greatlearning
Power Ahead

Recell Business Case - Analysis

SARAVANAN RANGARAJAN (SR)

Contents

- Business Problem Overview
- Solution Approach
- Data Overview
- Distribution of Used phone prices
- RAM across Brands
- Used Phones having Android Phone
- Weight vary for phones with larger Batteries.
- Screen Size across brands
- Budget phones offering greater than 8 MP across Brands
- Highly Correlated with used phone price.
- Model Summary – Initial
- Assumptions for Linear Regression
- Model Summary – Final
- Model Performance Summary
- Conclusions
- Insights - Recommendations
- Questions

Business Problem Overview

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.

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.

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 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.

Solution Approach

Per the requirement is to perform Model building and perform Linear Regression and provide actionable insights.

Following are the steps to perform

1. Understand Dataset
2. Sanity Check
3. Missing values treatment
4. Outliers Detection and Treatment
5. EDA
6. Test & Train Model
7. Perform Linear Regression
8. Performance of model Validation
9. Checking Linear Regression Assumptions
10. Test for MultiCollinearity
11. Check VIF
12. Test for Linearity & Independence
13. Test for Normality
14. Test for Homoscedasticity

Data Overview

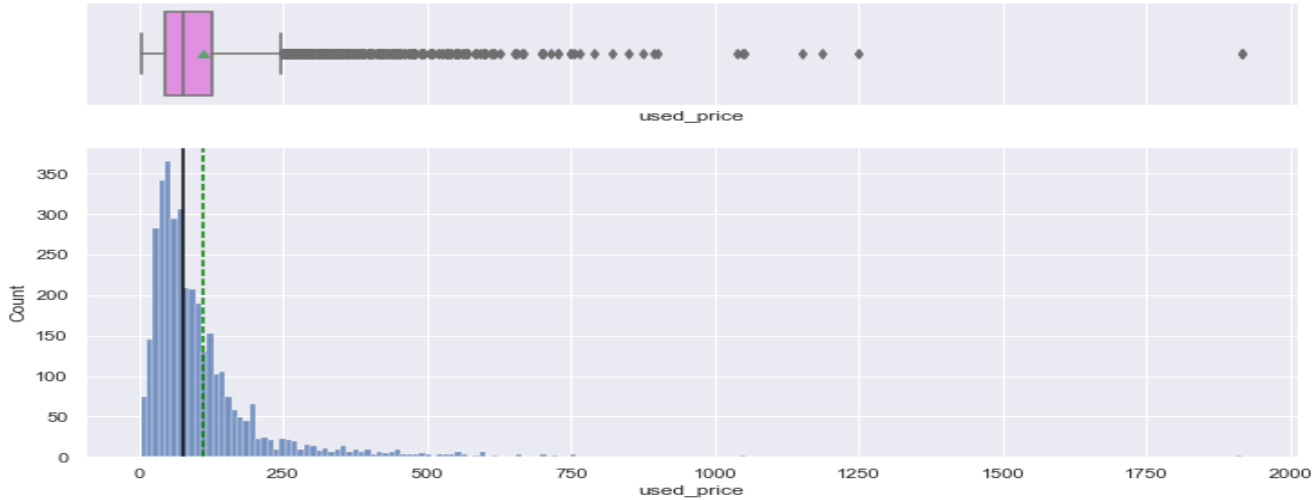
- **Data Dictionary**

- 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

Data Overview – Contd.

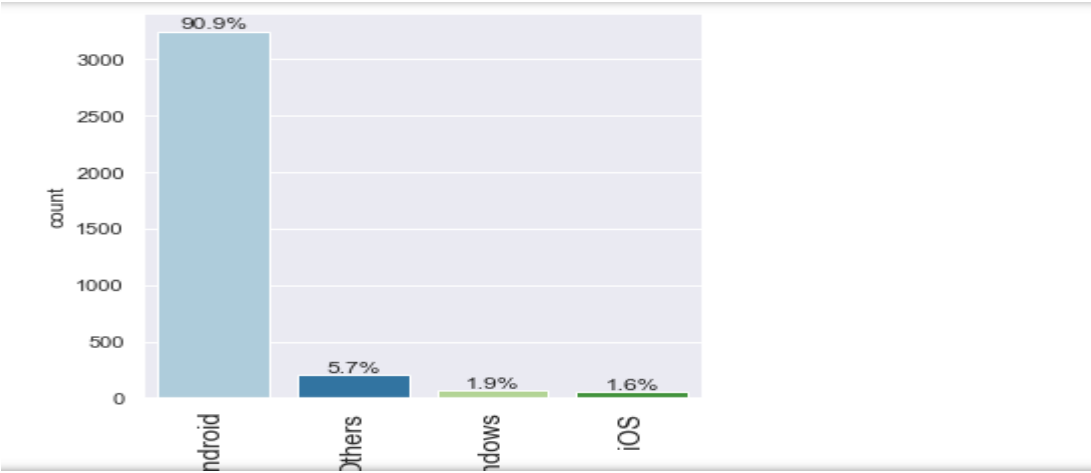
- It has 11 Numerical values & 4 Categorical values (Brand Name, OS,4G, 5G)
- Null checks have been done to check the data. Null values exist in Main Camera, Selfie Camera, Internal Memory, RAM, Battery & weight.
- Duplicate checks have been done. No Duplicate values exist.
- Check the distinct values in the Categorical columns and check the data spread for general information.
- Columns have, numerical value as start Character has to be renamed
- Missing Values has been replaced by Mean values, In some cases Last/Next Observation values have been imputed.
- One hot Encoding was used for 4G, 5G and OS columns.
- Due to Business requirements, Screen size was converted to Inches from CMs for model study.
- Outlier Treatment was done using IQR value, Capping & Flooring.

1.What does the distribution of used phone prices look like?



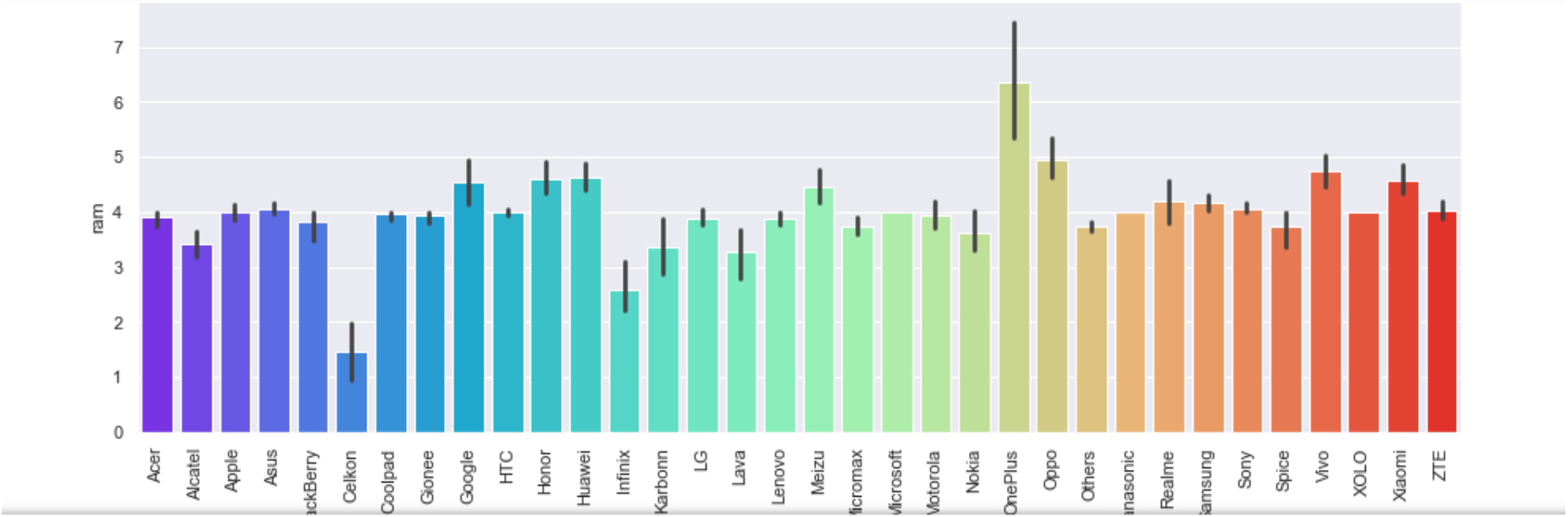
- Distribution is heavily right skewed
- Outliers present mostly towards right
- Mean is very much greater than Median
- Most values (50%) are between 45 and 126

2.What percentage of the used phone market is dominated by Android devices?



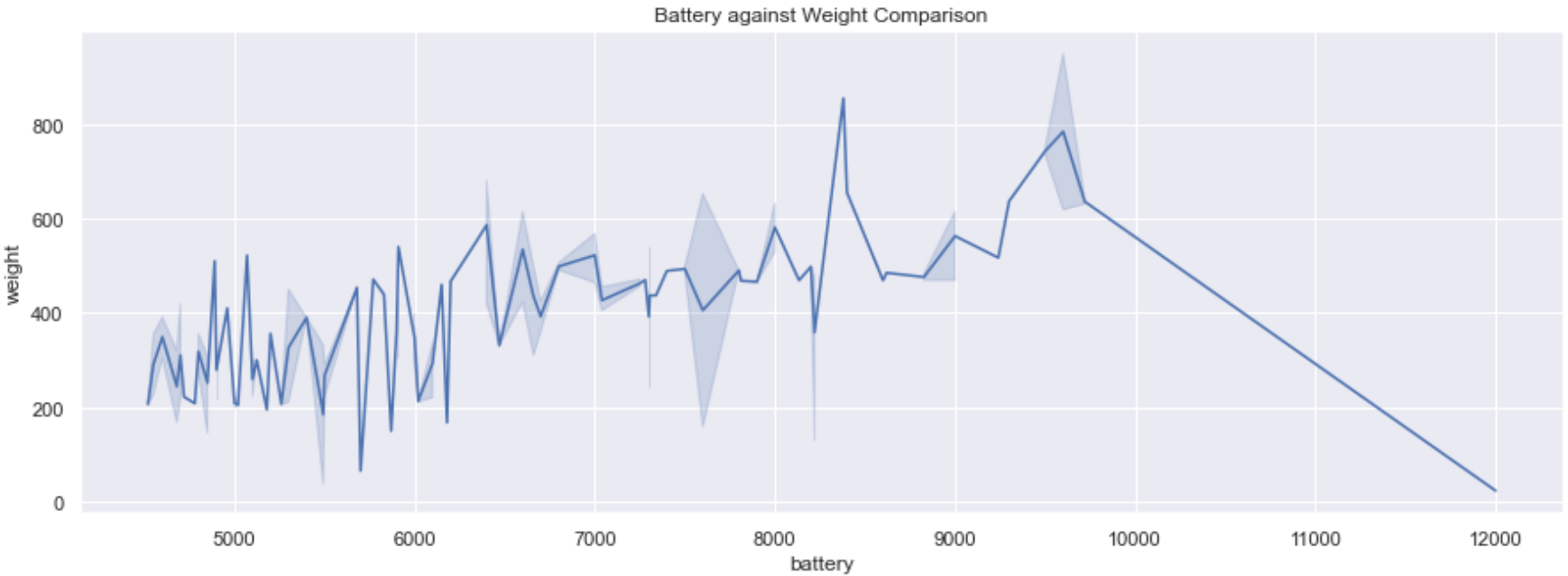
- Question can be answered by Bar plot with percentage label.
- Chart explicitly shows Android as major player with 90.9%

3. How does the amount of RAM vary with the brand?



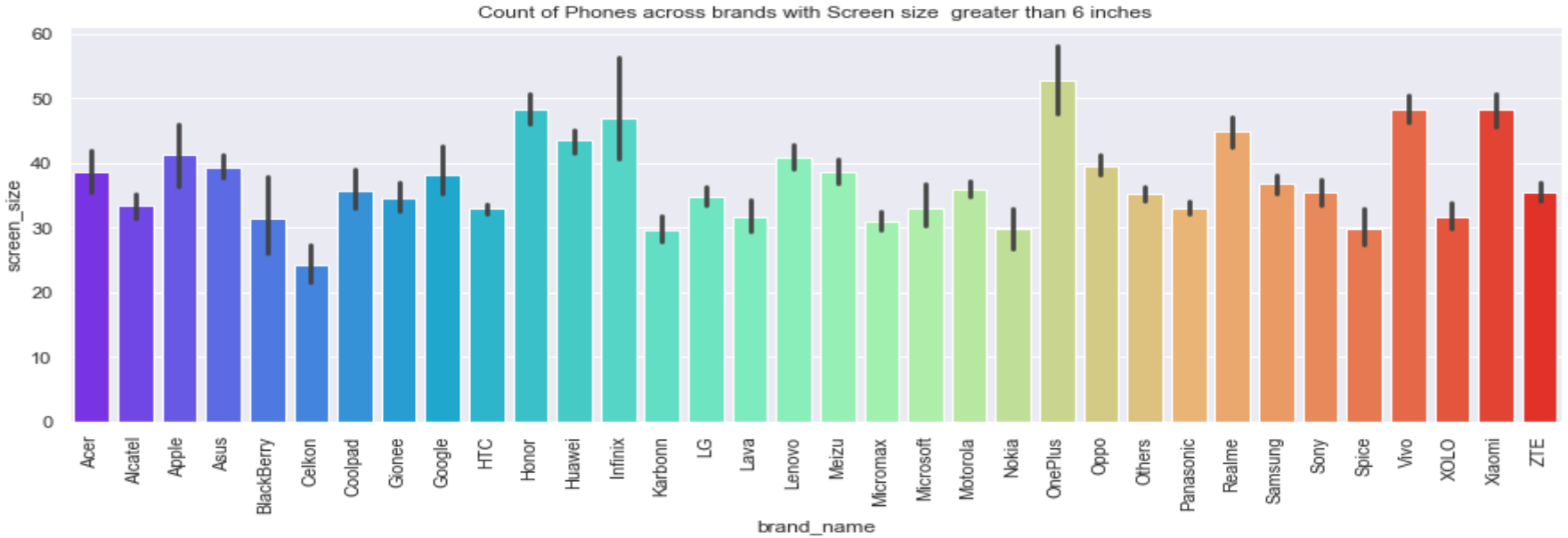
- Business Question can be answered by plotting Bar plot with Brand name in X axis and RAM value in Y-axis.
- RAM values varies between 1.5 to 6 GB, 4 GB being the most common value.

4. How does the weight vary for phones offering large batteries (more than 4500 mAh)?



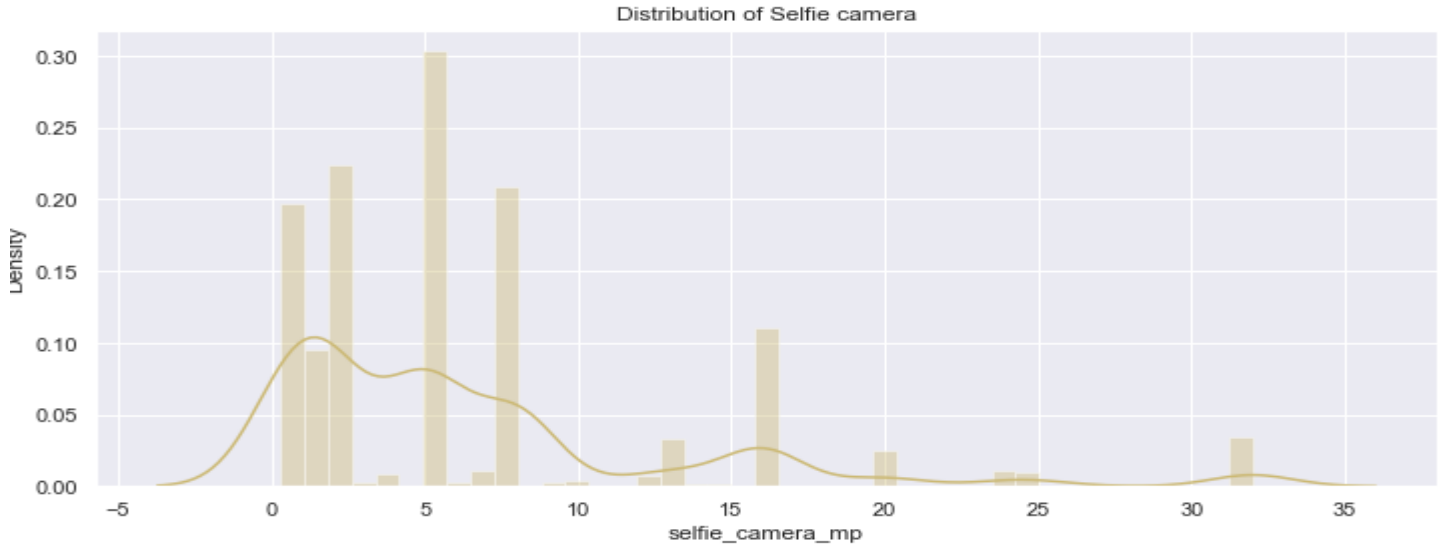
- Phone battery does influence Battery weight as we observe some chart
- Too heavier phones are difficult to carry, some outlier data may be present.

5. How many phones are available across different brands with a screen size larger than 6 inches?



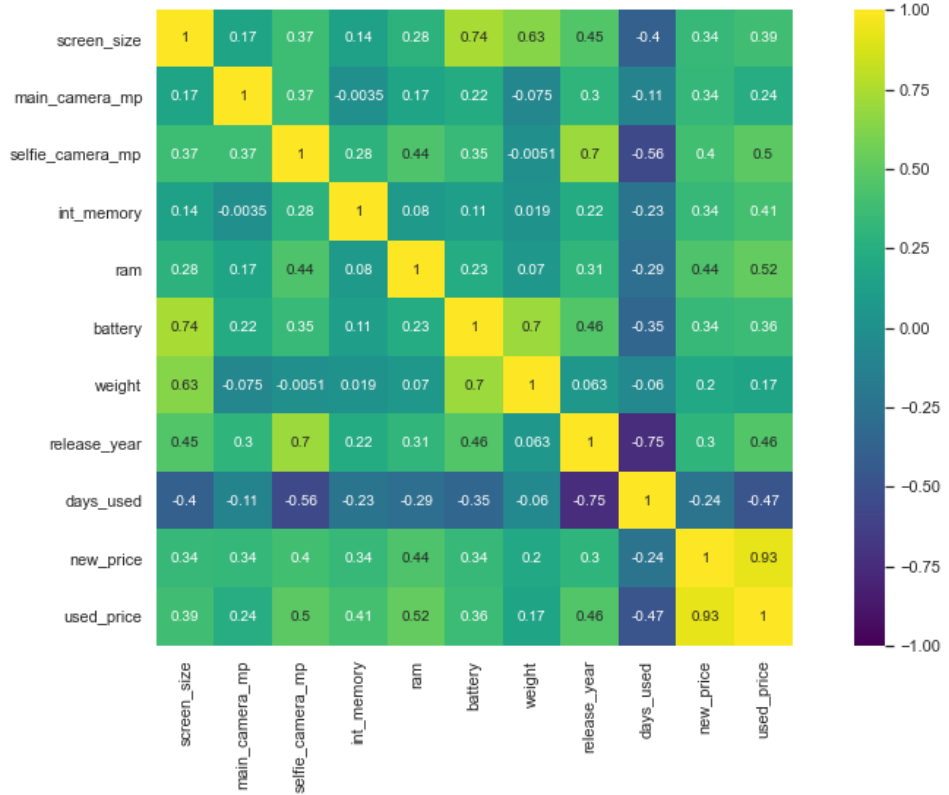
- Oneplus seems to have big screen size
- Samsung has more phones, followed by Huawei and LG
- Inifinix, Google seems to have lesser
- Screen size on average between 6" and 8"
- Some Brands have larger screen like from Chart we can infer Microsoft, Blackberry and Oppo etc.

6. What is the distribution of budget phones offering greater than 8MP selfie cameras across brands?



- Mean is 18.68 and Max value is 32. Most of Phones have selfie pixel as 16 Mp

7. Which attributes are highly correlated with the used phone price



- As expected used_price has -ve corelation with days_used
- Screen size and battery has high correlation
- Weight and screen size has high correlation as well as weight and battery.
- Price values new and Used have highest correlation
- Ram and Used price also seems to have correlation

Model Analysis Summary - Initial

- Initial Data size 3571 rows with 15 Columns
- After preprocessing 49 columns (due to dummy values for Categorical values) and 3571 rows was the size taken for model analysis.
- None of rows were dropped.
- Model was split in to 70:30 ration for train and testing.
- The Target variable in this case was “used_price”
- Number of rows in train data 2499 and in test data 1072
- Initial Model Performance
- Training Data

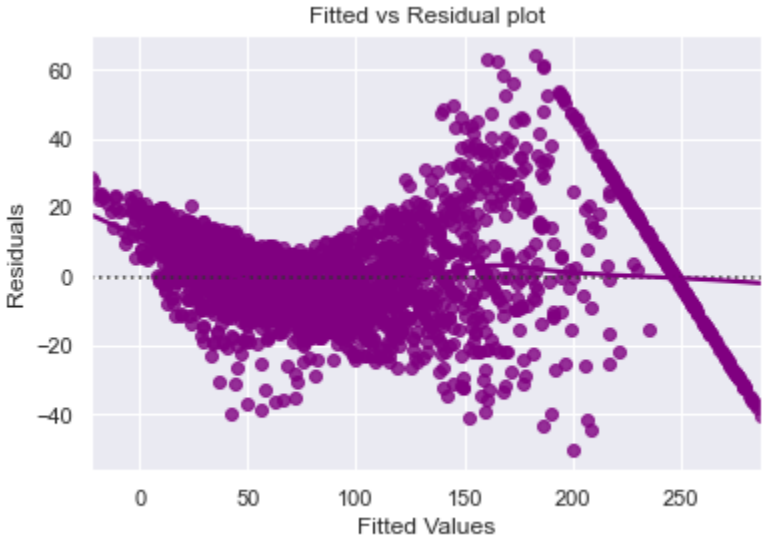
	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	13.96	10.22	0.96	0.95	18.49

- Testing Data

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	13.75	10.17	0.96	0.96	16.42

- The training $R2$ is 96%, indicating that the model explains 96% of the variation in the train data. So, the model is not underfitting.
- MAE and RMSE on the train and test sets are comparable, which shows that the model is not overfitting.
- MAE indicates that our current model is able to predict used phone prices within a mean error of ~10 currency units on the test data.
- MAPE on the test set suggests we can predict within ~16.42% of the used phone prices.

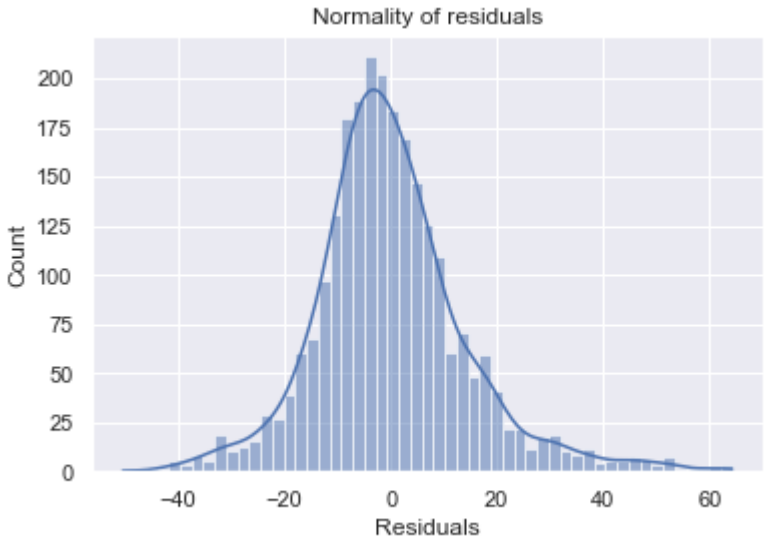
Test for Linearity & Independence

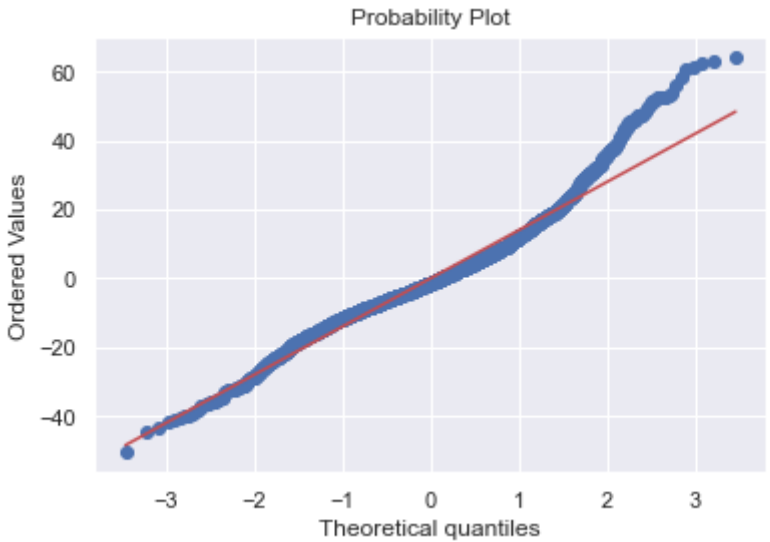


	Actual Values	Fitted Values	Residuals
844	100.48	99.74	0.74
1539	111.68	117.18	-5.50
3452	113.89	111.38	2.51
1727	64.09	70.54	-6.45
1926	67.95	68.20	-0.25

- The scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values).
- If there exist any pattern in this plot, we consider it as signs of non-linearity in the data and a pattern means that the model doesn't capture non-linear effects.
- We see no pattern in the plot above. Hence, the assumptions of linearity and independence are satisfied.

Test for Normality





- The histogram of residuals does have a bell shape.
- Let's check the Q-Q plot.
- The residuals more or less follow a straight line except for the upper tail.

Test for Homoscedasticity

- Testing Homoscedasticity using goldfeldquandt test.
- Stating Hypothesis and finding the p value.
- If p value >0.05, then Residuals are Homoscedasticity satisfying assumption.
- [('F statistic', 1.0570907089704333), ('p-value', 0.1640663013568954 3)]
- It states clearly all Linear Regression model assumptions are satisfied,

Model Analysis Summary - Final

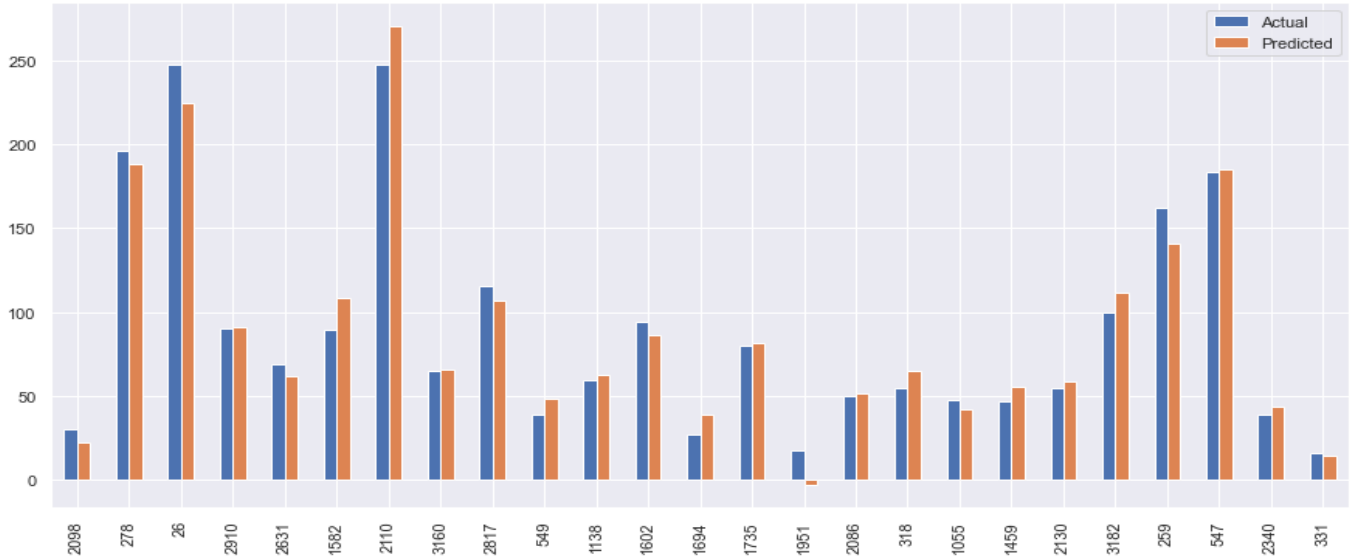
- Final Performance Comparison, Training vs Testing.

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	13.96	14.26
MAE	10.22	10.41
R-squared	0.96	0.95
Adj. R-squared	0.95	0.95
MAPE	18.49	18.81

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	13.75	13.90
MAE	10.17	10.18
R-squared	0.96	0.96
Adj. R-squared	0.96	0.96
MAPE	16.42	16.30

- The model is able to explain 95% of the variation in the data, which is very good.
- The train and test RMSE and MAE are and comparable. So, our model is not suffering from overfitting.
- The MAPE on the test set suggests we can predict within 16.3% of the used phone prices.
- Hence, we can conclude the model olsmod2 is good for prediction as well as inference purposes.
- Final Predictor Columns are 'main_camera_mp', 'selfie_camera_mp', 'int_memory', 'ram', 'days_used', 'new_price', 'os_Others', 'fourG_yes'

Model Performance Summary



	Actual	Predicted
2098	30.52	22.69
278	195.67	188.04
26	247.19	224.45
2910	89.97	91.09
2631	69.20	62.19
1582	89.58	108.67
2110	247.19	270.62
3160	65.34	65.55
2817	115.77	106.77
549	39.29	48.13

- We can observe here that the model has returned good enough prediction results, and the actual and predicted values are comparable.
- We can also visualize comparison result as a bar graph.

Model Insights

* If phone has better selfie camera it will increase the price of used phone
* If phone has better RAM, it will increase the price
* If phone has been used for more days, it will decrease the price
* 4G might decrease the value of phone, 5G would be better
* Other OS, apart from Droid, iOS would decrease the price.
* Internal Memory increase will increase the price of phone

Business Recommendations

Price factor of less than or equal to 600 Euros with RAM Size of 4GB or more, with at-least 4G would be a better features for the used phone to increase sales.

Questions

- Questions give different Perspective, most of the time result in idea.

greatlearning
Power Ahead

Happy Learning !