RE-CELL

ML- Pricing Strategy Analysis

Machine Learning (Linear Regression): Module 3 – February 2024

AGENDA

- Executive Summary
- Overview & Solution Approach
- EDA Results
- Data Preprocessing
- Model Summary – Performance
- Insights & Recommendations

Executive Summary

Over the past decade rising sales and marketing of used and refurbished electronic devices specifically cellular phones and tablets has grown exponentially. Forecasted data predicts the used cell phone market would be worth $52.7 billion by 2023 with a Compound Annual Growth Rate (CAGR) of 13.6% between 2018 and 2023. The cause for this growth is considerable savings on refurbished products versus the high price of new models.

Cost-effective benefits of refurbished devices for personal and business needs:

- ✓ Items can be sold with warranties and can be insured
- ✓ Third-party vendors provide incentives and special offers/pricing
- ✓ Lowers environmental impact to help reduce waste
- ✓ Consumers cut back on discretionary spending

Overview & Approach

ReCell a start-up company sees the potential of this rapidly growing niche market and would like to perform data analysis and look for Machine Learning solutions to develop a pricing strategy for used and refurbished devices. The data set contains the different attributes of used/refurbished phones and tablets. The data was collected in 2021.

Data Analysis Approach:

- Data Set: Check for Duplicate Values, Missing Value Treatment, Outliers, EDA Analysis
- Machine Learning: Model Building, Linear Regression Assumptions, Final Model Summary

- OBJECTIVE: Predict Price & Factors that Significantly Influence Purchases

Data Definitions:

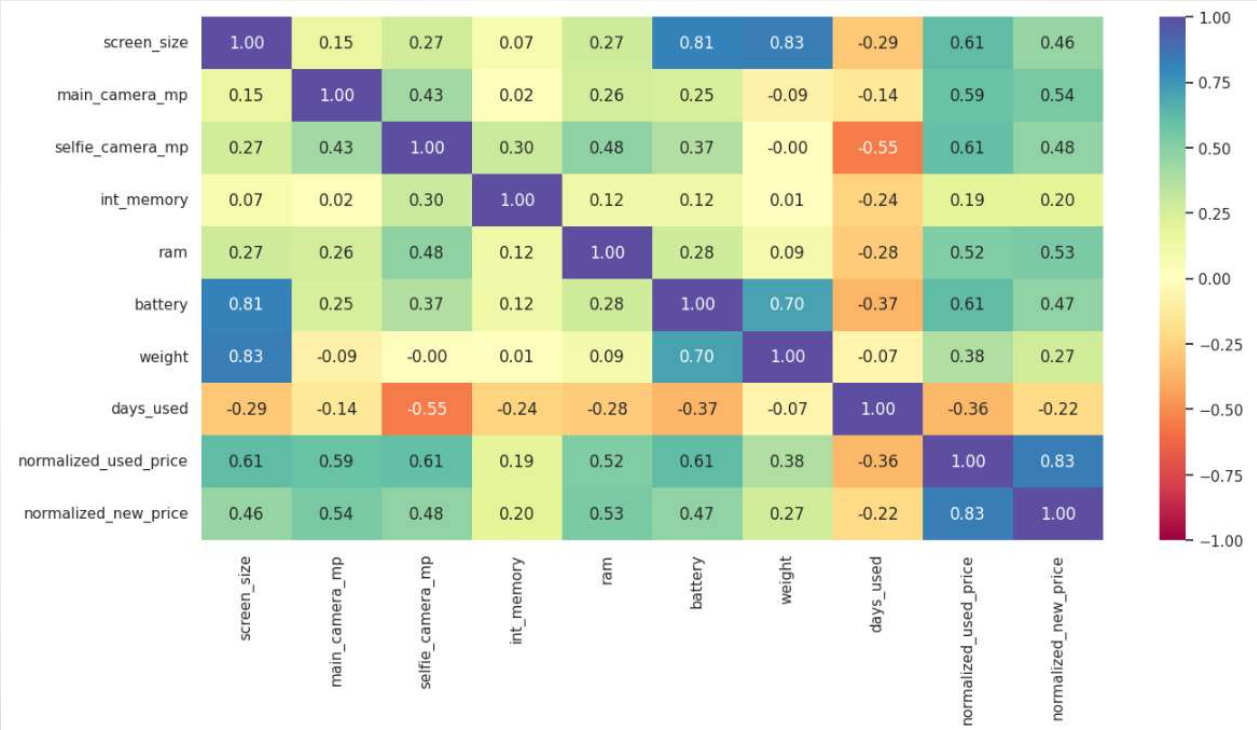
- brand_name: Name of manufacturing brand
- os: OS on which the device runs
- screen_size: Size of the screen in cm
- 4g: Whether 4G is available or not
- 5g: Whether 5G is available or not
- main_camera_mp: Resolution of the rear camera in megapixels
- selfie_camera_mp: Resolution of the front camera 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

Dataset Details:

- Size: 3,454 Products – 15 Attributes
- Products: Used Cellphones & Tablets
- Missing or Duplicate Values Strategy: Treated or Imputed
    - Check for Duplicate Values
    - Central Tendency Measures (mean, median, mode)
    - Drop the Missed Values
- EDA Analysis – Univariate, Bivariate, Data Preprocessing, Feature Engineering, Outlier Check, Data Modeling Preparation
- Linear Regression Model Build, Test Models, Final Model Summary

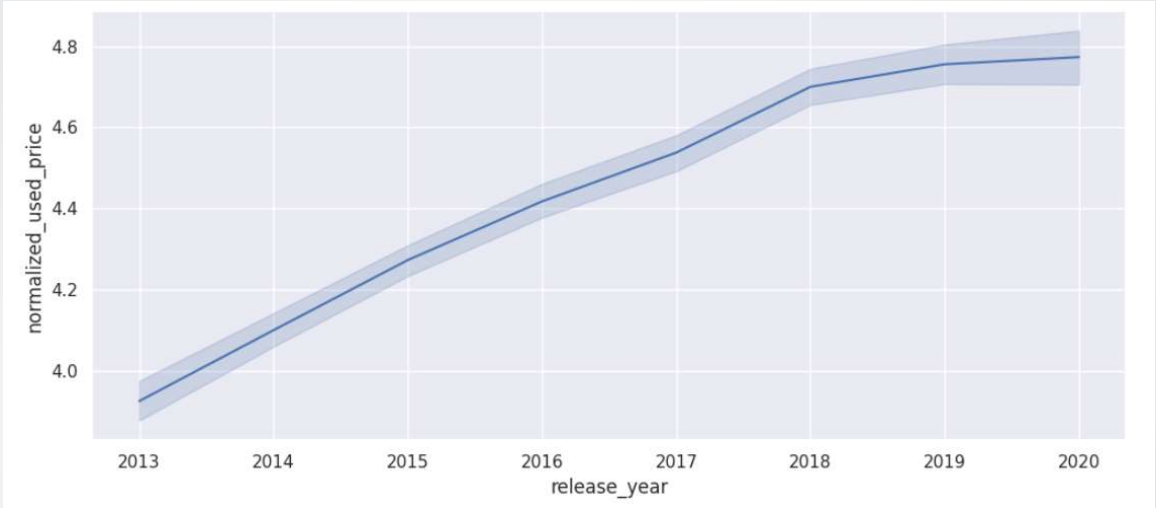
EDA Results

- Univariate Analysis observations were made on each attribute of the products
- Bivariate Analysis and Correlation Checks show which attributes are connected

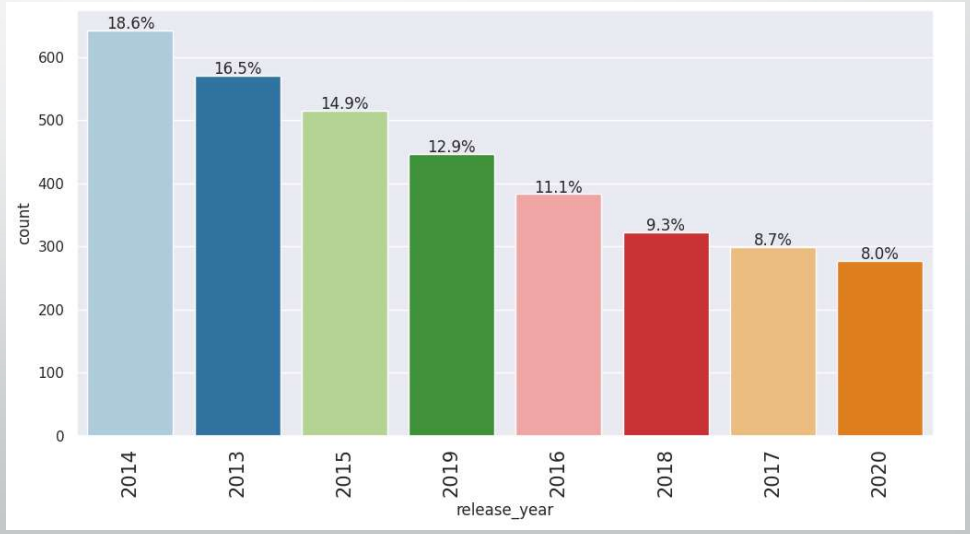


- High Correlations: Screen Size & Weight – Screen Size & Battery – Battery & Weight
- Used Price High Correlation: Screen Size, Main Camera mp, Battery, RAM & New Phone Pricing

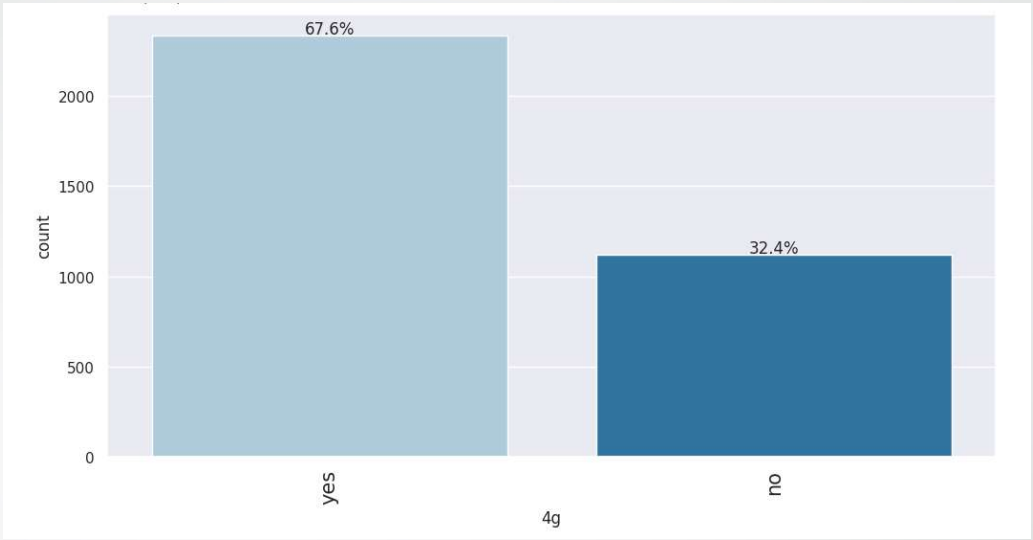
- 2013 – 2020 Release Year
- Used Device Pricing increased at a normal rate, starting to plateau in 2019

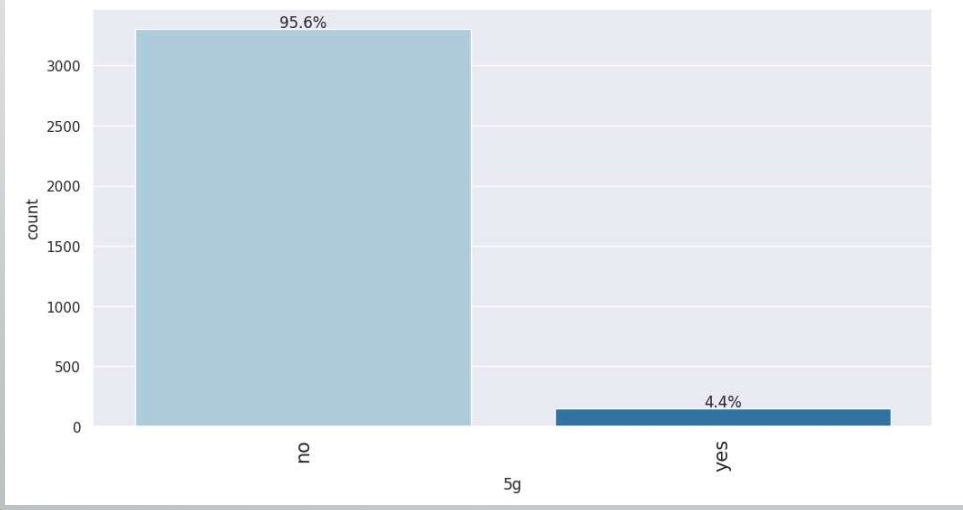


- Consumers purchased older phones for there value



- Consumers could still get 4g service even on older phones
- 67.9% of used Phones and Tablets purchased offered 4g service





- Older used devices did not come with 5g, since it is an emerging technology
- 95.6% of devices did not offer 5g service

EDA Results: Cont'

Bivariate Analysis Statistics compared to Brand Name of Product

- RAM:

  90% of Brands have at least 4GB, with a majority of outliers having less than 4GB

- Weight (larger battery):

  Median weight was between 100 & 200 grams. Apple had the highest median weight of 300 grams and also the heaviest weight of 700 grams

- Screen Size:

  12.83cm is the median screen size and was preferred by 16.6% of consumers. Nokia & Apple had the largest range of sizes. Nokia starting on the smaller end and Apple on the larger end

- Selfie Camera mp:

  Oppo, Xiaomi & Huawei offered the highest megapixels

  Oppo had the highest with a median of just over 15mp and going up to 30mp

- Main Camera mp:

  Threshold of 16mp – 30% of sales showed 13mp was preferred with 8.0mp coming in second at 21.9%. 16mp only comprised 4.5% of sales. Majority of brands offered less than 16mp – Sony had a median of 16mp and went up to just over20mp

Data Preprocessing:

- Dataset had no Duplicate Values
- Missing Values were treated
- Median Values updated for all attributes post-treatment

Missing 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
normalized_used_price   0
normalized_new_price    0
dtype: int64
```

Median Values

```
screen_size              12.830000
main_camera_mp            8.000000
selfie_camera_mp          5.000000
int_memory               32.000000
ram                       4.000000
battery                3000.000000
weight                  160.000000
release_year           2015.500000
days_used               690.500000
normalized_used_price     4.405133
normalized_new_price      5.245892
dtype: float64
```

Feature Engineering:

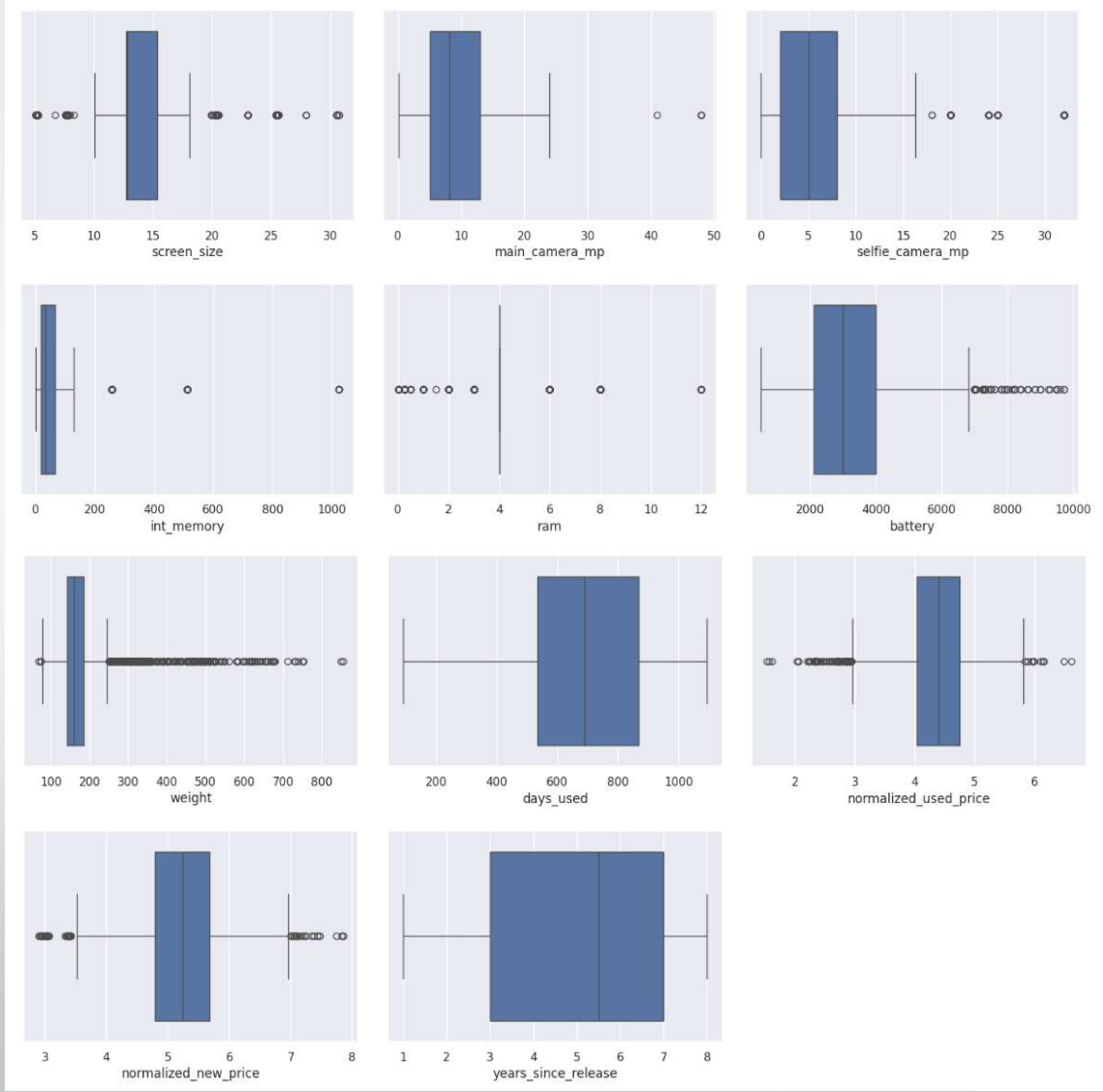
New Column Created 'Years Since Release'

- Years Since Release Median Values

```
count    3454.000000
mean        5.034742
std         2.298455
min         1.000000
25%         3.000000
50%         5.500000
75%         7.000000
max         8.000000
Name: years_since_release, dtype: float64
```

OUTLIERS

- There were many outliers within the Dataset
- They are actual values and were not treated
- We dropped the column ‘release_year and added the column ‘years_since_release’
- No more columns will be dropped, they are all numeric and are of significant influence to our objective



Data Preparation for Modeling

- Normalized Price Prediction of Used Devices
- Encode Categorical features prior to Model Building
- Split the Data into TRAIN & TEST for Evaluation
- Build a Linear Regression Model on TRAIN Data – Check Performance
    - Define Independent (x-axis) and Dependent Variables (y-axis)
    - Create a Constant to Intercept the Data
    - Create Dummy Variables
    - Split the Data 70% - 30%, Total Dataset 3,454
        - Train Data Rows = 2,417
        - Test Data Rows = 1,037

Model Build – Linear Regression

Train Model

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.845
Model:	OLS	Adj. R-squared:	0.842
Method:	Least Squares	F-statistic:	268.7
Date:	Tue, 30 Jan 2024	Prob (F-statistic):	0.00
Time:	23:45:36	Log-Likelihood:	123.85
No. Observations:	2417	AIC:	-149.7
Df Residuals:	2368	BIC:	134.0
Df Model:	48		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.3156	0.071	18.454	0.000	1.176	1.455

- Adjusted R-squared reflects the fit of the model. Conditions were met & our model shows a .842 which is very good.
- Constant Coefficient: Y- intercept is 1.3156, this is our predictor value
- Coefficient of a Predictor Value: We had none of these within our model

Model Performance Check – OLS1

Model Metrics:

RMSE
MAE
MAPE
R-squared

Training Performance

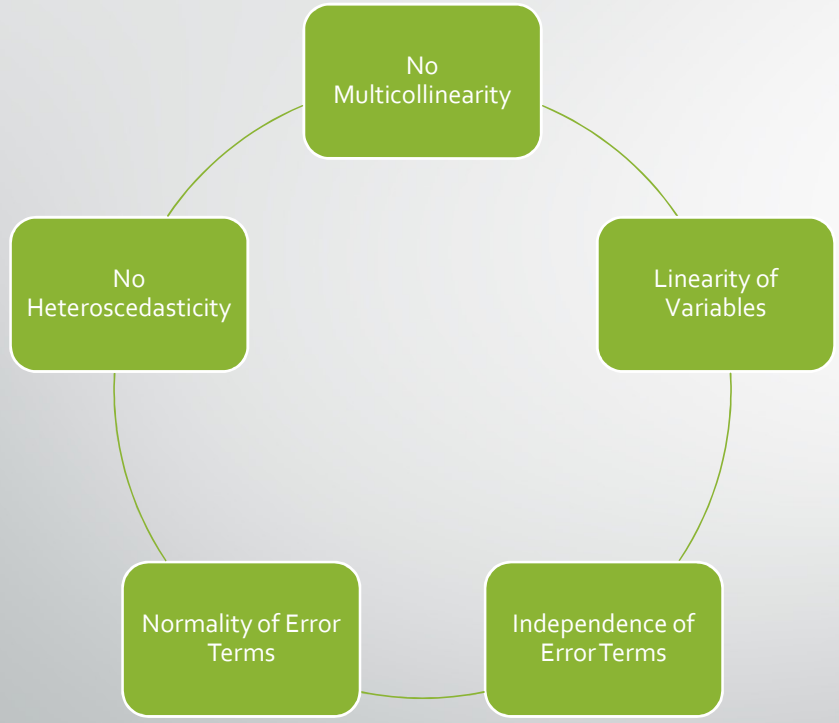
	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.229884	0.180326	0.844886	0.841675	4.326841

- Train R-squared is 0.834 or 83% – Model is not Underfitting
- Train and Test RMSE & MAE are comparable – Model is not Overfitting
- MAE suggests the Model can Predict Pricing – Within Mean Error of 0.18 on the Test Data
- MAPE of Test Data – Price Prediction within 4.5%

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.238358	0.184749	0.842479	0.834659	4.501651

Linear Regression Assumptions



- No Multicollinearity
  - Reliable Residuals without Correlation
  - VIF between 1 and 5
- Linearity of Variables
  - Straight line relationship with Dependent Variable
- Indep. Of Error Terms (residuals)
  - Statistical Significance of Confidence Intervals
  - Plot, No Pattern, Model is Linear & Independent
- Normality of Error Terms
  - Normal Distribution, Bell-Curve, Q-Q Plot, Hypothesis Testing
- No Heteroscedasticity
  - Constant Variance in Error Terms (limited Outliers)
  - Symmetrical Distribution across Regression Line

Treating Multicollinearity & VIF

- VIF > 5: Removed one column at a time then retest Model Performance on Train Data

  Columns Removed: brand_name_Apple, brand_name_Others & screen_size

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.841
Model:	OLS	Adj. R-squared:	0.838
Method:	Least Squares	F-statistic:	279.6
Date:	Wed, 31 Jan 2024	Prob (F-statistic):	0.00
Time:	00:31:00	Log-Likelihood:	97.446
No. Observations:	2417	AIC:	-102.9
Df Residuals:	2371	BIC:	163.5
Df Model:	45		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.4519	0.055	26.421	0.000	1.344	1.560

- Adjusted R-Value dropped from 0.842 to 0.838
- Dropping these columns did not hardly affect the model performance
- Multicollinearity eliminated – look at P-value Predictor variables to check significance

  Each Independent Feature: Null & Alternate Hypothesis
  P-value < 0.05 is considered to be Statistically Significant

High P-value Variables

- Dropped columns one at a time with P-values > 0.05
- OLS Model2 – P-Value Drop

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.839
Model:	OLS	Adj. R-squared:	0.838
Method:	Least Squares	F-statistic:	965.2
Date:	Wed, 31 Jan 2024	Prob (F-statistic):	0.00
Time:	00:44:50	Log-Likelihood:	80.857
No. Observations:	2417	AIC:	-133.7
Df Residuals:	2403	BIC:	-52.65
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.5150	0.047	32.089	0.000	1.422	1.608

- Adjusted R-Value remained at 0.838
- All P-values are now < 0.05

Model Performance Check – OLS2

Model Metrics:

RMSE
MAE
MAPE
R-squared

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.23401	0.18301	0.839268	0.838331	4.398555

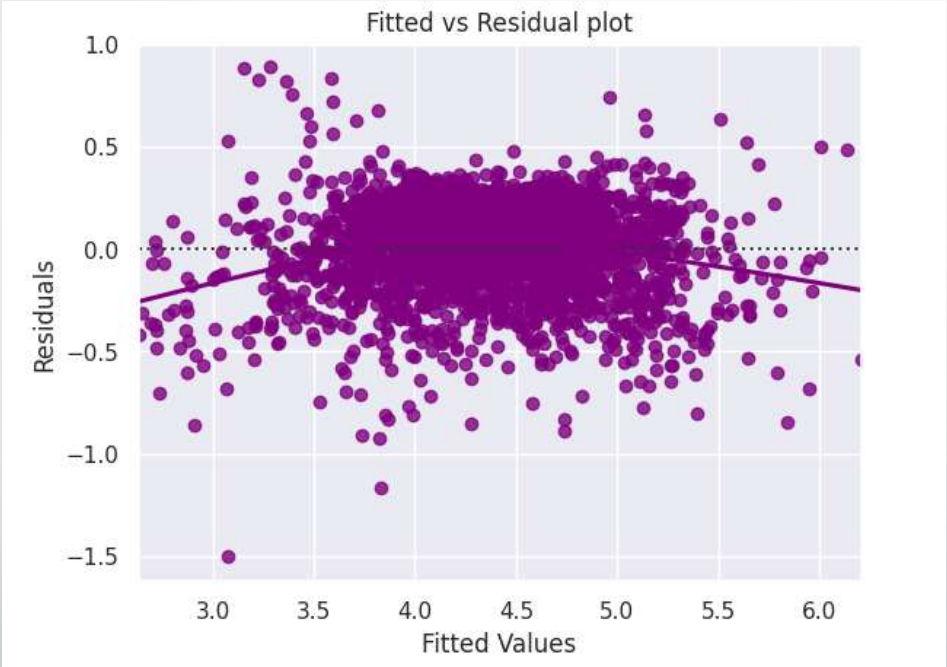
- No feature has P-value > 0.05
- X-Train3 will be used as the final set of Predictor Variables in OLS Model2
- Adjusted R-squared remained the same as OLS Model1 and shows no Multicollinearity and columns dropped did not affect the model
- RMSE and MAE values are comparable for Train & Test Sets, indicating the model is not Overfitting

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.241343	0.187535	0.838509	0.836297	4.570255

Linearity and Independence

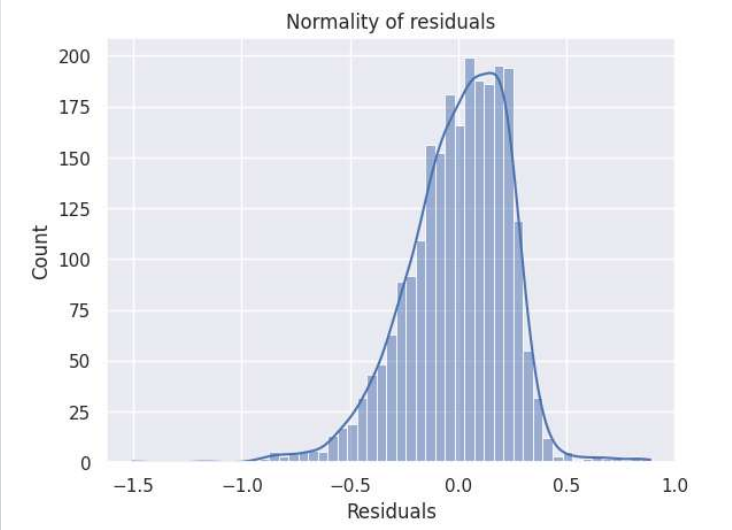
- Test: Plot Fitted Values vs. Residuals and check for Patterns
- No Pattern = Linear Model & Independent Residuals

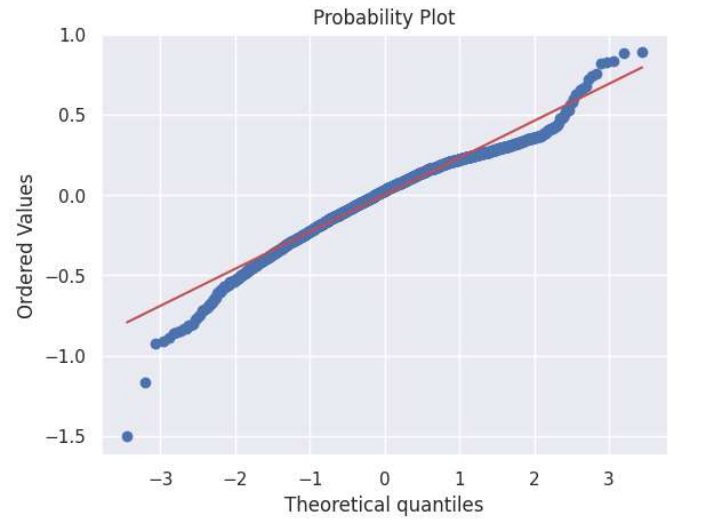


- Scatterplot does not show a pattern
- Linearity and Independence are satisfied

Test for Normality

- Check distribution of Residuals (Normality) - Q-Q Plot -Shapiro-Wilk Test






```
ShapiroResult(statistic=0.9683110117912292, pvalue=1.1352272619449774e-22)
```

- Normality Distribution has a Bell Shape – slight Left & Right Tail
- Probability Plot follows a straight line, except for the tails
- Residuals are not normal, but this can be considered close to normal
- The Assumption is Satisfied

Homoscedasticity vs. Heteroscedasticity

- Outliers remained within the Dataset = No Heteroscedasticity
- Goldfeldquandt Test: P-value > 0.05 Residuals are Homoscedastic

  Null Hypothesis: Residuals are Homoscedastic

  Alternate Hypothesis: Residuals had Heteroscedacity

Goldfeldquandt Test:

```
[('F statistic', 1.048458241344956), ('p-value', 0.130479678454497)]
```

- P-value = 0.130, which is greater than 0.05
- Residuals are Homoscedastic
- Assumption is Satisfied (Null Hypothesis)

Test Data Predictions

- All Assumptions and Linear Regression achieved
- Prediction on OLS Model 2 – Test3

	Actual	Predicted
1995	4.566741	4.375834
2341	3.696103	3.996268
1913	3.592093	3.638329
688	4.306495	4.090478
650	4.522115	5.178461
2291	4.259294	4.386189
40	4.997685	5.440825
1884	3.875359	4.044567
2538	4.206631	4.060787
45	5.380450	5.216557

OBSERVATIONS:

- Actual vs. Predicted Results are comparable
- We have returned a very good Model
- Re-Create into a Final Model and Summarize our findings

Model – Linear Regression

Final Model: train3 & test3 / olsmodel_final

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.839
Model:	OLS	Adj. R-squared:	0.838
Method:	Least Squares	F-statistic:	965.2
Date:	Wed, 31 Jan 2024	Prob (F-statistic):	0.00
Time:	01:08:38	Log-Likelihood:	80.857
No. Observations:	2417	AIC:	-133.7
Df Residuals:	2403	BIC:	-52.65
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.5150	0.047	32.089	0.000	1.422	1.608
main_camera_mp	0.0213	0.001	15.255	0.000	0.019	0.024
selfie_camera_mp	0.0143	0.001	13.404	0.000	0.012	0.016
ram	0.0227	0.005	4.511	0.000	0.013	0.033
weight	0.0016	6.04e-05	27.255	0.000	0.002	0.002
normalized_new_price	0.4342	0.011	40.026	0.000	0.413	0.455
years_since_release	-0.0287	0.003	-8.423	0.000	-0.035	-0.022
brand_name_Karbonn	0.1223	0.055	2.233	0.026	0.015	0.230
brand_name_Lenovo	0.0538	0.022	2.486	0.013	0.011	0.096
brand_name_Nokia	0.0624	0.031	2.020	0.043	0.002	0.123
brand_name_Xiaomi	0.0897	0.026	3.496	0.000	0.039	0.140
os_Others	-0.1429	0.028	-5.037	0.000	-0.198	-0.087
4g_yes	0.0456	0.015	3.031	0.002	0.016	0.075
5g_yes	-0.0645	0.031	-2.109	0.035	-0.125	-0.005

- Adjusted R-squared 0.838 or 83%
- P-values < 0.05
- Constant 1.5150

Model Performance Check – Final

Model Metrics:

RMSE
MAE
MAPE
R-squared

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.23401	0.18301	0.839268	0.838331	4.398555

- No feature has P-value > 0.05
- Final Model able to explain 84% of variation in Data
- Train & Test RMSE and MAE are low and comparable
- Model is not overfitting
- MAPE suggests we can Predict within 4.5%
- We can conclude that olsmodel_final is good for Prediction as well as Significant Influencing Factors

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.241343	0.187535	0.838509	0.836297	4.570255

Conclusion

Insights:

The Final Model:

- Explains approx. 84% of variation within the Data
- Predictive Pricing and Significant Influencing Factors are within 4.5% proving this is a very good model
- Main Camera, Selfie Camera, RAM & Years Since Release (new column) increase by one unit while all other variables remain the same
- Weight if the used device was the least impacting factor

Recommendations:

- As technology in new devices continues to increase, there will always be a need for used devices because of the significant savings in price
- The final model will provide future predictions as new data is provided. Updated datasets with similar attributes can be plugged into this final model to provide Predictive Pricing and Influencing Factors important to the consumer.
- Data shows there are many factors that go into purchasing a phone, attributes are different for each consumer, leading to many outliers. These outliers were actual numbers and were left within the dataset.
- Highest Correlation of Attributes: Screen Size, Main Camera mp, Battery, RAM & New Device Price