**Data Analyst Project Documentation**

**Project Overview**

This project involves analysing **Diwali Sales Data** to identify patterns in customer demographics, purchasing behavior, and sales trends. The objective is to extract insights that could inform marketing strategies and improve customer targeting during Diwali sales.

**Dataset Description**

The dataset has 11,251 entries with 15 columns:

* **User\_ID**: Unique identifier for each customer.
* **Cust\_name**: Customer name.
* **Product\_ID**: Identifier for each product purchased.
* **Gender**: Customer’s gender (F or M).
* **Age Group** and **Age**: The age group and specific age of the customer.
* **Marital\_Status**: Marital status (0 = Single, 1 = Married).
* **State** and **Zone**: The state and region of the customer.
* **Occupation**: Customer’s occupation type.
* **Product\_Category**: Category of the purchased product.
* **Orders**: Number of orders placed by the customer.
* **Amount**: Total purchase amount.
* **Status** and **unnamed1**: These columns contain only null values and are removed as they do not contribute to the analysis.

**Key Observations on Data Quality**

* The **Amount** column has a few missing values, which need to be handled for accurate analysis.
* **Status** and **unnamed1** are empty columns and can be dropped.

**Project Steps and Methodology**

**1. Data Loading and Initial Inspection**

Data was loaded using pandas to inspect the structure and content. Initial inspection included:

* Checking the **shape** and basic structure of the data.
* Using df.info() and df.describe() to understand data types, missing values, and key statistics.
* Displaying **sample rows** to preview the values and patterns in columns.

**2. Data Cleaning**

* **Handling Missing Values**: Filled or removed missing values in the Amount column to maintain data integrity.
* **Column Removal**: Dropped Status and unnamed1, as these columns contain only null values.

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

EDA was conducted to explore relationships and trends within the data:

* **Demographic Analysis**:
  + **Gender** and **Age Group** distributions were analysed to identify predominant customer segments.
  + Analysed **Marital\_Status** alongside age and gender to understand if family-based marketing strategies are relevant.
* **Occupation and Product Category Analysis**:
  + Examined which occupations had higher purchase frequencies and spending levels.
  + **Product\_Category** insights helped reveal top-selling product types.
* **Regional Analysis**:
  + Analysed sales data across **State** and **Zone** to find high- and low-performing areas, useful for geographically targeted campaigns.
  + Visualized regional patterns to assess if any zones or states stand out in terms of order volume and spending.

**4. Data Visualization**

Visualizations were created to enhance the analysis and convey findings:

* **Gender and Age**: Used bar and pie charts to illustrate the distribution of sales across gender and age groups.
* **Regional Sales Patterns**: Heatmaps and bar plots depicted regional sales, helping to identify key sales zones.
* **Popular Product Categories and Occupations**: Visualized product category sales and occupational trends to inform product-specific marketing strategies.

**5. Key Insights and Recommendations**

From the analysis, several insights emerged:

* **Targeted Age Groups**: Specific age groups showed higher purchasing volumes. Marketing campaigns tailored to these groups could maximize sales.
* **Gender-Focused Promotions**: If a significant gender disparity is found, promotions could be adjusted to appeal more effectively to the predominant group.
* **Regional Targeting**: High-sales regions can be prioritized for promotions, while special offers could boost engagement in lower-sales areas.
* **Occupation-Specific Campaigns**: Insights on occupations with high spending can guide personalized marketing strategies.