RECELL

USED CELL PHONE PRICE PREDICTION

BY: SYEDA AMBREEN KARIM BOKHARI

CONTENTS

1. Core business idea:
2. Business Problem Overview and Solution Approach
3. Data Overview:
4. Missing values treatment
5. Outliers treatment & Feature Engineering
6. Exploratory Data Analysis
7. Model Performance Summary
8. Assumptions of Linear regression
9. Final Model Performance Summary
10. Business Insights and Recommendations

CORE BUSINESS IDEA:

ReCell, a start-up aiming to tap the potential in the market, of used and refurbished phones which has grown considerably over the past decade, and as predicted by IDC (International Data Corporation), has a compound annual growth rate (CAGR) of 13.6% from 2018 to 2023.

BUSINESS PROBLEM OVERVIEW AND SOLUTION APPROACH

- Problem to tackle

  To analyse the provided data and build a linear regression model for prediction.

- Financial implications

  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.

- How to use ML model to solve the problem

  The linear regression model is trained and tested on the available data and can be used to predict with ~95% accuracy, the future price of a used phone and identify factors that significantly influence it.

Data Overview: The dataset file contains the used phone data with following specifications:

Variable name	Data types	Description	Missing values	Unique values
1. brand_name:	categorical	Name of manufacturing brand	0	34
2. os:	categorical	OS on which the phone runs	0	4
3. screen_size:	numeric	Size of the screen in cm	0	127
4. 4g:	categorical	Whether 4G is available or not	0	2
5. 5g:	categorical	Whether 5G is available or not	0	2
6. main_camera_mp:	numeric	Resolution of the rear camera in megapixels	180	44
7. selfie_camera_mp:	numeric	Resolution of the front camera in megapixels	2	37
8. int_memory:	numeric	Amount of internal memory (ROM) in GB	10	16
9. ram:	numeric	Amount of RAM in GB	10	14
10. battery:	numeric	Energy capacity of the phone battery in mAh	6	354
12. weight:	numeric	Weight of the phone in grams	7	613
12. release_year:	numeric	Year when the phone model was released	0	8
13. days_used:	numeric	Number of days the used/refurbished phone has been used	0	930
14. new_price:	numeric	Price of a new phone of the same model in euros	0	3099
15. used_price:	numeric	Price of the used/refurbished phone in euros	0	3044

DATA OVERVIEW: MISSING VALUES TREATMENT

- Brief description of significant manipulations made to raw data

Observations	Variables	Missing	Dependent variable
3571	15	215	used_price

Variable name	Missing data description	Treatment
main_camera_mp: selfie_camera_mp: battery weight:	205 cells have missing values which didn't show any pattern.	Mean of these fields were imputed in the empty cells.
int_memory, ram: battery	Block of 20 values were missing, from the same rows in columns int_memory and ram and 2 were missing in battery as well.	Last Observation Carried Forward (LOCF)were imputed as it was a block of 10 rows with same brand and similar specifications. Next Observation Carried Backward (NOCB)were imputed in battery in these rows

DATA OVERVIEW

- Brief description of significant manipulations made to raw data

Observations	Variables	Missing	Dependent variable
3571	15	215	used_price

Categorical Variables: Encoding into numeric.	
Variable name	**Treatment**
brand_name	Frequency encoding was used to as 34 unique values were present.
4g, 5g, os	One Hot Encoding was used as there were fewer unique values.

Variable name	Feature Engineering
screen_size	Converted from cm to inches

DATA OVERVIEW: OUTLIERS TREATMENT & FEATURE ENGINEERING

- Brief description of significant manipulations made to raw data

Variable name	Outliers detection and treatment	Standardization of variable
• int_memory, • main_camera_mp: • selfie_camera_mp: • battery • Weight • new_price • used_price	• IQR was used to detect outliers in all the numeric fields. • Outliers in the data were treated by flooring and capping.	• StandardScaler of sklearn library was used on all independent numeric variables to standardise all the variables to similar scale. • Normalization was done by using MinMaxScaler of sklearn library, so that distributions would be more Gaussian shaped.
• Screen_size • Weight • ram:	• Screen size had very big values so the data may be erroneous so they were removed. • Weight had very big values so the data may be erroneous so they were removed. • ram had majority values around 4 so outliers >8 were removed.	

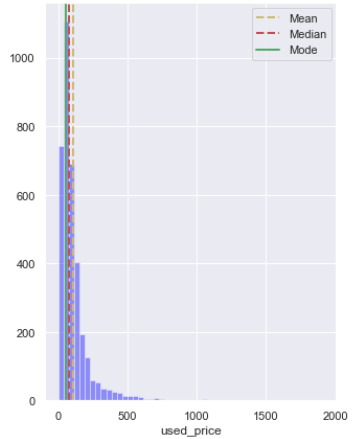
EXPLORATORY DATA ANALYSIS

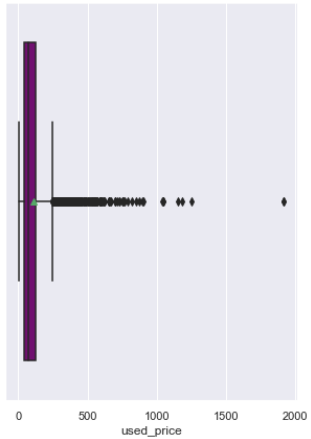
- Graphs and observation about the target attribute:

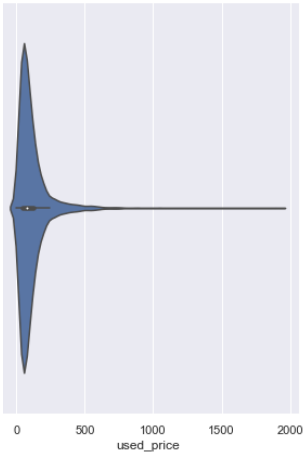
Observations

- It is the target variable.
- The distribution of used_price is heavily skewed to the right.
- The outliers to the right indicate that many cell phones, though used, have a very high Prices.
- Mean is 109.9 much greater than median due to extreme values towards higher end.
- There are a lot of outliers towards right, larger side.Majority of values 50% are between 45 and 126

SPREAD OF DATA FOR USED_PRICE







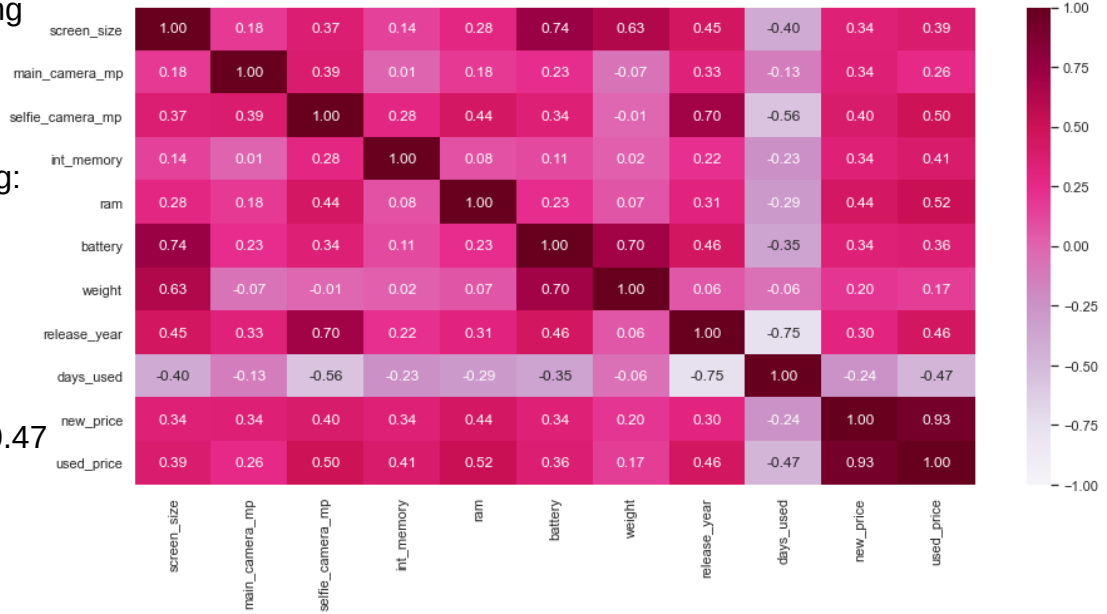
EDA

- Graphs showing the factors most heavily impacting the target attribute

Observations:

used_price has significant correlation with the following:

- high positive correlation with new_price 0.93
- Positive Correlation with ram 0.52
- Positive Correlation with selfie_camera_mp: 0.50
- somewhat negative correlation with days_used: -0.47

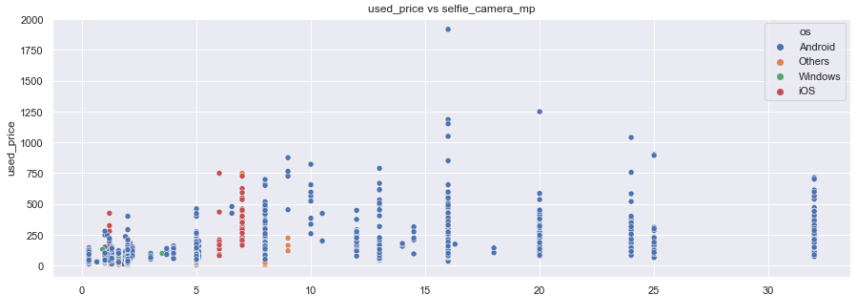


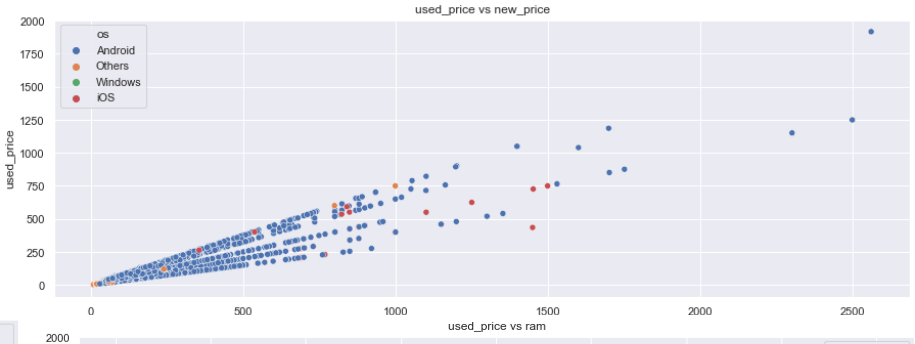
These fields also have some correlation:

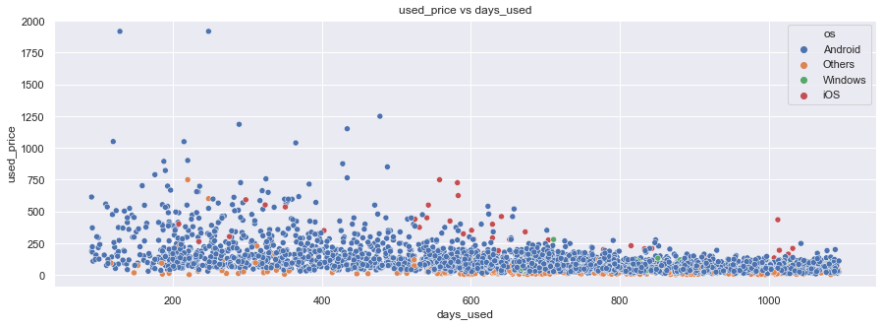
- release_year and selfie_camera_mp have positive correlation: 0.7
- weight and screen_size: 0.63,
- weight and battery: 0.7
- battery and screen size also have positive high correlation: 0.74
- days_used has high negative correlation with selfie_camera_mp: -0.56

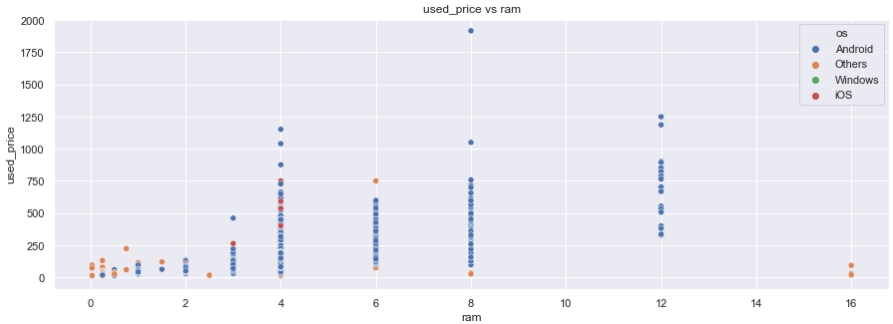
EDA

- Graphs showing the factors most heavily impacting the target attribute





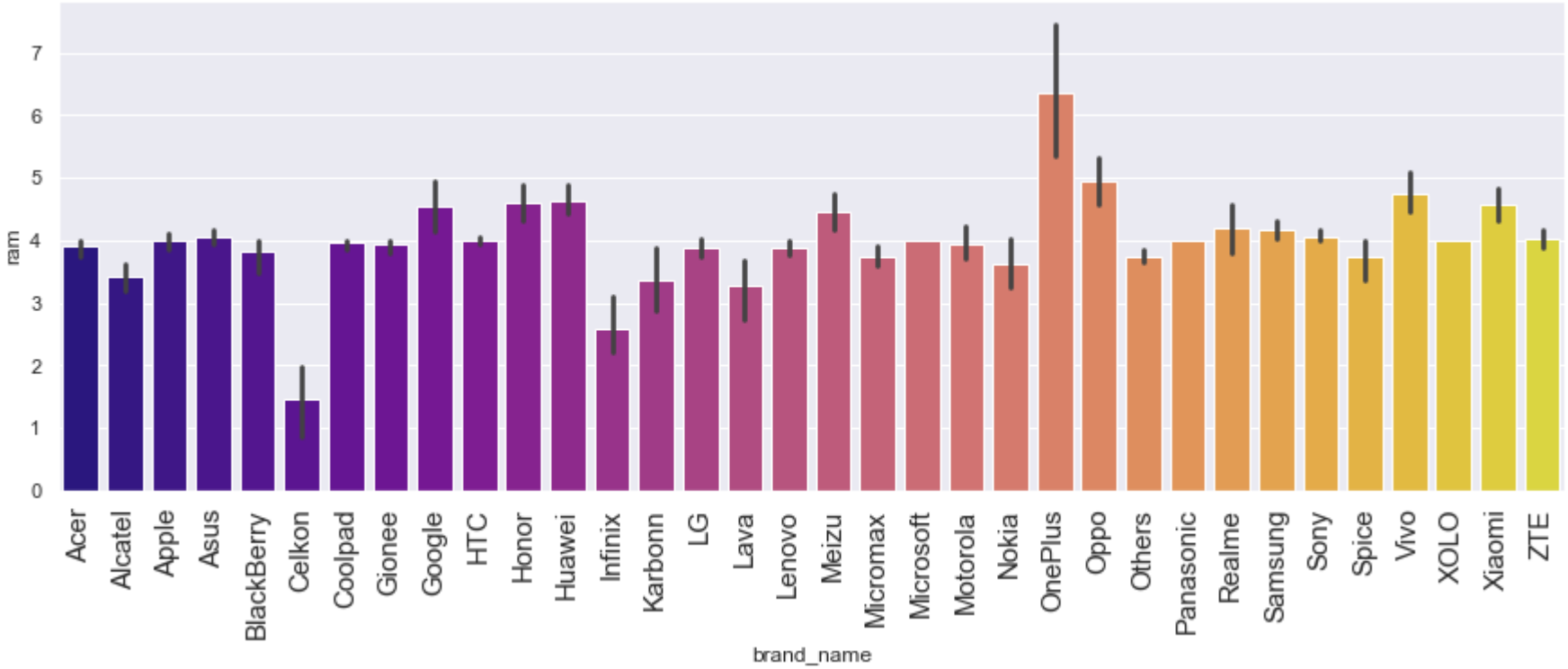




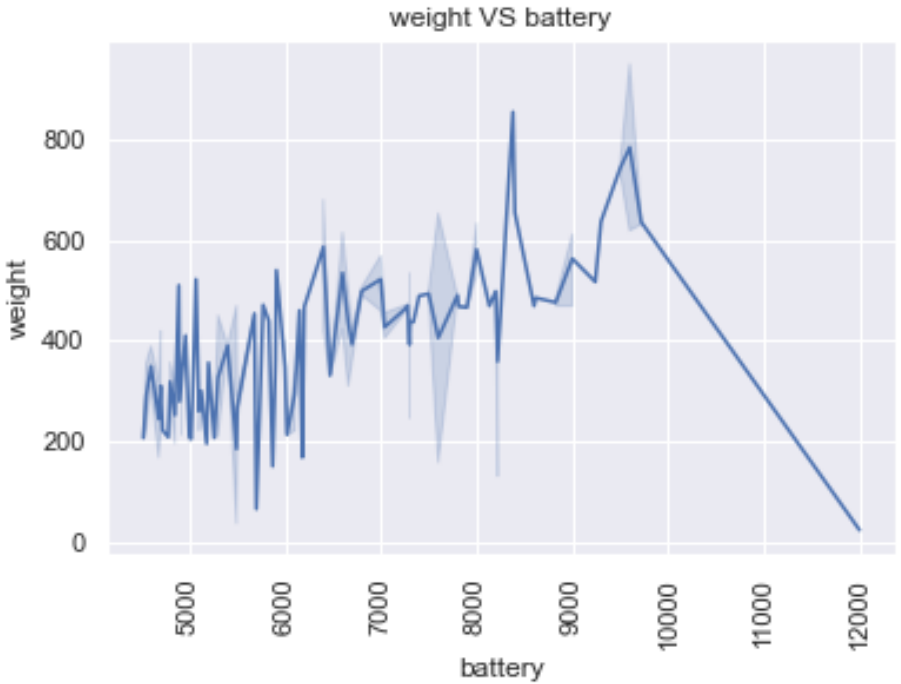
- Used_price increases if new_price is higher
- Used_price increases if ram is higher
- Used_price increases if selfie_camera_mp is better
- Used_price decreases is days_used is higher

HOW DOES THE AMOUNT OF RAM VARY WITH THE BRAND?

- brands vary in ram in range 1.5-6GB or 4gb ram on average



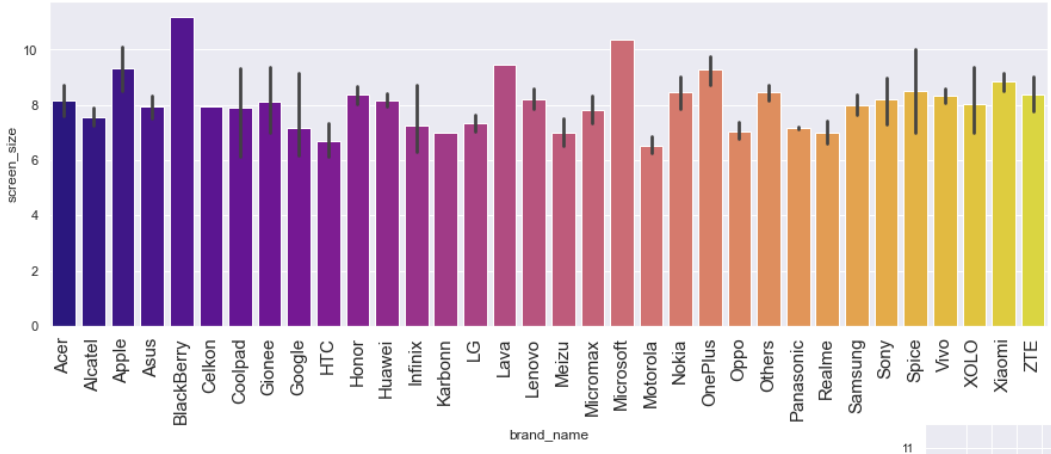
HOW DOES THE WEIGHT VARY FOR PHONES OFFERING LARGE BATTERIES (MORE THAN 4500 MAH)?

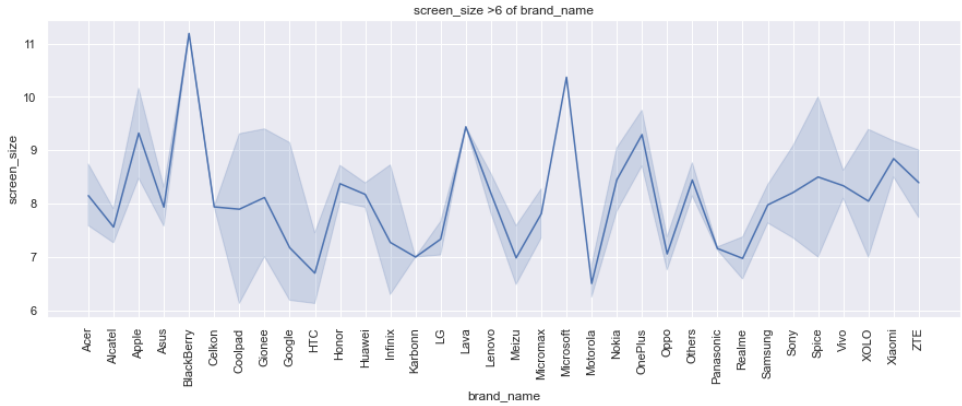


- On average phones with higher battery also have higher weight.
- Some phones have weight too high to be a cell phone weight. Cell phone heavier than 600 grams would be too heavy, maybe their units are wrong or data is erroneous.

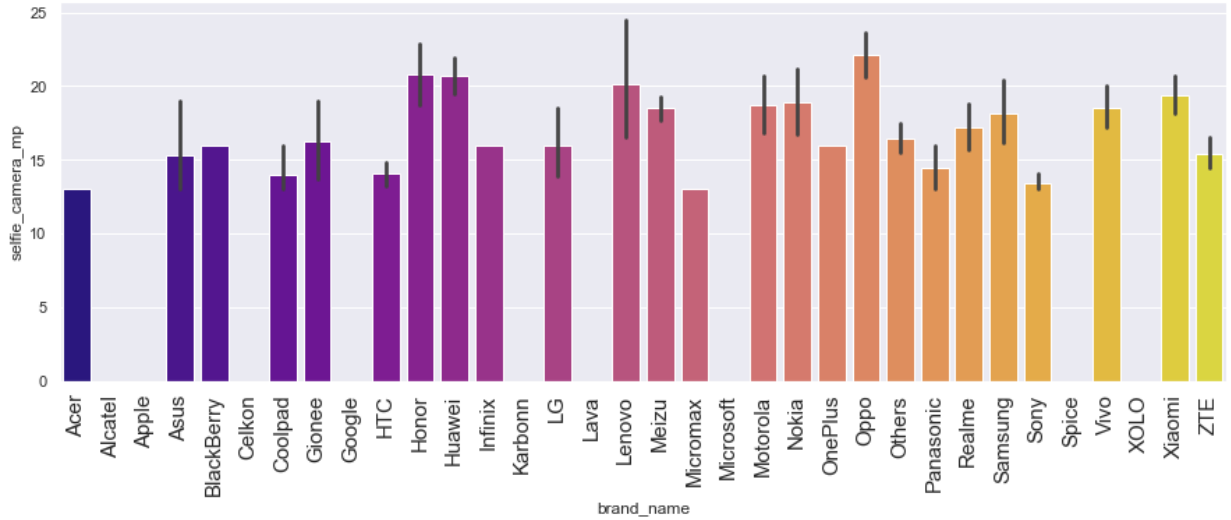
HOW MANY PHONES ARE AVAILABLE ACROSS DIFFERENT BRANDS WITH A SCREEN SIZE LARGER THAN 6 INCHES?

- Screen size on average are between 6"and 8".
- There are also some very big screen size present, like blackberry 11"+ and Microsoft 10"+, which may be some error in the data.

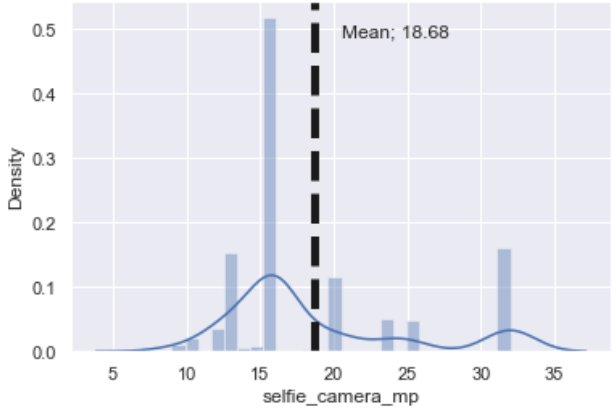




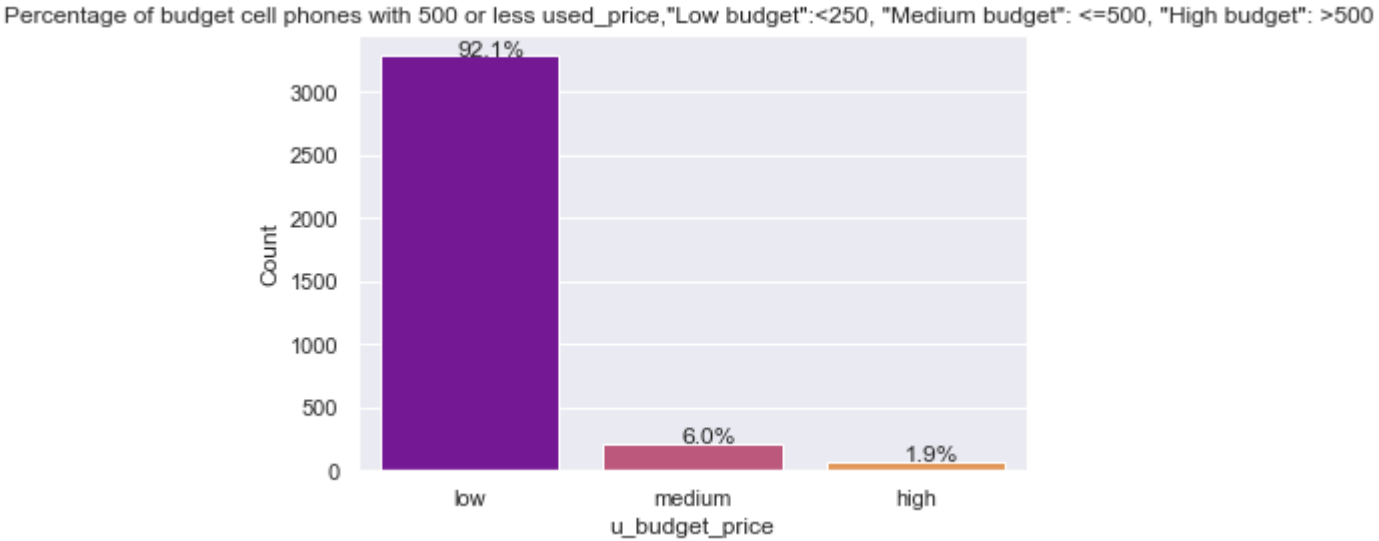
WHAT IS THE DISTRIBUTION OF BUDGET PHONES OFFERING GREATER THAN 8MP SELFIE CAMERAS ACROSS BRANDS?



- Specifications of selfie_camera_mp > 8mp
- max is 32mp
- mean is 18.7mp
- 50% of these phones have under 16mp



WHAT PERCENTAGE OF THE USED PHONE MARKET IS DOMINATED BY LOW BUDGET PHONES?



- Majority:92.1% (3289 out of 3571) phones come under the low budget phone: (have price 250 or less)
- Majority:6.0% (215 out of 3571) phones come under the medium budget phone: (have price 250 or less)
- Minority:only 1.9% (67 out of 3571) phones come under the high bidget phone: (have price above 500)

MODEL PERFORMANCE SUMMARY

Overview of ML model and its parameters:

- Multiple Linear Regression model was built to
  - find dependency of target variable: used_price on predictors and
  - Predict fitted values and compare them to actual values
- **Total rows and columns** after data preprocessing: 3188 x 17
- **Target variable:** used_price
- **Predictors:** 'screen_size', 'main_camera_mp', 'selfie_camera_mp', 'int_memory', 'ram', 'battery', 'weight', 'release_year', 'days_used', 'new_price','brand_name_encode', 'four_g_yes', 'five_g_no', 'os_Android', 'os_Others', 'os_iOS'
- Data was divided into Train and Test at 70:30 ratio.
- Number of rows in train data = 2231
- Number of rows in test data = 957

MODEL PERFORMANCE SUMMARY

The first ML model was tested using the following performance matrices:

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	13.430	9.972	0.956	0.956	18.639

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	14.057	10.285	0.956	0.955	20.586

	Coefficients
screen_size	0.473
main_camera_mp	-0.355
selfie_camera_mp	0.695
int_memory	0.021
ram	2.258
battery	-0.001
weight	0.028
release_year	0.003
days_used	-0.085
new_price	0.391
brand_name_encode	0.905
four_g_yes	-2.226
five_g_no	1.816
os_Android	1.670
os_Others	-0.120
os_iOS	12.772
Intercept	42.629

Observations

- The testing R^2 is 95.6%, indicating that the model explains 95.6% 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 Euros on the test data.
- MAPE on the test set suggests we can predict within ~20.6% of the used phone prices.

ASSUMPTIONS OF LINEAR REGRESSION

We will be checking ML model on Linear Regression assumptions:

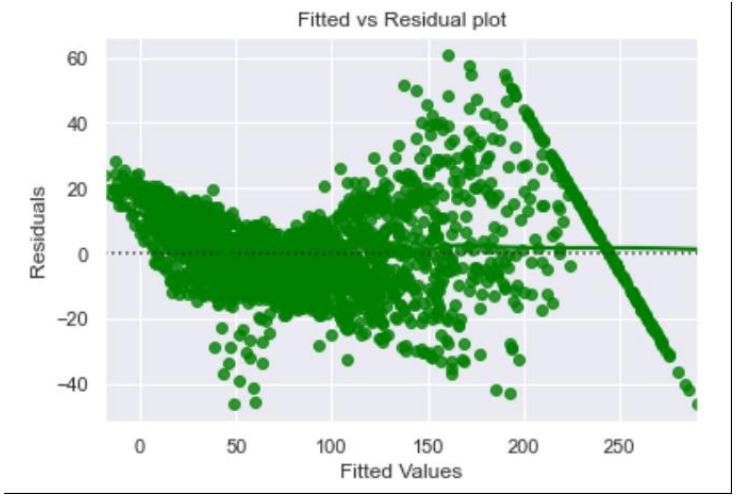
1. No Multicollinearity: final ML model features did not have VIF> 5

	feature	VIF
0	const	31.207
1	selfie_camera_mp	2.113
2	int_memory	1.147
3	ram	1.492
4	days_used	1.488
5	new_price	1.737
6	four_g_yes	1.464
7	os_iOS	1.077

- No significant multicollinearity

ASSUMPTIONS OF LINEAR REGRESSION

2. Linearity of variables & Independence of error terms

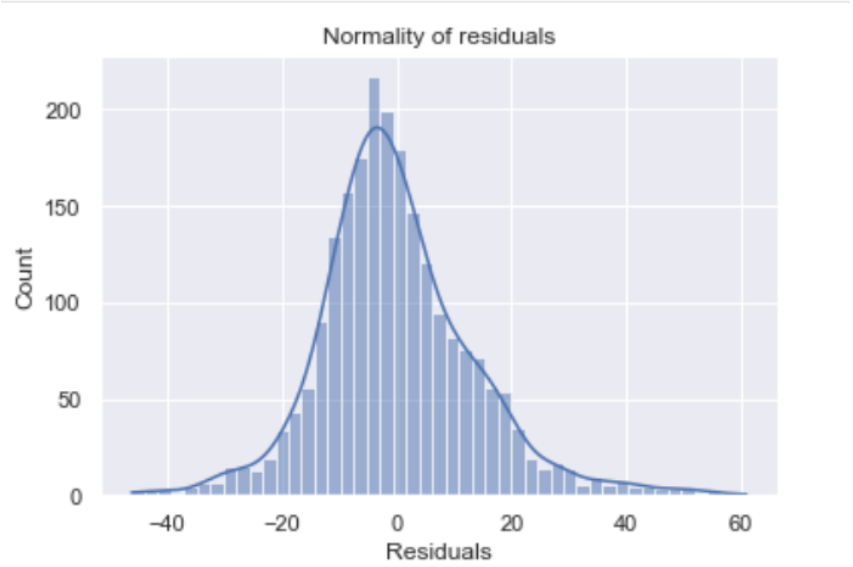


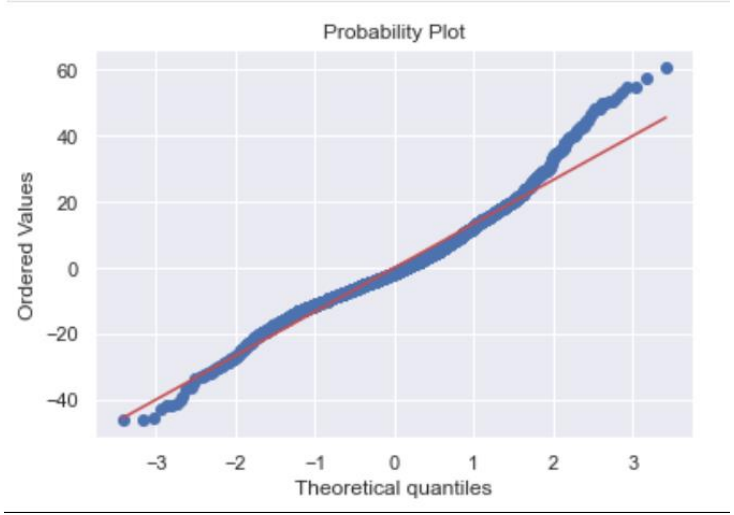
	Actual Values	Fitted Values	Residuals
776	103.940	94.076	9.864
1451	60.250	55.093	5.157
1812	65.690	60.086	5.604
2970	232.260	218.623	13.637
2741	132.110	156.819	-24.709

- The scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values).
- We see no pattern in the plot above.
- **Hence, the assumptions of linearity and independence are satisfied.**

ASSUMPTIONS OF LINEAR REGRESSION

3. Normality of error terms





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

ASSUMPTIONS OF LINEAR REGRESSION

4. No Heteroscedasticity
Goldfeldquandt test

- H_0= Null hypothesis: Residuals are homoscedastic
- H_1= Alternate hypothesis: Residuals have heteroscedasticity

shows p > 0.05,
Residuals are homoscedastic

```
[('F statistic', 0.9465894859592179), ('p-value', 0.8193959059441539)]
```

- The p-value is ~0.82 > 0.05, hence we can say that the residuals have constant variance.
- **Hence we can say that all the assumptions of our linear regression model are satisfied.**

FINAL MODEL PERFORMANCE SUMMARY

The final ML model was tested using the following performance matrices:

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	13.520	10.043	0.955	0.955	18.883

Test Performance

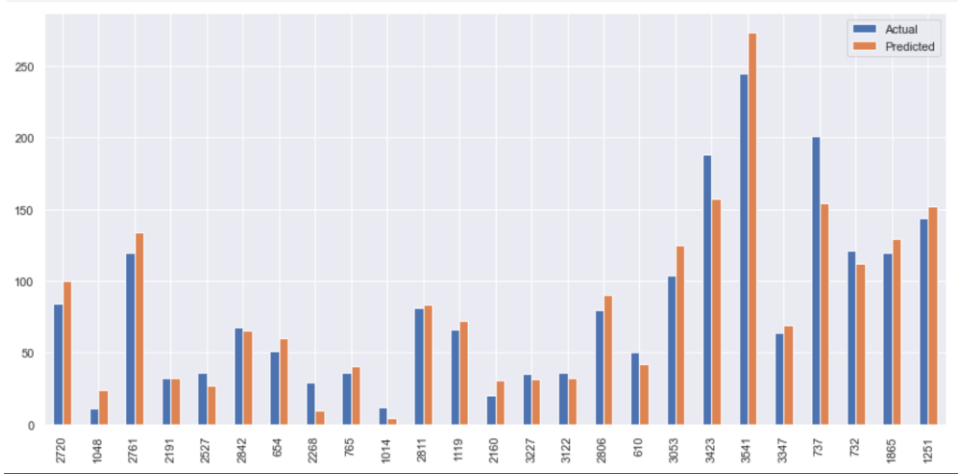
	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	14.080	10.304	0.956	0.955	20.657

Observations

- The testing R^2 is 95.6%, indicating that the model explains 95.6% 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 Euros on the test data.
- MAPE on the test set suggests we can predict within ~20.6% of the used phone prices.

FINAL MODEL PERFORMANCE SUMMARY

The final ML model performance comparison: Actual used_price vs Predicted used_price



	Actual	Predicted
2720	84.540	100.380
1048	10.980	23.895
2761	119.710	133.976
2191	31.980	32.024
2527	35.800	27.008
2842	67.760	65.123
654	50.920	59.858
2268	29.090	9.554
765	36.120	40.868
1014	11.960	4.605

Observations

- 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.
- Note: As the number of records is large, for representation purpose, we are taking a sample of 25 records only

- Training model performance comparison with stats model

Training performance comparison:

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	13.430	13.520
MAE	9.972	10.043
R-squared	0.956	0.955
Adj. R-squared	0.956	0.955
MAPE	18.639	18.883

- The performance of both the models is very similar.

FINAL MODEL PERFORMANCE SUMMARY

Summary of most important factors used by the ML model for prediction

The final predictor variables for the model are following:

All these variables have probability $p<0.05$ indicating these are all significant variables.

1. 'int_memory': increase in memory increases the used phone price.
2. 'ram': increase in memory increases the used phone price.
3. 'new_price': increase in memory increases the used phone price.
4. 'os_iOS': if phone is having ioS, it increases the used price of phone
5. 'selfie_camera_mp':if phone is having better mp selfie camera, it increases the used price of phone
6. 'days_used': if phone is used for longer time, it decreases the used price of phone
7. 'four_g_yes':if phone is having 4-g technology, it decreases the used price of phone, may be because newer phones have 5-g.

BUSINESS INSIGHTS AND RECOMMENDATIONS

Recommendations based on interpretation of the model input variables

- Company can check used phones with good conditions for
    - new_price <= 600
    - ram>=4,
    - selfie_camera>=8mp
    - days_used is not more than 700,
    - with at least 4-g technology present.
      and then they can sell these phones.
- Most used cell phones are Android devices, they can put them on deal or bundle offers to increase sales.
- ioS is mostly in Apple devices which can be sold as a higher end devices due to its durability and higher new_price.

BUSINESS INSIGHTS AND RECOMMENDATIONS

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

- Screen_size, weight variables should accept numbers which are within the correct range of mobile specification.

- It should have a limit check.

THE END

BY: SYEDA AMBREEN KARIM BUKHARI