**Technical Document: Exploratory Data Analysis (EDA) on Marketing Dataset**

**1. Introduction**

This document outlines the step-by-step approach taken to perform **Exploratory Data Analysis (EDA)** on the given **PS Marketing Dataset**. The goal of this analysis is to understand the data structure, detect missing values, analyze key variables, and extract insights using visualizations and statistical summaries.

---

**2. Setup and Environment**

**Libraries Used**

To perform EDA, the following Python libraries were imported:

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

import numpy as np
```

These libraries help in data manipulation, visualization, and statistical analysis.

---

**3. Data Loading and Inspection**

**Loading the Dataset**

```
# Load the dataset

marketing_data = pd.read_csv("C:\\Users\\navya\\Downloads\\Marketing - recordid_name_gender_age_location_email_phone_product_category_amount.csv")
```

This step loads the marketing dataset into a Pandas DataFrame for further analysis.

**Previewing the Data**

```
# Display the first few rows of the dataset

marketing_data.head(10)
```

This provides an overview of the dataset's structure, including column names, data types, and sample records of first ten rows.

**Checking Data Types and Missing Values**

```
# Display information about dataset

marketing_data.info()


# Check for missing values

marketing_data.isnull().sum()
```

- .info() provides insights into data types and non-null values.
- .isnull().sum() helps identify missing values in each column.

---

**4. Data Cleaning and Preprocessing**

**Handling Missing Values**

# Fill missing numerical values with median

marketing_data.fillna(marketing_data.median(), inplace=True)

# Fill missing categorical values with mode

marketing_data.fillna(marketing_data.mode().iloc[0], inplace=True)

This replaces missing values with the median (for numerical data) and mode (for categorical data), ensuring data consistency.

**Removing Duplicates**

# Remove duplicate rows

marketing_data.drop_duplicates(inplace=True)

Duplicates are removed to ensure clean and unique records.

---

**5. Exploratory Data Analysis (EDA)**

**Summary Statistics**

# Generate summary statistics

marketing_data.describe()

This provides key statistical insights, such as mean, standard deviation, and percentiles for numerical features.

**Distribution of Key Variables**

# Visualize distribution of numerical variables

marketing_data.hist(figsize=(10, 6), bins=30)

plt.show()

This histogram visualizes the distribution of numerical features, helping identify skewness and outliers.

**Correlation Analysis**

# Generate a correlation heatmap

plt.figure(figsize=(10,6))

```
sns.heatmap(marketing_data.corr(), annot=True, cmap='coolwarm')
plt.show()
```

A heatmap helps identify relationships between numerical variables.

**Customer Segmentation Analysis**

```
# Boxplot to compare revenue across customer segments
sns.boxplot(x='customer_segment', y='revenue', data=marketing_data)
plt.xticks(rotation=45)
plt.show()
```

Boxplots help compare revenue distributions across different customer segments.

---

**6. Key Findings & Next Steps**

**Findings:**

- Identified missing values and handled them appropriately.
- Detected correlations between numerical features.
- Analyzed revenue distribution across customer segments.
- Examined potential trends in revenue over time.

**Next Steps:**

- **Feature Engineering**: Create new meaningful variables.
- **Predictive Modeling**: Use machine learning to predict customer behavior.
- **A/B Testing**: Analyze different marketing campaign performances.

This concludes the EDA process for the marketing dataset. Further analysis can be performed based on business objectives.