Mobile Price Insights: Unveiling Trends through Exploratory Data Analysis

Introduction

Within the realm of data science, grappling with real-world datasets frequently introduces hurdles like missing values, outliers, and data noise. The initial step in data analysis often holds the key to unravelling hidden insights and shaping the trajectory of subsequent analyses. The accuracy and dependability of insights hinge on the adept application of diverse techniques in handling and visualizing data. This blog post navigates through the intricacies of the Mobile Price Detection dataset, unveiling an array of methods encompassing counting, imputing missing values, outlier detection, noise reduction, and data visualization.

Dataset Generation

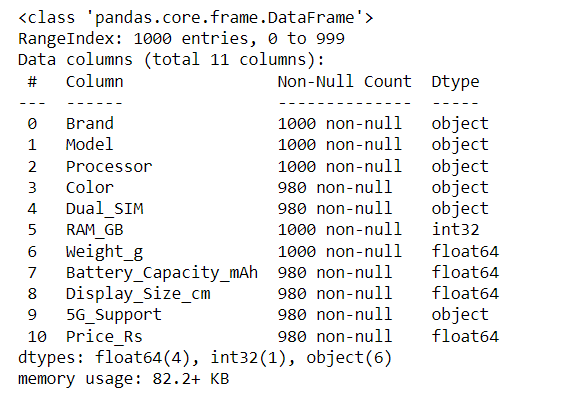
In crafting a comprehensive understanding of healthcare dynamics through Exploratory Data Analysis (EDA), our journey commences with a meticulously crafted dataset, ensuring authenticity and diversity. This synthetic mobile dataset, comprising 1000 rows, has been curated to simulate a broad spectrum of mobile scenarios, featuring a blend of categorical and numerical attributes that mirror the dynamic facets inherent in the mobile landscape.

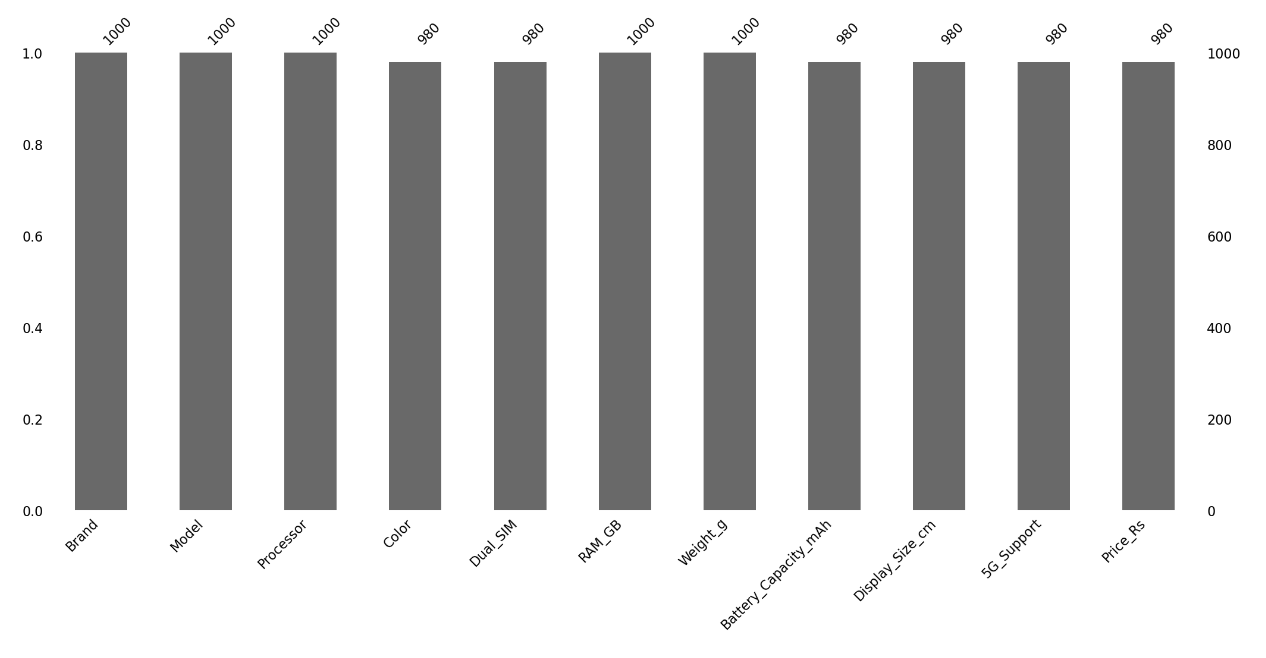
AttributesBrand: Mobile brand names (e.g., Samsung, Apple, Xiaomi)Model: Specific mobile models within each brandProcessor: Mobile processor details (e.g., Snapdragon 888, Kirin 9000)Colour: Varied colour options for mobile devicesDual SIM Support: Indicating whether a mobile supports dual SIM cards (Yes/No)RAM Capacity (in GB): Randomly chosen capacities (4GB, 6GB, 8GB)Weight (in grams): Rounded weights ranging from 120 to 190 gramsBattery Capacity: Randomly generated capacities between 3000 and 5000  
Display Size (in cm): Rounded display sizes between 12 and 17 cm5G Support: Indicating whether a mobile supports 5G connectivity (Yes/No)Price (in Rs): Rounded prices between 5000 and 50000 Indian Rupees

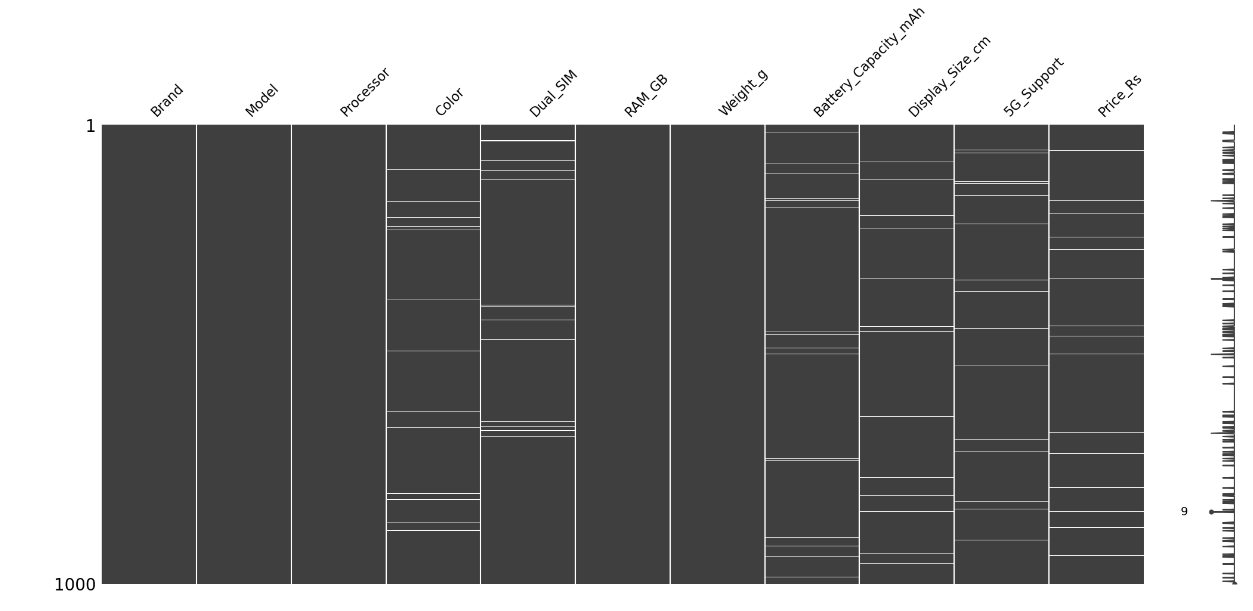
The incorporation of both categorical and numerical attributes, such as Brand, Colour, and RAM Capacity, enriches the mobile dataset. This diversity provides a multifaceted perspective on the intricate dynamics that define the mobile technology landscape.

Visualizing Missing values

Missing values often pose a tantalizing puzzle. Like elusive clues in a detective story, they can both hinder and illuminate our understanding of the bigger picture. In the dataset we're exploring, missing values play a significant role, demanding careful consideration and strategic handling. missing values in our mobile price dataset offer both a challenge and an opportunity. Through exploratory data analysis (EDA), we can uncover their patterns, understand their implications, and ultimately choose strategies to address them effectively.



The Above figures for mobile information reveals varying degrees of completeness across its attributes. Notably, key details such as 'Brand,' 'Model,' and 'Processor' exhibit full data coverage, indicating a comprehensive representation of mobile brands, models, and processors. However, the dataset faces challenges in terms of missing values, with 20 instances each in columns like 'Color,' 'Dual\_SIM,' 'Battery\_Capacity\_mAh,' 'Display\_Size\_cm,' '5G\_Support,' and 'Price\_Rs.' These gaps in information may impact analyses, particularly in areas where visual or quantitative assessments are integral. Addressing missing data is crucial, and data scientists must employ suitable strategies, such as imputation or removal, depending on the nature of the analysis.



In the above Image we can see white lines in colour, Dual sim, Battery Capacity, Display size, 5G\_support, Price columns. While other columns doesn’t have missing values.

Handling Missing values

Various strategies were employed to address missing values in the mobile dataset.

Price: Missing values were filled using the mean price of the respective mobile device model, leveraging the consistency of pricing within specific models for realistic imputations.

Colour: Imputation was performed with the mode of each mobile device model, capturing the most frequent colour and providing a practical estimate for missing values.

Battery Capacity and Display size: Utilizing the K-Nearest Neighbours (KNN) imputation method, missing values in these numerical attributes were accurately estimated based on the values of the nearest neighbours, enhancing precision in the imputed data.

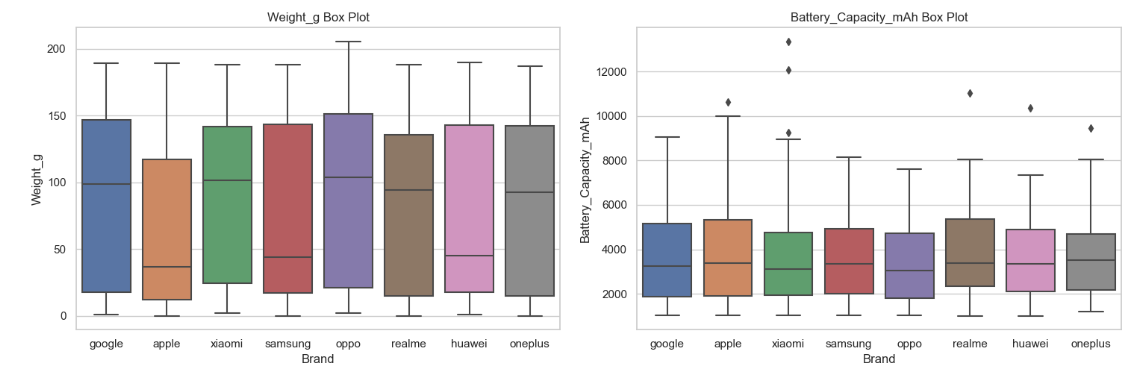
By applying these strategies, missing values were successfully handled in the mobile dataset, ensuring that the dataset is complete and suitable for further analysis.

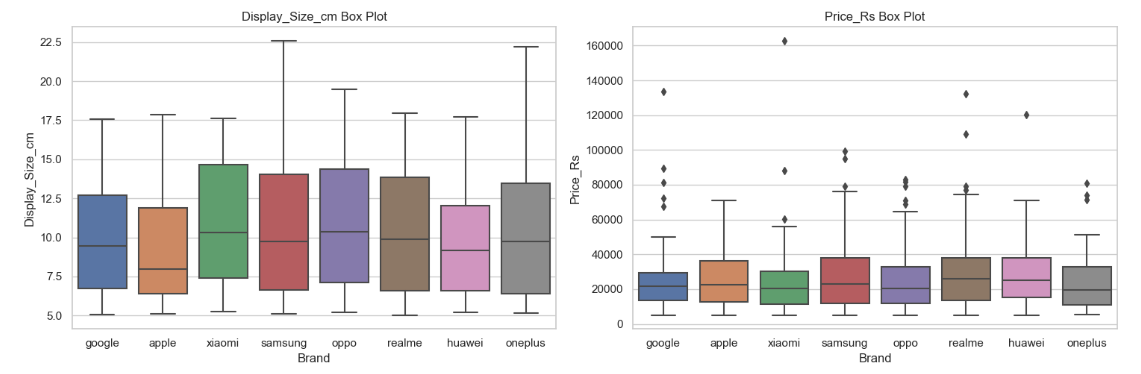
**Noise data**

We don’t have any noise in the mobile price detection dataset. As it is synthetic dataset but in real world there can be noise so we need to handle them by using techniques like binning.

**Detecting Outliers**

In the exploratory analysis of the mobile dataset, box plots were employed to visually inspect the distribution of key attributes across different mobile brands. Notably, the Battery Capacity and Price columns exhibit outliers, indicative of data points significantly deviating from the typical range within these attributes. Price demonstrates a higher prevalence of outliers compared to other attributes. These outliers effect while training a model so we need to remove them.

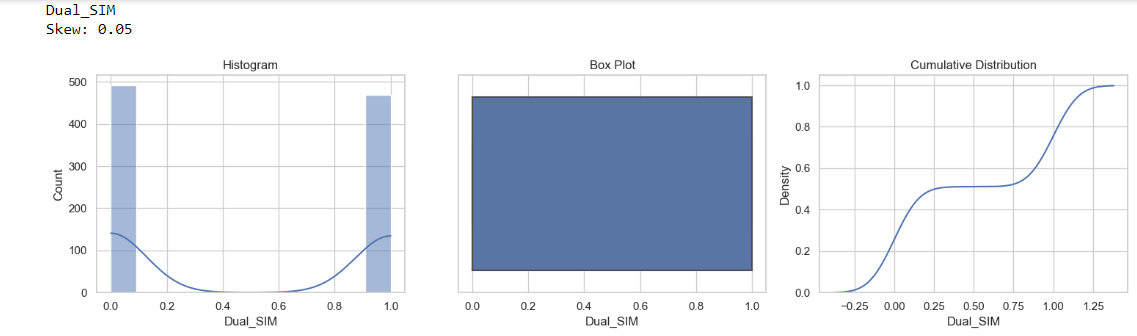


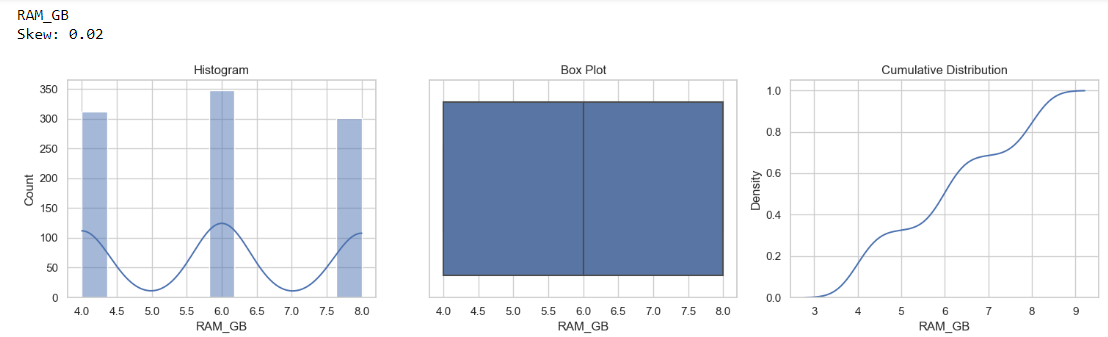


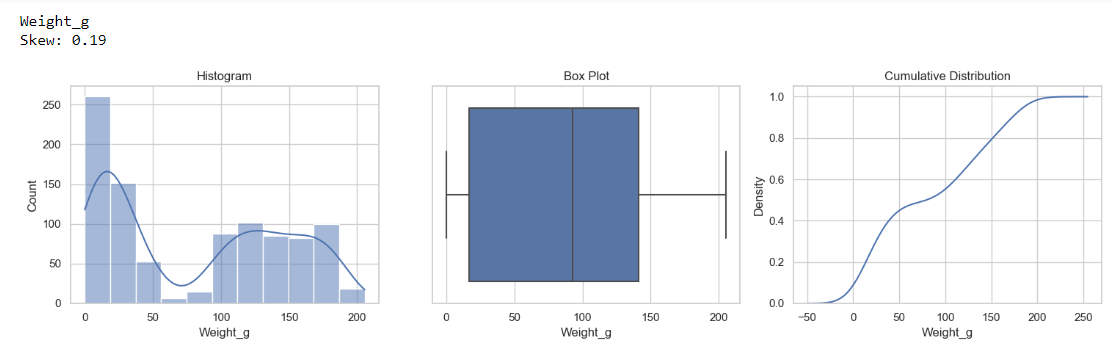
In the effort to address outliers in the mobile dataset, a systematic approach was implemented. Utilizing the Interquartile Range (IQR) method, a dedicated function, remove outliers, was devised to identify and eliminate outliers for each numeric column. This function calculated the first and third quartiles, computed the IQR, and established lower and upper bounds to determine the acceptable range of values. Subsequently, outliers falling beyond these bounds were systematically removed from the dataset for each numeric attribute. By iteratively applying this process to all numeric columns, the mobile dataset was effectively cleansed of outliers, ensuring a more reliable foundation for subsequent analyses.

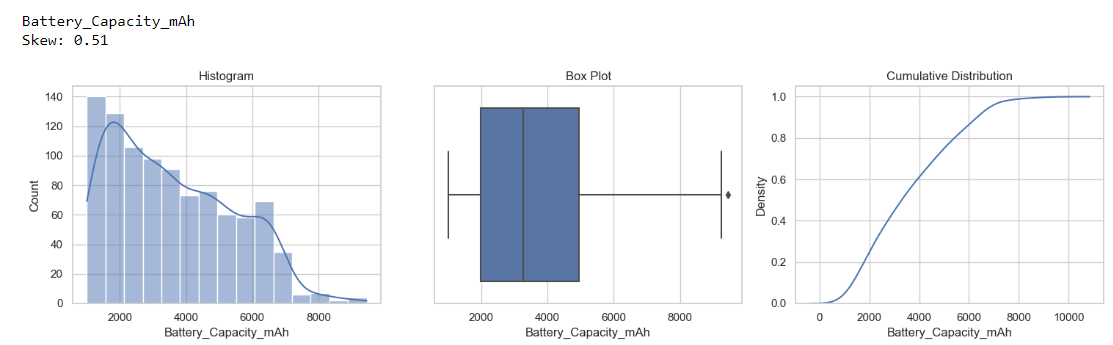
**Exploratory Data Analysis**

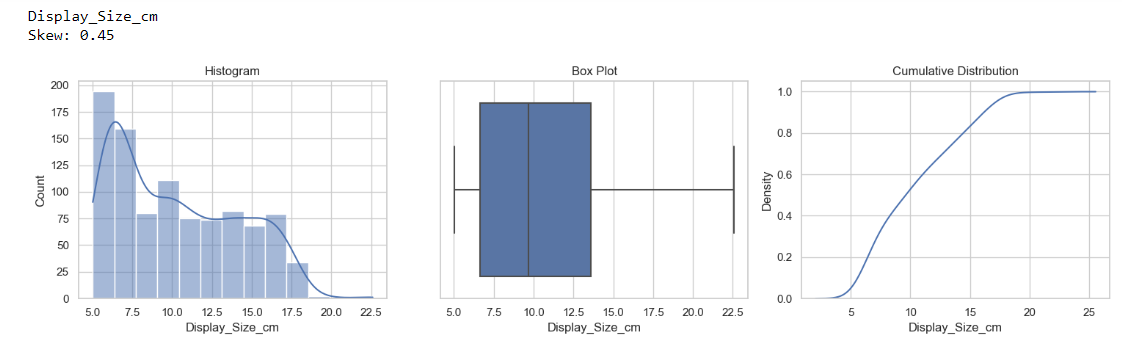
**Univariate Analysis**

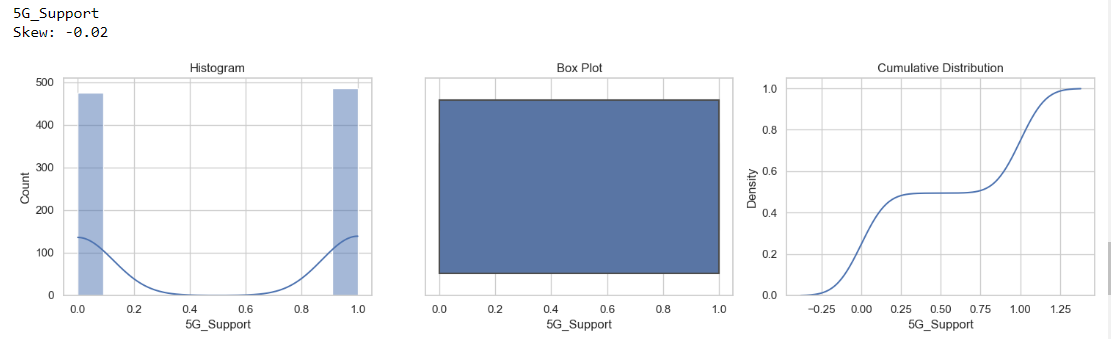
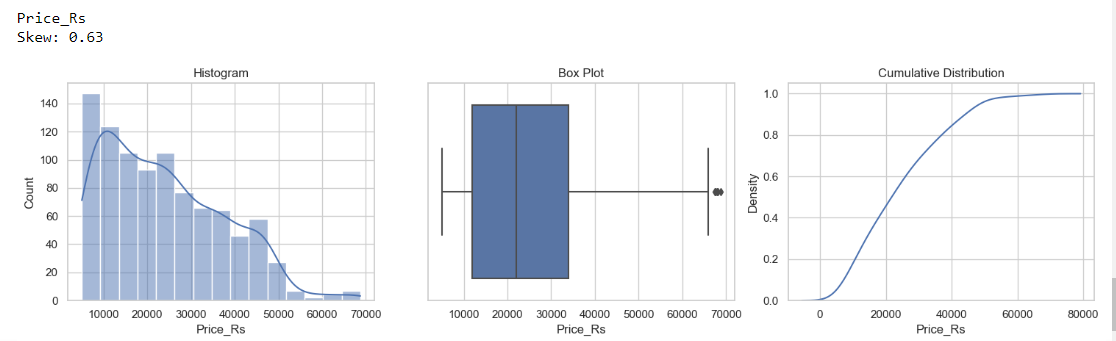








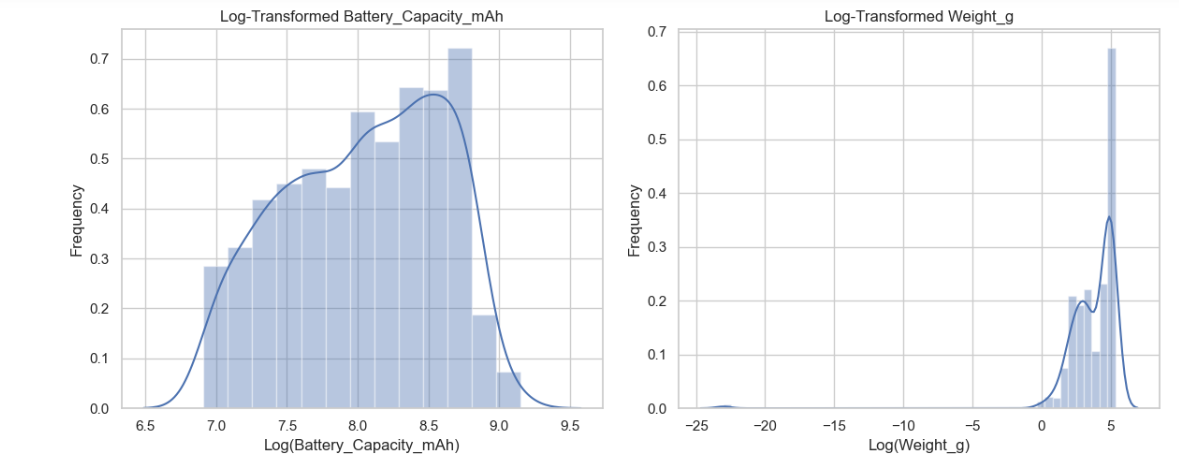


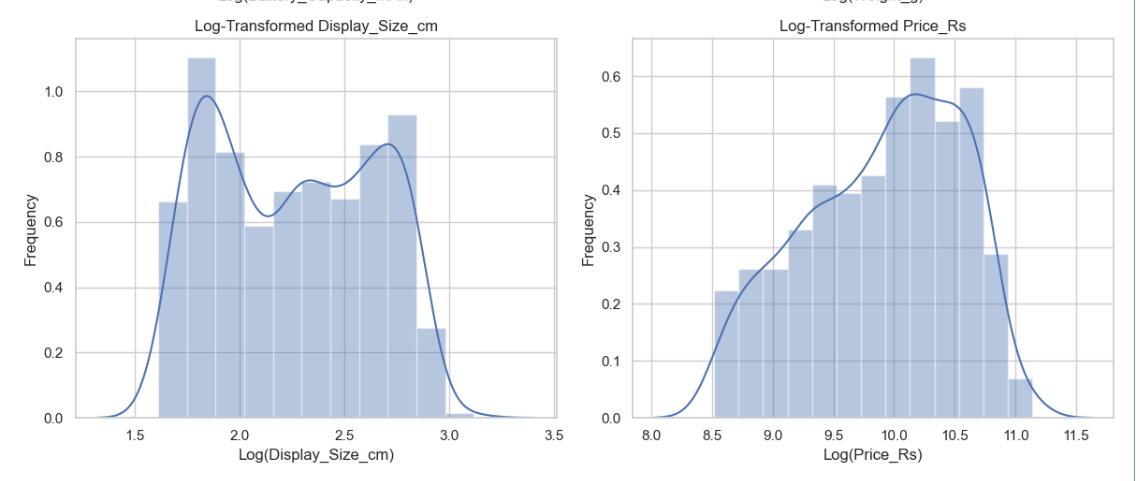
 

The skewness analysis and visualizations of key numerical columns in the mobile dataset offer nuanced insights into their distributional patterns. Dual\_SIM and RAM\_GB exhibit nearly symmetrical distributions with minimal skewness, reflecting a balanced spread of dual SIM support and RAM capacities. Weight\_g displays a slightly positive skew, suggesting a concentration of lighter mobile devices. Battery\_Capacity\_mAh and Display\_Size\_cm exhibit positive skewness, indicating concentrations towards higher battery capacities and larger display sizes. 5G\_Support maintains a symmetrical distribution. The Price\_Rs column reveals a right-skewed pattern, suggesting a prevalence of lower-priced mobile devices. As we can see outliers are decreased because of using Iqr method.

Data Transformation

Let’s look how to remove skeewness from data.

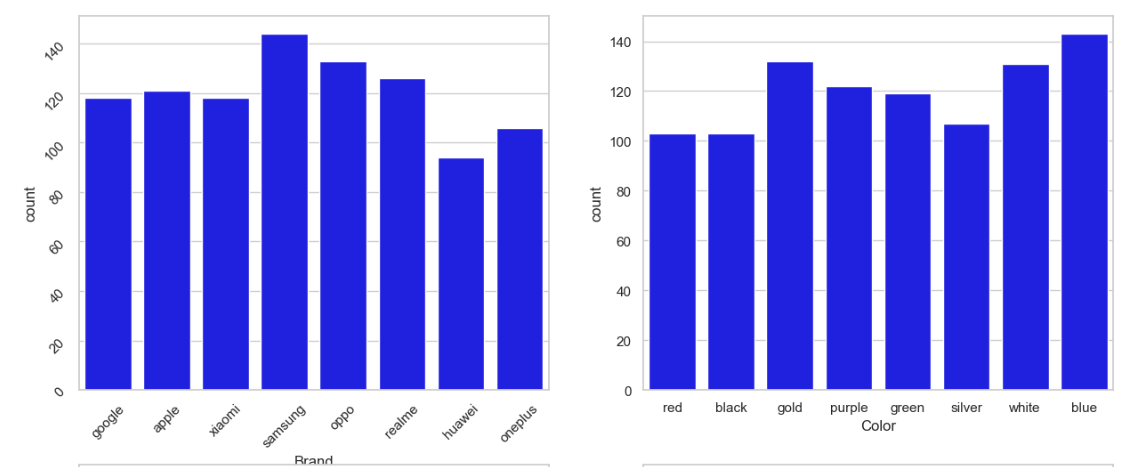


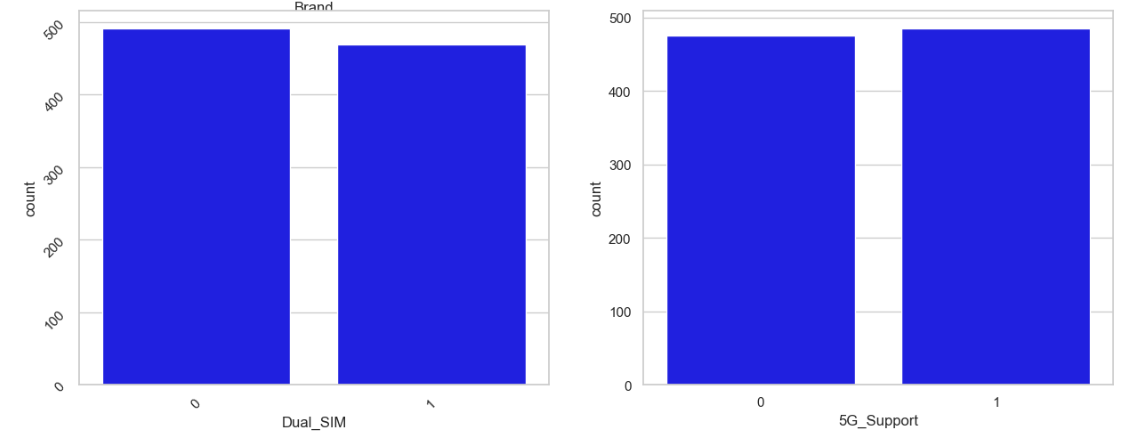


We did Log transformation to remove skewness.Log transformation can help in normalization, so this variable can maintain standard scale with other variables.

By applying log transformation on selected numerical columns ('Weight\_g', 'Battery\_Capacity\_mAh', 'Display\_Size\_cm', and 'Price\_Rs'). This transformation mitigates skewness, promoting a more symmetrical distribution and facilitating improved data normalization. The resulting visualizations depict the impact of the log transformation on the distributions of these variables, highlighting a more balanced representation conducive to statistical analyses.

Let’s explore categorical features





These are bar plots for categorical features from this we can infer that samsung and oppo uers are more in number than any other company and many people are not interseted in buying huawei mobiles.most people like gold,blue and white coloured phones.almost half of the phones don’t contain dual sim and don’t come up with 5G support.

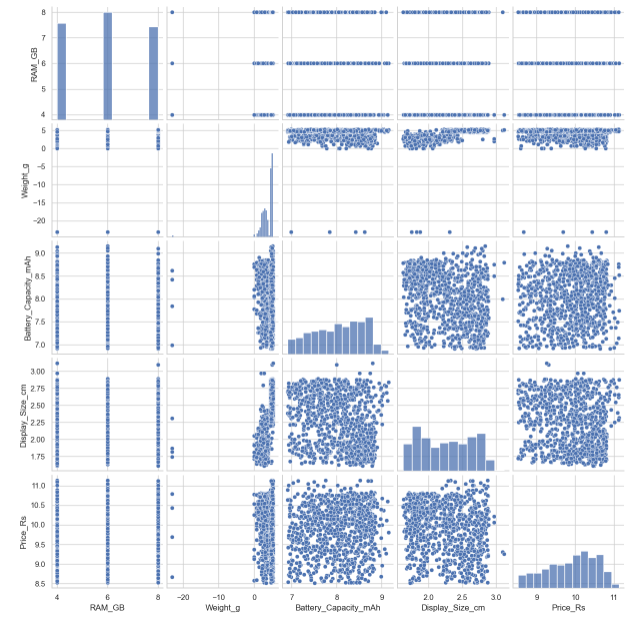
Bivariate Analysis

Now, let’s move ahead with bivariate analysis. Bivariate Analysis helps to understand how variables are related to each other and the relationship between dependent and independent variables present in the dataset.

For Numerical variables, Pair plots and Scatter plots are widely been used to do Bivariate Analysis.

Let’s look at pair plot between numerical data

a pair plot provides a visual overview of relationships and distributions among selected features in a dataset. It reveals patterns, correlations, and potential outliers through scatterplots and histograms, helping analysts quickly understand multivariate interactions and identify key insights within the data.



The pair plot involves 5 features: RAM\_GB, Weight\_g, Display\_Size\_cm, Battery\_Capacity\_mAh, and Price\_Rs.By Examining the diagonal plots to understand the individual distributions of each feature and Looking for symmetry, skewness, outliers, and potential data quality issues. We infer that

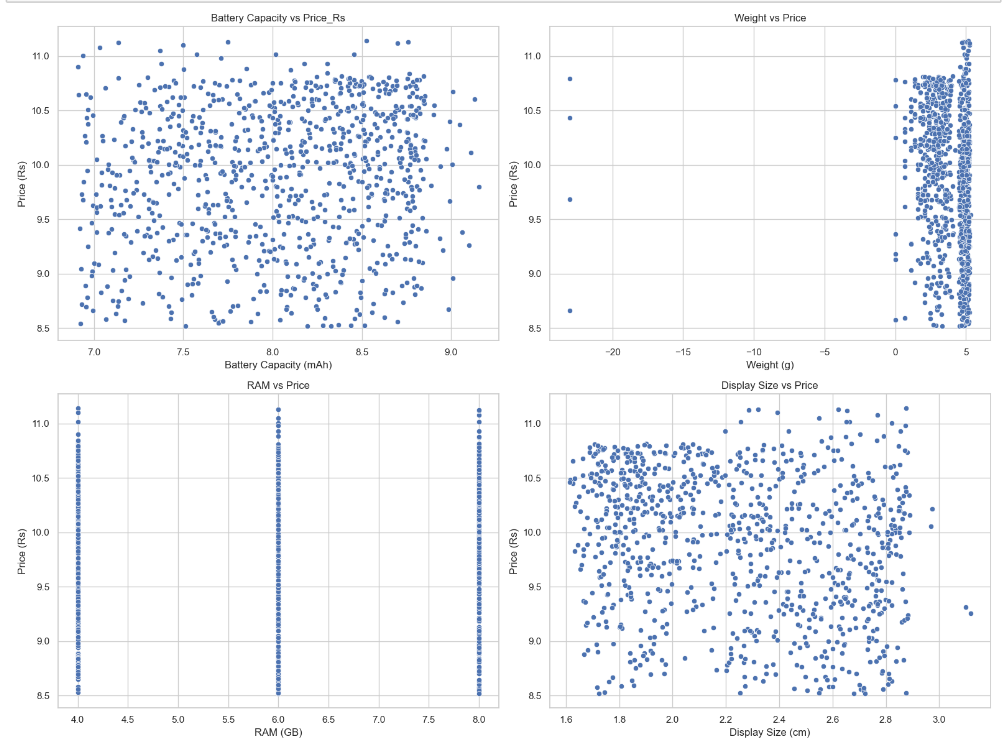
Battery capacity and Weight suggests that heavier devices tend to have larger batteries.

Display size and Weight suggest that Larger displays often contribute to greater weight.

Display size and Price suggest that Bigger displays might be associated with higher prices.

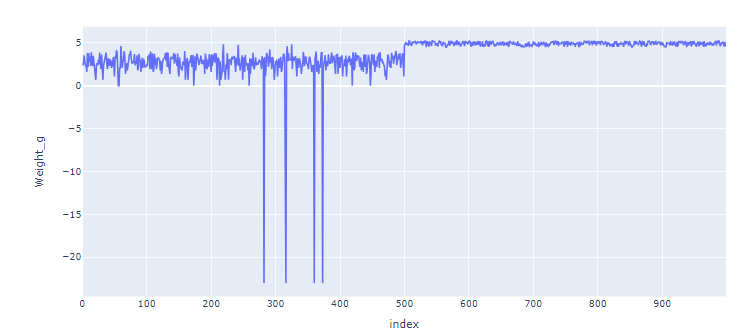
RAM and Price suggest that Increased RAM typically comes with a higher price tag.

Let’s look at price to some important features scatter plot

a scatter plot is a graphical representation of individual data points in a two-dimensional space, where each point represents the values of two variables. It provides a visual depiction of the relationship or pattern between the variables, helping to identify correlations, clusters, or trends in the data. 

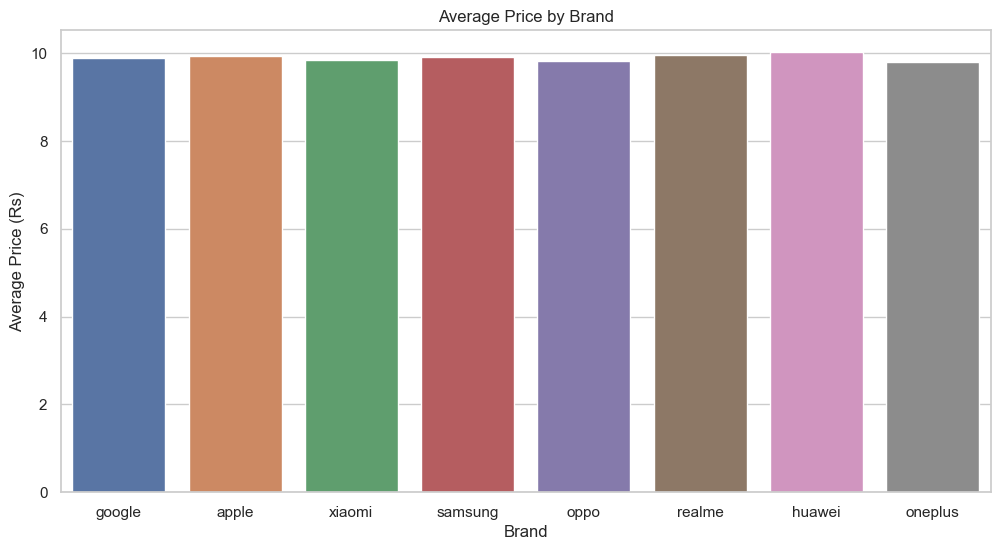
The scatter plot compares Battery Capacity (mAh) on the x-axis to Price (Rs) on the y-axis.The points generally form a positive linear trend, suggesting a positive correlation between the variables. The points are somewhat loosely scattered, indicating a moderate correlation.There are a few potential outliers in the upper right corner, representing devices with high battery capacity and price.

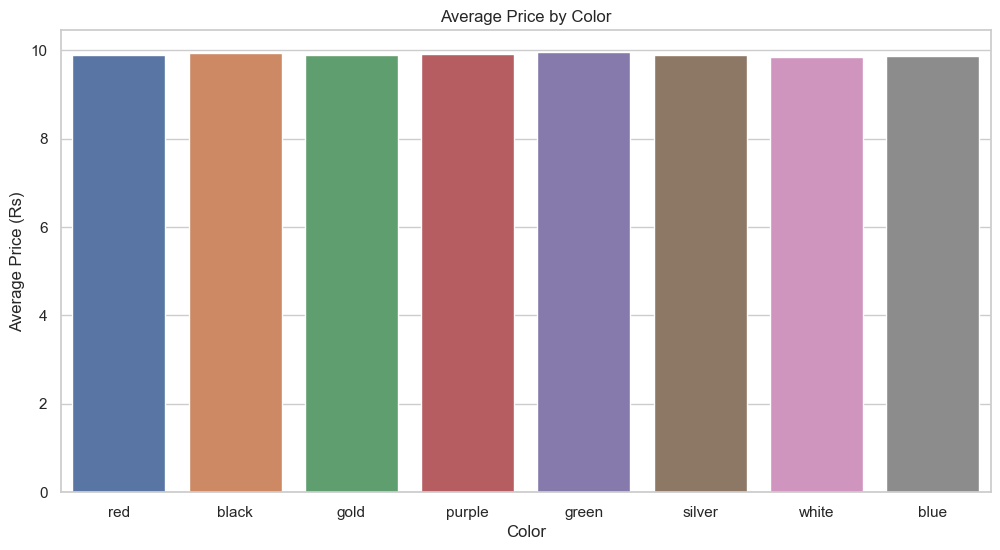
Line plot for weight of mobile.

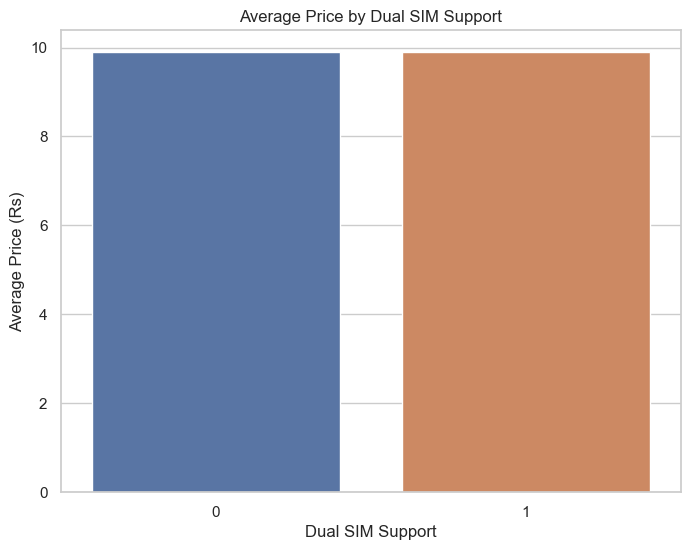


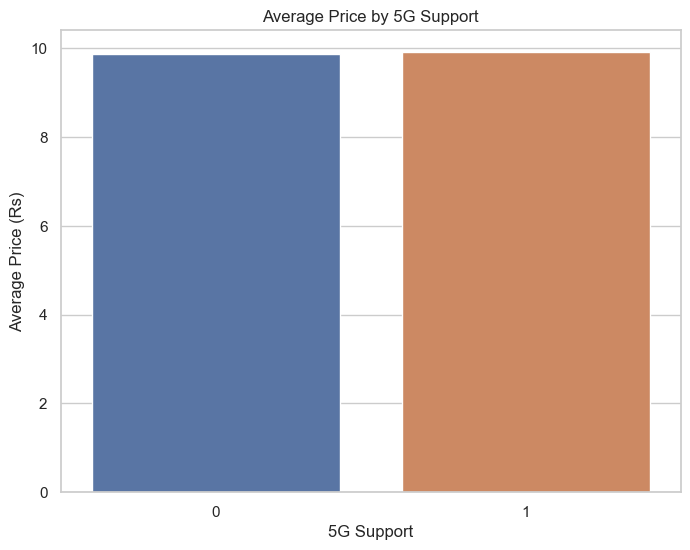
The weight starts at around 5 grams.It then decreases steadily to around -15 grams by index 400.After a slight increase to around -10 grams, it continues to decrease to around -20 grams by index 900.The general trend is a decrease in weight over time. The decline is steeper between indices 0 to 400, suggesting a more rapid weight loss in that phase. There's a temporary increase in weight around index 400, but the downward trend resumes. The line exhibits some minor fluctuations, indicating natural variations in weight.

Let’s analyze categorical data to price







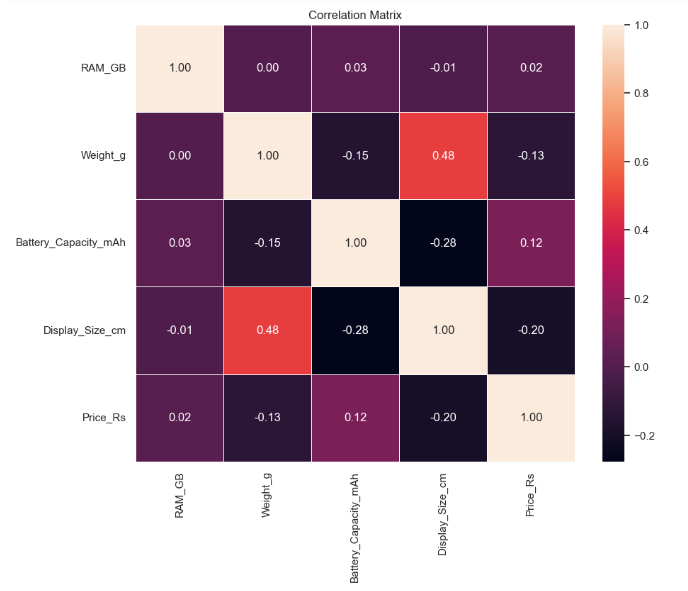


As it is synthetically generated data we cannot get many inferences from barplot but generally we can know which company has higher mobile prices and which colour has high price or low price does 5g mobile price is high or low or by adding dual sim can the price go high or low these infernces we can know.

Let’s look at correlation matrix and heat map of numerical data

A correlation matrix reveals the pairwise relationships between variables in a dataset, displaying correlation coefficients that indicate the strength and direction of linear connections. Positive values signify direct relationships, while negative values indicate inverse associations.

A heat map visually represents a correlation matrix using colors, where darker shades depict stronger correlations, and lighter shades suggest weaker or no correlations. This graphical tool aids in quickly identifying patterns and relationships within a set of variables, facilitating efficient data exploration and analysis.



From the above heat map we can infer that The correlations, in general, are relatively weak, indicating no strong linear relationships among the variables.'Display\_Size\_cm' and 'Weight\_g' show a moderate positive correlation, suggesting that heavier devices may have larger display sizes.'Battery\_Capacity\_mAh' and 'Display\_Size\_cm' exhibit a moderate negative correlation, indicating that devices with larger display sizes may have slightly smaller battery capacities.The correlation with 'Price\_Rs' is weak for most variables, suggesting that the selected features have limited direct linear relationships with the device prices.

Up till now we get insights on many numerical and categorical features but we did not concentrated much on brand and colour. So lets work on tree map between brand and colour.



The tree map is divided into eight top-level categories, each representing a major smartphone brand: Samsung, Realme, Xiaomi, OnePlus, Oppo, Apple, Google, and Huawei. Each top-level category is further divided into subcategories, likely representing different smartphone models or color options.

Samsung: Offers a wide variety of models and colors, with a focus on gold, silver, and white options.

Realme: Has a smaller overall presence but shows a strong emphasis on silver and purple phones.

Xiaomi: Presents a diverse range of colors, including blue, silver, green, gold, and purple.

OnePlus: Offers a more limited selection of models and colors, with a focus on blue, silver, green, and purple.

Oppo: Has a similar distribution of colors to OnePlus, with blue, silver, purple, and green being prominent.

Apple: Stands out with a unique color scheme, primarily featuring white, black, red, and gold options.

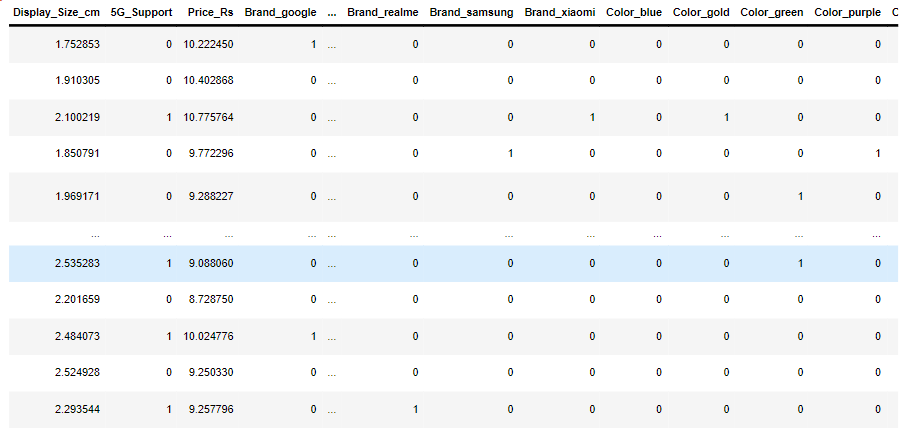
Google: Offers a limited range of colors, focusing on black, white, and green.

Huawei: Has a more balanced distribution of colors, including red, black, white, silver, and green

Coming to final part one hot encoder

Since we cannot deal with string for training we convert string to numerical by One Hot encoder.

One-hot encoding expands the dimensionality of the dataset by creating binary columns for each category. While this method introduces sparsity in the data, it ensures that categorical information is accurately represented. The resulting one-hot encoded matrix is particularly useful in scenarios where machine learning models require numerical input, enabling them to learn from and make predictions based on categorical features.



Conclusion:

In conclusion, our in-depth Exploratory Data Analysis (EDA) of the mobile dataset provided comprehensive insights into various crucial facets of the industry. We meticulously explored patterns and trends related to mobile features, brand preferences, and pricing dynamics through an array of visualizations and statistical analyses. Employing techniques such as one-hot encoding and handling missing values, we ensured the dataset's readiness for advanced analytics. Visualizations including box plots, histograms, and pair plots allowed us to uncover nuanced relationships and distributions within the mobile data. Utilizing correlation matrices and heatmaps, we deciphered intricate connections between attributes, shedding light on factors influencing pricing and device specifications. Our analytical journey through this synthetic mobile dataset exemplifies the power of EDA in unraveling complexities, providing valuable insights for industry stakeholders and data enthusiasts alike.