ReCell Presentation

Contents

Business Problem Overview and Solution Approach

- Core business idea
  - To optimize ReCell buying and selling strategy in order to maximize market share in the booming used and refurbished cell phone market that is expected to approach $52.7 Billion dollars by 2023 and a compound annual growth rate of 13.6% from 2018 to 2023.
- Problem to tackle
  - Use gathered data to optimize a business strategy for building a successful and dynamic pricing stategy
- Financial implications
  - Take advantage of the rapidly growing market of used and refurbished cell phones and the global impact of Covid-19 affecting customer expenditure on new phones
- How to use ML model to solve the problem
  - Build a predictive model that will allow ReCell to infer a buying and selling strategy

Data Overview

Data:

Contains different attributes of used/refurbished phones

There are 3571 rows and 15 Columns

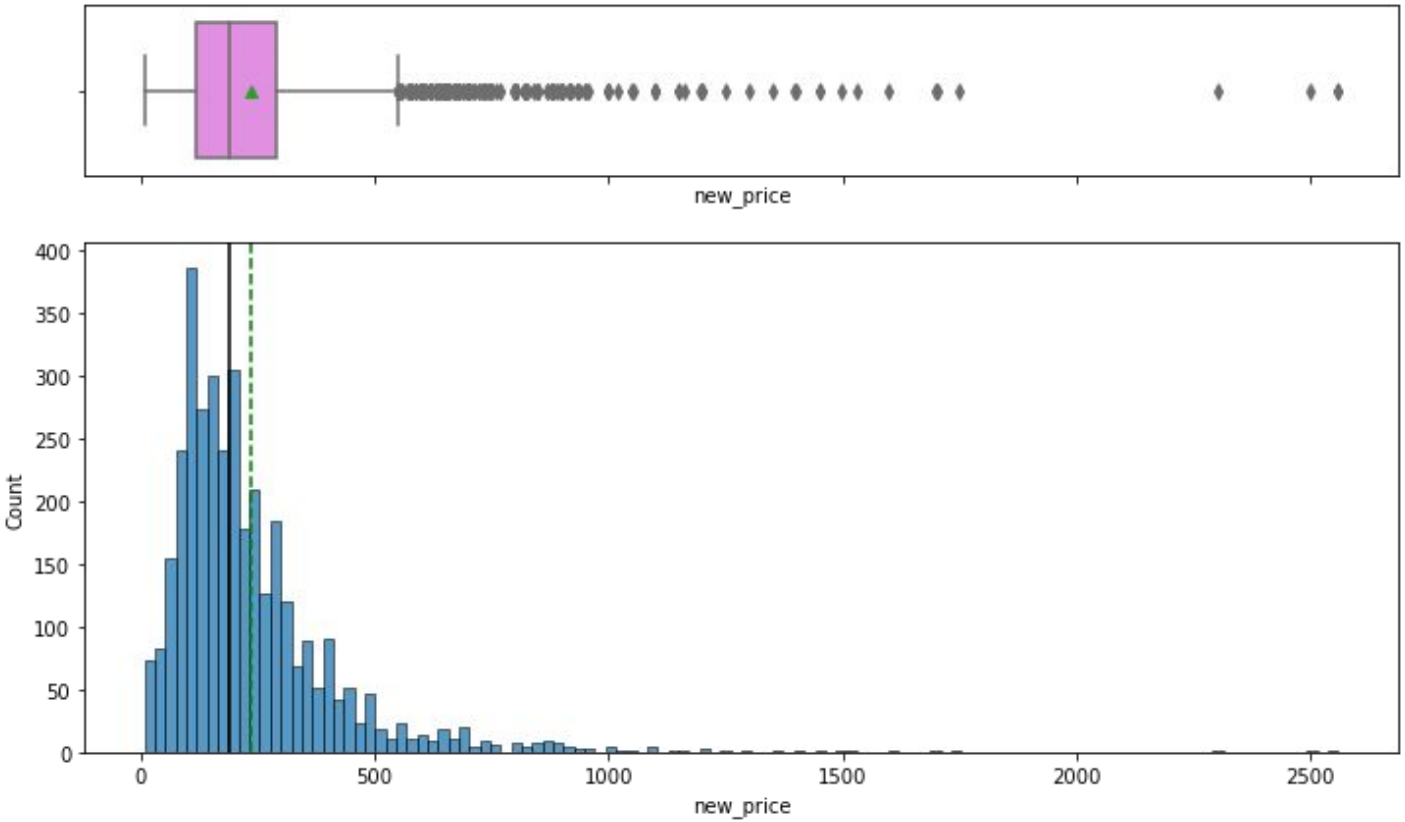
Variable	Description
brand_name	Name of MFG brand
os	OS on which phone runs
screen_size	Size of screen in cm
4g	Whether 4g is available
5g	Whether 5g is available
main_camera_mp	Resolution of rear camera in megapixels
selfie_camera_mp	Resolution of front camera in megapixels
int_memory	Internal memory in GB
ram	Amount of RAM in GB
battery	Energy capacity of battery in mAh
weight	Weight of phone in grams
release_year	Year phone model was released
days_used	Numbers of days phone has been used
new_price	Pierce of new phone in Euros
used_price	Price of used/refurbished phone in Euros

Data Overview Ctd:

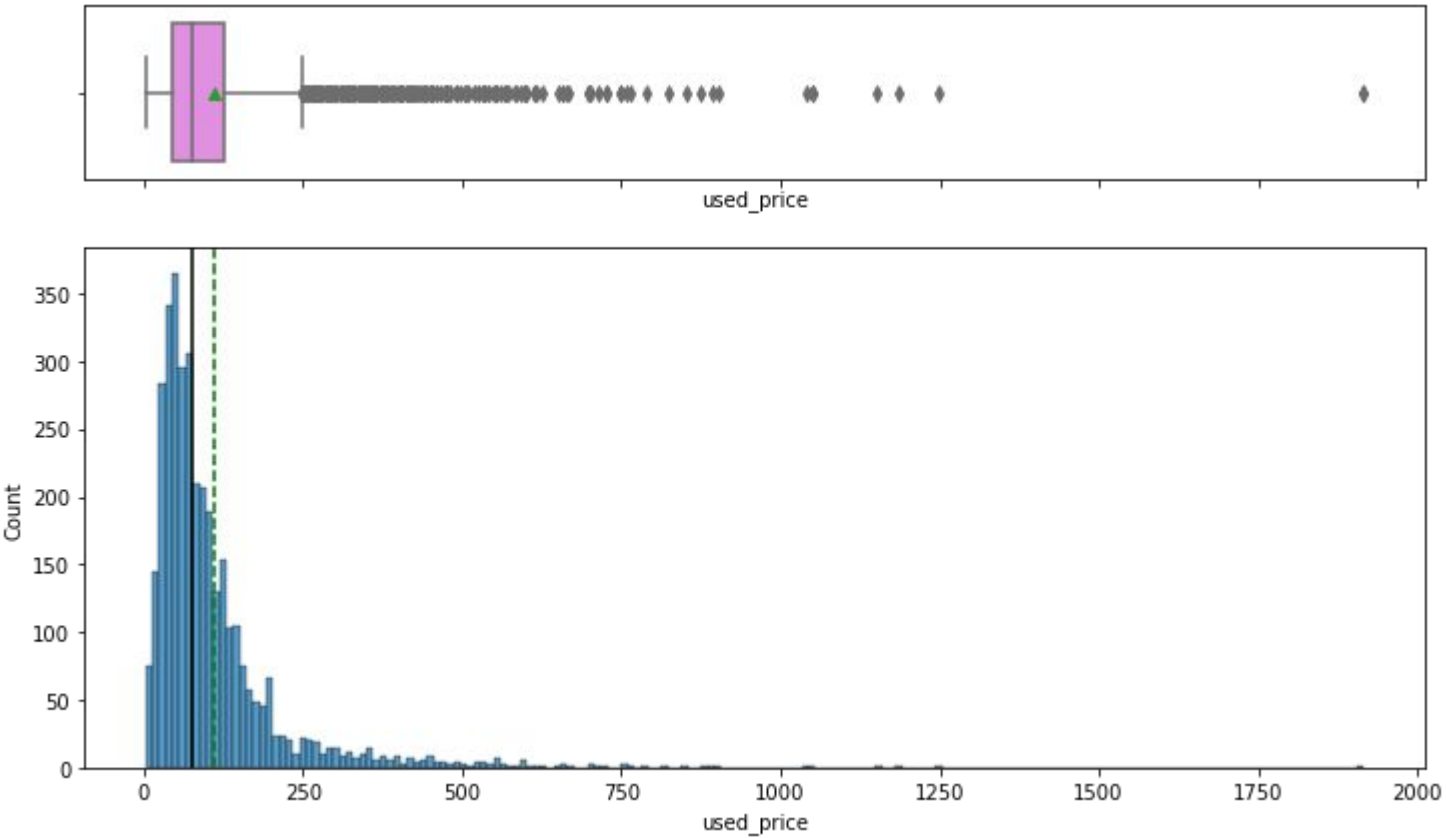
- Data contains object strings, floats and integer data types

- Columns with null values:
  - Main_camera_mp has the most null values (180)
  - Selfie_camera_,mp (2)
  - Int_memory (10)
  - Ram (10)
  - Battery (6)
  - Weight (7)

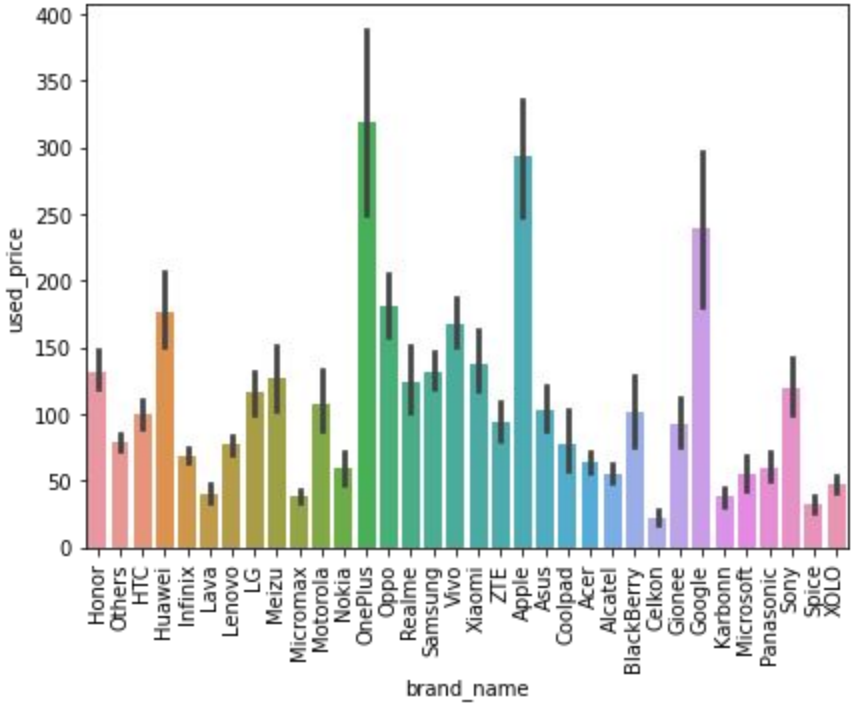
EDA: Univariate Analysis: New Phone Prices

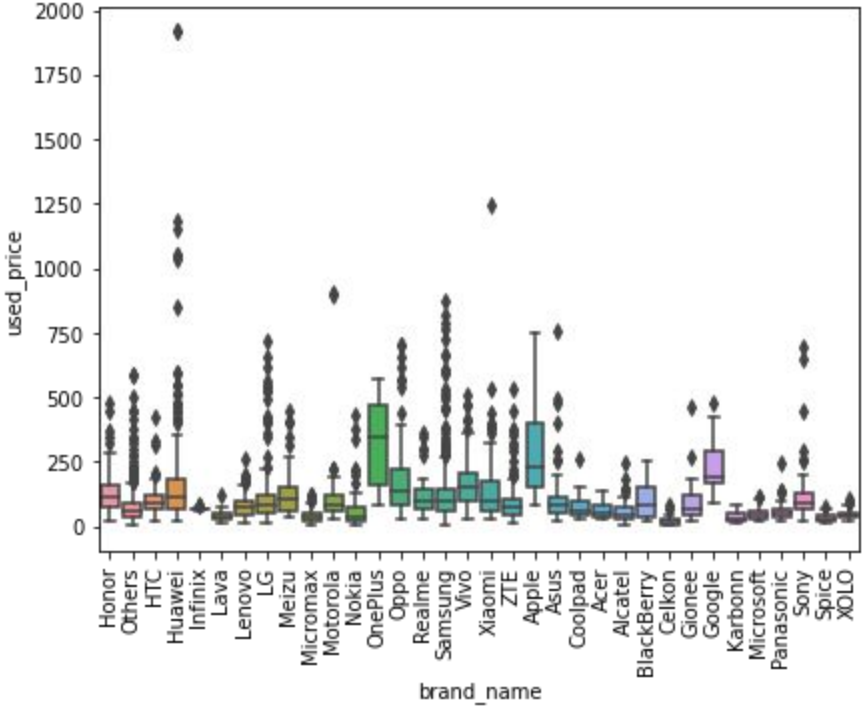


EDA: Univariate Analysis: Used Phone Prices



EDA: Used Price by Brand





OnePlus, Apple and Google have the highest used phone prices. Phone price across all brands show a lot of extreme outliers

EDA Analysis: Phone Price Observations

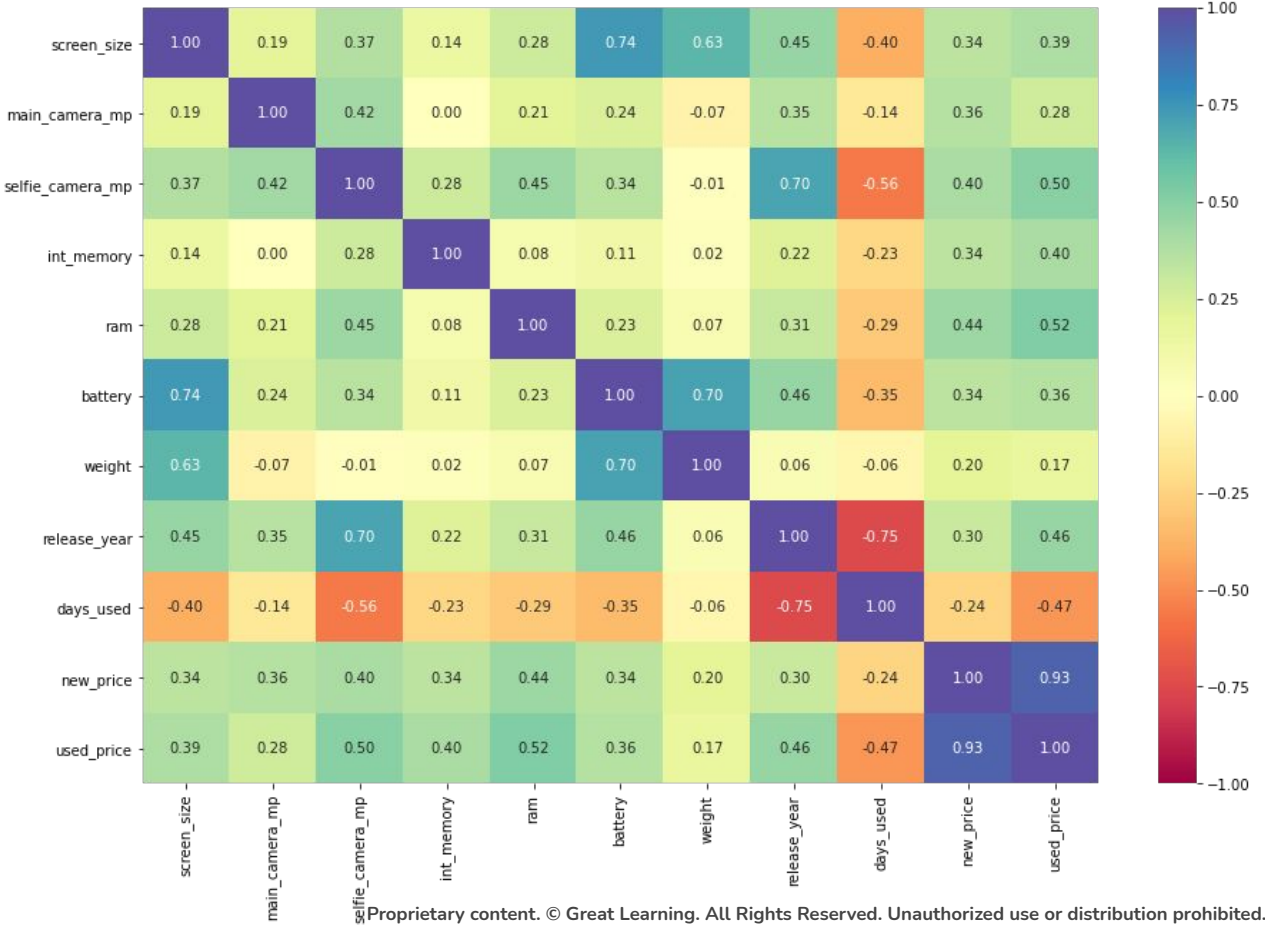
- Both new and used phones prices are very heavily right skewed

- Both new and used phone prices have a wide range of prices
  - Used phones Range from 2.51 Euros to 1916.54 Euros with a mean of 109.88 Euros
  - New Phones range from 9.13 Euros to 2560.2 Euros with a mean of 237.4 Euros

- Sheer number of outliers on both distributions cause the skewness, although a majority of the phones are priced below 250 Euros for used and below 500 Euros for new phones

- Both distributions follow the same trend, which means the used pricing is not too outlandish, nor too under priced

EDA: Correlation Analysis



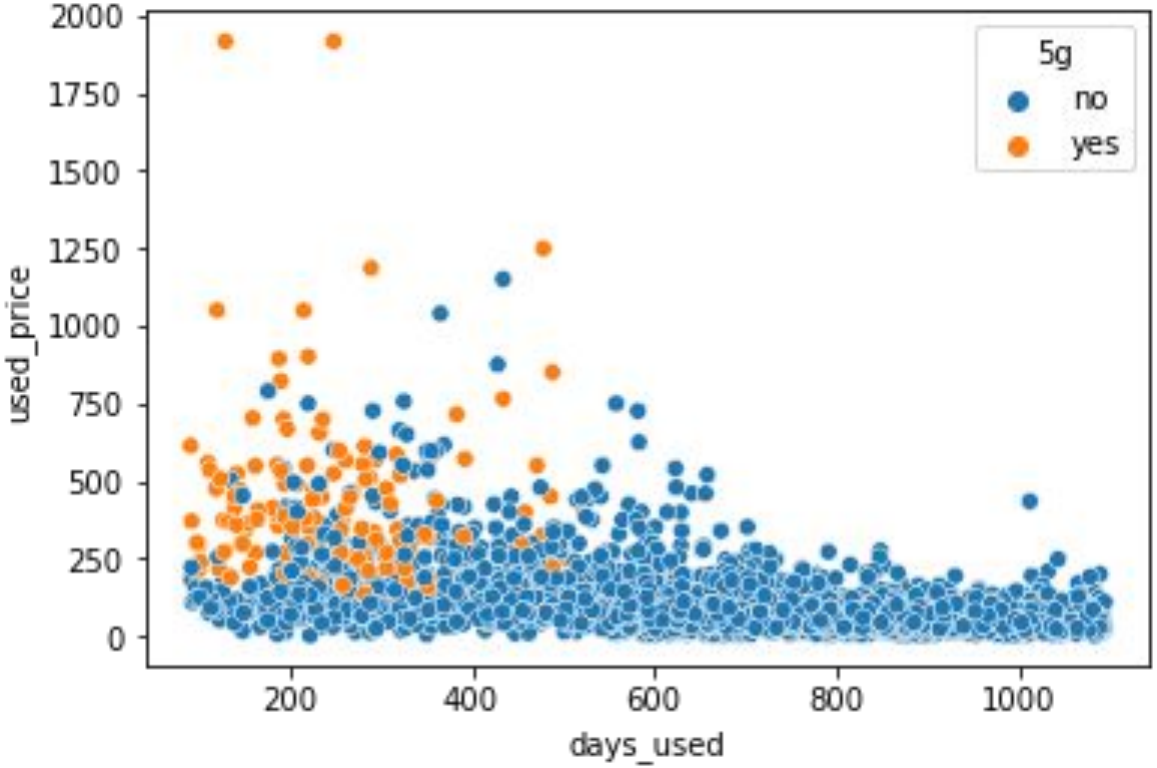
EDA: Correlation Analysis Observations

- Used_price correlation to independent variables are all positive except days_used which means they all can affect the price positively, save how many days the used device has been used. This correlation makes sense as a device with more use will demand less money when re-selling.

- The highest correlation on used_price is new_price which makes sense as sellers of used/refurbished phones must still make money and so the pricing scales for new and used will be very similar due to customer demand.

- Days_used and release_year also have very high correlation with used_price as the newer and less used the phone is the more the reseller can price it at.

EDA: Days Used vs Used Price with 5G



The older the phone the lower the price

5G phones have higher used prices

EDA: Operating Systems

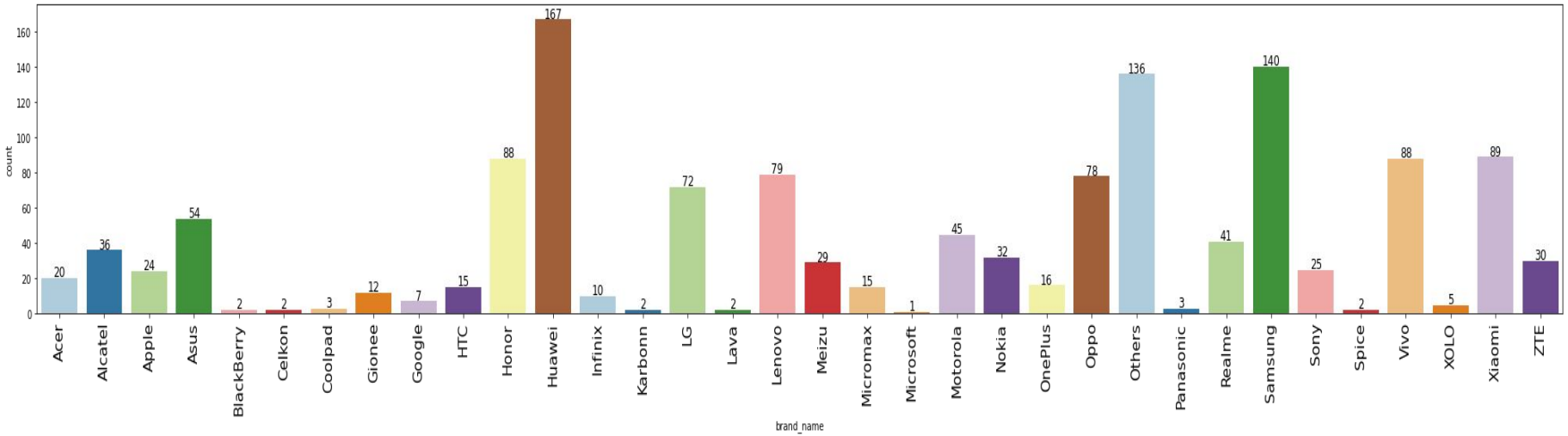
Operating System	Mean Price (Euros)
Android	112
Others	43.8
Windows	65.1
iOS	280.8

Operating System	Count
Android	3246
Others	202
Windows	67
iOS	56

The most expensive used phones have iOS operating systems

Android phones make up 90.9% of the used market and have the 2nd highest prices

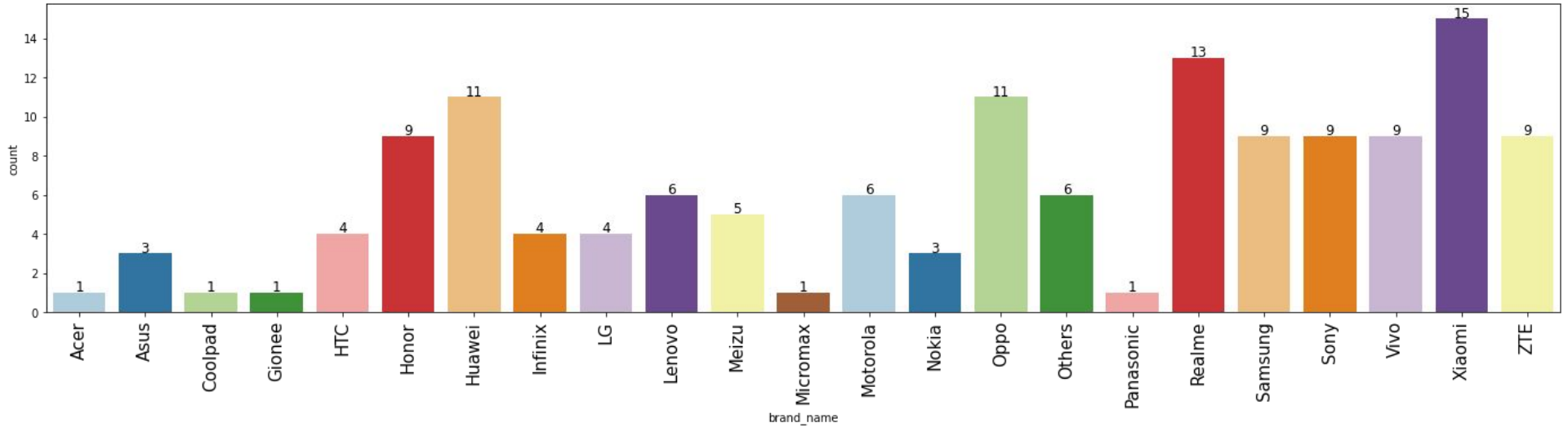
EDA: Phones with Screen size over 6 inches



Top three brands with screens over 6 inches:

1) Huawei with 167 phones
2) Samsung with 140
3) Others with 136

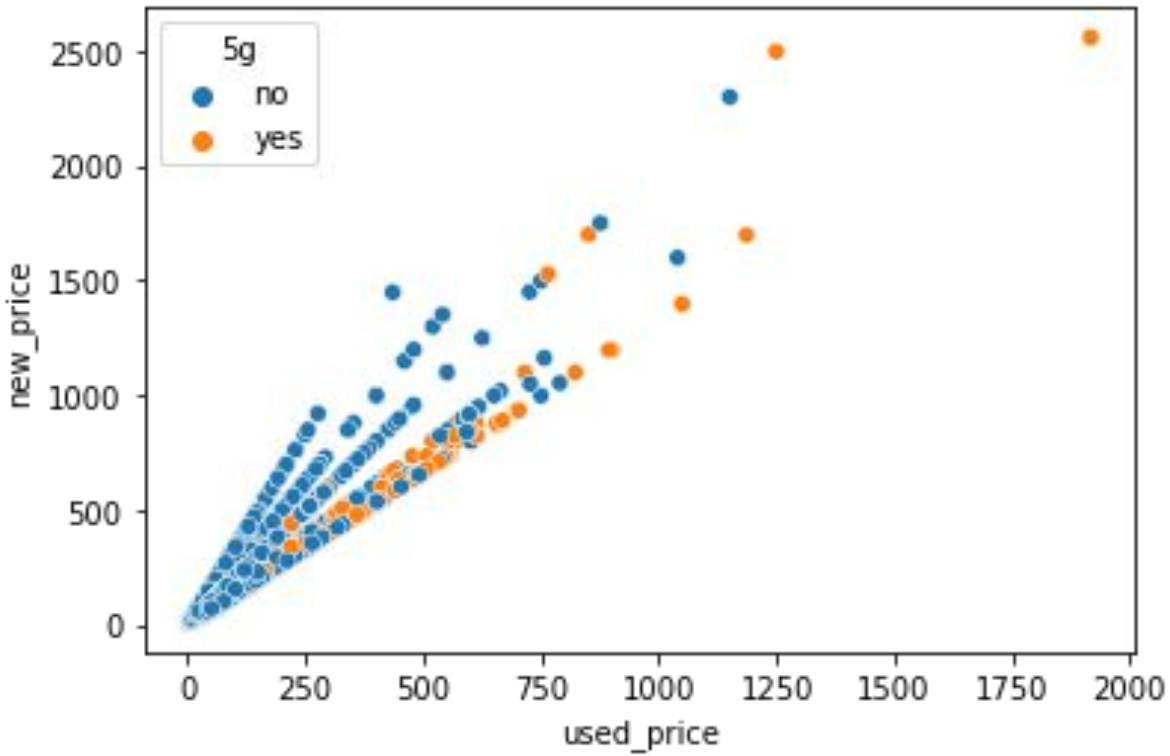
EDA: Budget Phone Camera Analysis



"Budget" phones will be defined as phones under the used price mean of 110.0 Euros

The leader in "Budget" Phones with 6 inch screens is Xiaomi (15), followed by Realme(13)

EDA: New vs Used Phone Prices with 5G

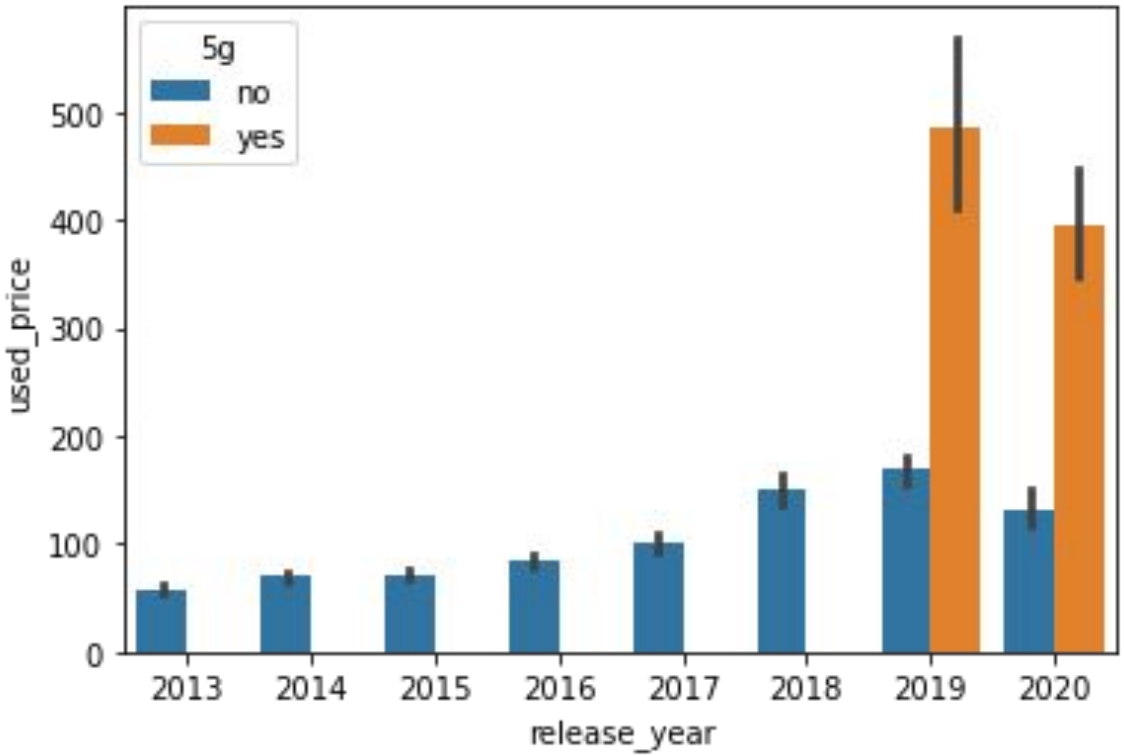


Used price correlates with new price as sellers of used and refurbished phones sell the phones almost 50% off from new price.

Phones with 5G have higher new and used prices.

EDA: Used Price vs Release Year and 5G

- 5G phones have a much higher used price compared to non 5G phones.
- Phones with 5G started to be offered in 2019 and are 50% more expensive than phones without 5G from the same year.



Data Preprocessing

Null values from data will need to be treated to continue processing. The column median values will be used to replace any null values.

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

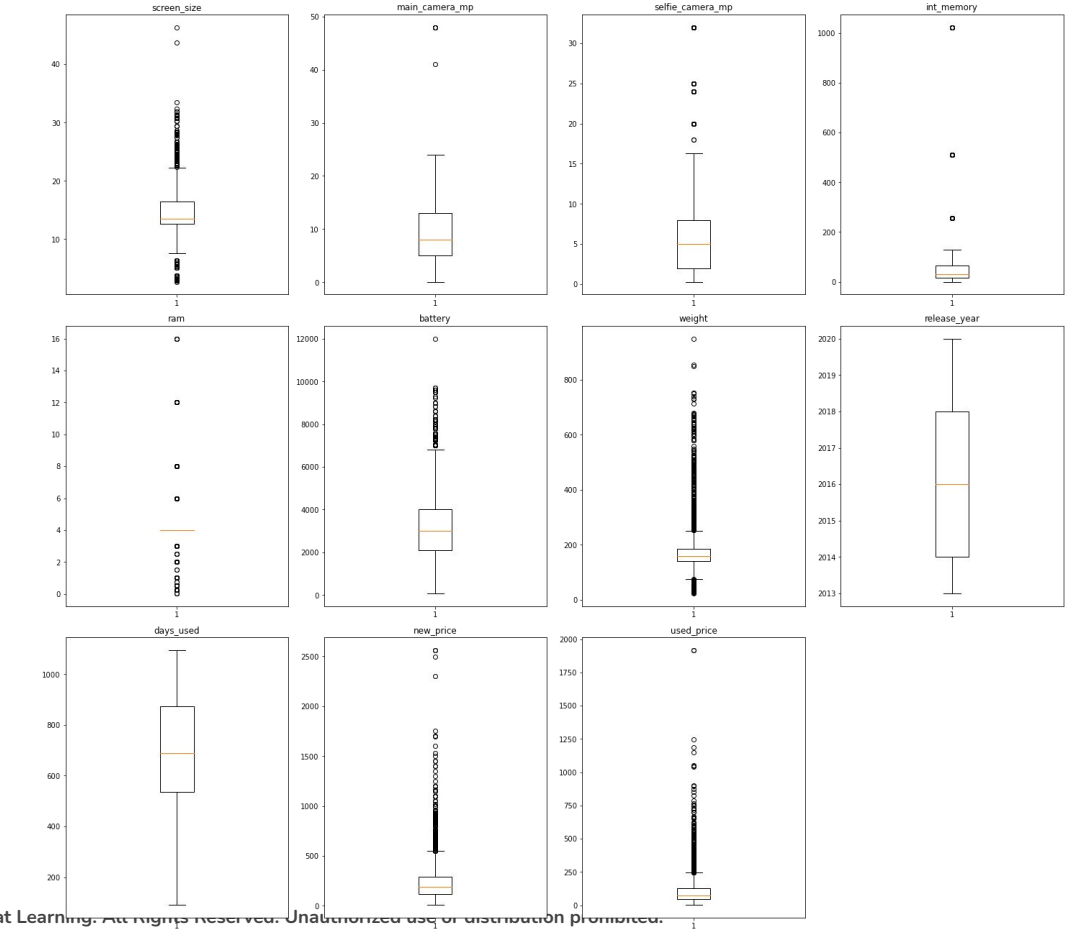
With null values

brand_name	0
os	0
screen_size	0
4g	0
5g	0
main_camera_mp	0
selfie_camera_mp	0
int_memory	0
ram	0
battery	0
weight	0
release_year	0
days_used	0
new_price	0
used_price	0

With column median replacing null

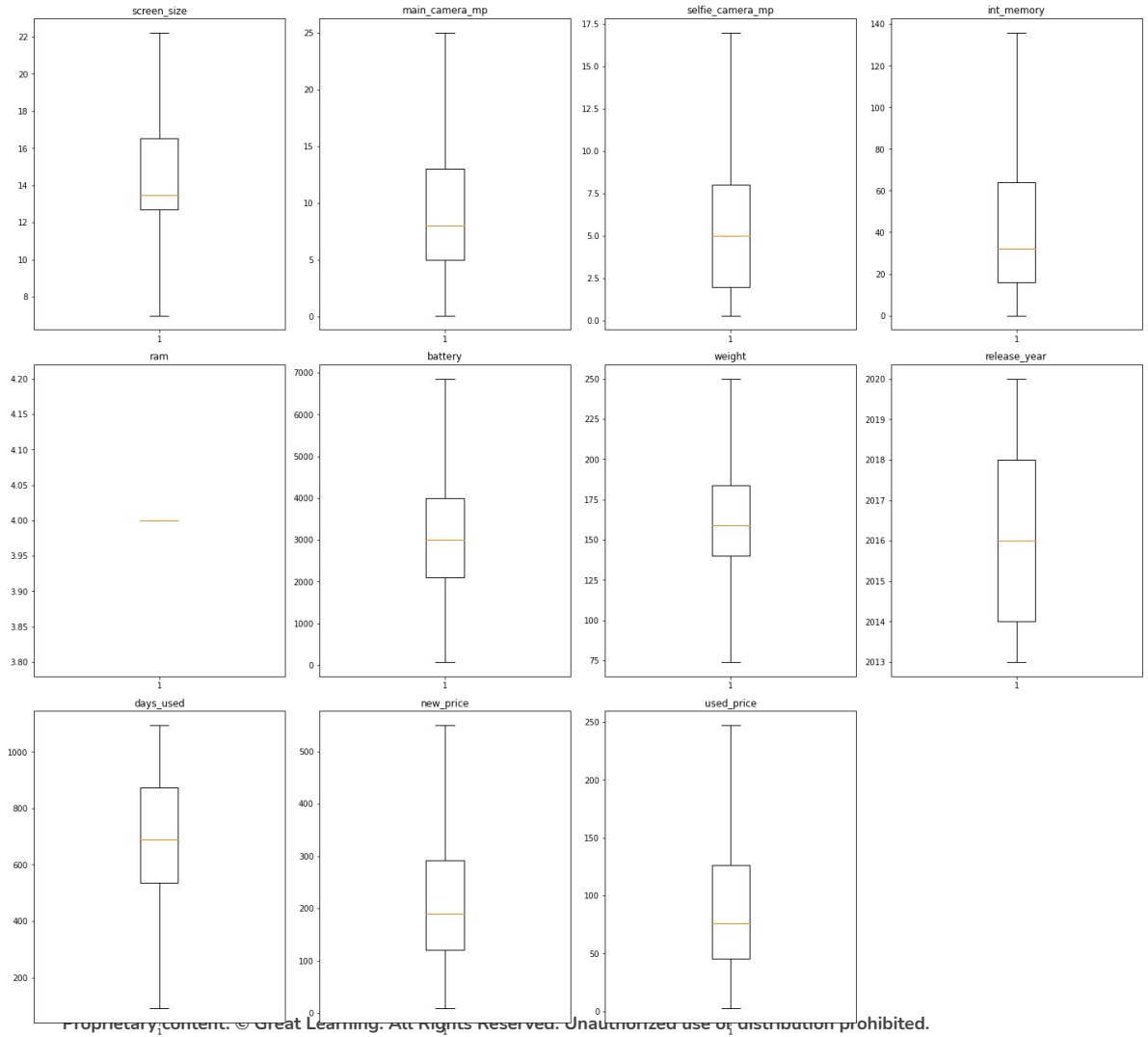
Data Preprocessing Continued

- Data is very heavy on outliers in all categories
- We will treat the outliers by assigning all values below lower whisker the value of lower whisker. Same formula will be assigned to upper whisker values.



- Outliers adjusted.

- RAM will be dropped from modeling going forward as values are at median of 4GB



Model Performance Summary

- Overview of ML model and its parameters
  - We want to build a predictive model for used phone prices.
  - Data will be split into training/test data at 70:30
  - X variables will contain all variables except “used_price’ (y variable) , and ‘RAM’
  - Categorical variables will be split to allow their use in modeling going forward using one-hot encoding features which will be applied to object and categorical columns:
    - Brand_name
    - OS
    - 4g/5g
  - Number of rows in train data = 2499
  - Number of rows in test data = 1072
  - Random state applied to all modeling to ensure same data is used for any changes made to formula

- Summary of key performance metrics for training and test data in tabular format for comparison

Model Performance Summary

- Summary of most important factors used by the ML model for prediction
- Observations from training model:
    - Largest positive effects on used price is Apple, Google, Coolpad, Blackberry phones and phones with 5G and Windows operating system.
    - Largest negative variables to used price is brand Infinix, OnePlus and phones with iOS operating systems and not having 5G.

Model Performance Summary

- Summary of key performance metrics for training and test data in tabular format for comparison
- Training model vs Test performance using RMSE, MAE, R-Squared, Adjusted R-Squared and MAPE:

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	13.960	10.222	0.955	0.954	16.489

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	13.745	10.171	0.957	0.955	16.418

The model shows similar results across both training and test data. No signs of underfitting or overfitting. Model is at 95% predictive value of price. The RMSE/MAE are a little high due to the wide range in price of used phones

Checking model linear regression assumptions

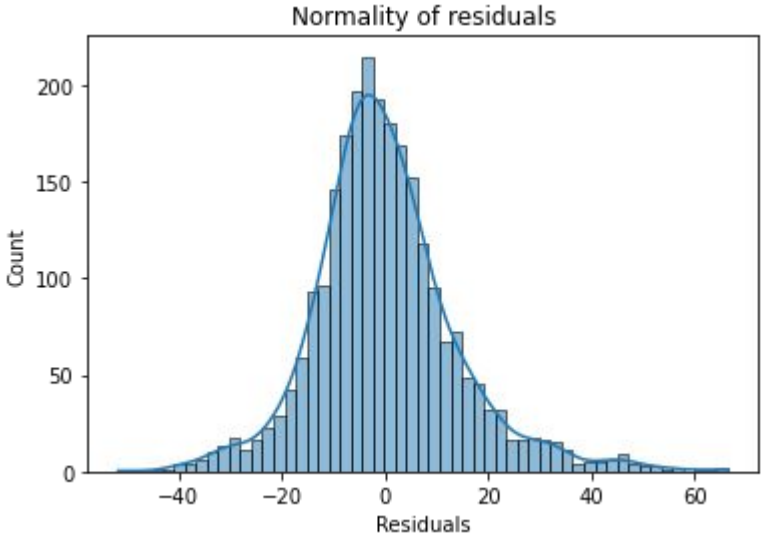
- In order to make statistical inferences from a linear regression model it is important to ensure that the assumptions of linear regression are satisfied.
  - No Multicollinearity (VIF)
    - Most of the VIF's are within the acceptable range of below 10. The ones that are above 10 are brand names and operating systems which we will keep as they offer good insight into the market and we do not want to make generalizations without keeping factors that could drive our model.
  - Linearity of variables
    - Plot residuals vs fitted values to check scatter plot for a pattern. If there is a pattern it is a sign of non-linearity in the data. Chart shows some non linearity.

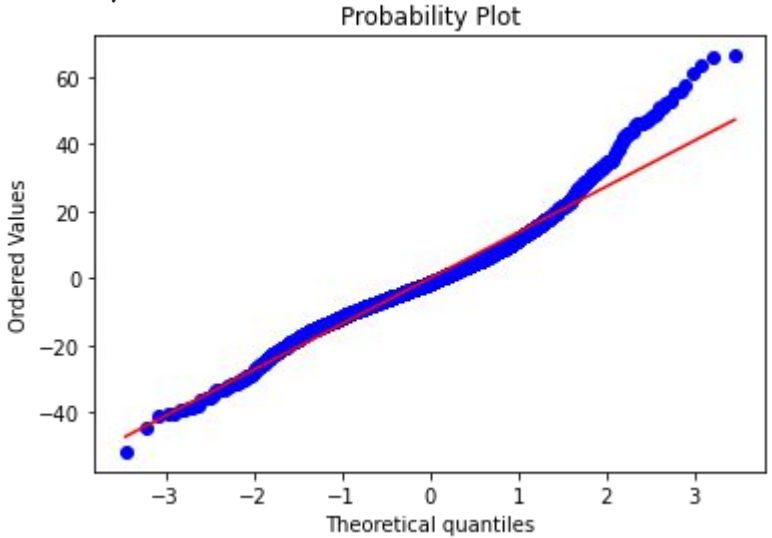
Checking model linear regression assumptions

Testing for normality using Q-Q plot of residuals:

Null Hypothesis: Residuals are normally distributed

Alternate Hypothesis: Residuals are not normally distributed





Residuals follow a bell curve and the Q-Q plot follows a relatively straight line. Satifies our test for normality.

Checking model linear regression assumptions

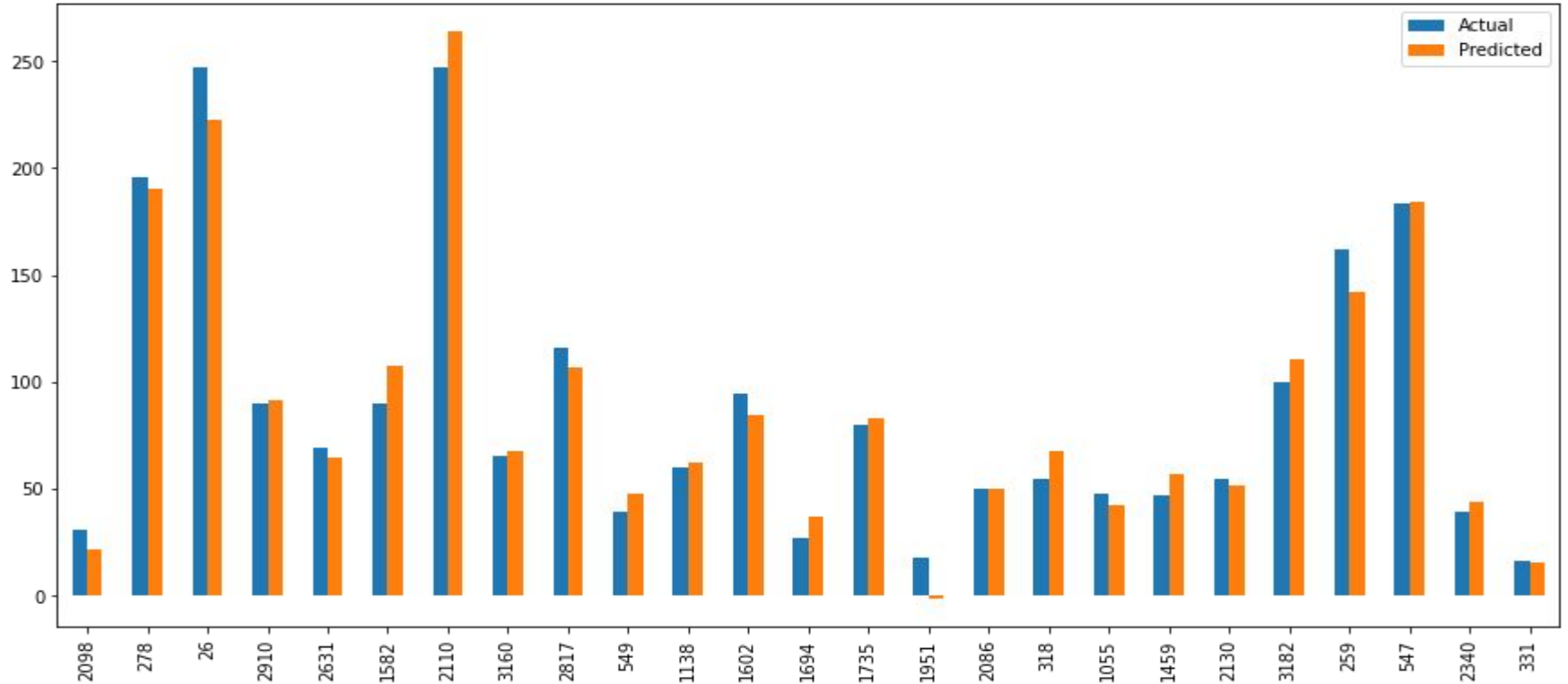
- Shapiro-Wilk test for normality of residuals.
- P-value of less than 0.05 will prove Null Hypothesis is correct.
- P-value is much less than 0.05, so the residuals are normal and Null Hypothsis is correct.
- We will accept the normality of the data based upon this approximation and the results of the Q-Q test.

```
ShapiroResult(statistic=0.9623290896415 71, pvalue=5.260188610962256e-25)
```

- Homoscedasticity:
- P-value is > 0.05 so the residuals are homoscedastic

```
: [('F statistic', 1.0345810562814 13), ('p-value', 0.2779043679249576)]
```

Actual vs Predicted on Test Set



Our model shows very well in predicting prices

Final Model Summary

- Testing our model using sklearn and statsmodels result in very similar metric values.
- Our model is very good for predicting prices

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	13.745	13.980
MAE	10.171	10.222
R-squared	0.957	0.955
Adj. R-squared	0.955	0.954
MAPE	16.418	18.489

Business Insights and Recommendations

- Recommendations based on interpretation of the model input variables
  - The largest impact on pricing is from release year, days used, brand, operating system and 5g.
  - Apple/Google phones with 5G and Windows operating systems will have the highest price
  - OnePlus and Infinix phones with iOS operating system and 4G will cause the most negative pricing impact.
  - Majority of the brands will cause minimal impact, but there are some that ReCell would do well to stay away from until the model is updated if necessary.

- Comments on additional data sources for model improvement, model implementation in real world, and potential business benefits from model.

  - Going forward to build a more robust predictive model it may behoove ReCell to consider filtering the columns/data to attempt to get a better value.
  - Model can be aggressively altered by dropping columns with higher multicollinearity and retesting model on training/test data. Current model show impactful intelligence as brand and operating systems cause meaningful impact on pricing and should be maintained.

greatlearning
Power Ahead

Happy Learning !