ReCell Market Analysis

Supervised Learning – Foundations Project
by Ryan J. Stambaugh

10 Jan 2025

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

ReCell aims to capitalize on the burgeoning used and refurbished cell phone and tablet market by leveraging data-driven insights to optimize pricing. A robust analysis using linear regression model provides the insights needed to determine normalized used prices, guiding actionable strategies for business growth.

The regression model used had a test performance with the following:

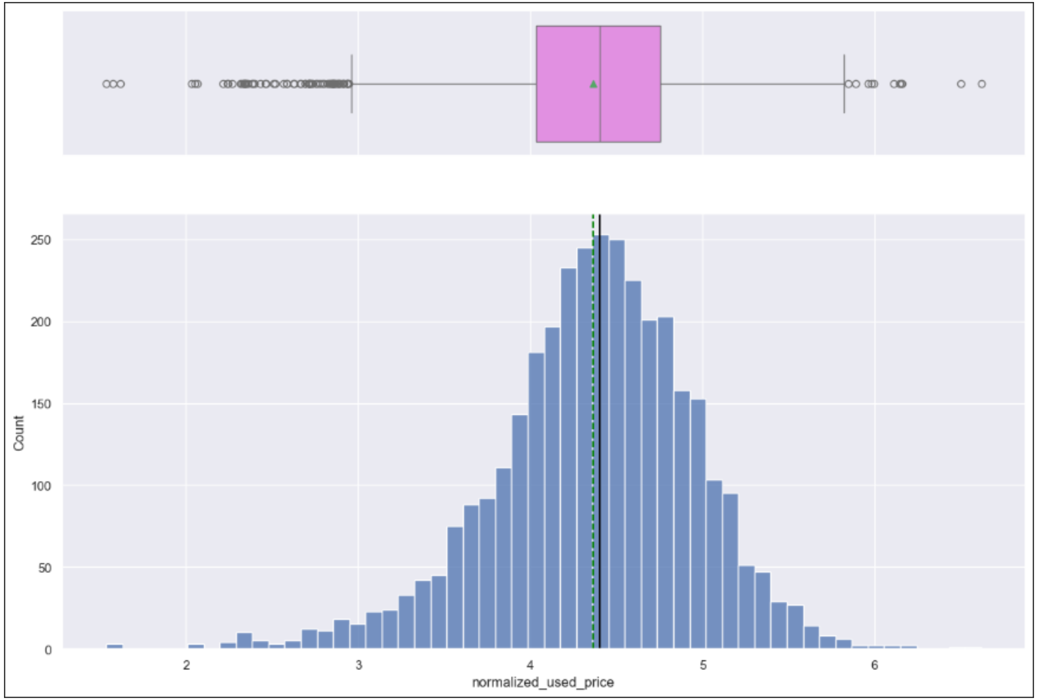
- **Low RMSE** (0.2347) indicating *predictive accuracy.*
- **Low MAE** (0.1865) indicating *avg absolute difference* between observed and predicted values.
- **Low MAPE** (4.313%) indicating *high precision* in predictions.

Business Problem Overview and Solution Approach

- The problem is to tap the potential market of used and refurbished cell phones and tablets.
- The solution is to find an **ML-based dynamic pricing strategy** for those used and refurbished devices.
  - In the following, we analyzed data to build a linear regression model that predicts the price of used phones and tablets with the major factors that significantly influence it.

EDA Results

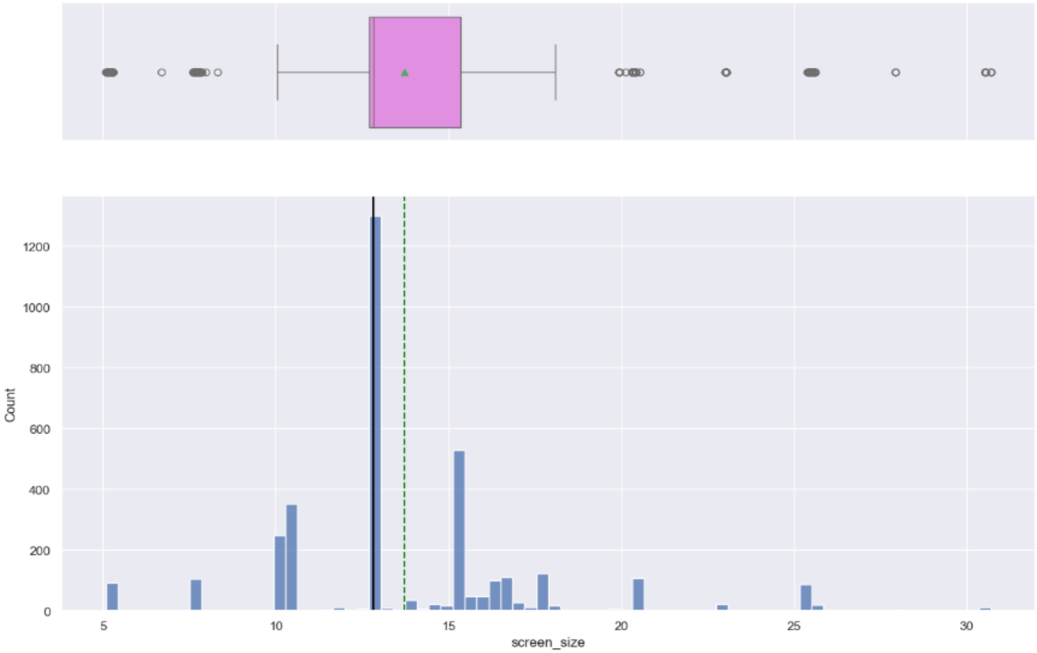
Used Prices fall into a predictable price pattern.



Link to Appendix slide on data background check

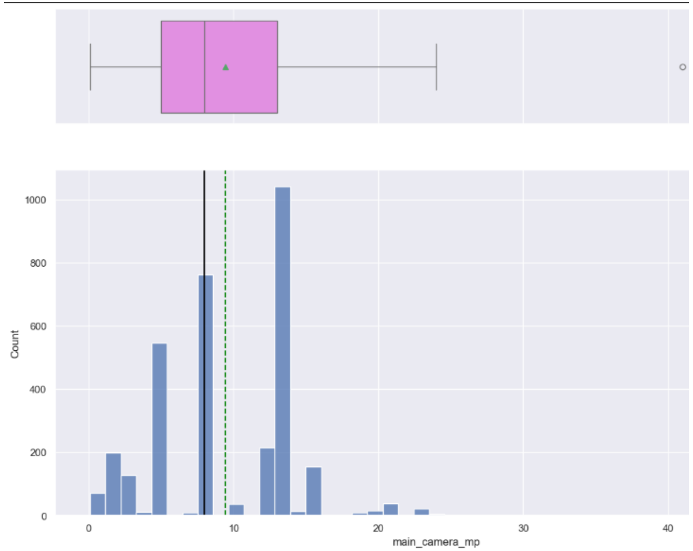
EDA Results

Screen size is mostly found in a tight range but there are outliers – more than likely tablet like devices.

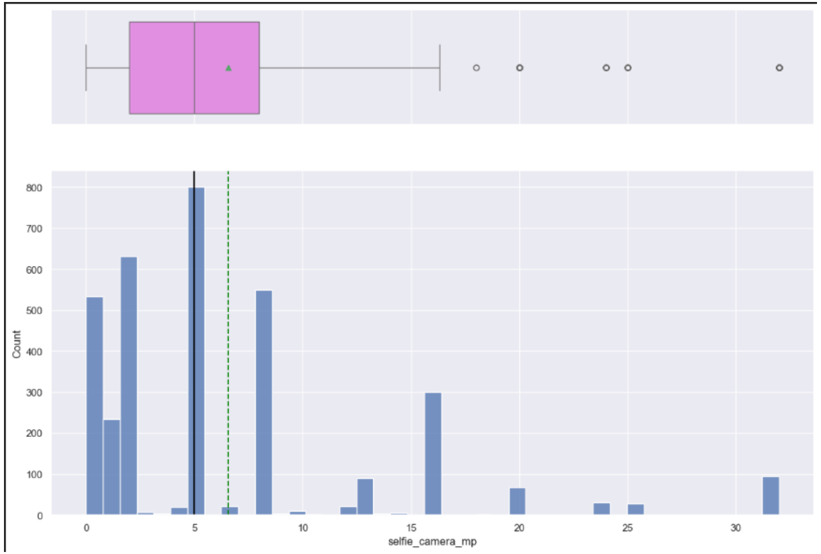


EDA Results

Main Camera Mega Pixels are in a similar range with upside outliers.



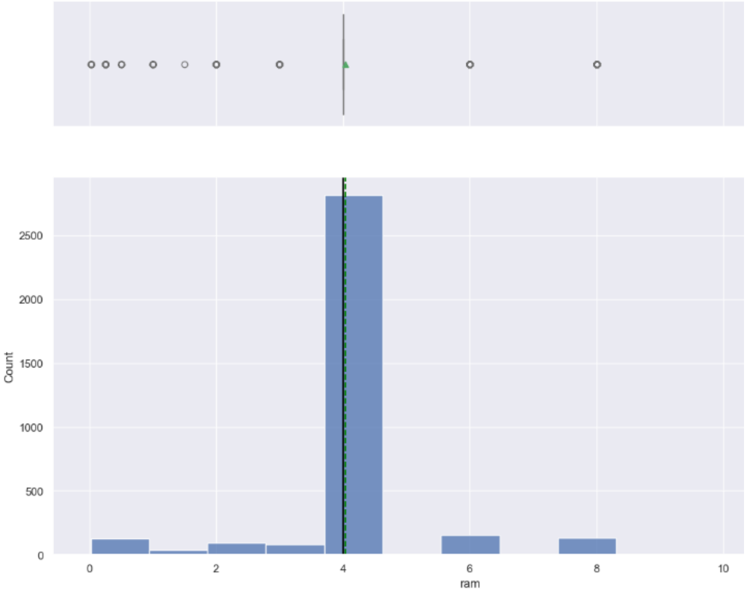
Selfie Camera Mega Pixels have a wider range with a higher quantity of outliers.

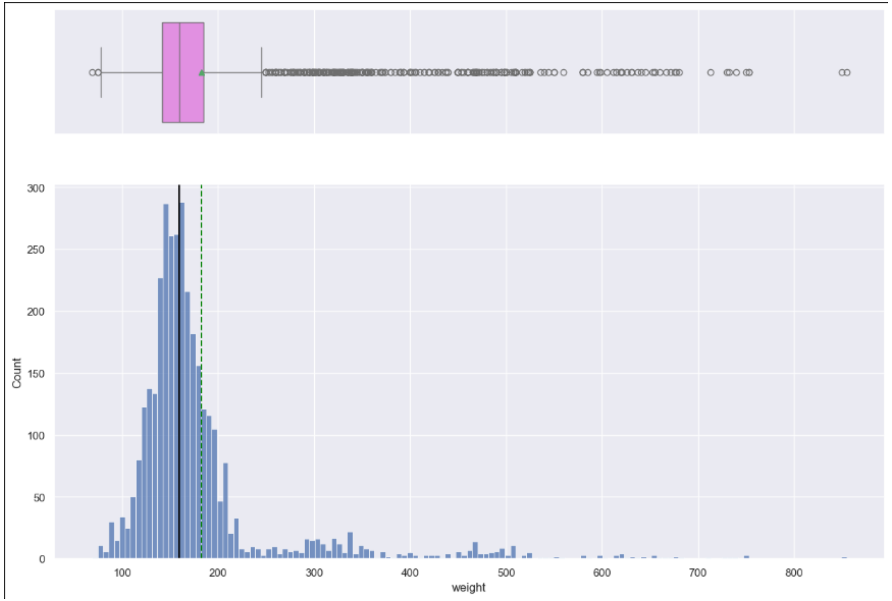


EDA Results

The RAM for mobile devices is mostly at 4Gb with very few above or below that level.

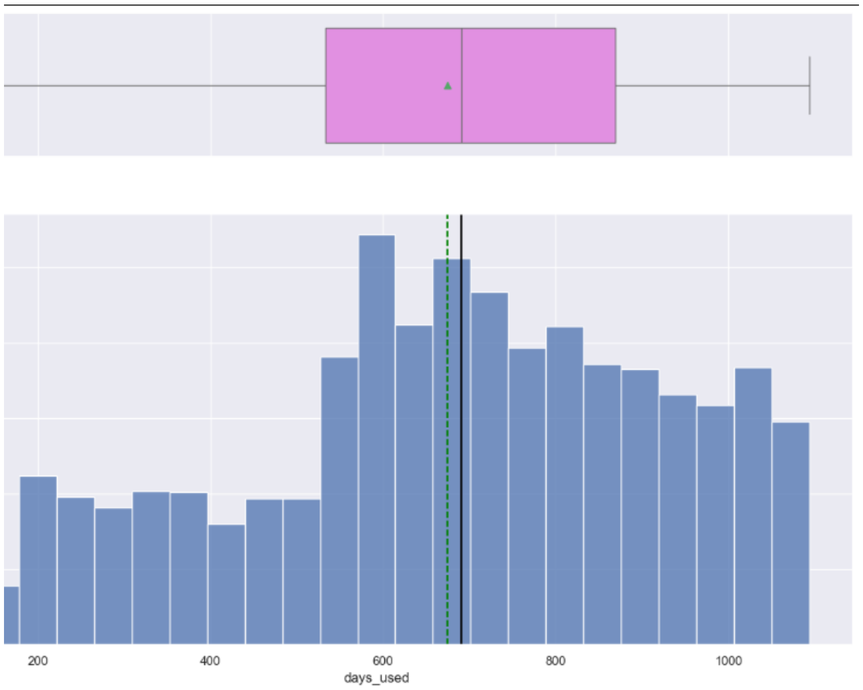
The weight for devices has a dense distribution between 100-200 grams but is very right-skewed with many outliers.





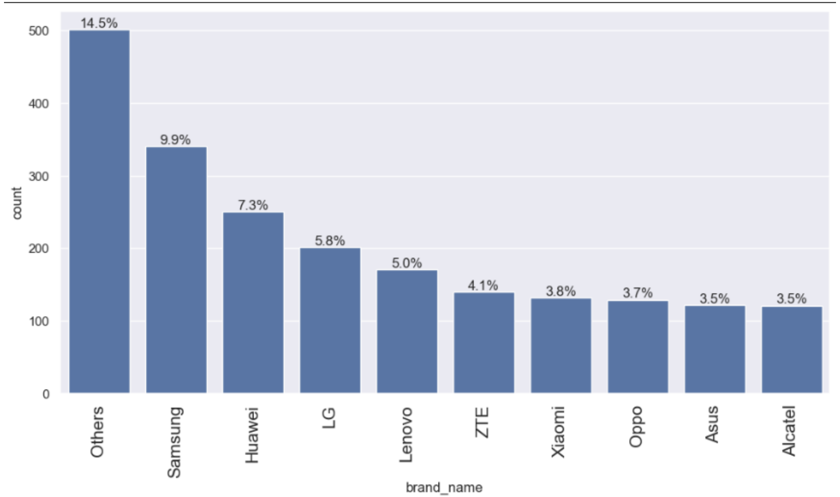
EDA Results

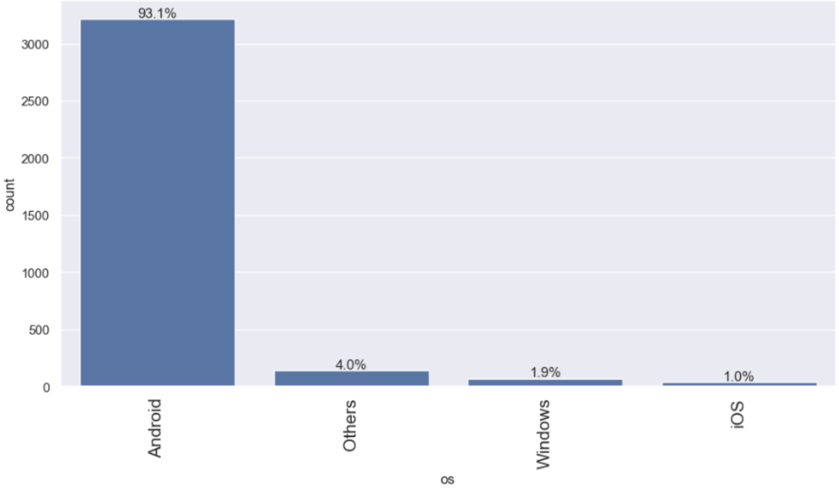
The number of days used spans a wide range with 75% of phone users owning their phone for 575 – 900 days.



EDA Results

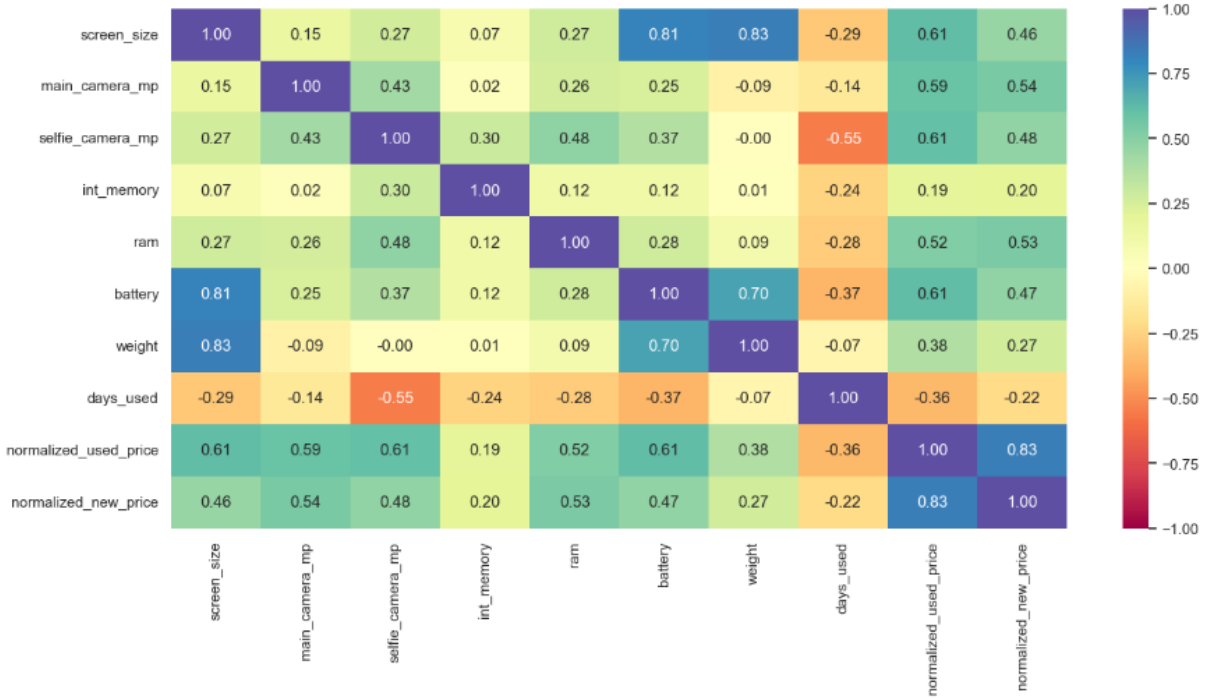
Brand name is as expected with an even distribution of lesser know names and the vast majority using Android OS.





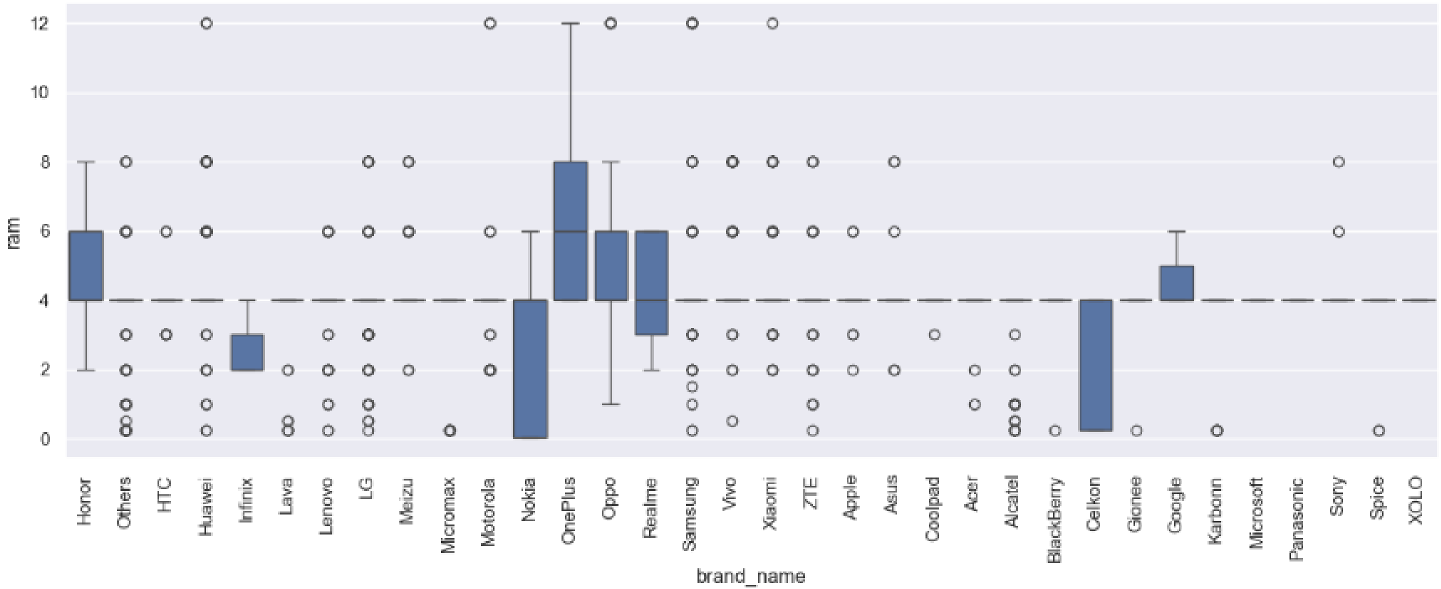
EDA Results

For correlation, only the most obvious show any significant correlation and that is battery size compared to screen size and weight compared to screen size.



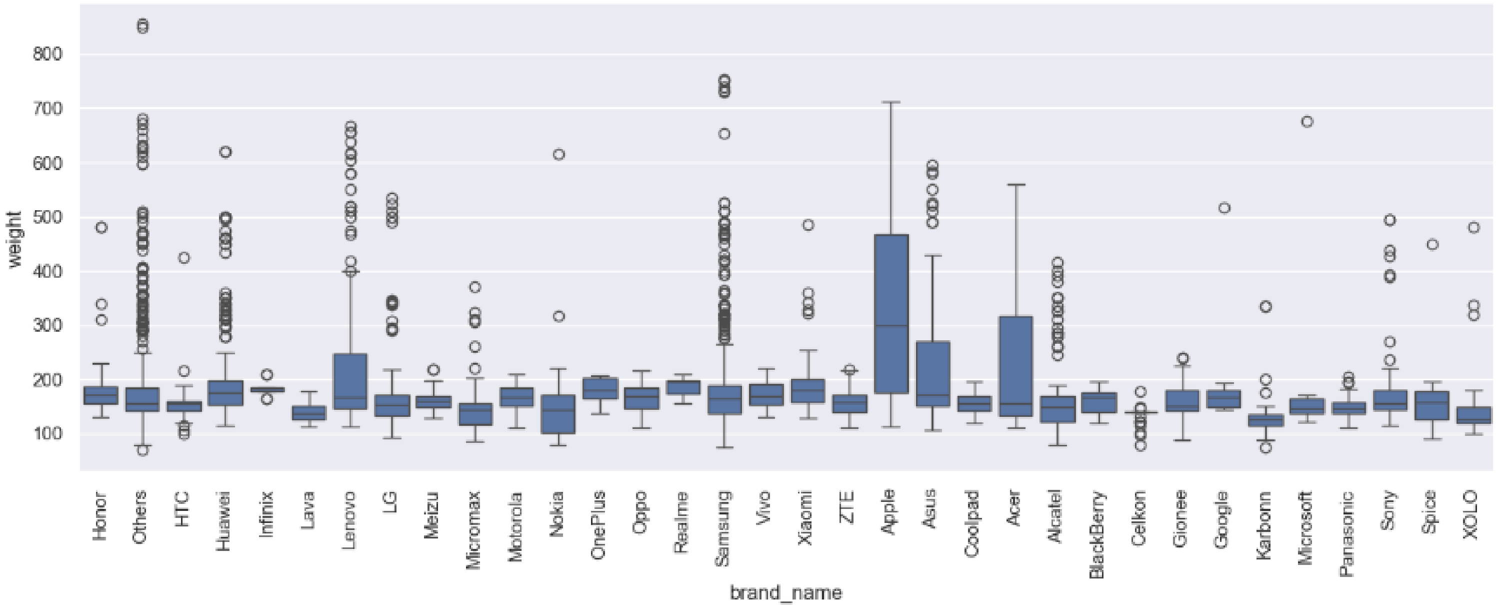
EDA Results

This chart displays the amount of RAM measured for each type of device. 4Gb is the standard across all devices but some have large variances.



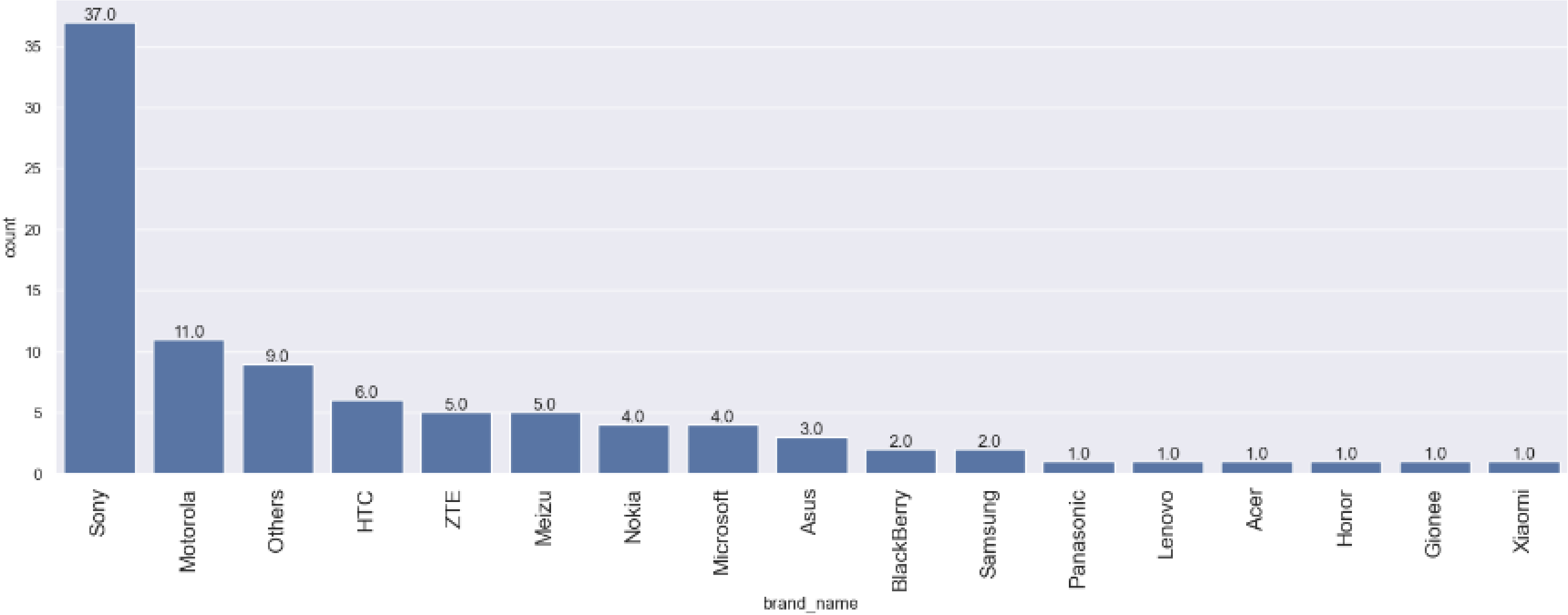
EDA Results

This chart displays the weight range for each manufacturer with the vast majority residing between 150 and 200 grams.



EDA Results

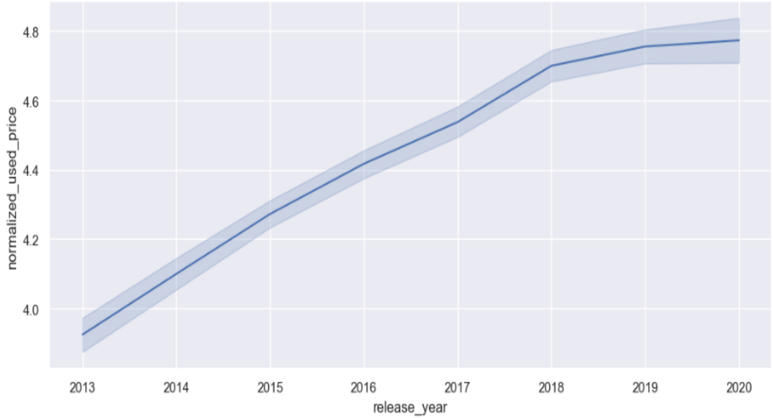
This chart displays the count of rear facing cameras with 16MP or higher by manufacturer.

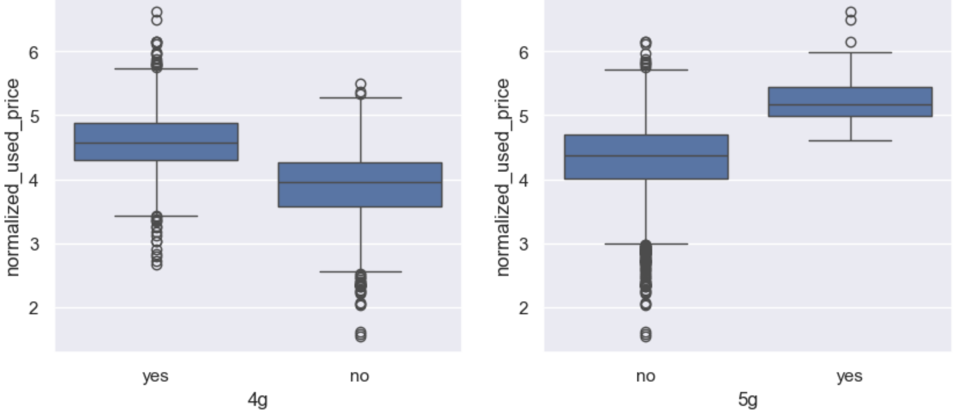


EDA Results

This chart displays the price of used devices across the years.

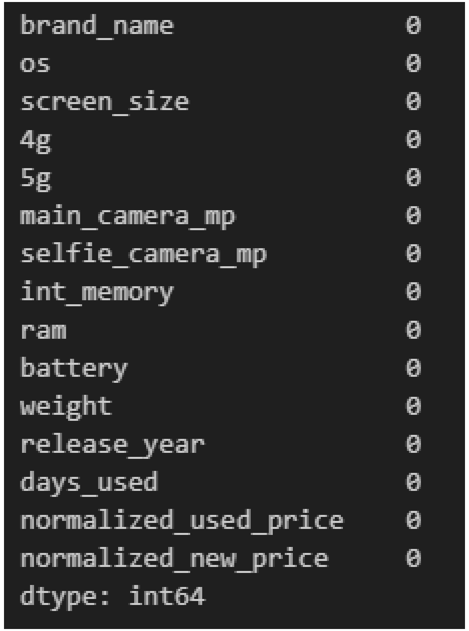
This chart shows price variance for used phones offering 4G and 5G networks.





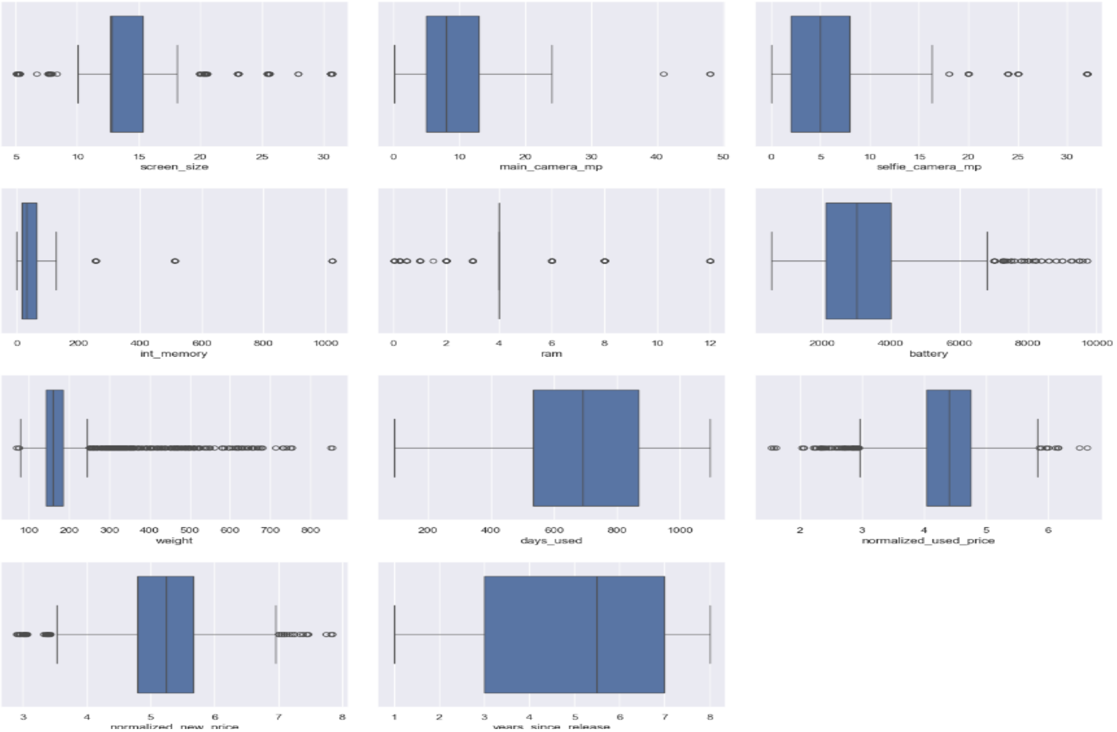
Data Preprocessing

Duplicate value check



```
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
```

Outlier check



Model Performance Summary

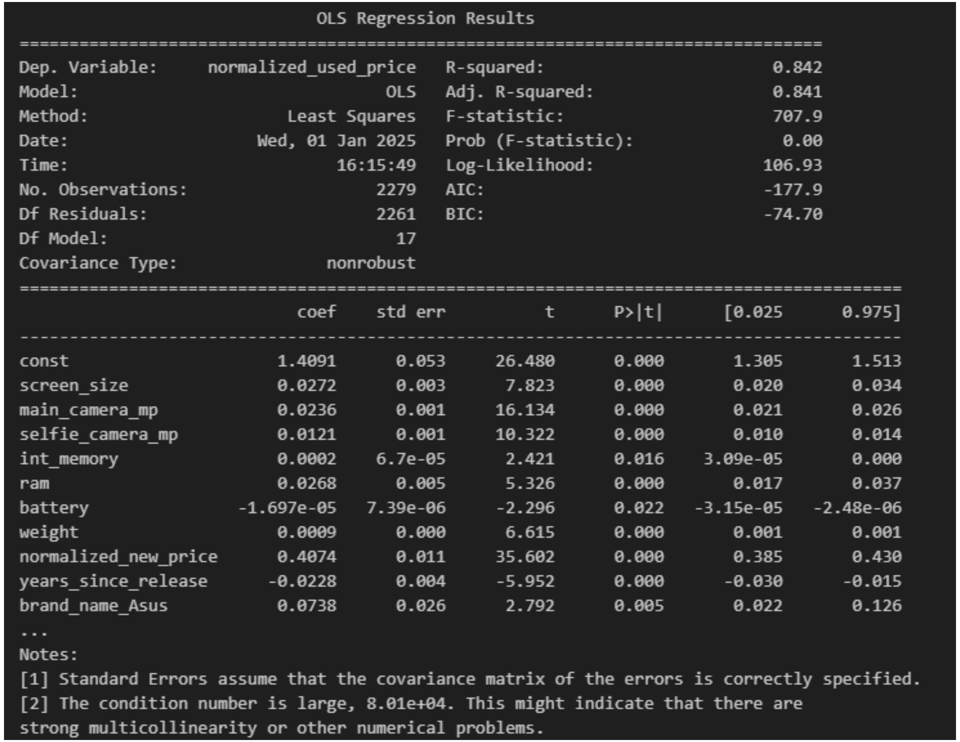
- Linear Regression Model built using Ordinary Least Squares(OLS) method.
  - Dependent variable – Normalized used price of cell phones
- Metrics
  - Observations – **2,279 data points** in the data set
  - Degrees of Freedom (Residual) – **2,231 of independent observations** after accounting for model
- Statistical Metrics
  - Nonrobust covariance – no adjustments for heteroscedasticity or serial correlation
  - **R-squared 0.843** – 84.3% of variance in the *normalized used prices* is explained by Ind. Var.
  - **Adjusted R-Sq** – 0.840 accounts for number of predictors, reflecting model fit adjusted for complexity
  - **F-statistic 255.6** – the high value suggest *strong predictive ability* of ind. Var. collectively
  - **Log-likelihood 117.97** – measure of model fit with higher values indicating a better fit

Model Performance Summary

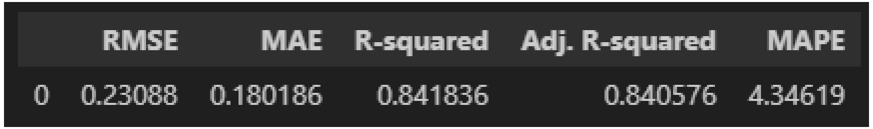
- Root Mean Squared Error (RMSE) **0.229764**
  - Measures avg magnitude of prediction errors
- Mean Absolute Error (MAE) **0.178979**
  - Represents avg absolute difference between observed and predicted values
- Mean Absolute Percentage Error (MAPE) **4.313467%**
  - Indicated avg prediction error as a percentage of actual values
- R-squared **0.843361** & Adj R-squared **0.839918**
  - Showing 84% of variance is explained by the model with the adjusted output reflecting model complexity and explanatory power

Model Performance Summary

Summary of key performance metrics for training data

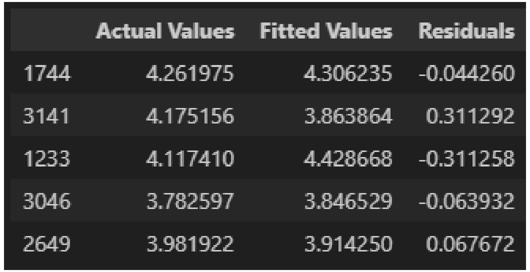


```
                            OLS Regression Results
==============================================================================
Dep. Variable:     normalized_used_price   R-squared:                       0.842
Model:                               OLS   Adj. R-squared:                  0.841
Method:                    Least Squares   F-statistic:                     707.9
Date:                   Wed, 01 Jan 2025   Prob (F-statistic):               0.00
Time:                           16:15:49   Log-Likelihood:                 106.93
No. Observations:                   2279   AIC:                            -177.9
Df Residuals:                       2261   BIC:                            -74.70
Df Model:                             17
Covariance Type:               nonrobust
======================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                     1.4091      0.053     26.480      0.000       1.305       1.513
screen_size               0.0272      0.003      7.823      0.000       0.020       0.034
main_camera_mp            0.0236      0.001     16.134      0.000       0.021       0.026
selfie_camera_mp          0.0121      0.001     10.322      0.000       0.010       0.014
int_memory                0.0002    6.7e-05      2.421      0.016    3.09e-05       0.000
ram                       0.0268      0.005      5.326      0.000       0.017       0.037
battery               -1.697e-05   7.39e-06     -2.296      0.022   -3.15e-05   -2.48e-06
weight                    0.0009      0.000      6.615      0.000       0.001       0.001
normalized_new_price      0.4074      0.011     35.602      0.000       0.385       0.430
years_since_release      -0.0228      0.004     -5.952      0.000      -0.030      -0.015
brand_name_Asus           0.0738      0.026      2.792      0.005       0.022       0.126
...
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.01e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

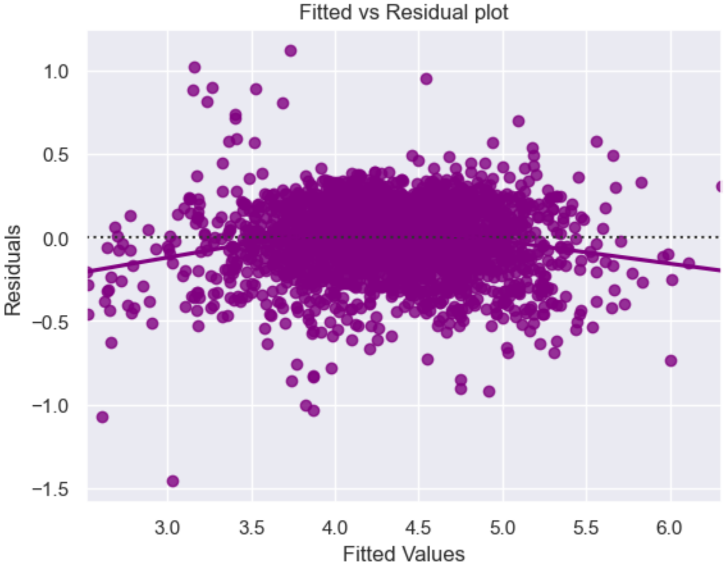


	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.23088	0.180186	0.841836	0.840576	4.34619

Model Performance Summary



	Actual Values	Fitted Values	Residuals
1744	4.261975	4.306235	-0.044260
3141	4.175156	3.863864	0.311292
1233	4.117410	4.428668	-0.311258
3046	3.782597	3.846529	-0.063932
2649	3.981922	3.914250	0.067672

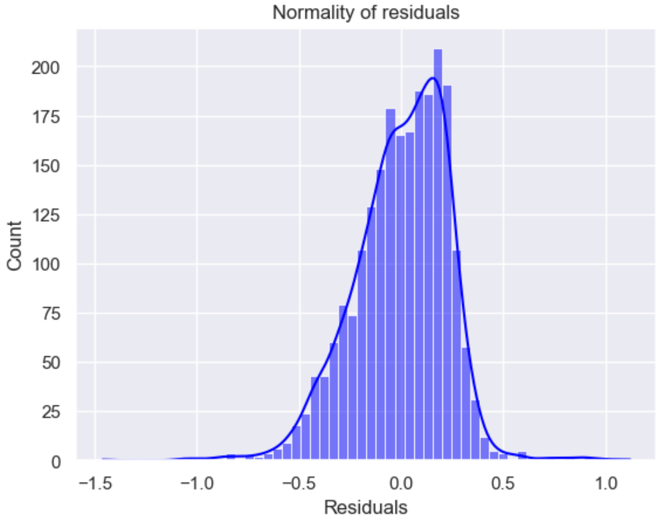


- Assumption of linearity in the relationship between independent variables and dependent variable holds
- Residuals are a small deviation of actual prices from predictions

---

- Homoscedasticity test by using goldfeldquandt test.
- P-value = 0.554434
- >0.05 means that residuals are homoscedastic.

Model Performance Summary

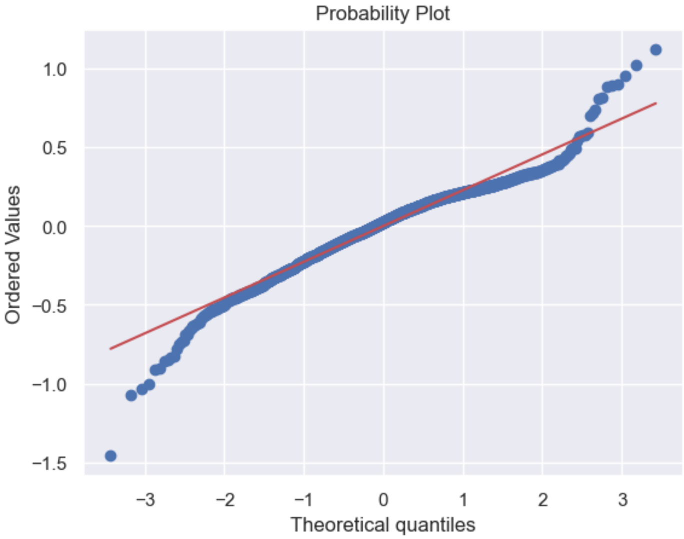


Testing for normality with a normal distribution.

Normal distribution by following straight line.

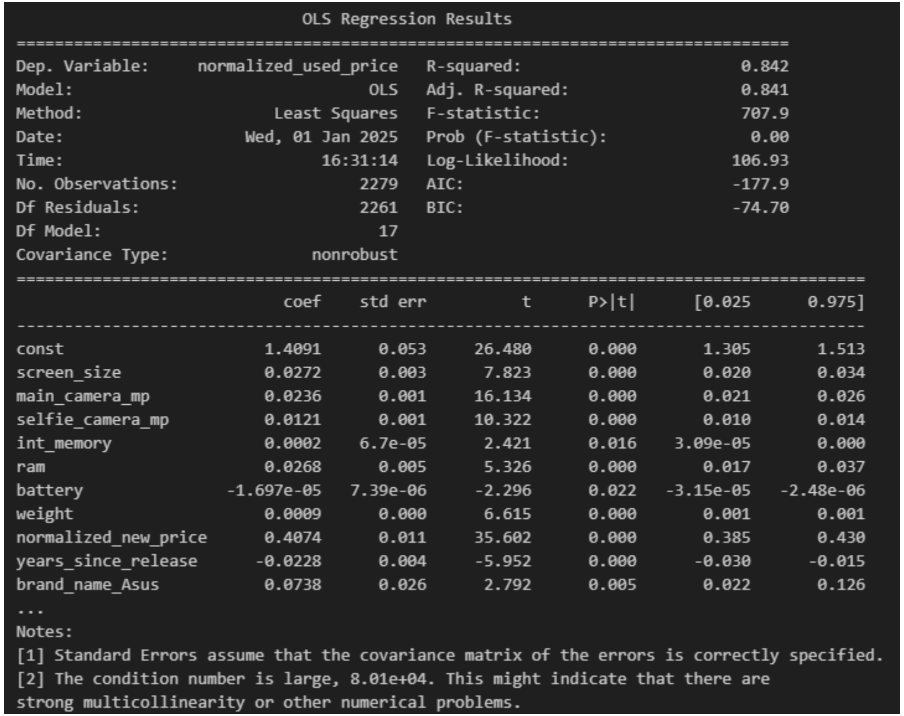
Confirmed by p-value of < 0.05 at

3.03872e-22.

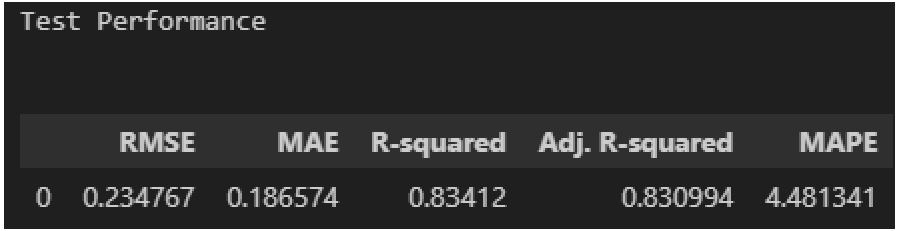


Model Performance Summary

Summary of key performance metrics for test data



```
                            OLS Regression Results
=================================================================================
Dep. Variable:     normalized_used_price   R-squared:                       0.842
Model:                               OLS   Adj. R-squared:                  0.841
Method:                    Least Squares   F-statistic:                     707.9
Date:                   Wed, 01 Jan 2025   Prob (F-statistic):               0.00
Time:                           16:31:14   Log-Likelihood:                 106.93
No. Observations:                   2279   AIC:                            -177.9
Df Residuals:                       2261   BIC:                            -74.70
Df Model:                             17
Covariance Type:               nonrobust
=========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                    1.4091      0.053     26.480      0.000       1.305       1.513
screen_size              0.0272      0.003      7.823      0.000       0.020       0.034
main_camera_mp           0.0236      0.001     16.134      0.000       0.021       0.026
selfie_camera_mp         0.0121      0.001     10.322      0.000       0.010       0.014
int_memory               0.0002    6.7e-05      2.421      0.016    3.09e-05       0.000
ram                      0.0268      0.005      5.326      0.000       0.017       0.037
battery              -1.697e-05   7.39e-06     -2.296      0.022   -3.15e-05   -2.48e-06
weight                   0.0009      0.000      6.615      0.000       0.001       0.001
normalized_new_price     0.4074      0.011     35.602      0.000       0.385       0.430
years_since_release     -0.0228      0.004     -5.952      0.000      -0.030      -0.015
brand_name_Asus          0.0738      0.026      2.792      0.005       0.022       0.126
...
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.01e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```



Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.234767	0.186574	0.83412	0.830994	4.481341

APPENDIX

Data Background and Contents

- The data contains different attributes of used/refurbished phones and tablets.
- Collected in 2021 with the following data:
    - Brand name, OS, Screen size, 4G or 5G
    - Main and Selfie Camera Mega Pixels
    - Internal Memory, RAM, Battery, Weight
    - Release year, Days used, new price and used price

Model Assumptions

Checking Linear Regression Assumptions

- Test for Multicollinearity using VIF
  - Remove columns with VIF score of > 5
  - Drop variable that makes least change in adjusted R-squared
  - Get all VIF scores under 5
- Test for linearity of variables
  - Make a plot of fitted values vs residuals – no pattern indicated that model is linear, and residuals are independent

Model Assumptions

Checking Linear Regression Assumptions Cont'd

- Test for normality of error terms
  - Check the distribution of residuals with Q-Q plot of residuals and the Shapiro-Wilks test
  - Residuals followed a normal distribution and made a straight-line plot
  - P-value of Shapiro-Wilk test was > 0.05, showing residuals are normally distributed
- Test for Heteroscedasticity
  - Used GoldFeldQuandt test – if p-value is > 0.05, residuals are homoscedastic
  - P-value = 0.554434

Great Learning

Happy Learning !