Analysis of data and building of a linear regression model to predict the price of a used phones

By Enoch Adogla

10-23-2021

Overview

➢ This data was collected by ReCell a startup company aiming to leverage the potential in the used phone market

➢ Used phone sales used to be limited to a few isolated online portals

➢ The market has however, seen considerably growth over the last few years with major companies such as Verizon and amazon joining in the used phone retail

➢ The market is expected to see an annual growths of 13% and reach over $ 50 billion by 2023

➢ This analysis will look at factors that affect used phone prices and how to leverage those factors to increase revenue

Data-dictionary

- **Data Dictionary**
- 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

Overview of data

➢ In all 3571 used_phone sales were covered

➢ 33 phones brands and "others" were covered in the data including

➢ 15 attributes of the phone sales were covered, including the price of the phone

➢ There were few missing data, mostly the main_camera_mp

➢ The missing data were replace with the median

➢ No duplicates were observed

➢ The mean used_price was ~109, while the mean new_price was ~237

➢ Over 90% of the phone were android phones

Phone types
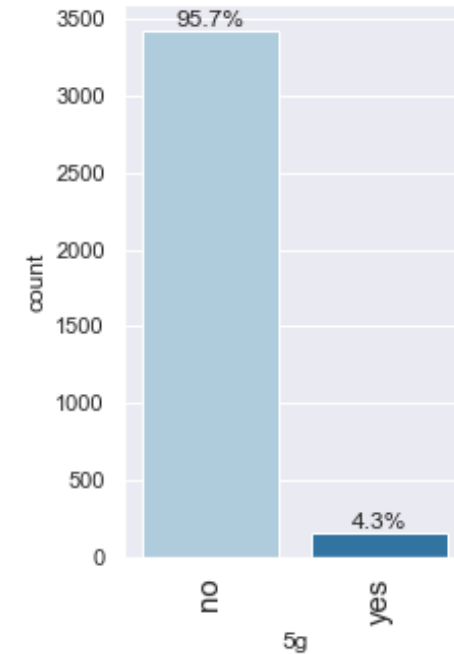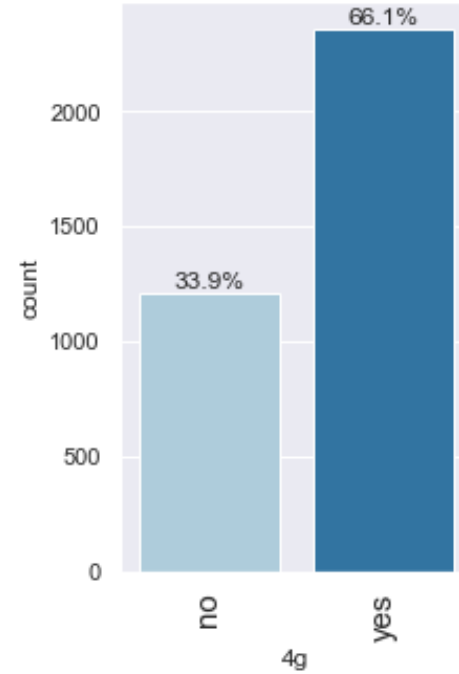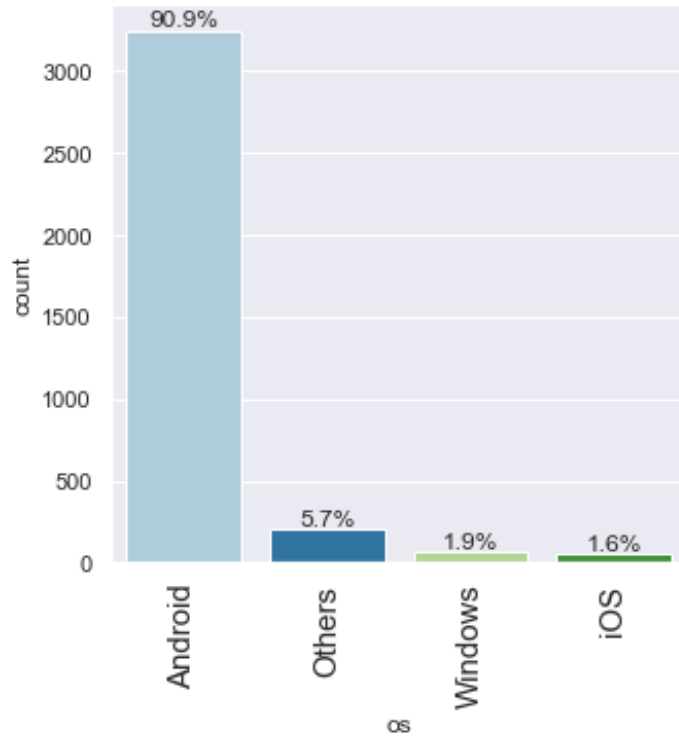
- ➤ The Os of 90.9% of the phones were android

- ➤ 4g availability was 66.1%
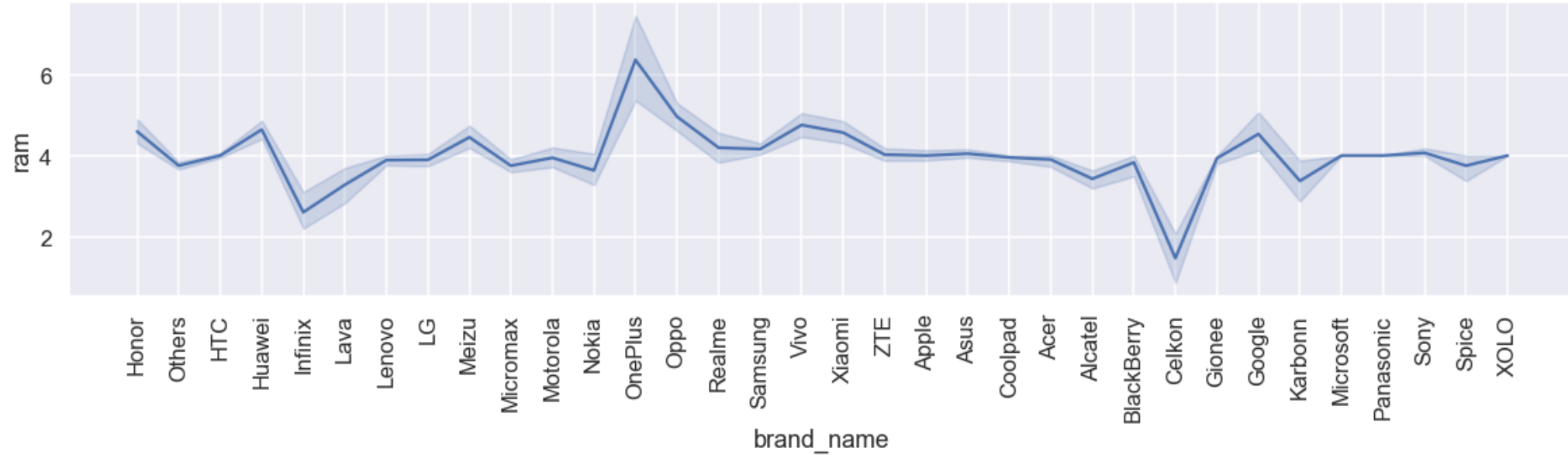
- ➤ 5g availability was 4.3%

Brand names
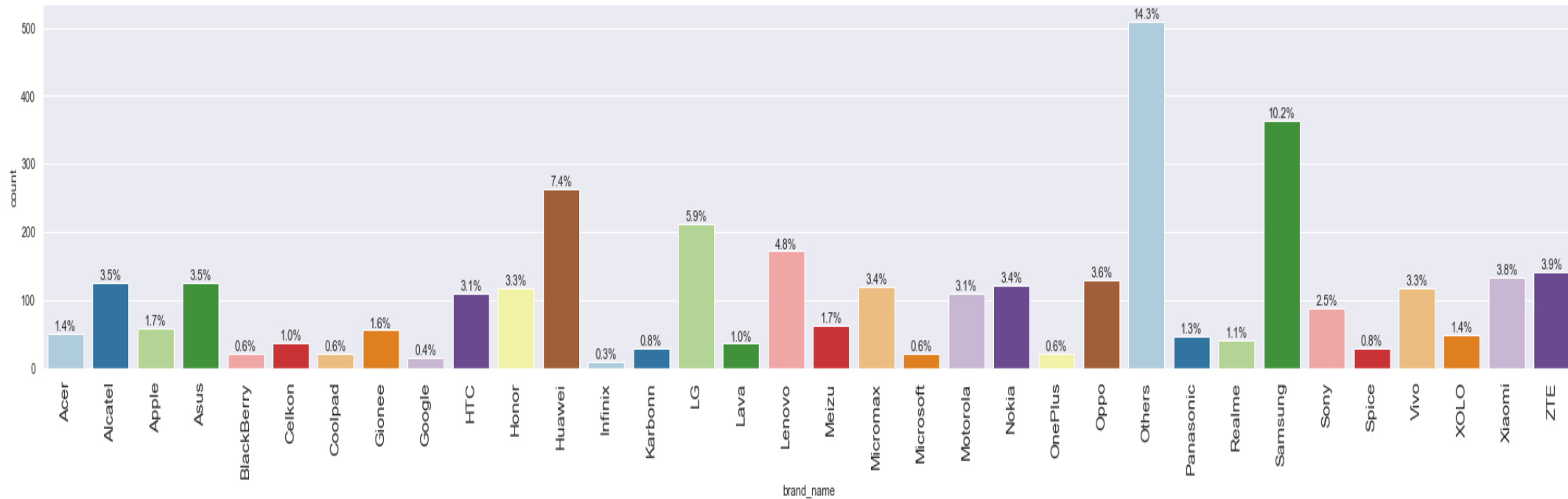
➢ There are 32 brand and 'Others'.

➢ Others make up 14.3% of the phones,

➢ The most popular brands are

➢ Samsung 10.2% and Huawei 7.4%.

➢ The least popular brand is Infinix 0.3%

Distribution of used phone price

The used phone price is right skewed

Variation of ram with brand_name

The brands with the top 5 highest rams

Onplus ~ 6.4

Oppo   ~ 5.0
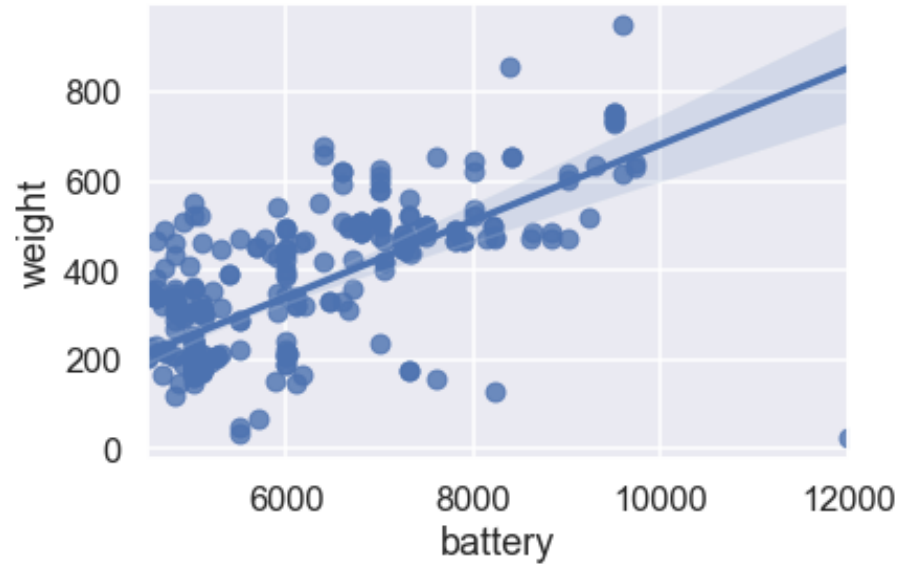
Vivo   ~ 4.8

Huawei ~ 4.7

Honor   ~ 4.6

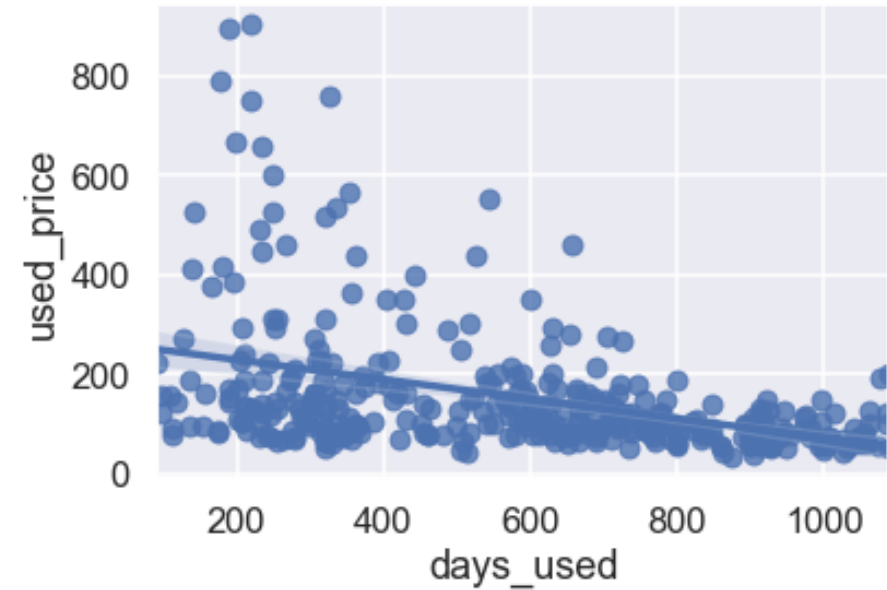Celkon has the lowest ram ~ 1.5 of all brand names

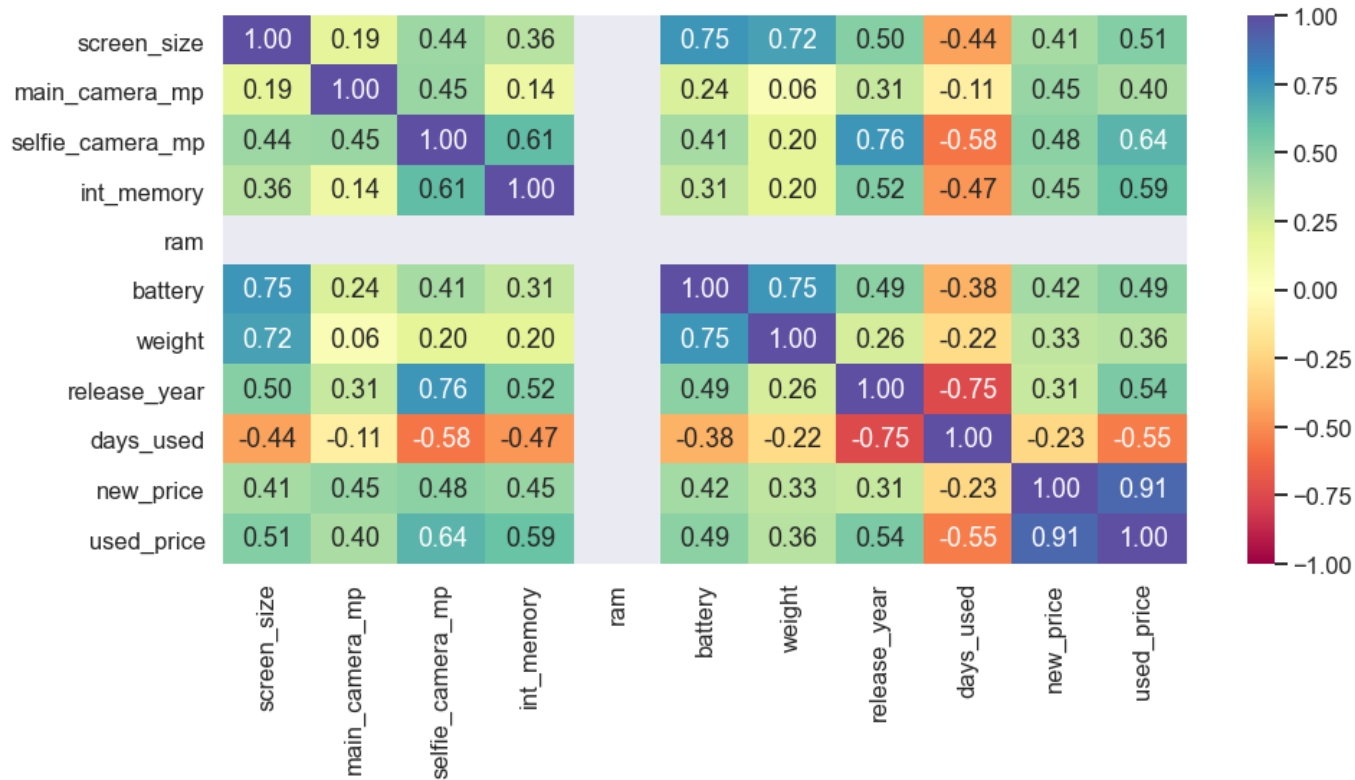# Relationship between some attributes

Positive correlation

Weight of phone and battery energy capacity in mAh

Negative correlation

use_price and days_used
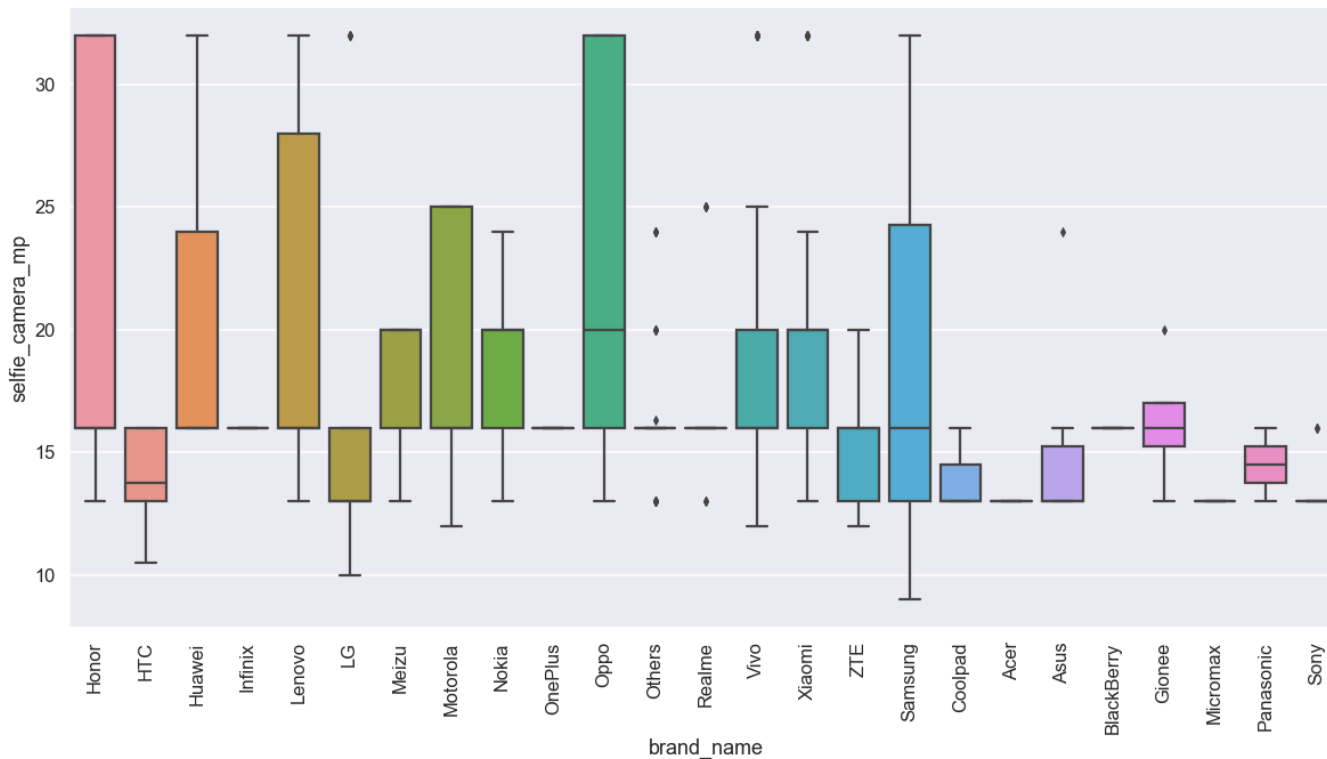
Correlation map of attributes after outlier treatment

Attributes that show high positive correlation to used_price:

➢ new_price (0.91),

➢ selfie_camera_mp (0.64),

➢ release_year (0.54),

➢ int_memeory(0.59)

➢ selfi_camera_mp (0.5)

**Attribute that showed negative correlation to used_price:**

➢ days_used

# Distribution of budget phones offering greater than 8MP selfie_camera_mp

Key Insights

➢ Oppo has the the highest average selfie_camera_mp and it's also right skewed.

➢ The selfie_camera_mp are skewed for most of the brands

➢ Only Panasonic has a uniform distribution of selfie_camera_mp

➢ Samsung is right-skewed Huawei is right-skewed with outliers

Fitter vs Residual plot

Fitted vs Residual plot
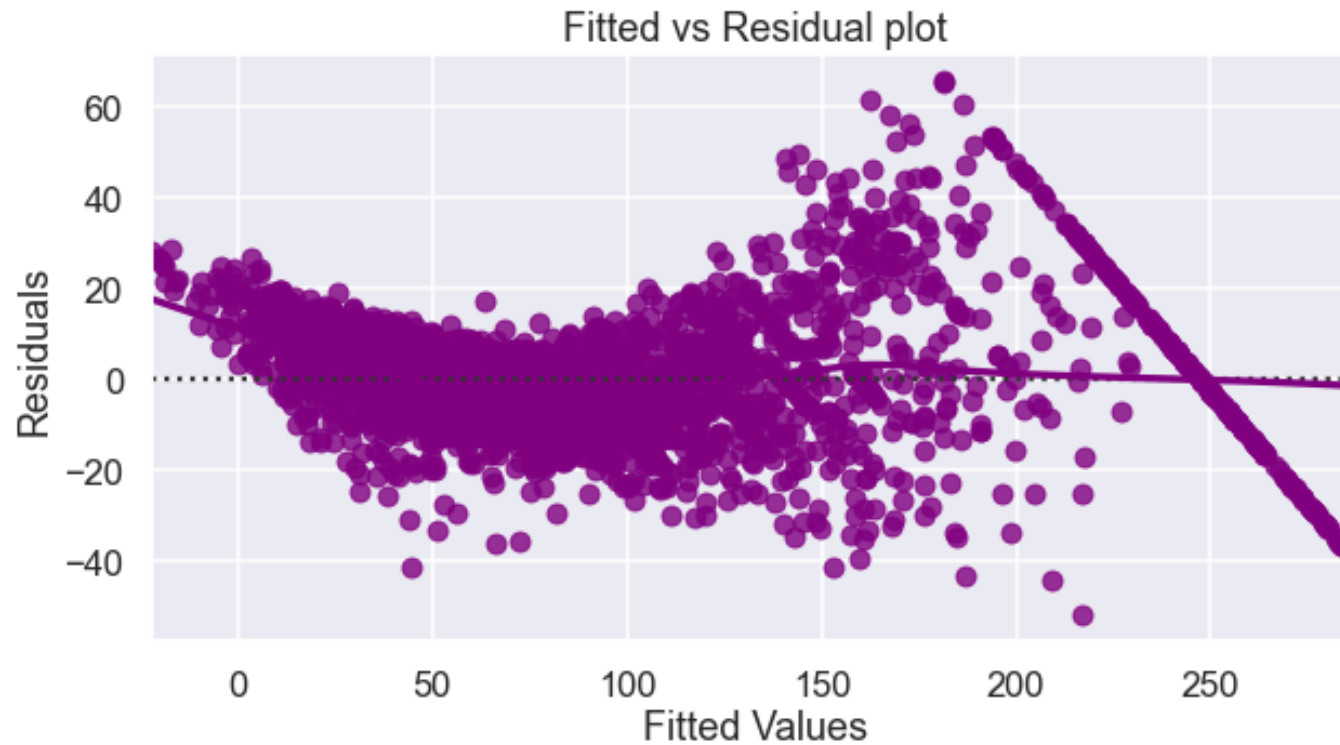
➢ No pattern in the plot above

➢ Therefore assumptions of linearity and independence are satisfied.

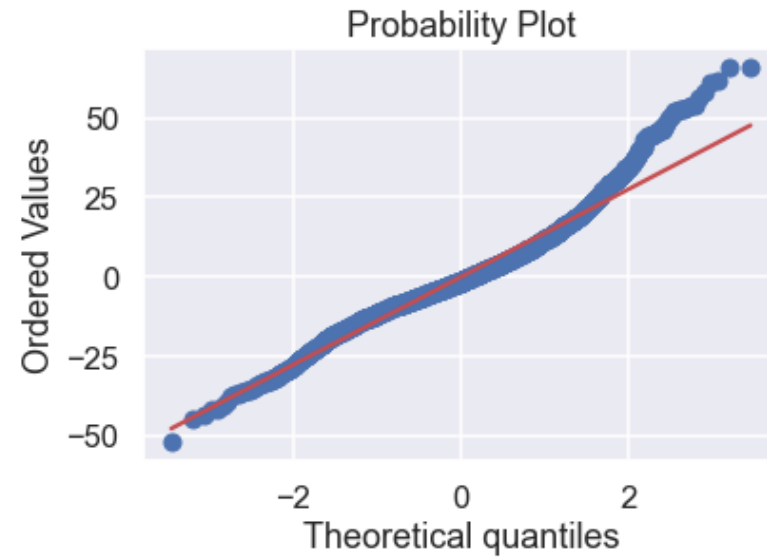# Test of Normality

Normality of residuals

Probability Plot
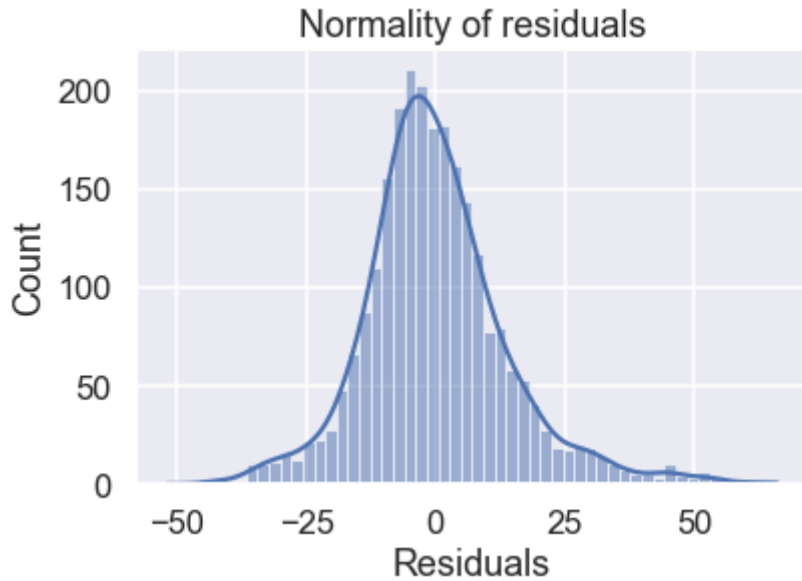
Bell shape histogram implies normal

Test of Homoscedascity returned p = 0.20
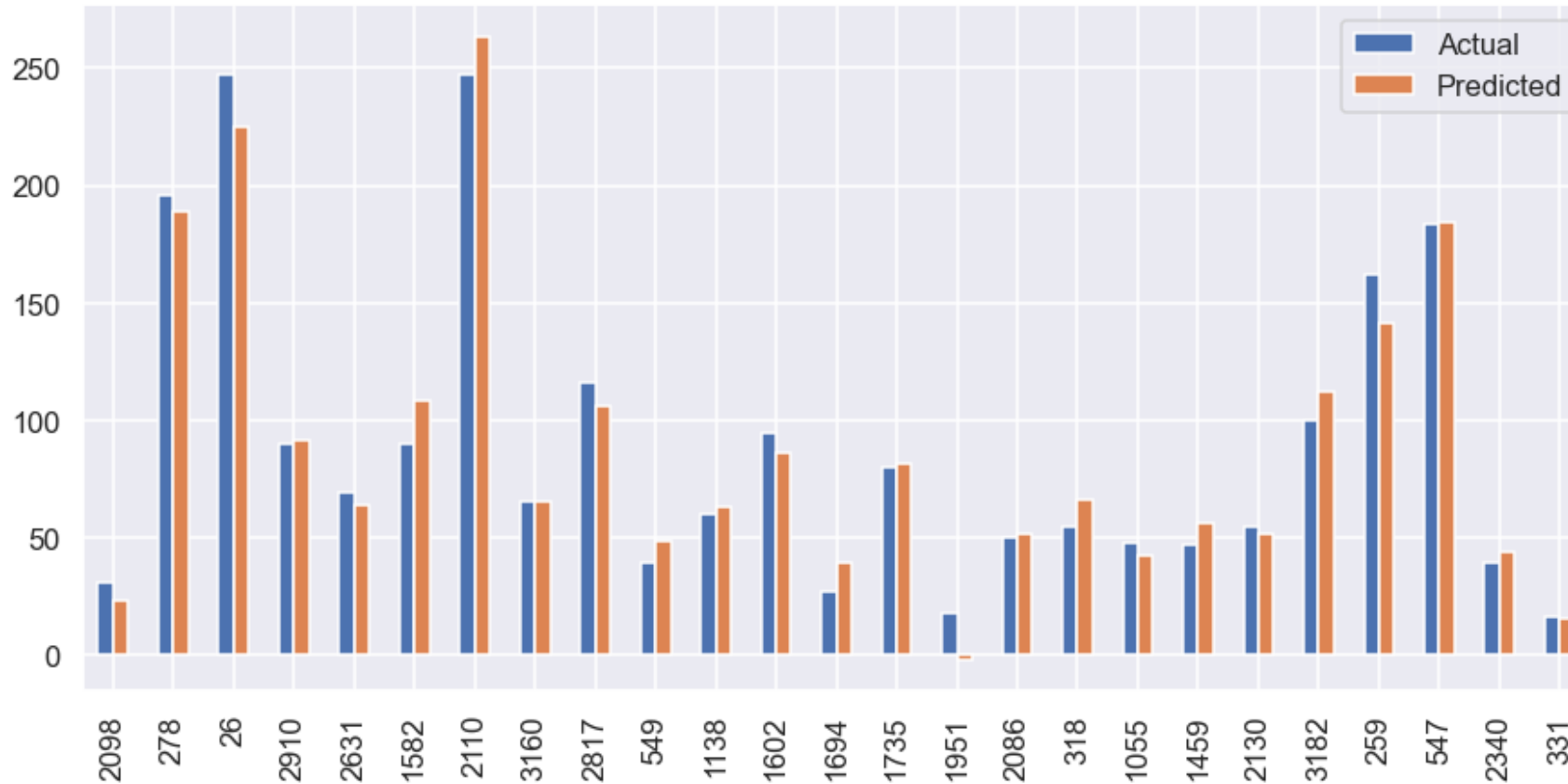
Since p > 0.05

The residual are homoscedastic

The residuals more or less follow a straight line except for the head

Implies normal distribution of residuals

Plot of Actual vs Predicted

The actual and predicted values are similar; insignificant differences
Good prediction by the model.

Model performance summary

Training Performance Comparison

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	13.960441	13.722107
MAE	10.222224	10.109717
R-squared	0.955136	0.957586
Adj. R-squared	0.954257	0.957025
MAPE	18.489055	16.300215

Test Performance Comparison

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	13.745320	13.722107
MAE	10.171443	10.109717
R-squared	0.957443	0.957586
Adj. R-squared	0.955446	0.957025
MAPE	16.417574	16.300215

➢ The performance of the Training and Test are similar for both the Sklearn and Statsmodels

➢ From the R-squared, model is able to explain ~95 of the data

➢ The train and test RMSE and MAE are comparable

Conclusion

Factors that negatively influence price of used phone, from the most impactful to the least:

➢ main_camera_mp ,days_used, 4g

➢ Brands : OnePlus, Infinix, Nokia, Gionee, and others(brands not listed)

Factors that positively influence price of used_phones:

➢ ram, selfie_camera_mp, new_price and int_memory

➢ Brands: Google, Apple

Recommendation:

To generate high revenues per phones company should deal in:

➢ Google and Apple used_phones

➢ High ram, high selfie_camera_mp, and high int_memory used_phone

➢ 5g phones.