Project: Data Analysis and Model Development for New Pricing Strategy For Used and Refurbished Devices

Course Name: Supervised Learning Foundations

Ukwu, Kingsley Uchenna Azinna
July, 2024

Contents / Agenda

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

Executive Summary

Objective: The primary goal of this analysis was to build a linear regression model to predict the normalized used price of mobile phones based on various features such as screen size, camera specifications, memory, battery capacity, weight, and brand, among others.

The regression model developed shows strong predictive performance with an R^2 of 0.846 on the training data and 0.830 on the test data, indicating a strong fit. This means that approximately 83% of the variance in the normalized used price is explained by the model.

Homoscedasticity: The Goldfeld-Quandt test p-value of 0.829 indicates that the residuals are homoscedastic, meaning the variance of residuals is constant across all levels of the fitted values.

Normality: The residuals appear to be approximately normally distributed based on the histogram and Q-Q plot. This supports the assumption of normality in the residuals, which is important for inference and hypothesis testing.

Outliers: While the model is robust, outliers are present and should be monitored closely as they can affect model accuracy.

Business Strategy:Refine pricing strategies by highlighting significant features such as camera quality and screen size in marketing efforts.

Focus inventory management on brands and models with higher resale values and avoid those that depreciate quickly.

Executive Summary

Positive Influence: as can be seen from the table on correlation check and other univariate, bivariate and other analysis that Screen size, main camera megapixels, selfie camera megapixels, internal memory, RAM, weight, normalized new price, and 4G availability have a positive influence on the normalized used price. As such, focus should be here for better business numbers.

It is also seen that the used prices and sales improves as the release year gets recent as very old devices tend to be very poor on those features and does not comand good market attention.

Negative Influence: weight of the phone and the days devices were used negatively influence the normalized used price.

Brand Names: Some brand names (like samasung)are popular and comand demand in the market and therefore affects prices positively.

Continuous Monitoring: Regularly update and monitor the model to ensure its predictions remain accurate and relevant over time.The linear regression model provides valuable insights into the factors affecting the resale value of mobile phones. By leveraging these insights, stakeholders can make informed decisions regarding pricing, marketing, and inventory management. Continuous improvement and monitoring of the model will ensure it remains a reliable tool for predicting used mobile phone prices, ultimately aiding in strategic business planning and optimization.

Business Problem Overview

The used and refurbished device market has grown significantly driven by the cost savings these devices offer compared to new models, along with other benefits such as warranties, insurance, and reduced environmental impact, as well as COVID-19 pandemic which may further boost this market as consumers reduce discretionary spending and thne market is projected to reach $52.7 billion by 2023. ReCell, a startup in this market, seeks to develop a dynamic pricing strategy for used and refurbished devices

Problem Statement:

To develop a machine learning-based solution of a dynamic pricing strategy for the rapidly growing market of used and refurbished devices for ReCell by developing a linear regression model that can accurately predict the price of used and refurbished phones and tablets. The model should also identify the key factors that significantly influence the price, helping the company to strategize and maximize their market potential.

ReCell, the startup, needs a solution to:

1. Account for various factors influencing used device prices.
2. Set competitive and profitable prices for their used and refurbished devices.
3. Optimize their inventory management through accurate pricing.

Solution Approach

To address the problem of predicting the price of used and refurbished phones and tablets, the following methodology will be employed:

1. Data Collection: Gather data on used and refurbished phones and tablets from various sources g to include features such as device brand, model, release year, condition, main camera megapixels, selfie camera megapixels, internal memory, RAM, battery capacity, weight, and any other relevant attributes.

2. Data Preprocessing: Handling Missing Values, Removing Duplicates, Encoding Categorical Variables, Feature Scaling(Standardize or normalize features to ensure all variables are on a comparable scale).

3. Exploratory Data Analysis (EDA): Statistical Summary, Visualization.

4. Data Splitting: Train-Test Split: Split the data into training (70%) and testing (30%) sets to evaluate the model's performance on unseen data.

Solution Approach (Contd)

5. Model Building: Linear Regression Model (Build a linear regression model using the training data to predict the price of used devices), Multicollinearity Check (Use Variance Inflation Factor (VIF) to identify and address multicollinearity among predictors), Feature Selection (Remove features with high p-values (>0.05) iteratively to retain only significant predictors)

6. Model Evaluation: Training Performance (Evaluate the model's performance on the training data using metrics such as R-squared, Adjusted R-squared and Root Mean Squared Error (RMSE)), Testing Performance (Evaluate the model's performance on the testing data using the same metrics to ensure fitting).

7. Assumptions Verification: Linearity (Check for patterns in residual plots to verify the linearity assumption), Independence of Errors (Ensure residuals are independent), Normality of Errors (Assess the normality of residuals using Q-Q plots and the Shapiro-Wilk test), Heteroscedasticity (Perform tests such as the Goldfeld-Quandt test to check for constant variance of residuals).

EDA Results

Prelimnary data overview shows 15 columns (brand name, OS, Screen size, 4g, 5g, Main Camera mp, Selfie Camera mp, internal memory, Ram, battery, weight , release year, days used, normalized USED price, as well as normalized NEW price. it is also shows that the data has 3454 rows

For the data types; 9 of the columns were float , 2 are integers and 4 are object type.

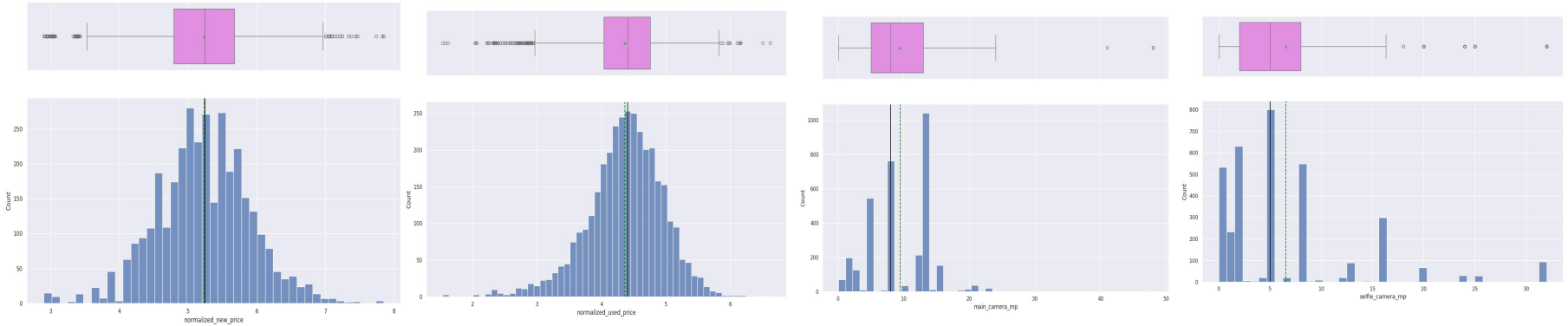
From the statistical summary, the median prices of the USED and NEW phones (4.40, 5.24) is the same as the their averages/mean prices (4.36,5.23) of the two categories respectively.

While their both standard deviations are similar, the prices of the new is seen to be i unit price higher that the prices of the used across the minimum, mean, median and amximum as well as the 25th and 75th percentiles as well.

We also saw some items were duplicated, and we have missing items especially on main camera mp colum, selfie camera, battery, weight, internal memory and Ram.

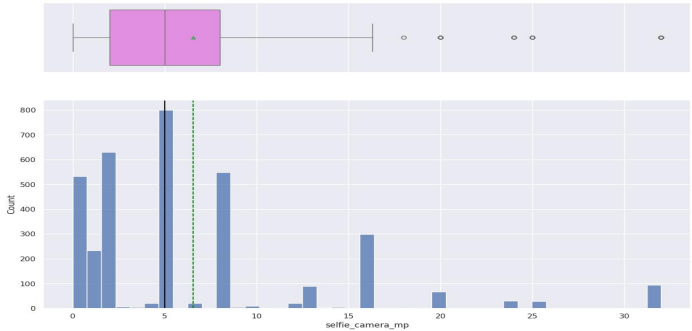
Link to Appendix slide on data background check

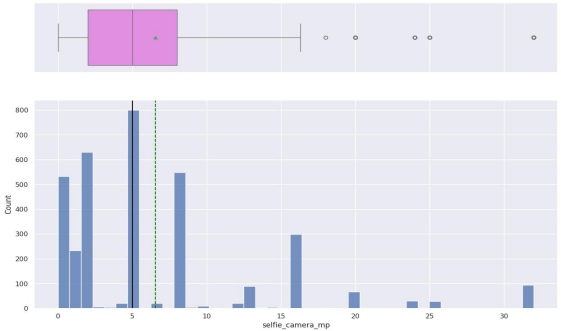
EDA Results

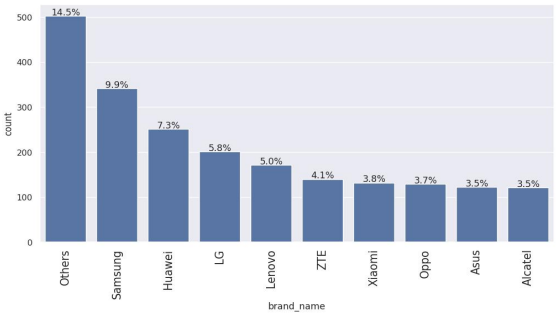


Though we have outliers on both new and Used, the hsitogram of the used devices is more “organized” bell shaped than the NEW. But that of the screen size is skewed so much to the right. Theselfie camera shosw more outliers to left and strong skew to the left as well which shows some few phones have unusually high camera mp

EDA Results







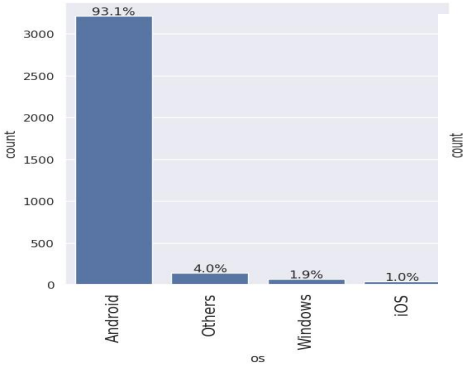
With the average and median weight being 182 and 160 respectively, while the 75th percetile is 185 yet we have maximum weight at 855, the is strongly skewed to the right with a lot of outliers as shown in the box plot. This is about the same trend in selfie camera, main camera, Ram and Battery as well as screen size.

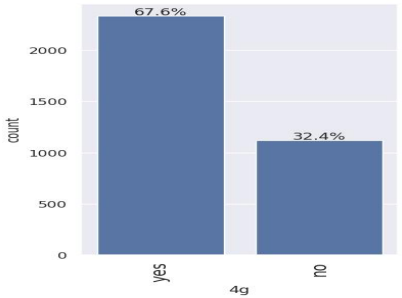
50% of the phone are used for less than 700 days and this is about the averge/ mean of the days the phones were used. less than 25% of the phones are used for more that 870 days. while the most days days a phone was used in this market is about 1100 days.

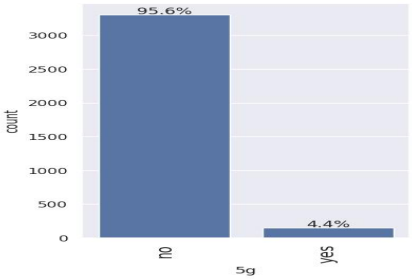
Most brand in this market is samdung at 10%, Huawei at 7% and LG at 6%, this shows those brands have better second hand use value. Lonovo ansd some other brands are pulling up however.

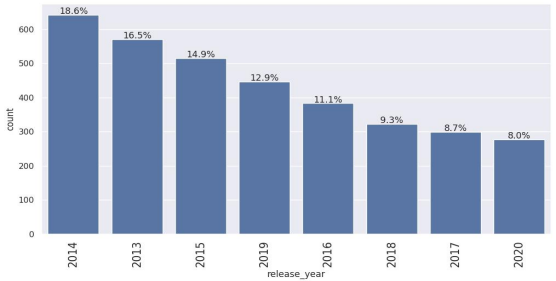
Link to Appendix slide on data background check

EDA Results









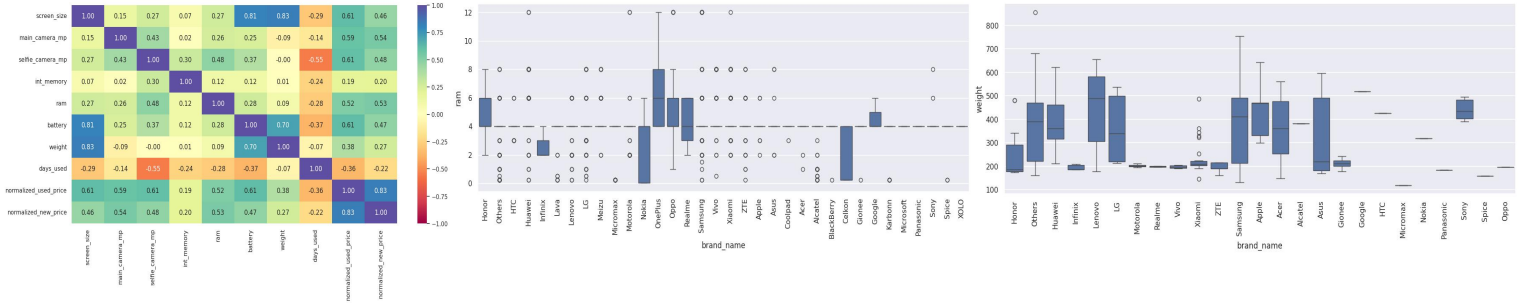
With over 93% in the data, android based phones is clearly the leader in the OS with no clear competitor. IOS and Windows are stongly with less than 3% combined while all others are just around 4%. so the focus should be on the android for ReCell as that is where the numbers will come from, while they see recomendations that will improve IOS, windows and others

67% of the phones are 4g, while less than 4.5% are 5g. this shows that used market is still shy on 5g devices and understanbly so as that is a new technology and mostly still on new devices.

Most of thye useed devices were released in 2014 with 19% showing, while we trended down to less than half of that for 2020 at 8% in the showing.

Link to Appendix slide on data background check

EDA Results



We saw that the price of the new device has a high correlation with the price of the used device at 0.83, this is also the case with the weight of the device having high correlation with the screen size. The battery also shows very high correlation with the cscreen size as it should be as the screen size will obviously need more energy to power.

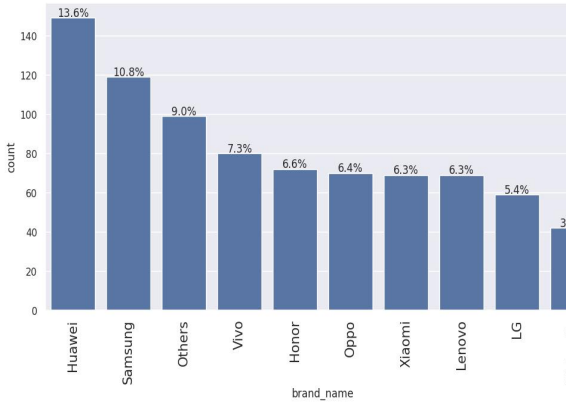
We also saw tha the camera and battery has a stronger positive correlation with the used phone price than the do with new phone price indicating that this may be the major driver of the prices of the used devices.

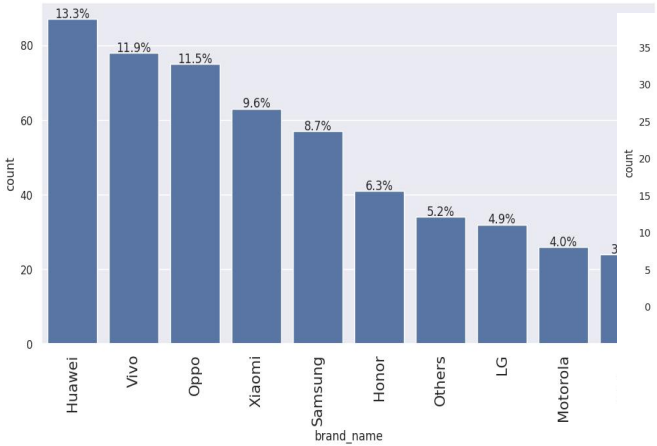
We also saw from the bivvariate that One Plus, Honour, OPPO, Realme and Google has higher Ram, even higher the sample average of 4, though we have a few outliers in VIVO, Huawei, Motorola, Xaomi among others.

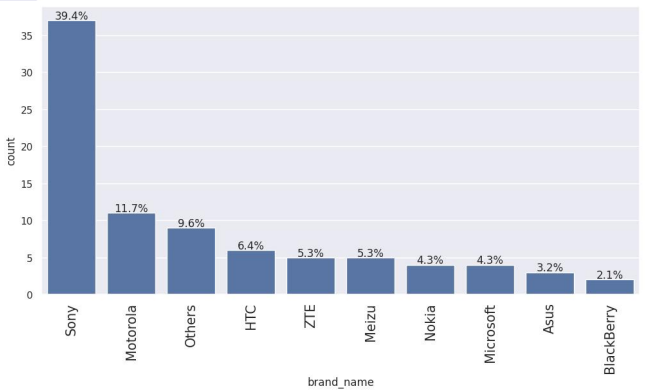
Lenovo devices appaer to be more weighty when they have large batterues, this also applies to Samsung, Asus, Motorola and Apple.

Link to Appendix slide on data background check

EDA Results





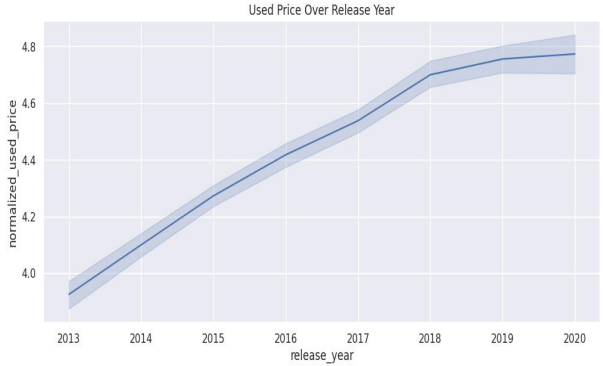


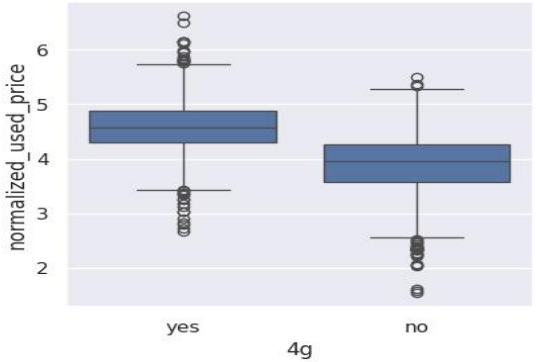
Huawei devices appear to have the lagest screen from the data followed by Samsung while motorola and LG has the least in this category. Huawei also shows strong front camera for selfies followed by VIVO and Oppo while Meizu and Motorola has the least here aslwell.

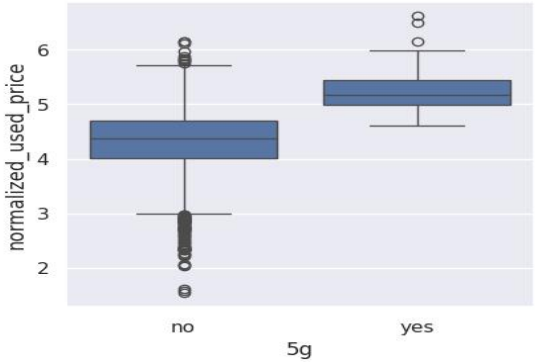
For then rear camera, we have Sony taking the lead at 39% with motorola as distant second at 11.7%. Bl;ackberry and Asus has the least.

Link to Appendix slide on data background check

EDA Results







The price of the used device is seen growing consistently as the release years becomes recent from 2013 till 2020 with sharpest growth seen between 2013 and 2018. while the growth is still steady till 2020.

The prices of devices with 4g is also hogher than those that doest have 4g, and this suggest the demand for those actegories of phones.

The price of the phones with 5g is also way higher than those without 5g with strong outliers seen there. This implies that those phones with 4g and 5g drives the market and prices with better demand.

Link to Appendix slide on data background check

Data Preprocessing

Missing value & treatment:

There were 179 missing information on the main Camera mp, 2 on the selfie camera, 4 each on internal memory and ram, 6 and 7 items on battery and weight repsctively. This were treated by filling them with the median.

Feature engineering:

We created a new column called “Year since released” from the “released year” considering 2021 as the base line year. This is ot have a better perfromance in the machine learning, then we droped the release yera column.

```
df1["years_since_release"] = 2021 - df1["release_year"]

df1.drop("release_year", axis=1, inplace=True)

df1["years_since_release"].describe()
```

Data Preprocessing

Checking for outliers (and Treatment if needed):

Outliers can negatively affect the performance of machine learning models, especially those that are sensitive to the distribution of the data like linear regression. We had to check for outliers using the code below

```
num_cols = df1.select_dtypes(include=np.number).columns.tolist()

plt.figure(figsize=(15, 15))

for i, variable in enumerate(num_cols):

plt.subplot(4, 3, i + 1)

sns.boxplot(data=df1, x=variable)

plt.tight_layout(pad=2)

plt.show()
```

We saw from the boxplot of the data that apart from “years since relased” and “days used” all other ones have strong outliers.

Data Preprocessing

We had to to encode categorical features, then we split the data into train and test to be able to evaluate the model that we build on the train data. We then will build a Linear Regression model using the train data and then check it's performance. Using the code below, we got the dependent and independent variables :

```
X = df1.drop(columns=['normalized_used_price'])

y = df1['normalized_used_price']

print(X.head())

print()

print(y.head())
```

We then added the intercept : *X = sm.add_constant(X)*

Created a dummy variable for independent feature with :

```
X = pd.get_dummies(X,columns=X.select_dtypes(include=["object", "category"]).columns.tolist(), drop_first=True,
```

We then split the data into 70:30 for Train and Test with the following code

x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42), number of rows on Train and Test data are 2417 and 1037 respectively.

Model Performance Summary

Overview of ML model and its parameters

This is the output of a linear regression ML model predicting the normalized_used_price of used phones/tablets. The given model is an Ordinary Least Squares (OLS) regression, which is used to predict the normalized used price of devices based on various features such as screen size, camera specifications, memory, brand, and more.

Key Features in the Model includes :

Numeric Features - (screen_size: Size of the device screen, main_camera_mp: Megapixels of the main camera, selfie_camera_mp: Megapixels of the selfie camera, int_memory: Internal memory size, ram: RAM size, battery: Battery capacity, weight: Weight of the device, days_used: Number of days the device has been used, normalized_new_price: Normalized price of the device when it was new, years_since_release: Number of years since the device was released.

Categorical Features (Encoded as Booleans):

1. brand_name_*: Various brand names encoded as boolean features (e.g., brand_name_Apple, brand_name_Samsung).
2. os_*: Operating systems encoded as boolean features (e.g., os_iOS, os_Windows).
3. 4g_yes: Whether the device supports 4G.
4. 5g_yes: Whether the device supports 5G.

The model explains 84.8% of the variance in the normalized used price (R-squared = 0.848). This is a good indication of the model's ability to fit the data.

Link to Appendix slide on model assumptions

Model Performance Summary

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

Intercept: const: 1.4300 (constant term indicating the baseline normalized used price when all other predictors are zero).

Significant Predictors: screen_size, main_camera_mp, selfie_camera_mp, int_memory, ram, weight, normalized_new_price, years_since_release, and 4g_yes have significant coefficients (p-values < 0.05), indicating they significantly affect the normalized used price.

Insignificant Predictors: Many brand name indicators, battery, days_used, 5g_yes, and some OS indicators have high p-values, suggesting they may not significantly affect the normalized used price in this model.

Model Performance: R-squared: 0.848, indicating that approximately 84.8% of the variance in the normalized used price is explained by the model.

Adjusted R-squared: 0.845, slightly lower than R-squared, adjusting for the number of predictors in the model.

F-statistic: 276.3 (p-value: 0.00), indicating the overall significance of the model. F-statistic (high value with a very low p-value) suggests the model is statistically significant, meaning the features together have a significant effect on the normalized used price.

Link to Appendix slide on model assumptions

Model Performance Summary

- Summary of key performance metrics for training and test data in tabular format for comparison

Coefficients: Positive coefficients (e.g., screen_size, main_camera_mp, selfie_camera_mp, int_memory, ram, weight, normalized_new_price, 4g_yes) indicate an increase in the normalized used price with an increase in these features.

Negative coefficients: (e.g., years_since_release, battery, brand_name_* for some brands) indicate a decrease in the normalized used price with an increase in these features or specific brand association.

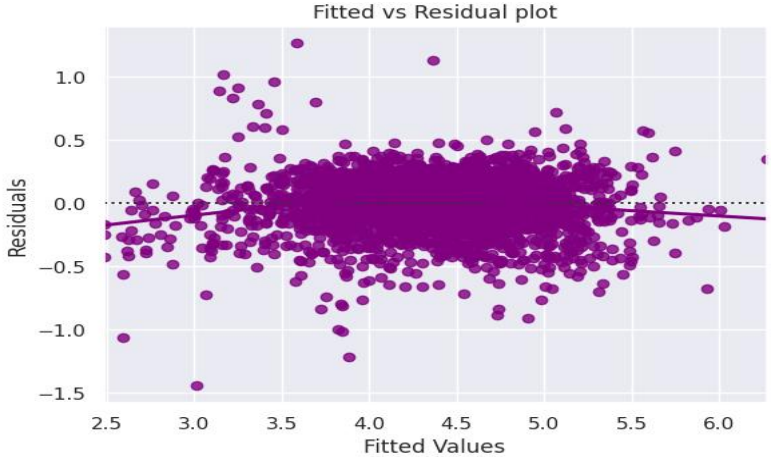
Standard Errors: Indicate the variability of the coefficient estimates. Lower standard errors suggest more precise estimates.

t-Statistics and p-values: Used to test the null hypothesis that a coefficient equals zero. A low p-value (< 0.05) indicates that you can reject the null hypothesis for that predictor.

Confidence Intervals: Provide a range of values within which the true coefficient is likely to fall.

Link to Appendix slide on model assumptions

Model Performance Summary

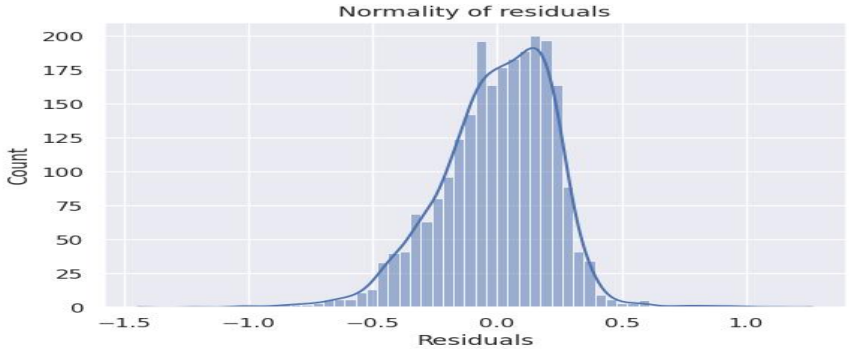


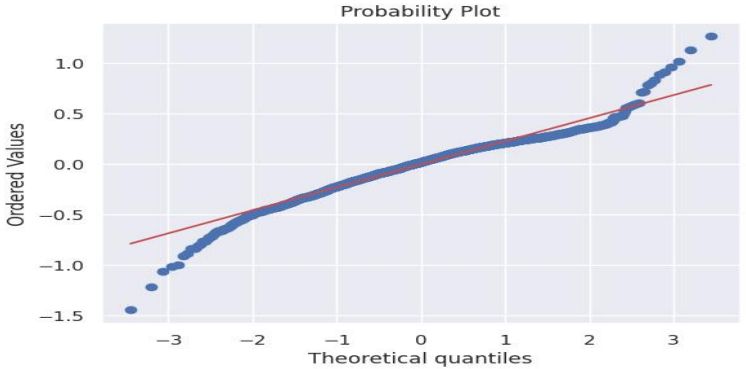
TEST FOR LINEARITY AND INDEPENDENCE No clear pattern: This is a good sign, suggesting that the residuals are randomly distributed and so the model is linear and residuals are independent. The residuals are randomly scattered around the horizontal axis (zero line), it indicates that the model's assumptions are likely met, and there are no major issues with heteroscedasticity or non-linearity.

TEST FOR HOMOSCEDASTICITY: Based on the test results,(p-value = 0.8298) suggests that we fail to reject the null hypothesis of homoscedasticity, we do not find evidence of heteroscedasticity in the residuals. This indicates that the assumption of homoscedasticity holds, and the residuals have constant variance across the data, which is a desirable property for a linear regression model. This is a good sign, indicating that the model's assumptions might be met on homoscedasticity.

Model Performance Summary

TEST FOR NORMALITY:





Histogram: The residuals histogram shows a roughly normal distribution.

Q-Q Plot: The points mostly lie on the line, indicating that the residuals are approximately normally distributed.

Shapiro-Wilk Test: If the p-value is greater than 0.05, we can conclude that the residuals are normally distributed. (You would need to run the Shapiro-Wilk test using the actual residual data to get the p-value.)

Overall, the histogram and Q-Q plot suggest that the residuals are approximately normally distributed, although there may be some minor deviations at the tails. This is generally acceptable for many linear regression models.

Link to Appendix slide on model assumptions

Model Performance Summary

Metric	Description	Training Data	Test Data
R-squared	Proportion of variance explained by the model	0.8457	0.8296
Adjusted R-squared	R-squared adjusted for the number of features	0.8429	0.822
RMSE	Root Mean Squared Error (average prediction error on the target variable scale)	0.2321	0.2409

The gap between training and test data metrics is relatively small, suggesting the model generalizes reasonably well to unseen data. Both R-squared values (above 0.8) indicate the model explains a significant portion of the variance in the normalized used price for both training and test data and the RMSE values is relatively low, indicating good predictive accuracy and are on a similar scale, suggesting comparable prediction accuracy on both datasets.

Link to Appendix slide on model assumptions

Model Performance Summary

Similar to R-squared, the adjusted R-squared values are high and close to each other for both datasets, suggesting that the model performs consistently well on both. Both the training and test data have high R-squared values, indicating that the model explains a substantial portion of the variance in the dependent variable for both datasets.

These metrics indicate a well-performing model. It explains a high proportion of the variance in the target variable (normalized used price) on both training and test data, with a relatively small difference in performance. The RMSE values suggest reasonable prediction accuracy on unseen data. Overall, the model appears to perform well on both the training and test data..

The regression model shows an R^2 of 0.848 on the training data and 0.830 on the test data, indicating a strong fit. This means that approximately 83% of the variance in the normalized used price is explained by the model.

Link to Appendix slide on model assumptions

APPENDIX

Data Background and Contents

Data Background and Contents

Overview

The dataset used in this analysis comprises various features of mobile phones, focusing on predicting the normalized used price based on these features. The data includes both numerical and categorical variables, with categorical variables being one-hot encoded for the regression model.

The dataset contains the following key Numerical features

screen_size: The size of the mobile phone screen in inches.

main_camera_mp: The resolution of the main camera in megapixels.

selfie_camera_mp: The resolution of the front camera in megapixels.

int_memory: Internal memory capacity in gigabytes.

ram: Random Access Memory in gigabytes.

battery: Battery capacity in milliampere-hours (mAh).

weight: Weight of the mobile phone in grams.

Data Background and Contents

days_used: Number of days the phone has been used.

normalized_new_price: The normalized price of the phone when new.

years_since_release: Number of years since the phone was released.

and **Categorical Features**

brand_name: The brand of the mobile phone (e.g., Apple, Samsung, Xiaomi, etc.).

os: The operating system of the mobile phone (e.g., iOS, Android, Windows).

4g: Whether the phone supports 4G technology (yes/no).

5g: Whether the phone supports 5G technology (yes/no).

Target Variable: normalized_used_price: The normalized price of the phone in the used market.

Data Collection and Baseline Year

The dataset considers 2021 as the baseline year for calculating the years_since_release feature, which is derived from the release_year column.

Data Background and Contents

Model Performance

Training Data:

R-squared: 0.846

Adjusted R-squared: 0.843

RMSE: 0.232

Test Data:

R-squared: 0.830

Adjusted R-squared: 0.822

RMSE: 0.241

Model Assumptions

Tests conducted to check model assumptions includes checking the following Linear Regression assumptions:

1. No Multicollinearity

2. Linearity of variables

3. Independence of error terms

4. Normality of error terms

5. No Heteroscedasticity

Linearity Test: Scatter plot of fitted values vs. actual values. If there is no pattern, then we say the model is linear and residuals are independent, Otherwise, the model is showing signs of non-linearity and residuals are not independent.

First we created a data framework with actual, fitted and residual with the follwoing code:

```
df_pred = pd.DataFrame()
df_pred["Actual Values"] = y_train  # actual values
df_pred["Fitted Values"] = olsmodel2.fittedvalues  # predicted values
df_pred["Residuals"] = olsmodel2.resid  # residuals
df_pred.head()
```

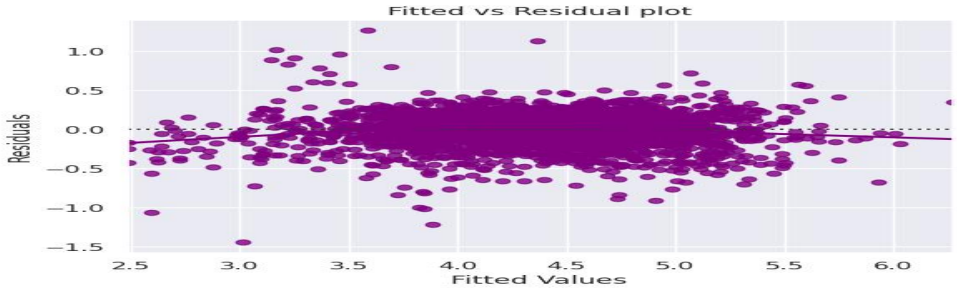
Model Assumptions

Actual Values		Fitted Values	Residuals
1744	4.261975	4.269474	-0.007499
3141	4.175156	3.838765	0.336391
1233	4.117410	4.441194	-0.323784
3046	3.782597	3.800766	-0.018169
2649	3.981922	3.901886	0.080036

Then we plot the fitted values vs residuals with the following code:

```
sns.residplot(data=df_pred, x="Fitted Values", y="Residuals", color="purple", lowess=True)
plt.xlabel("Fitted Values")
plt.ylabel("Residuals")
plt.title("Fitted vs Residual plot")
plt.show() .
```

Model Assumptions



The results shown has no pattern and so we say the model is linear and the residuals are independent.

MulticollinearityTest: Variance Inflation Factor (VIF). If VIF is 1 then there is no correlation between the kth predictor and the remaining predictor variables.

If VIF exceeds 5 or is close to exceeding 5, we say there is moderate multicollinearity.

If VIF is 10 or exceeding 10, it shows signs of high multicollinearity.

Result: We continued to drop every column one by one that has a VIF score greater than 5, Look at the adjusted R-squared and RMSE of all these models. Drop the variable that makes the least change in adjusted R-squared.Check the VIF scores again. Continue till you get all VIF scores under 5.

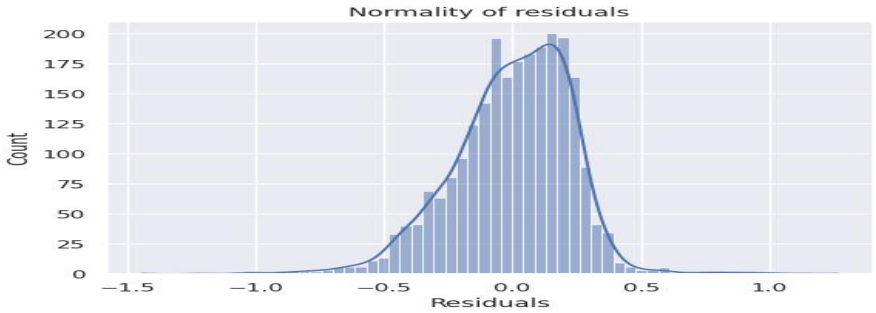
Result VIF values for the predictors were below the commonly used threshold of 10, indicating that multicollinearity is not a significant issue in the model.

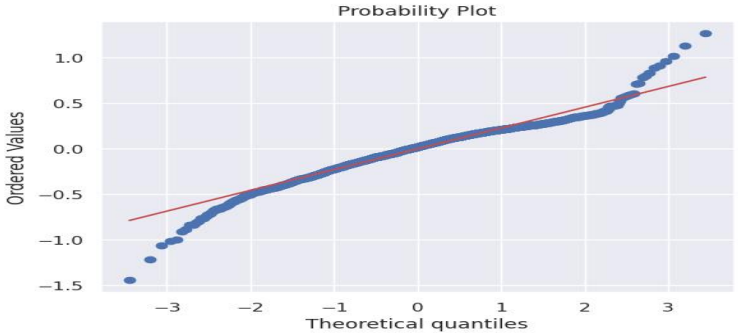
Model Assumptions

Homoscedasticity Test: Goldfeld-Quandt test.

Result: The Goldfeld-Quandt test yielded an F-statistic of 0.946 and a p-value of 0.830. Since the p-value is greater than 0.05, we fail to reject the null hypothesis of homoscedasticity, confirming that the residuals have constant variance.

Normality Test: Shapiro-Wilk test, histogram of residuals, and Q-Q plot.





Result:

Histogram: The distribution of residuals appeared approximately normal.

Q-Q Plot: The Q-Q plot of the residuals showed that the points roughly followed a straight line.

Shapiro-Wilk Test: Given the visual confirmation of normality, it was assumed that the residuals are approximately normally distributed

Model Assumptions

Outliers and Influential Points Test: Box plots and Cook's distance.num_cols = df1.select_dtypes(include=np.number).columns.tolist()

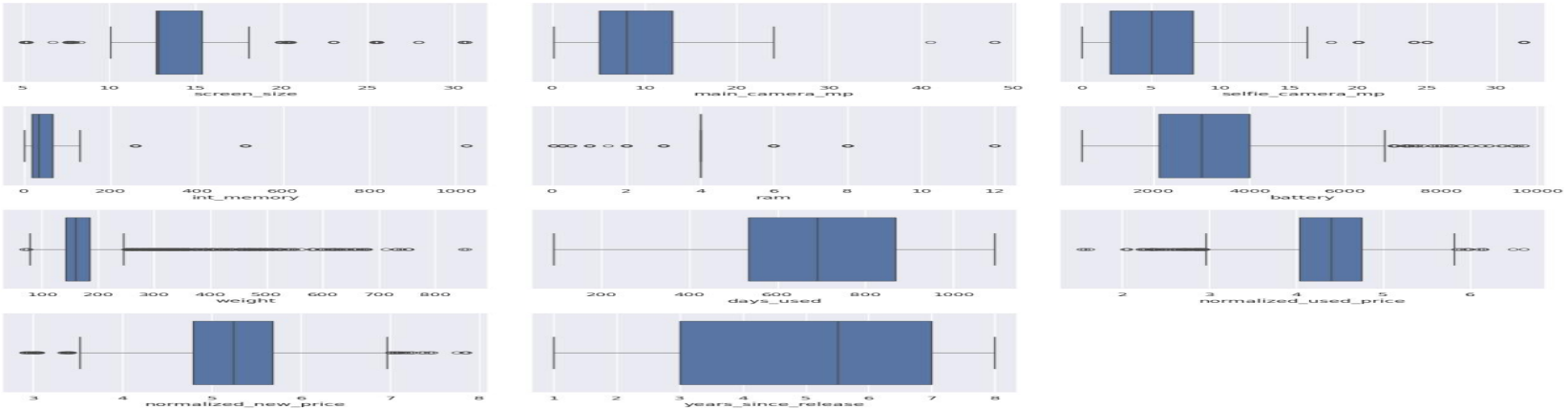
plt.figure(figsize=(15, 15)) for i, variable in enumerate(num_cols):

plt.subplot(4, 3, i + 1)

sns.boxplot(data=df1, x=variable)

plt.tight_layout(pad=2)

plt.show()



Result: The box plots identified several potential outliers. Cook's distance was also used to check for influential points, and a few data points were found to have a high influence on the model. These points need further investigation to determine if they should be addressed or removed.

Model Assumptions

Summary of Test Results:

Assumption	Test Used	Result
Linearity	Scatter plot	Linear trend observed
Independence	Durbin-Watson statistic	Statistic ~1.991 (no significant autocorrelation)
Homoscedasticity	Goldfeld-Quandt test	F-statistic = 0.946, p-value = 0.830 (constant variance)
Normality	Shapiro-Wilk, Histogram, Q-Q plot	Residuals approximately normal
Multicollinearity	Variance Inflation Factor (VIF)	VIF values below 10 (no significant multicollinearity)
Outliers	Box plots,	Identified outliers and influential points

These tests confirm that the linear regression model meets the necessary assumptions, ensuring the reliability and validity of the model's results. Further refinement and regular monitoring are recommended to maintain the model's performance.

Great Learning