Name : Vaishnavi Eknath Avhad
Class : D15C
Roll No. : 41

Practical 3

Aim : To perform Exploratory Data Analysis and visualization using python

Theory :

Exploratory Data Analysis (EDA) is the process of analyzing and summarizing datasets to understand their main characteristics, often visualizing them to uncover hidden patterns, trends, and relationships. This process is crucial in data science, as it helps identify data quality issues (like missing values), distribution characteristics, outliers, and relationships between variables. The aim of this practical was to perform EDA and visualize various aspects of the "luxury_cosmetics_fraud_analysis_2025.csv" dataset using Python libraries such as Pandas, Matplotlib, and Seaborn.

The key objectives of the EDA in this practical are :

1. Data Cleaning : Handling missing values and transforming data types.
2. Summary Statistics : Gaining insights into the data distribution.
3. Visualizing Data : Creating different plots to understand the relationships between variables and their distributions.

The following steps were taken :

1. Loading the Data : The dataset was read using pandas.read_csv() into a DataFrame.
2. Handling Datetime : Transaction date and time columns were converted to datetime and time objects for analysis.
3. Missing Values : Checked for any missing data using isnull() and sum() to assess the need for handling missing values.
4. Summary Statistics : We used df.describe() to gain a numerical summary of the dataset, including basic statistics like mean, standard deviation, and quartiles.
5. Visualizations : We created a variety of visualizations to analyze different aspects of the data :

- ○ Fraud vs Non-Fraud Transactions : A count plot to visualize the distribution of fraudulent vs non-fraudulent transactions.
- ○ Purchase Amount Distribution : A histogram with a Kernel Density Estimation (KDE) to observe the distribution of the "Purchase Amount".
- ○ Purchase Amount by Fraud Flag : A boxplot to compare the spread and outliers of purchase amounts for fraud and non-fraud transactions.
- ○ Transactions by Payment Method : A bar plot to analyze the number of transactions per payment method.
- ○ Transactions by Device Type : A bar plot to analyze the distribution of transactions across different device types.
- ○ Customer Age Distribution : A histogram with KDE to visualize the distribution of customer ages.
- ○ Age vs Purchase Amount : A scatter plot to examine the relationship between customer age and purchase amount, with fraud highlighted.
- ○ Correlation Heatmap : A heatmap to visualize correlations between numerical features in the dataset.

Code with output :

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv("/content/sample_data/luxury_cosmetics_fraud_analysis_2025.csv")
df['Transaction_Date'] = pd.to_datetime(df['Transaction_Date'], errors='coerce')
df['Transaction_Time'] = pd.to_datetime(df['Transaction_Time'], format="%H:%M:%S",
errors='coerce').dt.time

# --- Missing Values ---
print("\nMissing Values:\n", df.isnull().sum())
# --- Summary Statistics ---
print("\nSummary Statistics:\n", df.describe(include="all"))
```

```
Missing Values:
 Transaction_ID           0
Customer_ID              0
Transaction_Date         0
Transaction_Time         0
Customer_Age           106
Customer_Loyalty_Tier  106
Location                 0
Store_ID                 0
Product_SKU              0
Product_Category         0
Purchase_Amount          0
Payment_Method         106
Device_Type              0
IP_Address               0
Fraud_Flag               0
Footfall_Count           0
dtype: int64

Summary Statistics:
                               Transaction_ID  \
count                                    2133
unique                                   2133
top     83108c8e-2ede-4e86-a0f6-ea4875c9e523
freq                                        1
mean                                      NaN
min                                       NaN
25%                                       NaN
50%                                       NaN
75%                                       NaN
max                                       NaN
std                                       NaN

                                 Customer_ID               Transaction_Date  \
count                                   2133                           2133
unique                                  2133                            NaN
top     e94cecaf-9db8-49f4-97f6-a17bb5dd1187                            NaN
freq                                       1                            NaN
mean                                     NaN  2025-05-16 03:36:02.025316352
min                                      NaN            2025-02-14 00:00:00
25%                                      NaN            2025-04-01 00:00:00
50%                                      NaN            2025-05-17 00:00:00
75%                                      NaN            2025-07-01 00:00:00
max                                      NaN            2025-08-12 00:00:00
std                                      NaN                            NaN

       Transaction_Time  Customer_Age Customer_Loyalty_Tier Location  \
count              2133   2027.000000                  2027     2133
unique             2106           NaN                     5       20
top            21:38:38           NaN                Bronze   Sydney
freq                  2           NaN                   808      128
mean                NaN     41.684262                   NaN      NaN
min                 NaN     18.000000                   NaN      NaN
25%                 NaN     30.000000                   NaN      NaN
```

SIMPLE VISUALIZATION

1. Fraud vs Non-Fraud

plt.figure(figsize=(6,4))
sns.countplot(data=df, x="Fraud_Flag")
plt.title("Fraud vs Non-Fraud Transactions")
plt.show()

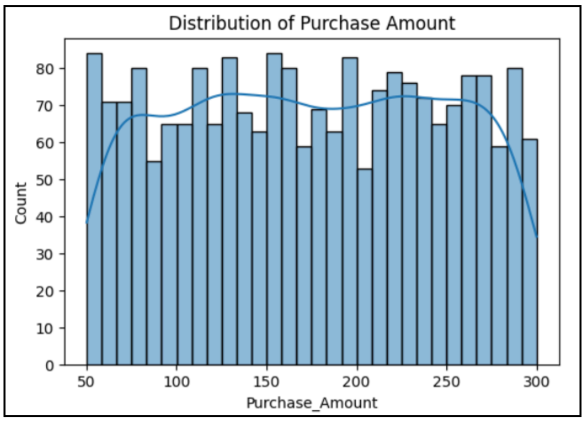
2. Histogram : Purchase Amount

plt.figure(figsize=(6,4))

sns.histplot(df['Purchase_Amount'], bins=30, kde=True)

plt.title("Distribution of Purchase Amount")

plt.show()



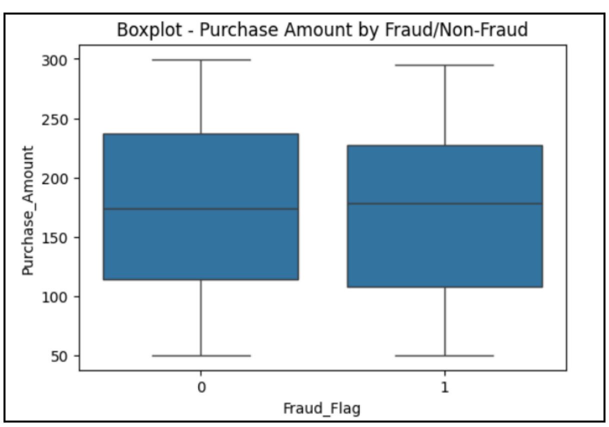
3. Boxplot : Purchase Amount by Fraud Flag

plt.figure(figsize=(6,4))

sns.boxplot(data=df, x="Fraud_Flag", y="Purchase_Amount")

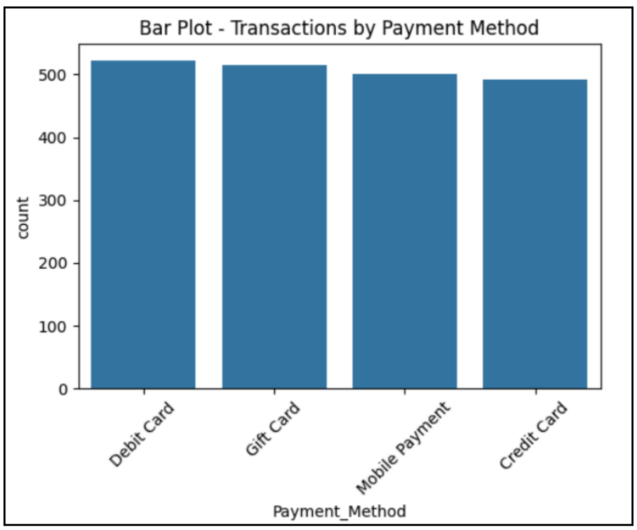
plt.title("Boxplot - Purchase Amount by Fraud/Non-Fraud")

plt.show()



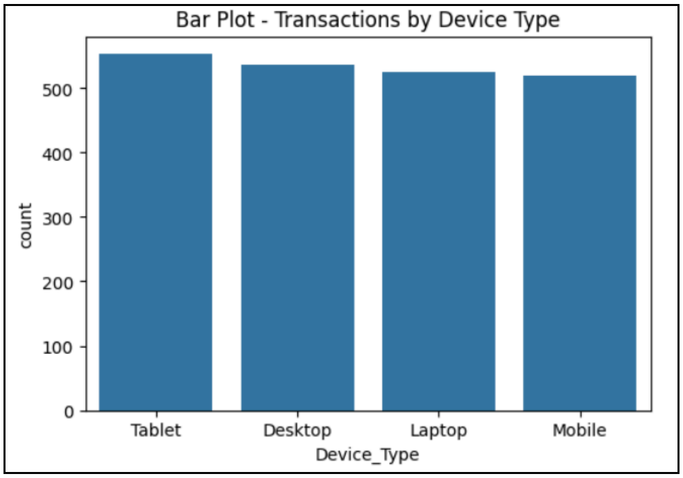
4. Bar Plot : Transactions by Payment Method

```
plt.figure(figsize=(6,4))

sns.countplot(data=df, x="Payment_Method",
order=df["Payment_Method"].value_counts().index)

plt.title("Bar Plot - Transactions by Payment Method")

plt.xticks(rotation=45)

plt.show()
```



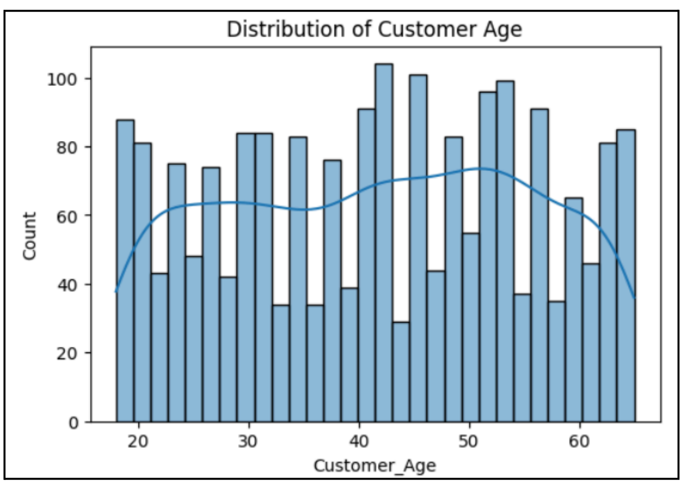
5. Bar Plot : Transactions by Device Type

```
plt.figure(figsize=(6,4))

sns.countplot(data=df, x="Device_Type", order=df["Device_Type"].value_counts().index)

plt.title("Bar Plot - Transactions by Device Type")

plt.show()
```



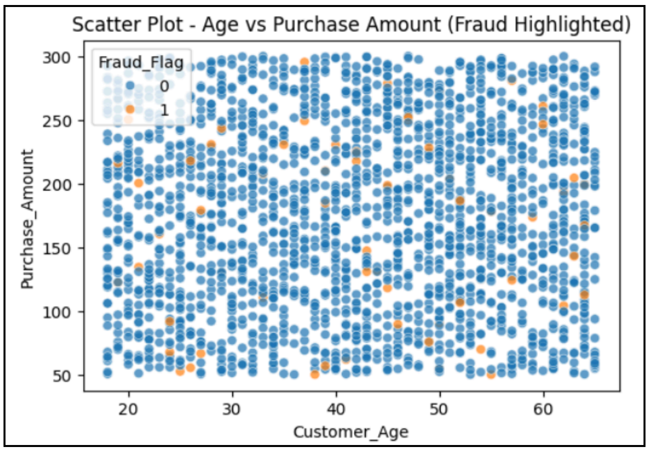
6. Histogram : Customer Age

```
plt.figure(figsize=(6,4))

sns.histplot(df['Customer_Age'].dropna(), bins=30, kde=True)

plt.title("Distribution of Customer Age")

plt.show()
```



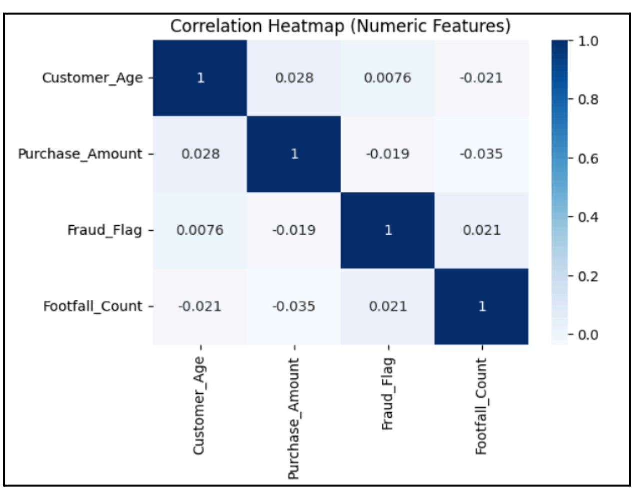
8. Scatter Plot : Age vs Purchase Amount

```
plt.figure(figsize=(6,4))
sns.scatterplot(data=df, x="Customer_Age", y="Purchase_Amount", hue="Fraud_Flag",
alpha=0.7)
plt.title("Scatter Plot - Age vs Purchase Amount (Fraud Highlighted)")
plt.show()
```



9. Correlation Heatmap

```
plt.figure(figsize=(6,4))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap="Blues")
plt.title("Correlation Heatmap (Numeric Features)")
plt.show()
```



Conclusion :

In this practical, we performed Exploratory Data Analysis (EDA) and visualized key patterns in the dataset. We identified missing values, analyzed the distribution of variables like "Purchase Amount" and "Customer Age," and compared fraud vs non-fraud transactions. Key insights included that fraud transactions had higher variability in purchase amounts, and certain payment methods and device types were more commonly associated with transactions. The correlation heatmap revealed weak relationships between numerical features, suggesting the need for additional features in future modeling. Overall, EDA helped uncover important trends and prepared the data for further analysis and predictive modeling.