DMBIPRACTICALNO:3

**Name:Bhumika Patil**

**Class:D15C**

**RollNo:61**

**Batch: C**

**Aim:**

* ToperformExploratoryDataAnalysisandvisualizationusingpython.

**Theory:**

* Exploratory Data Analysis (EDA) is an essential step in data science that focuses on investigatingdatasetstouncoverpatterns,detectanomalies,testassumptions,andgenerate insights using statistical and visual methods. It helps identify missing values, outliers, distributions, and relationships among variables, guiding featureengineering andhypothesis testing.
* In this experiment, EDA is applied to a smartphone dataset (Kaggle: *Mobiles Dataset 2025*), which includes company, model, RAM, battery capacity, camera specifications, screen size, and launch prices across different countries. The process involves data loading, cleaning, summary statistics, and visualization to extract insights such as brand distribution, specification/feature patterns, and correlations.
* Data visualization translates raw data into intuitive formats such as charts and graphs, enabling easier detection of trends, outliers, and relationships.

# Pythonlibrariesused:

* + **Pandas:**Datamanipulation(DataFrames,CSVreading,summarystatistics).
  + **NumPy:**NumericaloperationsandNaNhandling.
  + **Matplotlib.pyplot**:Basicplotting(figures,labels,visualization).
  + **Seaborn:**Statisticalvisualizations(histograms,barplots,heatmaps).
  + **Re:**Regularexpressionsforextractingnumericvalues.

**Histogram(sns.histplot):**Ahistogramrepresentsthefrequencydistributionofacontinuousvariable. In sns.histplot, the dataset is provided through data and the feature is specified using x. The bins parameter adjusts the number of intervals, kde adds a smooth density curve, and color sets the bar color.

**BarPlot(sns.barplot):**Abarplotcomparescategoricalandnumericalvariables.With sns.barplot,the category is placed on the x axis and the numeric values on the y axis. The dataset is defined through data, colors are controlled with palette, and error bars can be shown or removed using the errorbar argument.

**Box Plot (sns.boxplot):** A box plot summarizes data spread using median, quartiles, and outliers. In sns.boxplot,thecategoryisgivenasx,thenumericvariableasy,andthedatasetwithdata.Thepalette option customizes the box colors.

**Heatmap (sns.heatmap):** A heatmap visualizes correlations or values in a matrix using colors. With sns.heatmap, data such as df.corr() is plotted, annot displays values inside cells, cmap controls the color scheme, and fmt sets the number format.

**Scatter Plot (sns.scatterplot):** A scatter plot shows relationships between two continuous variables. In sns.scatterplot, features are set with x and y, groups can be highlighted with hue, transparency is adjusted using alpha, and the dataset is provided with data.

**PieChart(plt.pie):**Apiechartrepresentsproportionsofcategoriesasslices.Thefunctionplt.pietakes values as input, assigns category names through labels, shows percentages with autopct, adjusts rotation using startangle, and applies colors with the colors parameter.

# Syntax&Graphsappliedinthisexperiment:

* + **Histogram:**ShowsfrequencydistributionofcontinuousvariableslikeRAM,battery capacity, and screen size.
    - **Syntax:**sns.histplot(data=df,x="RAM",bins=10,kde=True,color="blue")
  + **Bar Graph:** Compares categorical data, e.g., top smartphone brands by model count or average battery capacity.
    - **Syntax:**sns.barplot(x="CompanyName",y="BatteryCapacity",data=df, palette="Set2")
  + **BoxPlot:**Highlightsspread,quartiles,andoutliers;usedforcomparingRAM,screensize, and cameras.
    - **Syntax:**sns.boxplot(x="Feature",y="Value",data=df,palette="Set3")
  + **Heatmap:** Displayscorrelationmatrixwith colourcoding, e.g., RAMvs priceorbatteryvs price.
    - **Syntax:**sns.heatmap(df.corr(),annot=True,cmap="coolwarm",fmt=".2f")
  + **ScatterPlot:**Exploresrelationshipsbetweentwocontinuousvariables,e.g.,screensizevs price with brand as hue.
    - **Syntax:**sns.scatterplot(x="ScreenSize",y="Price\_India",hue="CompanyName", data=df, alpha=0.7)
  + **PieChart:**Representsproportions,suchassmartphonedistributionbylaunchyearor brand share.
    - **Syntax:**plt.pie(values,labels=labels,autopct="%1.1f%%",startangle=90, colors=sns.color\_palette("pastel"))

**Conclusion:**

In this experiment, Exploratory Data Analysis (EDA) was performed on a smartphone dataset using Python. The dataset was cleaned and preprocessed to handle inconsistent formats for features such as RAM, battery capacity, and prices. Various visualization techniques including histograms, bar graphs, boxplots, scatter plots,heatmaps,and piecharts wereapplied toanalyzebrand distribution, price variations, and relationships among specifications such as RAM, battery, and screen size.

The analysis provided clear insights into the structure of the dataset and revealed patterns, trends, andcorrelationsthatwouldnothavebeenevidentfromrawdataalone.Challengessuchasencoding issues, missing values, and cluttered visualizations were addressed using appropriate data cleaning methods and plot customization. Overall, the experiment demonstrated the effectiveness of EDA in transformingrawdatasetsintomeaningfulinformationandreinforcedtheimportanceofvisualization in data analysis.

CODE&OUTPUT

