RE-CELL –
USED PHONES PRICING STRATEGY
MACHINE LEARNING(SUPERVISED LEARNING)

ARPAN DINESH

05/08/2023

CONTENTS

EXECUTIVE SUMMARY

The regression analysis conducted for Re-Cell, a key player in the used mobile phone market, revealed significant factors influencing the normalized used price.

Screen size, camera specifications, internal memory, RAM, years since release, brand names, and connectivity features were found to impact prices.

Leveraging these insights, Re-Cell can optimize pricing, develop feature-rich devices with competitive new prices, and capitalize on trusted brand associations. Integrating the latest connectivity options, including 5G, can enhance market competitiveness.

These findings empower Re-Cell to make informed decisions and maximize value in the used mobile phone market.

BUSINESS PROBLEM

ReCell, a startup operating in the used and refurbished device market, is the need to develop a dynamic pricing strategy for their products.

The objective is to leverage machine learning (ML) techniques, specifically building a linear regression model, to predict the price of used phones and tablets. Furthermore, ReCell aims to identify the key factors that significantly influence the pricing of these devices.

By solving this problem, ReCell can optimize their pricing decisions, improve profitability, and gain a competitive advantage in the growing market of used and refurbished devices.

SOLUTION APPROACH

The Re-Cell project involves developing a dynamic pricing strategy for used and refurbished devices.

The solution approach includes exploratory data analysis (EDA) to understand the data, data preprocessing for cleaning and feature engineering, and building a linear regression model to predict device prices.

Assumptions of the linear regression model are tested, and the model's performance is evaluated using various metrics.

Actionable insights and recommendations are provided based on the model’s results.

DATA OVERVIEW

The data is collected from 2021 and contains different attributes of used devices. The description of each variable is given below. The dependent variable is the new price whilst the rest are all independent variables.

- brand_name: Name of manufacturing brand
- os: OS on which the device 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 device battery in mAh
- weight: Weight of the device in grams
- release_year: Year when the device model was released
- days_used: Number of days the used/refurbished device has been used
- normalized_new_price: Normalized price of a new device of the same model in euros
- normalized_used_price: Normalized price of the used/refurbished device in euros

First 5 rows of data

	brand_name	os	screen_size	4g	5g	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	release_year	days_used	normalized_used_price	normalized_new_price
0	Honor	Android	14.50	yes	no	13.0	5.0	64.0	3.0	3020.0	146.0	2020	127	4.307572	4.715100
1	Honor	Android	17.30	yes	yes	13.0	16.0	128.0	8.0	4300.0	213.0	2020	325	5.162097	5.519018
2	Honor	Android	16.69	yes	yes	13.0	8.0	128.0	8.0	4200.0	213.0	2020	162	5.111084	5.884631
3	Honor	Android	25.50	yes	yes	13.0	8.0	64.0	6.0	7250.0	480.0	2020	345	5.135387	5.630961
4	Honor	Android	15.32	yes	no	13.0	8.0	64.0	3.0	5000.0	185.0	2020	293	4.389995	4.947837

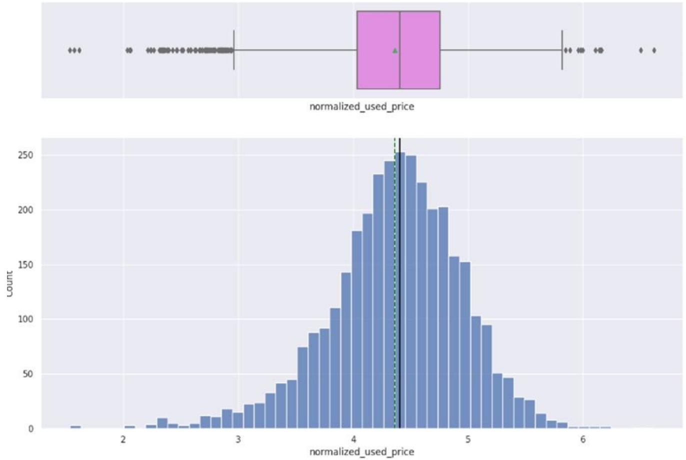
EDA - UNIVARIATE(PRICES)

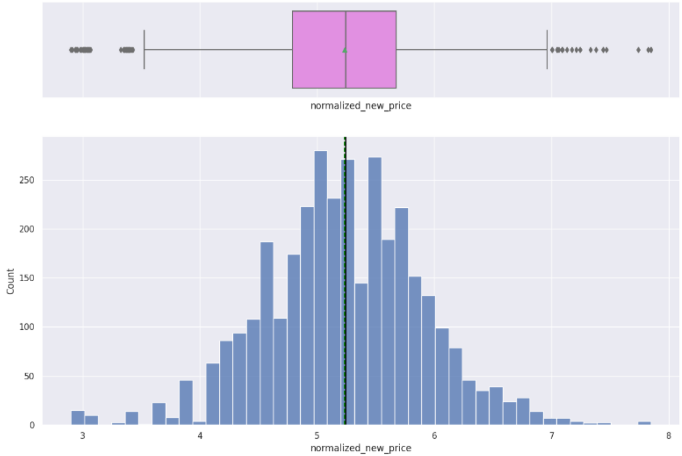
- The used price follows a normal distribution with a mean of roughly 4.5 euros.

- Almost 75% of used cells cost more than 4 euros and there are outliers mostly below the minimum.

- The new price seems to follow a bi- or tri-modal distribution.

- Most of the new cells cost more than the used cells.

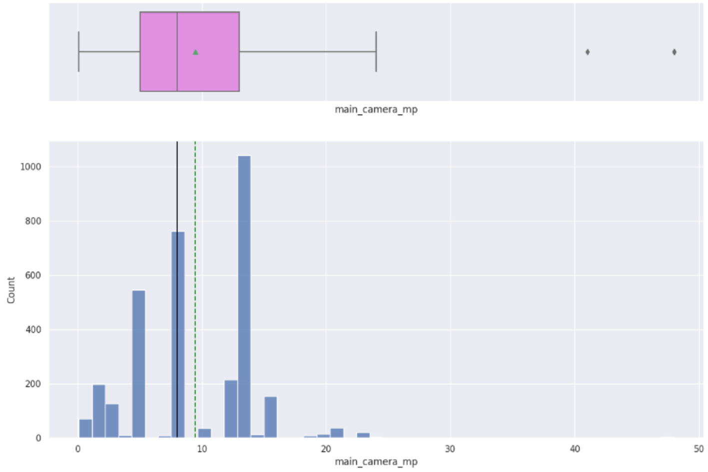


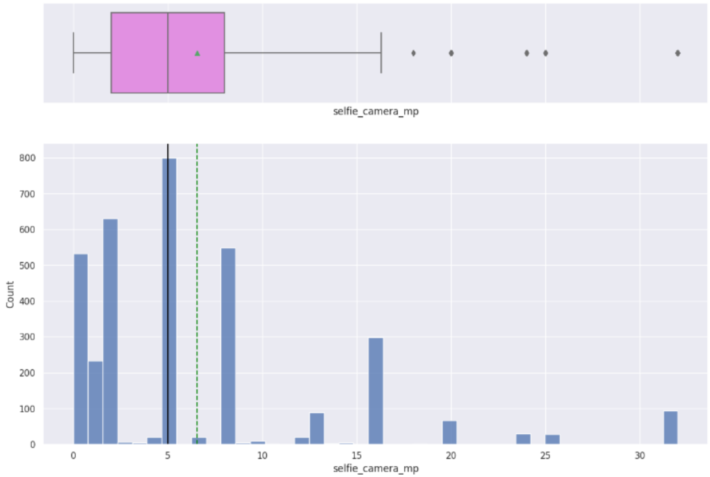


EDA - UNIVARIATE(CAMERA RESOLUTION)

- The main camera's resolution seems to have 3 common values.
- Most of the cell's have a main camera with a resolution less than 10 mega pixels.

- The selfie camera's resolution also seems to have a few commonly occurring values with some cells having very large resolutions on their selfie cameras.
- As expected, the resolution of a selfie camera is generally lesser than that of the main camera.

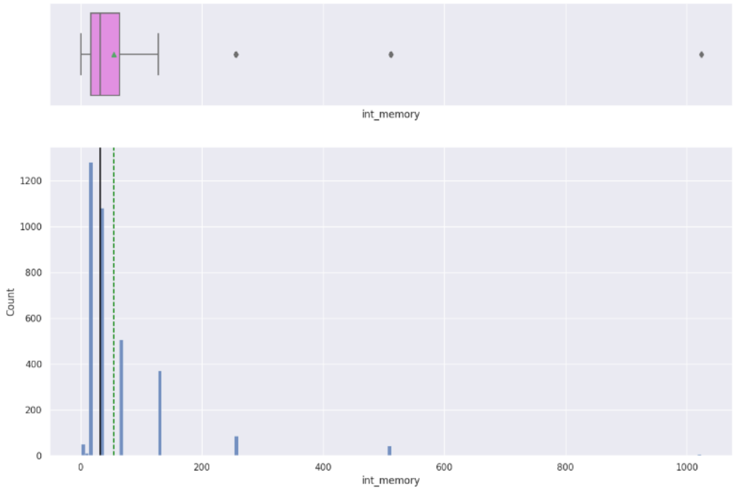


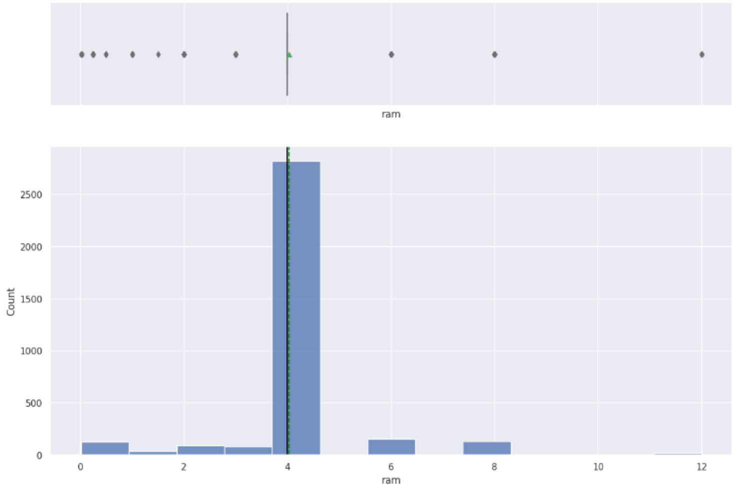


EDA - UNIVARIATE(MEMORY/RAM)

- Most cells have a low internal memory but a few have an incredibly large memory.
- There are 5 fixed values of internal memory a cell can take.

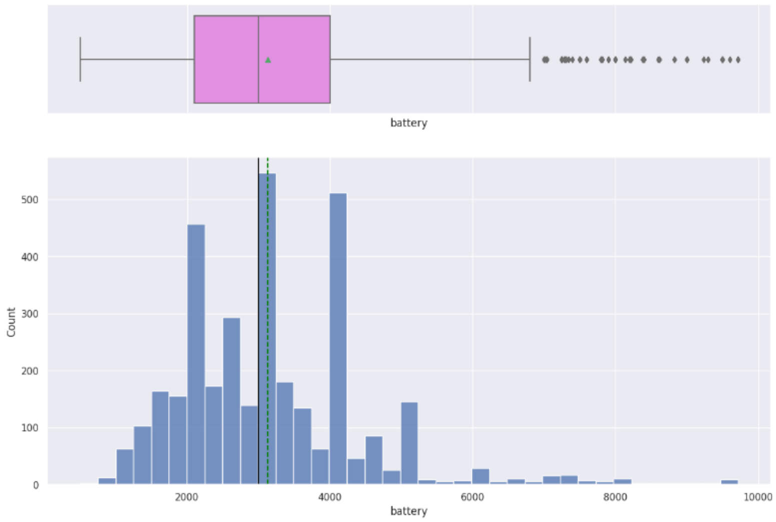
- Most cells have 4GB of RAM.
- A few cells have more than 6 GB of RAM.



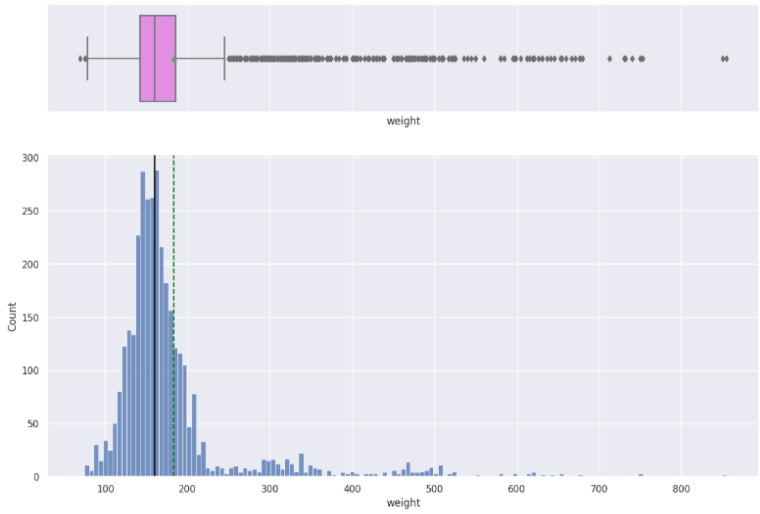


EDA - UNIVARIATE(BATTERY/WEIGHT)

- The cell's battery follows a tri-modal like distribution with a lot of outliers above the maximum.
- Most cells have at least 2000mAh of battery capacity.



- The weight of the cells has a right skewness to it with a lot of outliers above the maximum.
- Most cells weigh less than 200g.



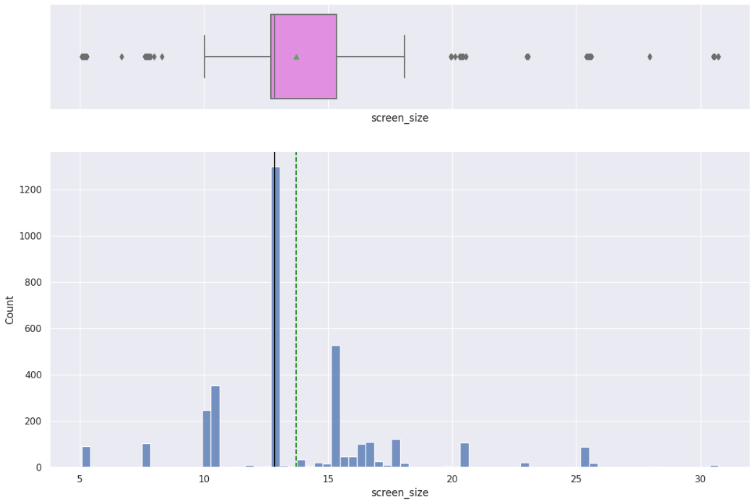
EDA - UNIVARIATE(SCREEN SIZE/DAYS USED)

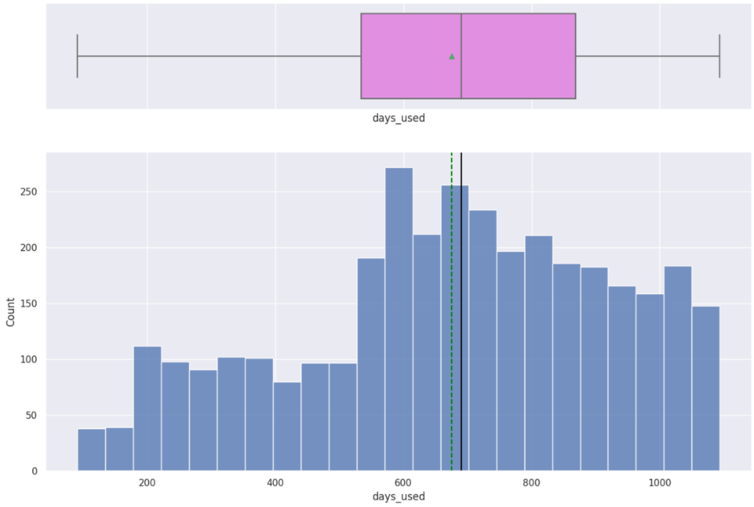
- The cell's screen sizes are segregated into different set of values.

- The most common screen size is around 13 cm.

- The days used follows a slightly left skew with an average of roughly 700 days.

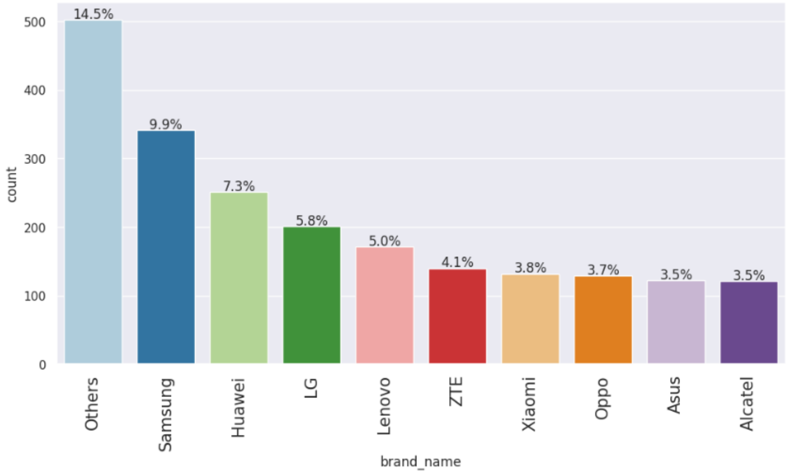
- People use their cells for mostly 600-800 days.

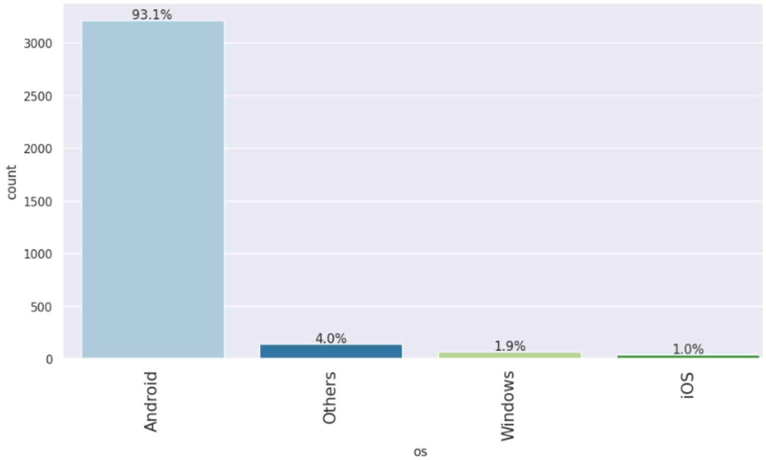




EDA - UNIVARIATE(BRAND/OS)

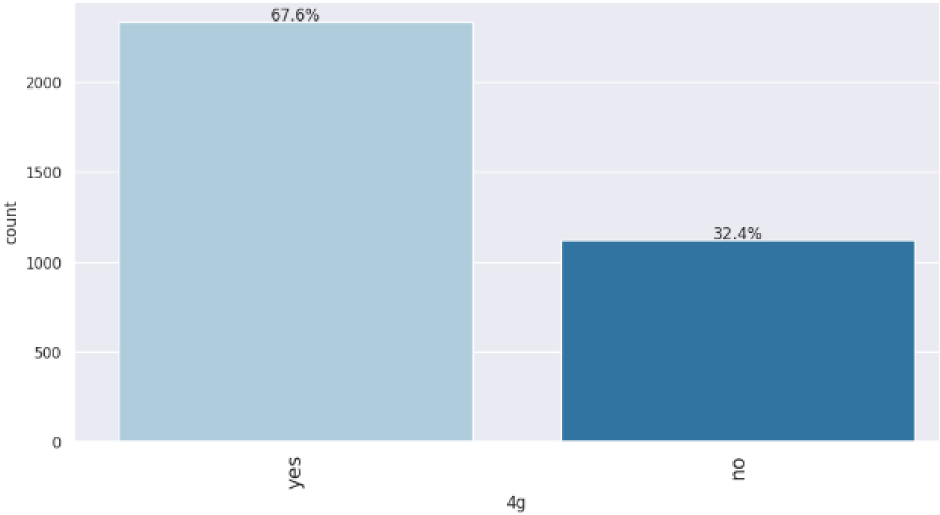
- A third of cells are either Samsung, Huawei, LG, Lenovo or ZTE.
- Surprisingly, Apple does not make up even 3.5% of all cells.
- Almost all cells use an Android OS, while only 1% of cells use an iOS.
- Other Operating Systems are used more by cells than Windows or iOS.

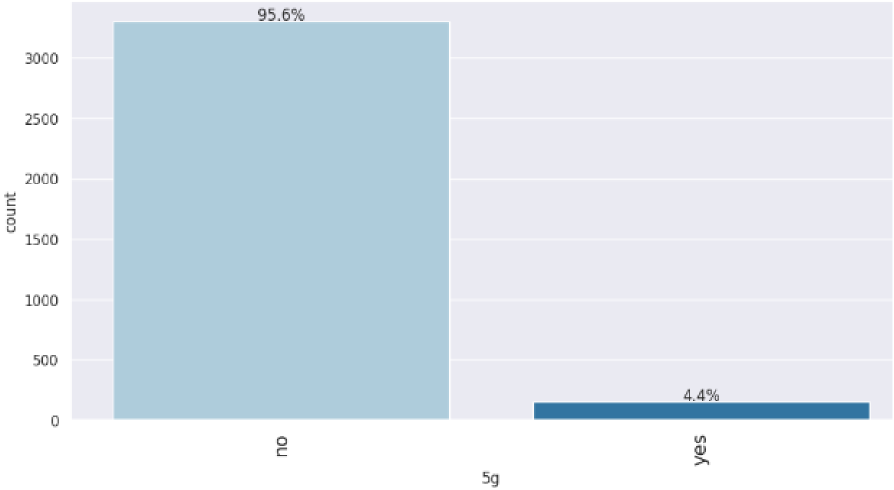




EDA - UNIVARIATE(4G/5G)

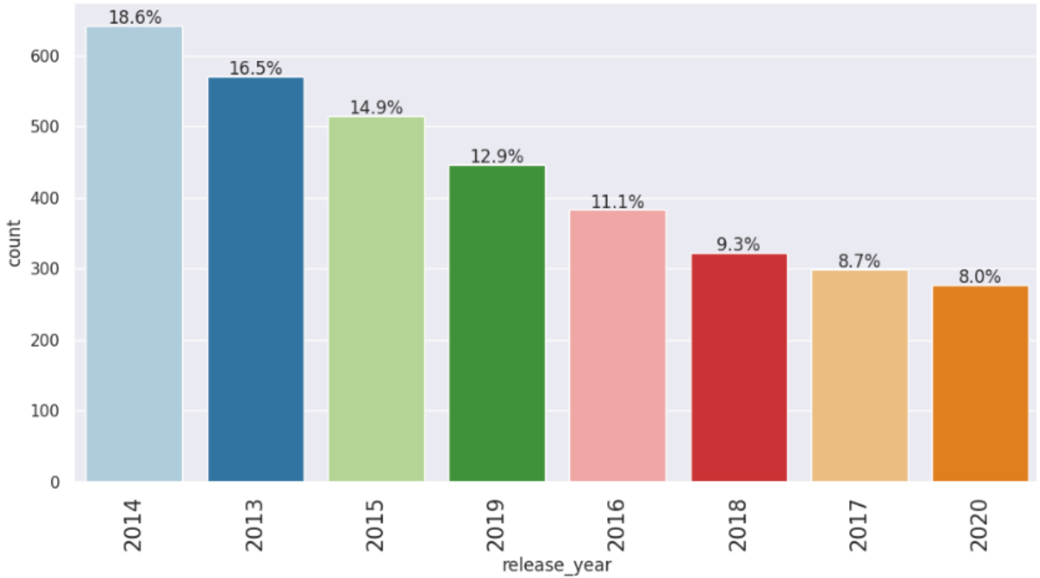
- More than two thirds of the cells have 4G capability.
- Less than 5% of cells have 5G capability.





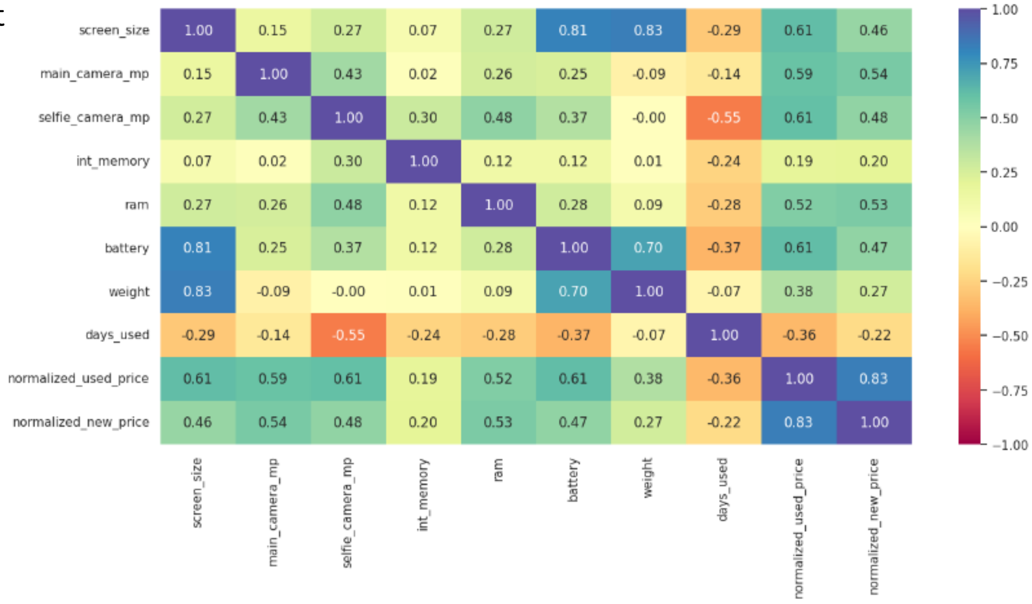
EDA - UNIVARIATE(RELEASE YEAR)

- Almost 50% of cells were released before 2015.
- 2014, 2013 had the most releases.
- Only 8% of cells were released in 2020.
- There seems to be no linear trend in the way cells were released as time passed.



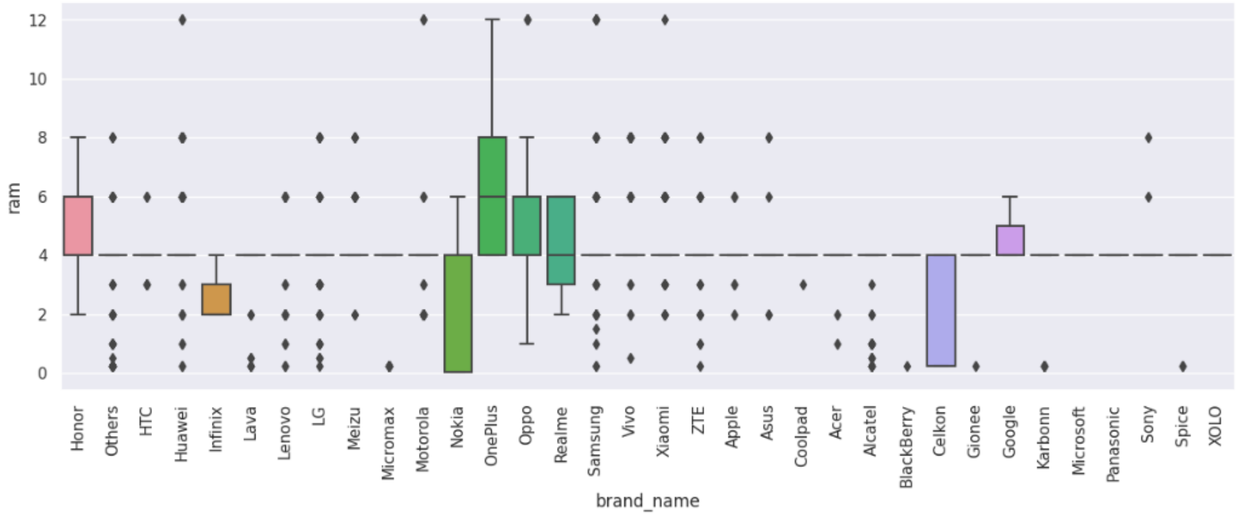
EDA – BIVARIATE(CORRELATION)

- Expectedly, Battery and Weight have the strongest positive correlation with screen size.
- The number of days a user uses the cell tends to affect selfie camera resolution more than the main camera’s resolution.
- The weakest correlation is between the screen size of a cell and it’s internal memory.
- The used and new prices tend to have strong positive correlation.



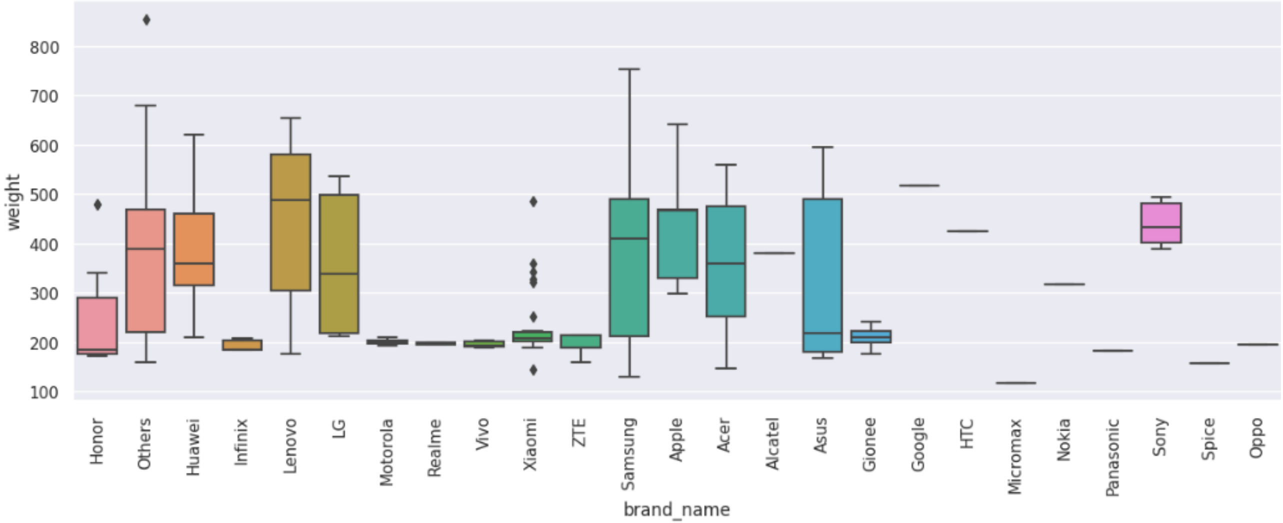
EDA – BIVARIATE(RAM OF BRANDS)

- OnePlus cells have a minimum RAM of 4GB but it varies all the way up to 12GB.
- All of Celkon's cells have less than 4GB of RAM.
- Honor cells have a large variation from 2GB to 8GB of RAM.
- Most of Nokia's cells have less than 4GB of RAM.
- Apple cells do not posses more than 6GB of RAM.
- Only 5 brands have cells that posses 12GB of RAM.



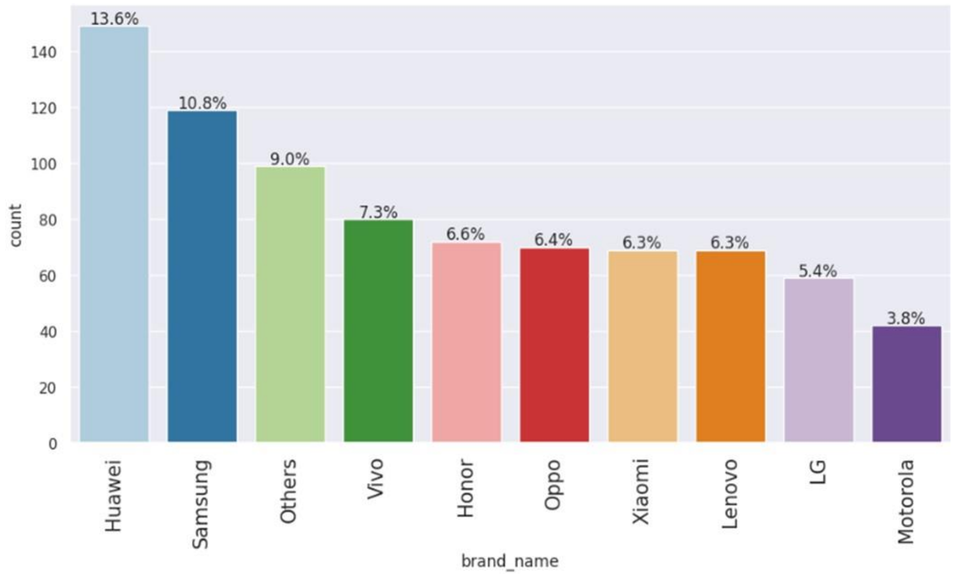
EDA – BIVARIATE(WEIGHT OF BRANDS FOR LARGE BATTERIES)

- Samsung cells have the largest variation in the weight of cells for large batteries.
- Apple's least weighing large battery cell is 300g.
- MicroMax has the lightest large battery cell.
- Sony cells with large batteries weigh at least 400g.
- The heaviest large battery cell is almost a kilo.



EDA – BIVARIATE(BRANDS BY LARGE SCREEN SIZE)

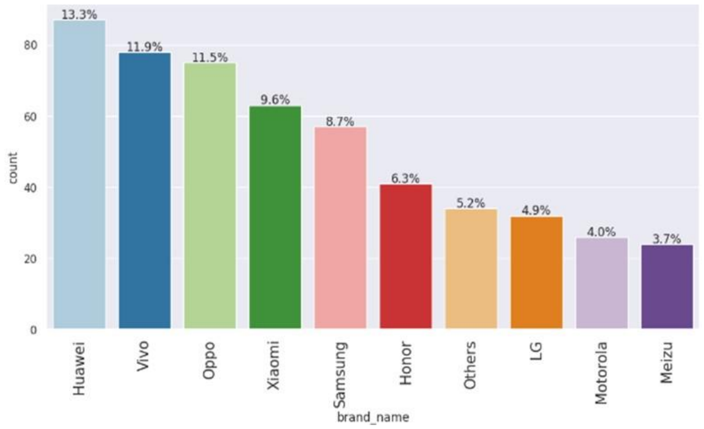
- Most cells with large screens are Huawei's and Samsung's.
- The other brands seem to have more small size screen than large ones.
- Apple does not feature on the list possible suggesting they posses mainly small screens.
- A third of cells with large screens are from 4 brands.



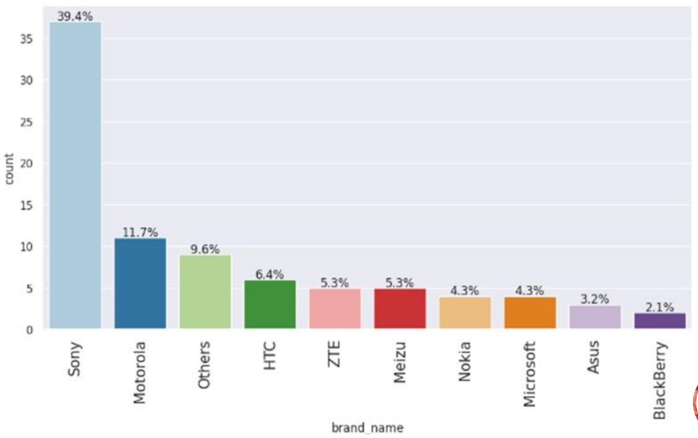
EDA – BIVARIATE(BRANDS BY CAMERA RESOLUTIONS)

- 13% of cells that have an 8 Megapixel or greater selfie camera are Huawei's.
- Vivo and Oppo also have many cells with selfie cameras with a resolution larger than 8 Megapixels.
- Surprisingly, brands with high selfie camera resolutions do not seem to have high main camera resolutions.
- Sony has the most cells with a main camera resolution of more than 16 Megapixels.
- Apple does not feature on the list of high selfie camera or high main camera resolutions.
- Motorola seems to be the only brand that has cells with both a large selfie and main camera resolution.

High Selfie Camera Resolution(>8Mp)

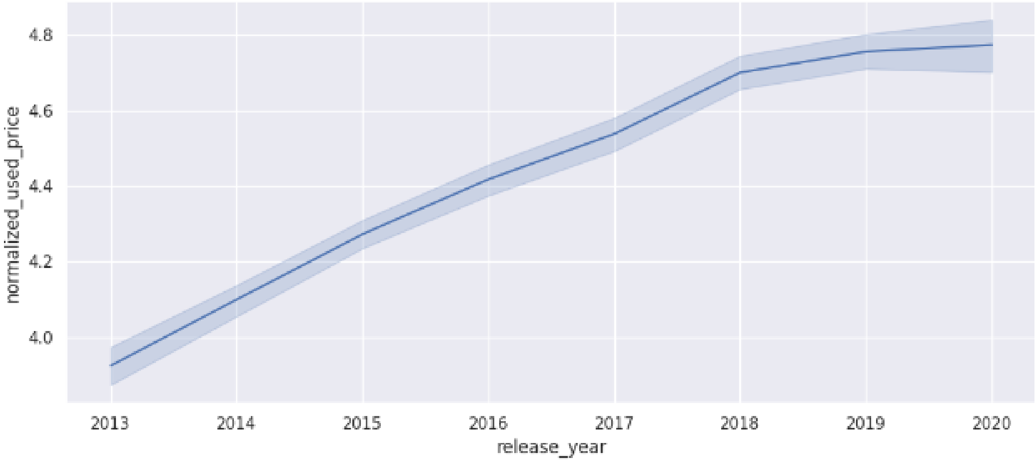


High Main Camera Resolution(>16Mp)



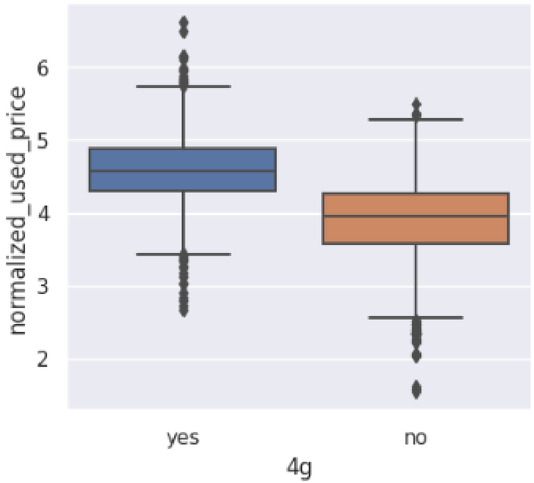
EDA – BIVARIATE(PRICE OVER TIME)

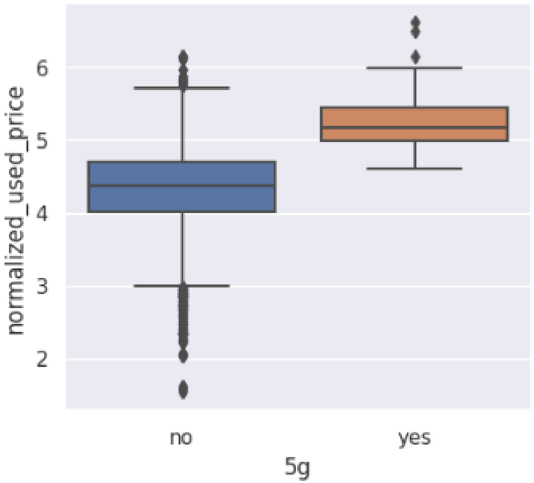
- The used price has tended to increase steadily from 2013 to 2018.
- The increase in prices slowed down post 2018.
- There has almost been a Euro increase in prices of used cells in just 7 years.
- Every year, the price of used cells has gone up by roughly 0.2 Euros.



EDA – BIVARIATE(PRICE BY 4G/5G NETWORK)

- Cells with 4G Networks cost almost a Euro more than cells with no 4G Networks.
- 50% of cells with 4G networks cost between 4 to 5 Euros.
- 75% of cells with no 4G network cost less than roughly 4.2 Euros.
- The median price of a cell with 4G network is the minimum price of a cell with 5G network.
- Most cells with 5G network cost more than 5 Euros.
- There is large variation in prices of cells that do not posses 5G network.



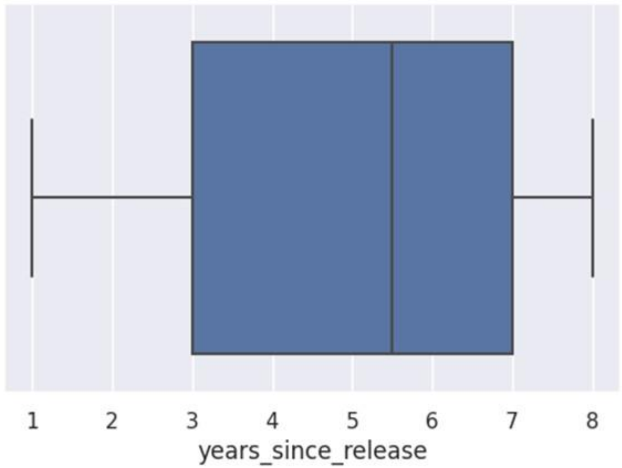


DATA PRE-PROCESSING – MISSING/DUPLICATE VALUE

- Missing values in data are first imputed with column medians grouped by release year and brand name, but the main and selfie camera, battery and weight still have missing values.
- The remaining missing values are imputed with column medians grouped by Brand Name. This time, only the main camera still has missing values in it.
- The remaining missing values in the main camera variable will be imputed with the median of values from the main camera column. Now, none of the variables have a missing variable.
- There are no duplicate records in the dataset.

DATA PRE-PROCESSING – FEATURE ENGINEERING

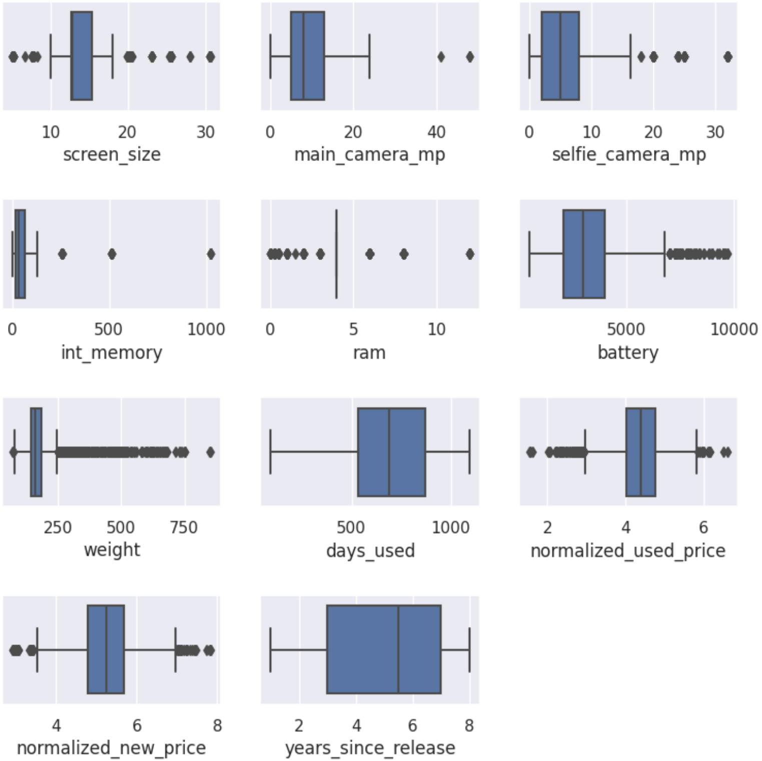
- A new column called ‘Years since Release’ is created from the release year column.
- The baseline for the year of data collection is taken as 2021.
- The year since release is 2021 minus the release year.
- The original release year column is dropped.
- The average is 5 years since release.



count	3454.000000
mean	5.034742
std	2.298455
min	1.000000
25%	3.000000
50%	5.500000
75%	7.000000
max	8.000000

DATA PRE-PROCESSING – OUTLIER CHECK

- Outliers can be seen in most of the variables.
- Only the release year and days used variables do not have outliers.
- Outliers will not be treated or excluded because it could lead to a loss of important information.



DATA PRE-PROCESSING – DATA PREP FOR MODELLING

- The normalized used price needs to be predicted for Re-Cell.
- The categorical features(Brand/OS/4G/5G) are encoded into binary units(1 for yes & 0 for no).
- The data will be split 70:30 into train and test to evaluate the model built on the train data.(2417 rows in train data & 1037 in test data).
- A regression model will be built on the train data and it's performance will be assessed.

Top 5 rows of data with encoded Categorical Variables

	const	screen_size	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	years_since_release	days_used	normalized_new_price	brand_name_Alcatel	brand_name_Apple	brand_name_Asus	brand_name_BlackBerry	brand_name_Celkon
0	1.0	14.50	13.0	5.0	64.0	3.0	3020.0	146.0	1	127	4.715100	0	0	0	0	0
1	1.0	17.30	13.0	16.0	128.0	8.0	4300.0	213.0	1	325	5.519018	0	0	0	0	0
2	1.0	16.69	13.0	8.0	128.0	8.0	4200.0	213.0	1	162	5.884631	0	0	0	0	0
3	1.0	25.50	13.0	8.0	64.0	6.0	7250.0	480.0	1	345	5.630961	0	0	0	0	0
4	1.0	15.32	13.0	8.0	64.0	3.0	5000.0	185.0	1	293	4.947837	0	0	0	0	0

MODEL BUILDING – OLS MODEL 1

- The coefficients of the initial model can be seen on the figure to the right.
- The model can explain approximately 85% of the variation in the dependent variable which is very good.
- The adjusted R-squared value of 0.846 further confirms that the model's fit is not artificially inflated by excessive predictors, providing a more reliable assessment of its performance.

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.849
Model:	OLS	Adj. R-squared:	0.846
Method:	Least Squares	F-statistic:	277.1
Date:	Thu, 11 May 2023	Prob (F-statistic):	0.00
Time:	17:48:55	Log-Likelihood:	125.15
No. Observations:	2417	AIC:	-152.3
Df Residuals:	2368	BIC:	131.4
Df Model:	48		
Covariance Type:	nonrobust		

	coef
const	1.4172
screen_size	0.0295
main_camera_mp	0.0232
selfie_camera_mp	0.0116
int_memory	0.0002
ram	0.0305
battery	-1.665e-05
weight	0.0008
years_since_release	-0.0255
days_used	3.376e-05
normalized_new_price	0.4104
brand_name_Alcatel	-0.0004
brand_name_Apple	-0.0438
brand_name_Asus	0.0068
brand_name_BlackBerry	0.0312
brand_name_Celkon	-0.2372
brand_name_Coolpad	-0.0308
brand_name_Gionee	-0.0130
brand_name_Google	-0.1192
brand_name_HTC	-0.0403
brand_name_Honor	-0.0478
brand_name_Huawei	-0.0599
brand_name_Infinix	0.1338
brand_name_Karbonn	-0.0592
brand_name_LG	-0.0608
brand_name_Lava	-0.0230
brand_name_Lenovo	-0.0364
brand_name_Meizu	-0.0860
brand_name_Micromax	-0.0645
brand_name_Microsoft	0.0749
brand_name_Motorola	-0.0686
brand_name_Nokia	0.0380
brand_name_OnePlus	-0.0384
brand_name_Oppo	-0.0294
brand_name_Others	-0.0679
brand_name_Panasonic	-0.0427
brand_name_Realme	-0.0359
brand_name_Samsung	-0.0617
brand_name_Sony	-0.0756
brand_name_Spice	-0.0356
brand_name_Vivo	-0.0644
brand_name_XOLO	-0.0781
brand_name_Xiaomi	0.0325
brand_name_ZTE	-0.0460
os_Others	-0.0604
os_Windows	-0.0374
os_iOS	-0.0141
4g_yes	0.0406
5g_yes	-0.0916

LINEAR REGRESSION ASSUMPTIONS

The following linear regression assumptions will be checked:

- Multicollinearity
- Linearity of Variables
- Independence of residuals
- Normality of residuals
- No heteroscedasticity

LINEAR REGRESSION ASSUMPTIONS - MULTICOLLINEARITY

Multicollinearity is checked via the VIF. Variables with a VIF > 5 will be dropped carefully. As the variables are dropped, the Adjusted R-squared and RMSE metrics of the models are assessed. The variables that cause these metrics on the model to change the least will be dropped, and the VIF scores for all variables will be checked. The final model will contain all variables with a VIF score < 5.

The first VIF test produced 9 variables with VIF scores > 5 as shown to the right.

	feature	VIF
0	const	232.676933
1	screen_size	8.262147
7	weight	6.417982
12	brand_name_Apple	11.195090
21	brand_name_Huawei	6.395779
24	brand_name_LG	5.354573
34	brand_name_Others	10.833481
37	brand_name_Samsung	8.013571
46	os_iOS	10.037221

The change in the Adj. R-squared and RMSE metrics on the model by dropping each of these columns is shown below. 'os iOS' is dropped because it is related to Apple. 'Weight' is dropped because it relates to screen size, and 'Other Brands' is dropped due to it's high VIF and less effect on Adj. R-squared and RMSE metrics of the model.

	col	Adj. R-squared after_dropping col	RMSE after dropping col
0	const	0.996765	0.250332
1	os_iOS	0.845888	0.232077
2	brand_name_Apple	0.845883	0.232081
3	brand_name_Huawei	0.845779	0.232159
4	brand_name_LG	0.845779	0.232159
5	brand_name_Samsung	0.845765	0.232169
6	brand_name_Others	0.845731	0.232195
7	weight	0.843587	0.233803
8	screen_size	0.841348	0.235470

LINEAR REGRESSION ASSUMPTIONS – P-VALUE VARIABLES

Variables with a high p-value(>0.05) indicate insignificance to the model and will be dropped.

The function shown below(left) will select the variables that have a significant p-value(<0.05) and the result is seen in the image below(right).

The resulting model(Model2) and it's performance evaluation is shown on the next few slides.

```
# initial list of columns
predictors = x_train2.copy()  ## Complete the code to check for p-values on the right dataset
cols = predictors.columns.tolist()

# setting an initial max p-value
max_p_value = 1

while len(cols) > 0:
    # defining the train set
    x_train_aux = predictors[cols]

    # fitting the model
    model = sm.OLS(y_train, x_train_aux).fit()

    # getting the p-values and the maximum p-value
    p_values = model.pvalues
    max_p_value = max(p_values)

    # name of the variable with maximum p-value
    feature_with_p_max = p_values.idxmax()

    if max_p_value > 0.05:
        cols.remove(feature_with_p_max)
    else:
        break

selected_features = cols
print(selected_features)
```

```
const
screen_size
main_camera_mp
selfie_camera_mp
int_memory
ram
years_since_release
normalized_new_price
brand_name_Celkon
brand_name_Nokia
brand_name_Xiaomi
4g_yes
5g_yes
```

MODEL BUILDING – OLS MODEL 2

- The coefficients of the model after removing multicollinearity and insignificant variables(p-value>0.05) can be seen on the figure to the right.
- The model can explain approximately 84.5% of the variation in the dependent variable which is very good.
- The adjusted R-squared value of 0.844 further confirms that the model's fit is not artificially inflated by excessive predictors, providing a more reliable assessment of its performance.

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.845
Model:	OLS	Adj. R-squared:	0.844
Method:	Least Squares	F-statistic:	1089.
Date:	Fri, 12 May 2023	Prob (F-statistic):	0.00
Time:	06:49:48	Log-Likelihood:	91.640
No. Observations:	2417	AIC:	-157.3
Df Residuals:	2404	BIC:	-82.01
Df Model:	12		
Covariance Type:	nonrobust		

	coef
const	1.2483
screen_size	0.0443
main_camera_mp	0.0218
selfie_camera_mp	0.0109
int_memory	0.0002
ram	0.0280
years_since_release	-0.0157
normalized_new_price	0.4121
brand_name_Celkon	-0.1852
brand_name_Nokia	0.0885
brand_name_Xiaomi	0.0779
4g_yes	0.0369
5g_yes	-0.0873

PERFORMANCE EVALUATION– OLS MODEL 2(TRAIN/TEST)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.232968	0.180776	0.844638	0.843797	4.351499

Test Performance

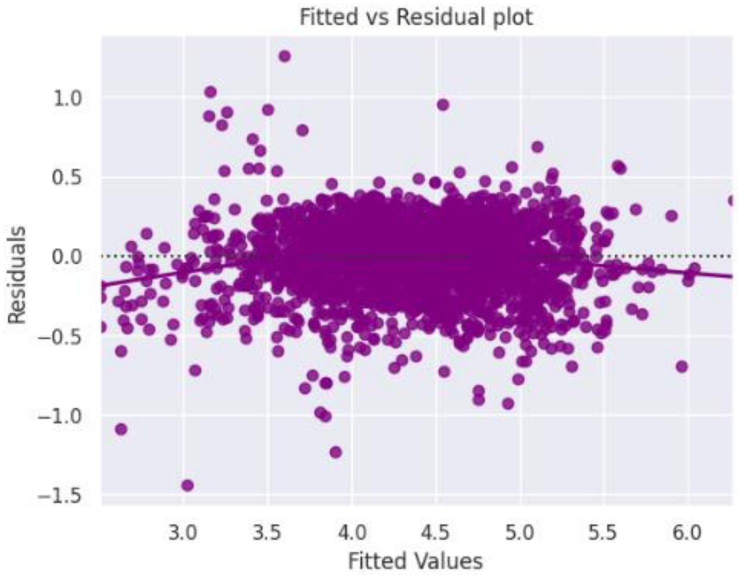
	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.241774	0.191237	0.828346	0.826164	4.572191

- Low RMSE & MAE.
- High R & Adj. R-Squared Values.
- Slightly High MAPE but good enough in the context.
- Training performance is extremely good.

- Low RMSE & MAE but higher compared to Training.
- High R & Adj. R-Squared Values but lower than Training’s.
- Slightly High MAPE but good enough in the context.
- Testing performance is not as good as the training one but still extremely good.

LINEAR REGRESSION ASSUMPTIONS – LINEARITY/INDEPENDENCE

A plot of fitted values vs residuals is created(seen below). No clear pattern is observed so it is a fair assumption to make that the model is a linear one and that the residuals are independent.

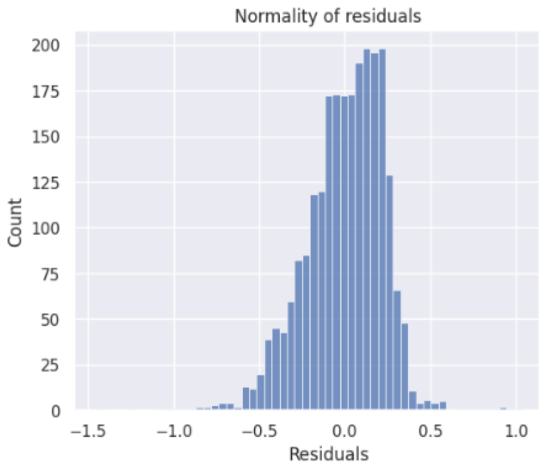


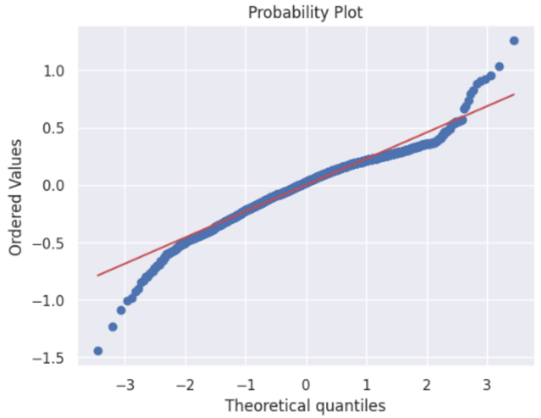
LINEAR REGRESSION ASSUMPTIONS – NORMALITY

Normality will be checked via the distribution of the residuals, the Q-Q plot of the residuals and the Shapiro-Wilk test.

The normality plot is slightly left skewed and the Q-Q plot is straight in the middle but slightly curved towards the start and the end. The Shapiro-Wilks result shows a very small p-value that suggests non-normality, but considering the other tests and the good performance metrics, it would be a fair assumption to make that the residuals of the model are normal enough to proceed.

```
ShapiroResult(statistic=0.9662010669708252, pvalue=2.2102587254686378e-23)
```





LINEAR REGRESSION ASSUMPTIONS – HOMOSCEDASTICITY

Homoscedasticity of the residuals is tested using the Goldfeld-Quandt test. A p-value > 0.05 indicates that the residuals are homoscedastic.

The function to compute this p-value and the result are shown below.

```
import statsmodels.stats.api as sms
from statsmodels.compat import lzip

name = ["F statistic", "p-value"]
test = sms.het_goldfeldquandt(df_pred["Residuals"], x_train3)
lzip(name, test)
```

```
[('F statistic', 0.9309497451236788), ('p-value', 0.8917770165243013)]
```

A p-value of 0.89 validates the assumption that the residuals are homoscedastic.

MODEL BUILDING – FINAL MODEL

The OLS regression analysis aimed to predict the normalized used price of mobile phones based on several independent variables. The model yielded a high R-squared value of 0.845, indicating that approximately 84.5% of the variance in the used price can be explained by the included independent variables. The adjusted R-squared value of 0.844 suggests that the independent variables collectively contribute significantly to explaining the variation in the used price, even after considering the number of predictors in the model. The F-statistic of 1089, with a corresponding p-value of 0.00, indicates that the regression model as a whole is highly significant. This suggests that the chosen independent variables have a strong overall relationship with the dependent variable.

Examining the individual coefficients, it is observed that variables such as screen size, main camera megapixels, selfie camera megapixels, internal memory, RAM, years since release, normalized new price, and certain brand names (Celkon, Nokia, Xiaomi) all have statistically significant relationships with the normalized used price. Additionally, the presence of 4G and absence of 5G connectivity also appear to have significant effects on the used price.

Overall, the regression analysis provides valuable insights into the factors influencing the normalized used price of mobile phones. These results can be used to understand the relative importance of different variables and make informed decisions in areas such as pricing, marketing, and product development within the mobile phone industry.

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.845
Model:	OLS	Adj. R-squared:	0.844
Method:	Least Squares	F-statistic:	1089.
Date:	Fri, 12 May 2023	Prob (F-statistic):	0.00
Time:	07:19:37	Log-Likelihood:	91.640
No. Observations:	2417	AIC:	-157.3
Df Residuals:	2404	BIC:	-82.01
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.2483	0.045	27.469	0.000	1.159	1.337
screen_size	0.0443	0.002	29.067	0.000	0.041	0.047
main_camera_mp	0.0218	0.001	15.403	0.000	0.019	0.025
selfie_camera_mp	0.0109	0.001	9.828	0.000	0.009	0.013
int_memory	0.0002	6.7e-05	2.774	0.006	5.45e-05	0.000
ram	0.0280	0.005	5.507	0.000	0.018	0.038
years_since_release	-0.0157	0.004	-4.384	0.000	-0.023	-0.009
normalized_new_price	0.4121	0.011	36.912	0.000	0.390	0.434
brand_name_Celkon	-0.1852	0.054	-3.453	0.001	-0.290	-0.080
brand_name_Nokia	0.0885	0.029	3.033	0.002	0.031	0.146
brand_name_Xiaomi	0.0779	0.025	3.083	0.002	0.028	0.127
4g_yes	0.0369	0.015	2.435	0.015	0.007	0.067
5g_yes	-0.0873	0.031	-2.775	0.006	-0.149	-0.026

Omnibus:	231.271	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	619.957
Skew:	-0.530	Prob(JB):	2.39e-135
Kurtosis:	5.244	Cond. No.	1.15e+03

PERFORMANCE EVALUATION – FINAL MODEL (TRAIN DATA)

The performance of the linear regression model based on the provided training metrics is quite promising. The model exhibits a relatively low Root Mean Squared Error (RMSE) of 0.233 and Mean Absolute Error (MAE) of 0.180, indicating small deviations between the predicted and actual values. These metrics suggest good accuracy and precision in the model's predictions.

The model shows strong explanatory power, with an R-squared value of 0.844, indicating that approximately 84.4% of the variability in the dependent variable can be explained by the model. The adjusted R-squared value of 0.844 further confirms that the model's fit is not artificially inflated by excessive predictors, providing a more reliable assessment of its performance.

While the Mean Absolute Percentage Error (MAPE) of 4.35 may appear relatively higher, it is important to consider the specific context and requirements of the problem at hand. Overall, based on these metrics, the linear regression model demonstrates favorable performance on the training data. However, it is essential to validate the model's performance on unseen or test data and compare it with other models or benchmarks to ensure its generalization ability and evaluate its relative performance.

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.232968	0.180776	0.844638	0.843797	4.351499

PERFORMANCE EVALUATION- FINAL MODEL(TEST DATA)

The linear regression model performs well on the test data, as indicated by the metrics. The model achieves a relatively low Root Mean Squared Error (RMSE) of 0.241 and Mean Absolute Error (MAE) of 0.191, indicating small deviations between the predicted and actual values. These metrics demonstrate the model's accuracy and precision in predicting the test data.

The R-squared value of 0.828 suggests that approximately 82.8% of the variability in the target variable can be explained by the linear regression model. This indicates a reasonably good fit between the model and the test data. The adjusted R-squared value of 0.823, which considers the number of predictors, provides a more accurate assessment of the model's performance and helps avoid overfitting.

Although the Mean Absolute Percentage Error (MAPE) of 4.513 is slightly higher compared to the training data, it is important to consider the specific context and requirements of the problem. Overall, based on these test data metrics, the linear regression model demonstrates favorable performance. However, it is recommended to compare these metrics with other models or benchmarks and evaluate the model's performance in the specific context to make a comprehensive assessment.

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.241774	0.191237	0.828346	0.826164	4.572191

BUSINESS RECOMMENDATIONS

- **Pricing Strategy**: Re-Cell should consider the influential factors identified in the regression model, such as screen size, camera specifications, internal memory, RAM, years since release, and brand names, when determining their pricing strategy for used mobile phones. By leveraging these factors, Re-Cell can justify higher price points for their devices. Emphasize the value proposition of features like larger screen sizes, higher camera megapixels, and ample memory to command higher prices and differentiate Re-Cell's offerings in the used mobile phone market.
- **Product Development**: Re-Cell should focus on developing high-quality, feature-rich mobile phones with competitive new prices. The regression model indicated a positive correlation between the normalized new price and the normalized used price. By creating mobile phones that offer a compelling combination of advanced features, durability, and affordability, Re-Cell can maximize the potential used price and overall market value of their devices.
- **Branding and Marketing**: Re-Cell should leverage the brand reputation and value associated with "Celkon," "Nokia," and "Xiaomi" to enhance the perceived value of their mobile phones in the used market. Emphasize Re-Cell's association with these trusted brands in marketing campaigns to instill confidence in customers and justify higher prices. Highlight the unique features, reliability, and quality of Re-Cell's devices to differentiate them from competitors and establish a strong brand identity in the used mobile phone industry.
- **Connectivity Features**: In the development of new mobile phone models, Re-Cell should carefully consider connectivity options. As the regression model revealed, the presence of 4G and the absence of 5G connectivity significantly influence the used price. Re-Cell should integrate the latest connectivity options, including 5G technology, in new models to enhance their market value. Furthermore, ensure backward compatibility with previous technologies to preserve the resale value of Re-Cell's devices and cater to a wider customer base.

THANK YOU

ARPAN DINESH