Others	-0.0734	0.009	-8.530	0.000	-0.090	-0.057
5g_yes	-0.0639	0.013	-4.755	0.000	-0.090	-0.038
ram_bins_medium	-0.0109	0.006	-1.716	0.086	-0.023	0.002

Omnibus:	776.001	Durbin-Watson:	1.958
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5628.372
Skew:	-1.349	Prob(JB):	0.00
Kurtosis:	10.046	Cond. No.	24.1

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Training and Testing Data Performance

- The model explains 98% of data indicating a good model
- The Adj R2 score for training and testing are close indicating no over fitting or underfitting
- MAE, MSE and RMSE lower and comparable

	MAE	MSE	RMSE	MAPE	R2	Adj. R2
Training	0.070	0.008	0.089	1.74	0.98	0.98
Testing	0.070	0.0079	0.089	1.75	0.98	0.98

Linear Model Assumptions

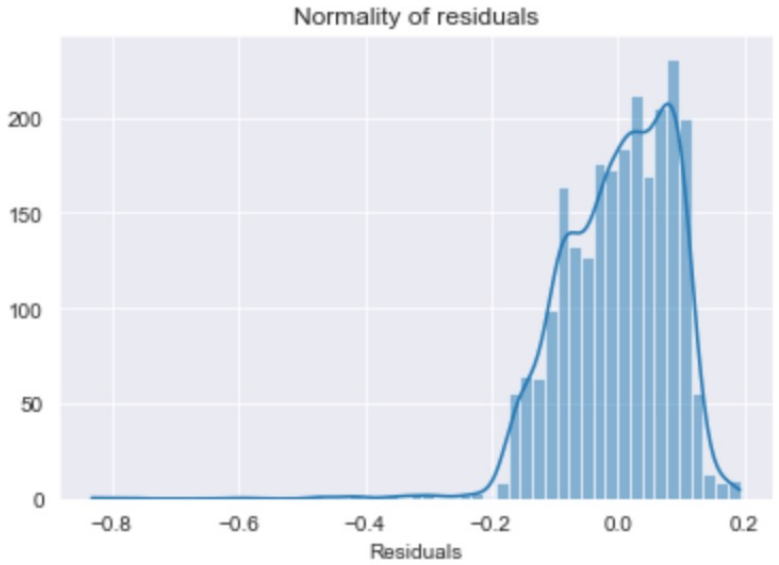
Multicollinearity – Satisfied, removed the following features

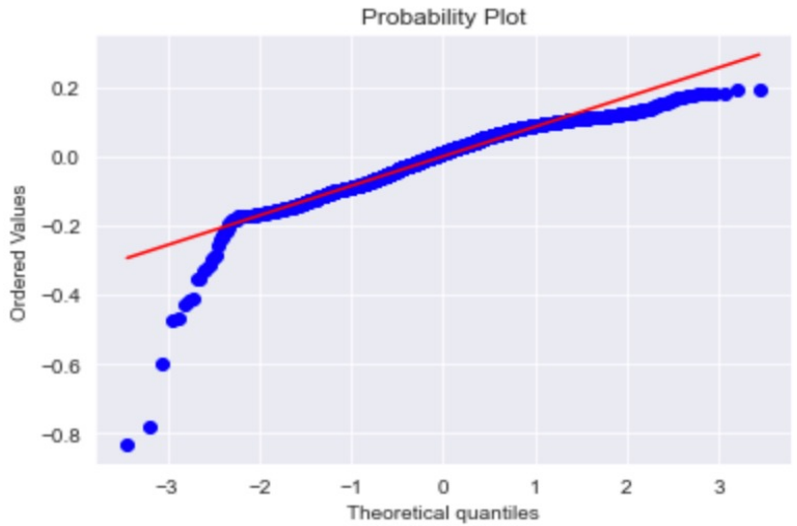
- brand_name_Apple
- brand_name_Huawei
- brand_name_Others
- os_iOS
- brand_name_Samsung

To improve model removed variables with p-value > 0.05

Linear Model Assumptions

- Normality






```
ShapiroResult(statistic=0.9268954396247864, pvalue=1.5487840429958254e-32)
```

p-value is <0.05, assumption is satisfied

Linear Model Assumptions

- Heteroscedasticity

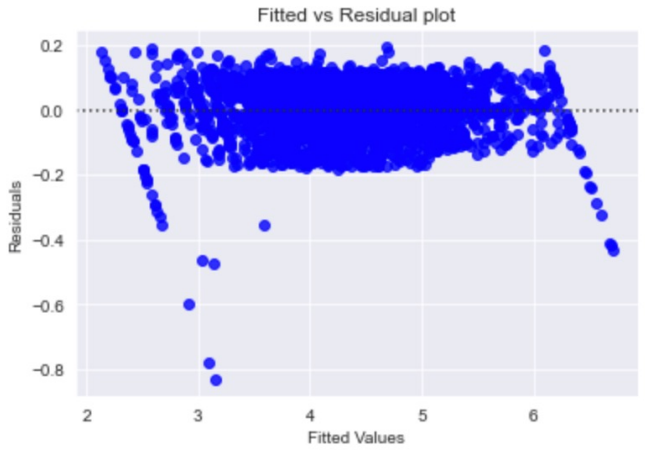
Goldfeld-Quandt Test

```
[('F statistic', 0.9545592180950387), ('p-value', 0.7876733944450137)]
```

p-value > 0.05, the assumption is satisfied

- Linearity

There's some linearity in residuals but most of it is non-linear



Predictions

Business Insights & Recommendations

- The features that have effect on the used price are the New price, Number of days used, whether phone is a BlackBerry, Other OS, 5g and Medium RAM
- New price has positive effect, higher priced phone have higher used price
- Brand OS other than iOS, Android and Windows have negative used price effect
- Smaller the RAM reduces used price for the phone