ReCell

Business Case

Background

Buying and selling used smartphones used to be something that happened on a handful of online marketplace sites. But the used and refurbished phone market has grown considerably over the past decade, and a new IDC (International Data Corporation) forecast predicts that the used phone market would be worth $52.7bn by 2023 with a compound annual growth rate (CAGR) of 13.6% from 2018 to 2023. This growth can be attributed to an uptick in demand for used smartphones that offer considerable savings compared with new models.

Refurbished and used devices continue to provide cost-effective alternatives to both consumers and businesses that are looking to save money when purchasing a smartphone. There are plenty of other benefits associated with the used smartphone market. Used and refurbished devices can be sold with warranties and can also be insured with proof of purchase. Third-party vendors/platforms, such as Verizon, Amazon, etc., provide attractive offers to customers for refurbished smartphones. Maximizing the longevity of mobile phones through second-hand trade also reduces their environmental impact and helps in recycling and reducing waste. The impact of the COVID-19 outbreak may further boost the cheaper refurbished smartphone segment, as consumers cut back on discretionary spending and buy phones only for immediate needs.

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, has hired you as a data scientist. They want you to 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.

Objective

To 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 Information

The data contains type of products, feature specifications and prices

Variable	Description	Type of Variable
brand_name	Name of manufacturing brand	Category
os	OS on which the phone runs	Category
screen_size	Size of the screen in cm	Float
4g	Whether 4G is available or not	Category
5g	Whether 5G is available or not	Category
main_camera_mp	Resolution of the rear camera in megapixels	Float
selfie_camera_mp	Resolution of the front camera in megapixels	Float
int_memory	Amount of internal memory (ROM) in GB	Float
ram	Amount of RAM in GB	Float
battery	Energy capacity of the phone battery in mAh	Float
weight	Weight of the phone in grams	Float
release_year	Year when the phone model was released	Integer
days_used	Number of days the used/refurbished phone has been used	Integer
new_price	Price of a new phone of the same model in euros	Float
used_price	Price of the used/refurbished phone in euros	Float

Observations	Variables
3,571	15

Manipulations to Raw Data:

1.Missing values were replaced with the median
2.Object variables were converted to Category
3. Outliers were replaced with
   - Values smaller than lower whisker replaced with value of lower whisker
   - Values greater than upper whisker replaced with value of upper whisker

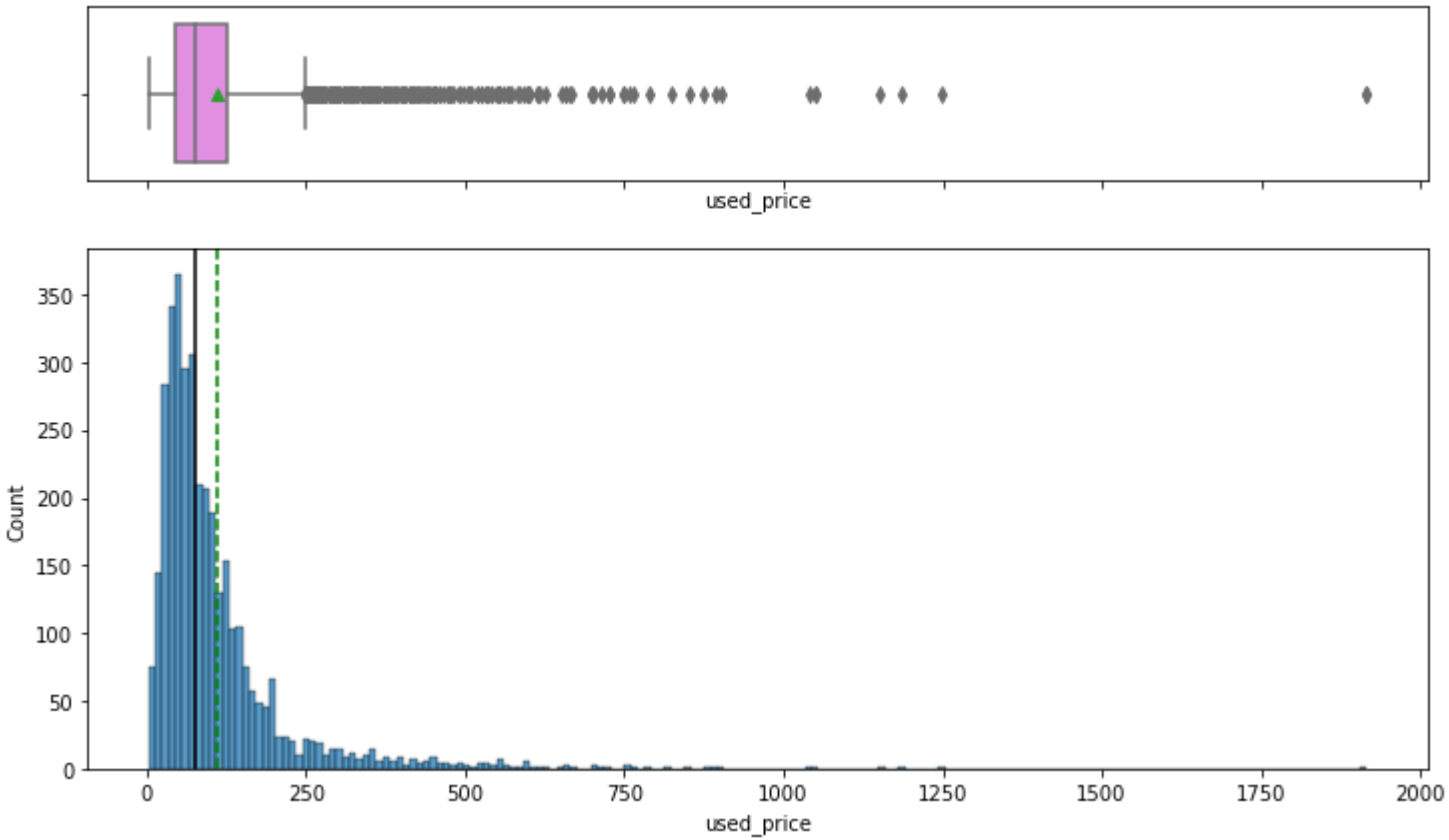
Exploratory Data Analysis – Used Price

This data contains the used price

Observations:

1.This is a very important variable as it is the dependent variable in this business case.
2. The mean and median are very close together
3. The data is heavily right-skewed
4. It is similar to the new price

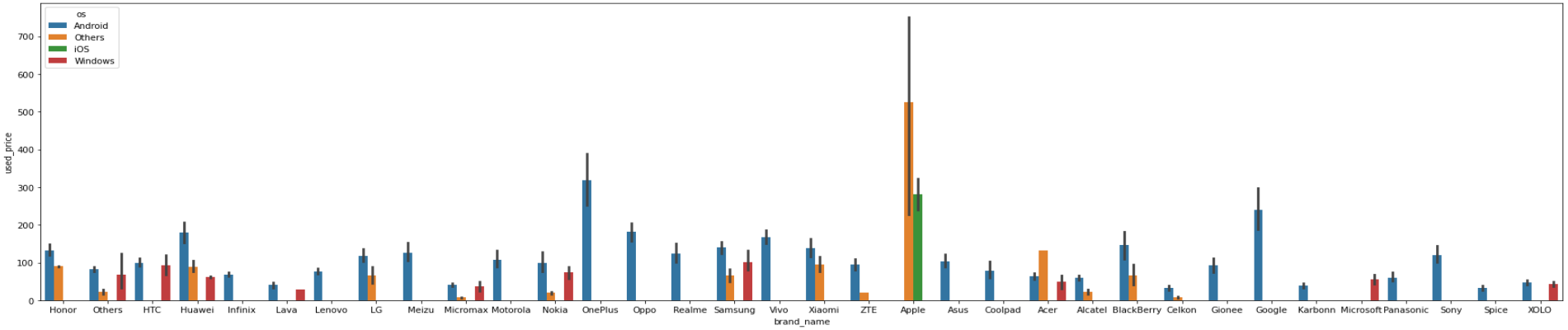
Used Price

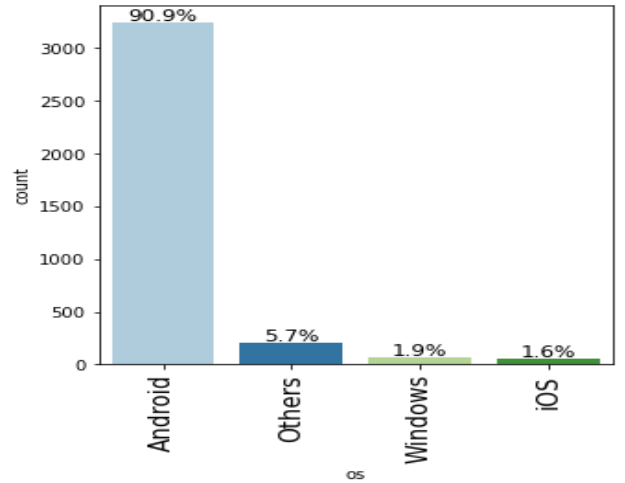


Exploratory Data Analysis – Used Price wrt Brand Name

This data contains the Brand Name wrt Used Price

Used Price wrt Brand Name and OS





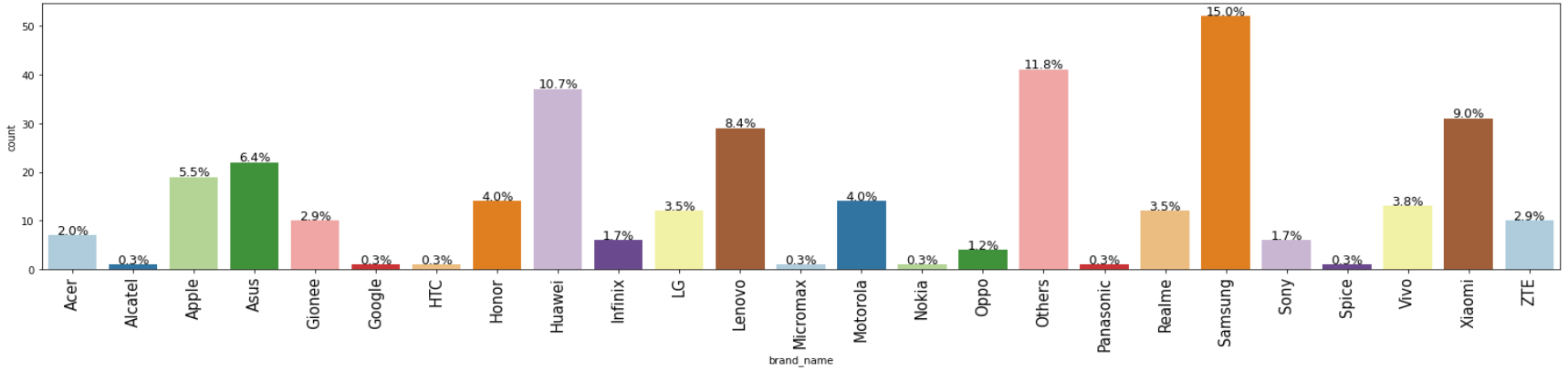
Observations:

1.There are 32 different brands
2.The prices varies significantly against the phones with some strong outliers namely Apple, Oneplus and Google
3.Android has the largest number of phones by a very wide margin, making up about ~91% of the phones

Exploratory Data Analysis – Brand Name

These are all significant factors that affect the prices of the phones

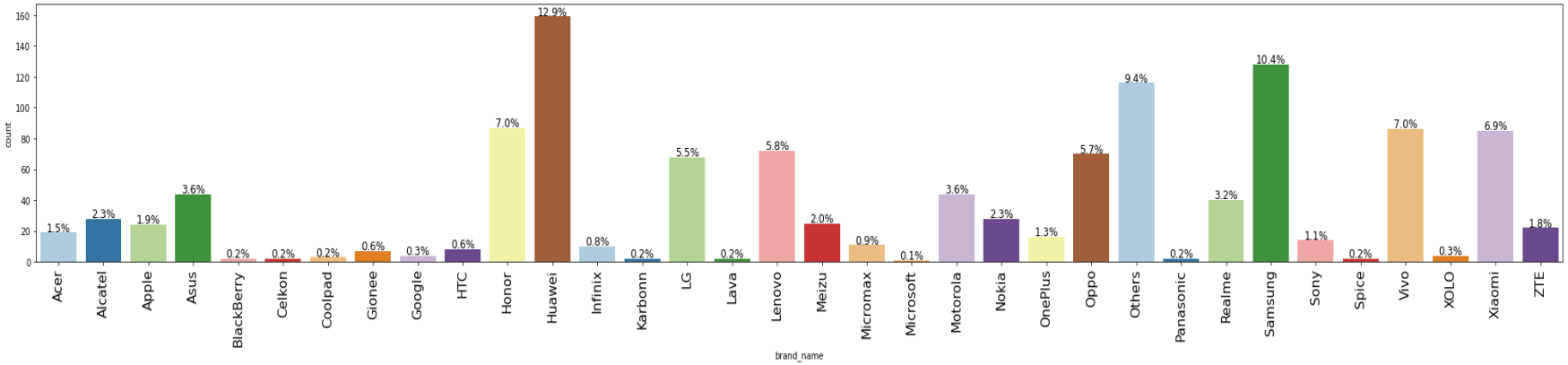
Phone Brand wrt to Large Battery Size (>4500 mAh)



Observations:

1. The number of phones under consideration is 346
2. For phones with a battery higher than 4500 mAh, Samsung, Lenovo, Huawei, Asus, Apple and Others make up ~58%

Phone Brand wrt to Large Screen Size (> 6 inches)



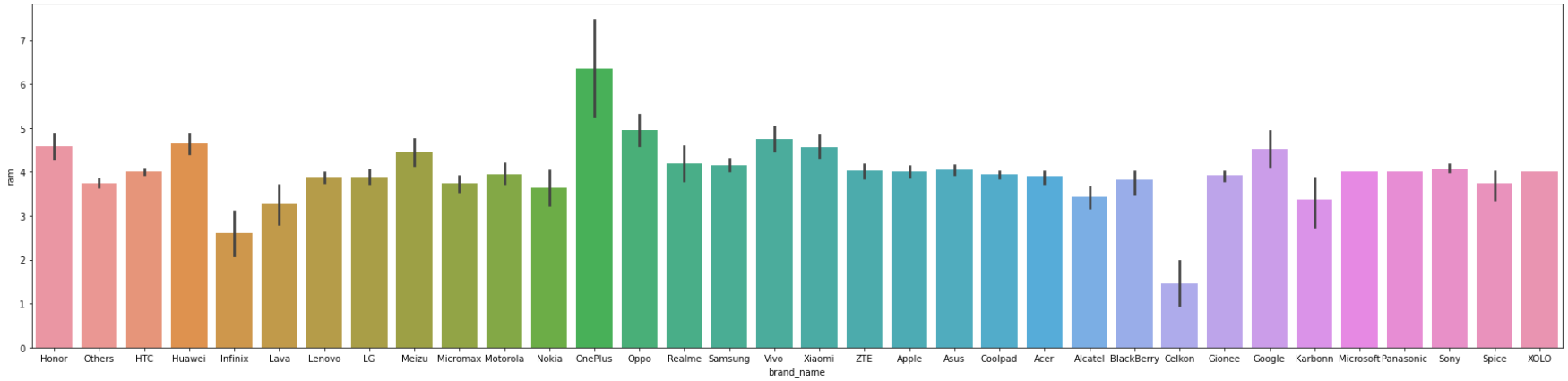
Observations:

1. The number of phones under consideration is 1,235
2. Huawei, Samsung, Others, Lenovo, Vivo, Honor make up ~53% of the phones

Exploratory Data Analysis – Brand Name

These are all significant factors that affect the prices of the phones

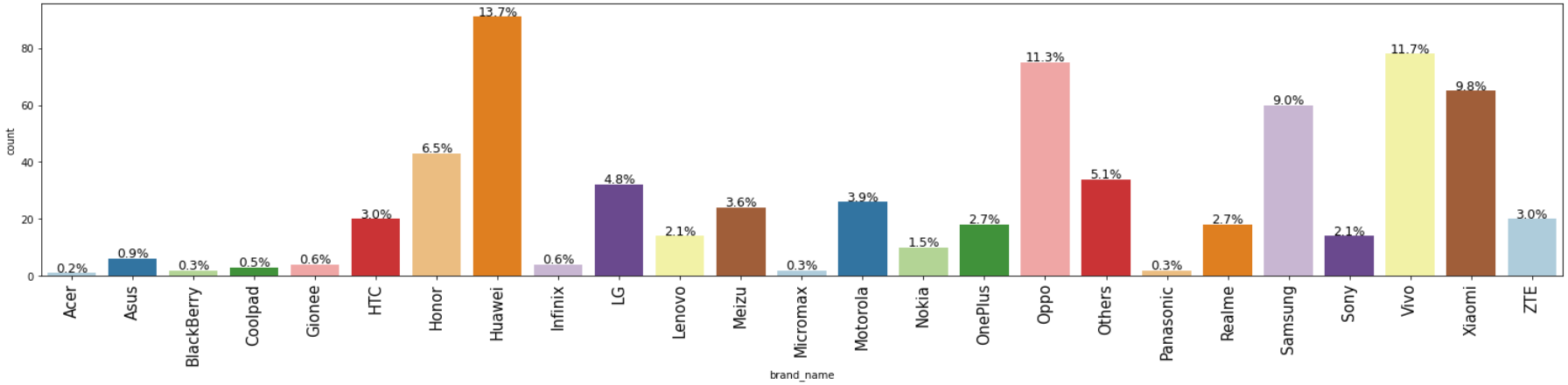
Phone Brand wrt to RAM



Observations:

1. Most phones have a ram size of ~4GB
2. However, some of them are much lower than the others with Oneplus being much higher than the average

Phone Brand wrt to High Quality Selfie Cameras (>8MP)

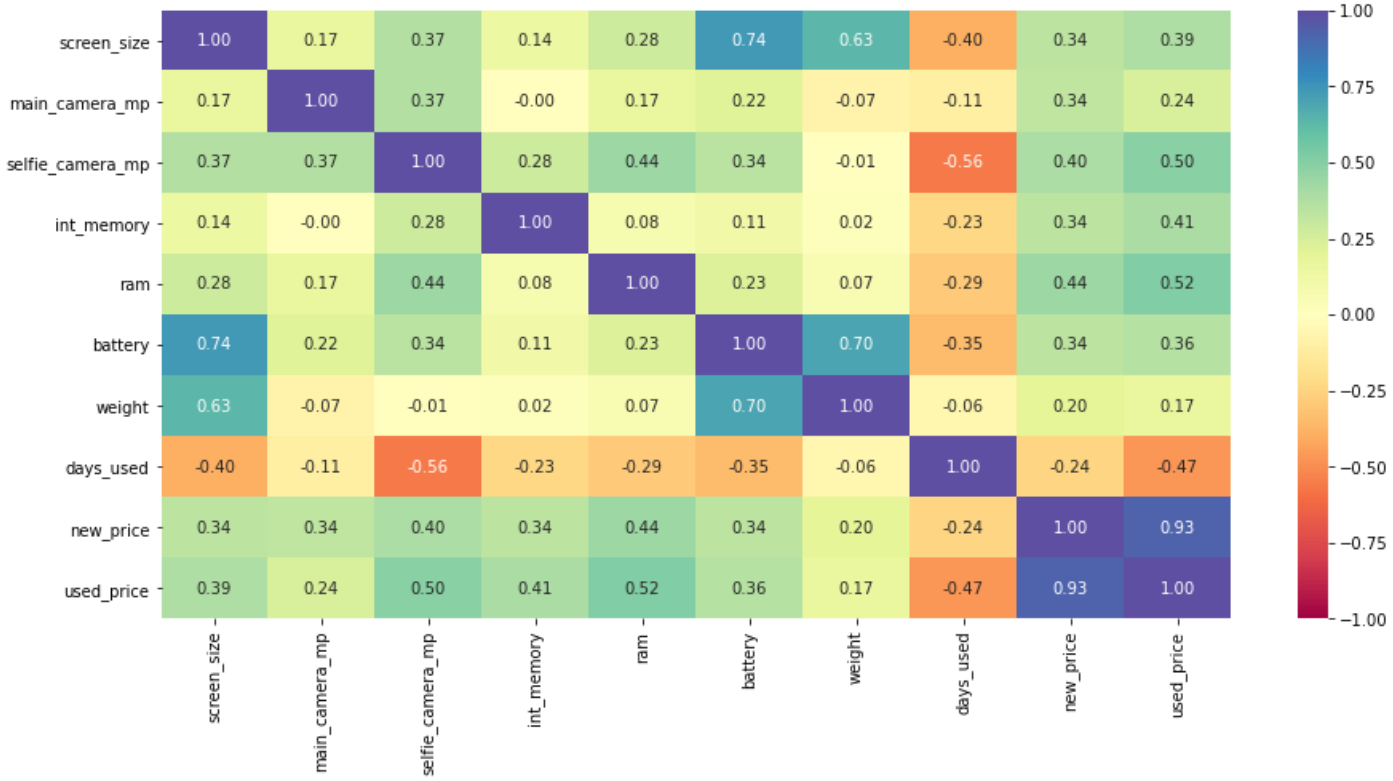


Observations:

1. The number of phones under consideration is 666
2. Huawei, Oppo, Samsung, Vivo, Xiaomi make ~55% of the phones

Exploratory Data Analysis – Correlation of Variables

These are all significant factors that affect the prices of the phones



Observations:

1. New_price and used_price are very highly positively correlated as is to be expected which means that the higher the new price the higher the used price
2. The used_price is negatively correlated with the number of days used
3. Battery, weight and screen size are all relatively highly correlated indicating that the physical features are all relatively highly correlated.

Prediction Model

Linear Regression was used to build the model due to the fact that it is able to find the relationship between dependent (used_price) variables and independent (all other variables) variables.

Overview of the Model

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	14.541195	14.269212
MAE	10.519978	10.337965
R-squared	0.951665	0.953457
Adj. R-squared	0.949397	0.952706
MAPE	18.455811	17.945185

The above table shows the comparisons on the test data between the two types of Linear Regression used in order to double check the model's accuracy.

The parameters:

- Constant (Intercept) - expected mean value of Y when all X=0: ~650
- Slope - ratio of change in Y per change in X:

```
The slopes of the various variables are  [ 4.03778757e-01 -4.38579325e-01  8.37275213e-01  1.06295165e-01
  4.81910623e-11 -1.11335948e-04 -1.58722550e-02 -2.92804911e-01
 -8.44522408e-02  3.79859007e-01  1.84475996e+00 -1.37681762e+01
  2.49083432e+00  1.07808106e+01 -3.54637133e+00  2.40512281e+00
 -3.80407947e+00  1.63763700e+01  7.51625629e-01 -4.02134693e-01
  3.23954287e-01 -1.59799236e+01 -6.27702482e-01  2.88197398e+00
  8.71318052e-01 -2.03483472e+00 -3.51848068e-01  3.43107397e+00
  1.85781114e+00 -1.76265004e+00 -7.41626413e+00 -1.53973232e+01
 -1.75094413e+00  9.51038846e-01 -1.41560992e+00 -2.06597048e+00
  1.39668540e+00  4.14670701e+00  4.96195237e+00  1.02322261e+00
  3.27185521e+00 -1.39161177e+00  9.37581206e-01 -4.10349266e+00
  9.18334575e-01  2.47206371e+01 -1.68620870e+00  2.46893263e+00]
```

Prediction Model

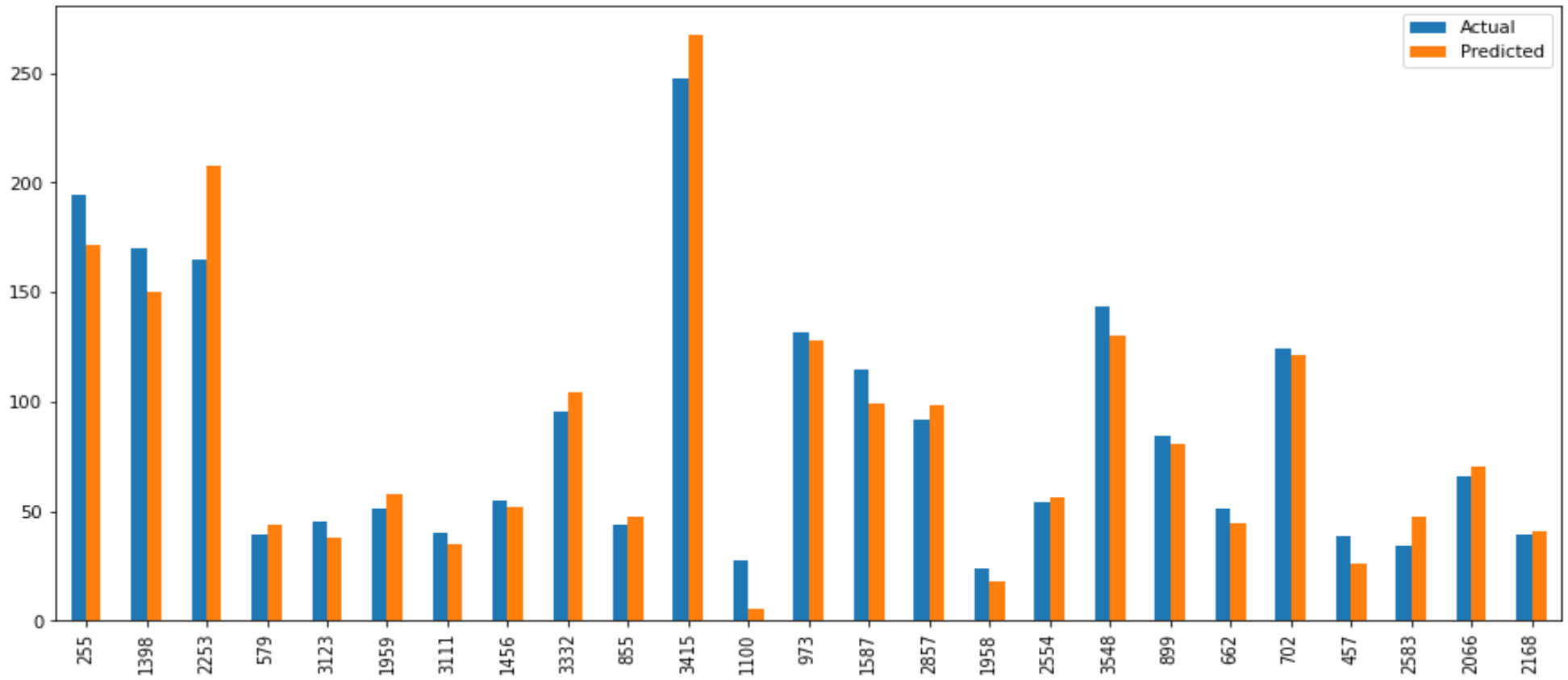
In order to build the model, the most important variables were:

- screen_size
- ram
- days_used
- new_price
- brand_name
- os
- int_memory

Comparison between Test and Training Data

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
Training Data	13.793	10.186	0.957	0.956	17.925
Test Data	14.269	10.338	0.953	0.953	17.945

Model Output



1.The model is able to explain ~95% of the variation in the data, which is very good.
2.The train and test RMSE and MAE are low and comparable. So, our model is not suffering from overfitting.
3.The MAPE on the test set suggests we can predict within 17.9% of the used price.
4.Hence, we can conclude the model *olsmod2* is good for prediction as well as inference purposes.

Conclusion

I analysed the variables, keeping *used_price* as the dependent variable:

- The *new_price* comes out to be very significant, as expected. As it increase, the *used_price* increases, as is visible in the positive coefficient sign.
- *days_used* has a negative coefficient meaning as the number of days used increases, the price decreases. This is as expected.
- Internal factors like *int_memory* and *ram* have a positive correlation with the the *used_price*.
- iOS is more ~2x higher than the next OS which is Android thus indicating that Apple has the highest used_price.
- Even though the highest iOS has the highest resale value, it has the smallest number of devices indicating that it is not a readily available phone for resale. This scarcity possibly contributes to its high price.
- *screen_siz*, *selfie_camera_mp*, *int_memory*, *ram* and *new_price* (referring to the original price of the phone) have positive coefficients. So, as they increase, *used_price* also increases.

Recommendation

Based on the analysis, there are following recommendations that can help the business grow:

- The refurbished phones should be split into buckets that address different segments of the community where they are to be sold
- Partnerships with telecom networks should be explored for deals that can be given to the customers
- Partnerships with phone makers should be explored so that older model phones can be bough directly from the supplier thus ensuring a certain standard. These can be marketed differently and sold at a higher price as the case may be
- This business can be done in places where there is high foot traffic in order to get people to come in and physically inspect the phone
- Online marketing is also very important because as soon as new phone model comes out, the older ones become more obsolete. However certain brands retain their values longer

However, for a more robust model, the following features can be considered

- Cleanliness of the phone body
- Model of the phone
- Region phone is manufactured for as different regions have different phone specifications
- Type and number of accessories that come with the phone e.g., earphones, chargers, screen protectors etc
- Original phone parts i.e., has the phone been repaired before.