Retail Analytics: Predicting Customer Purchase Behavior

Business Introduction:

Welcome to RetailVC's Retail Analytics Solutions, a leading provider of data-driven solutions for retail businesses. RetailVC specializes in using advanced analytics techniques to help retailers optimize their operations, enhance customer experiences, and drive business growth. Its mission is to empower retail organizations with actionable insights that enable them to make informed decisions, improve performance, and remain competitive in today's dynamic marketplace.

Problem Statement:

The goal of this capstone project is to develop a comprehensive understanding of customer behavior and preferences based on retail transaction data. The aim is to analyze customers' demographic information, purchasing patterns, and interactions with marketing campaigns to identify key insights. These insights will inform strategic decision-making and drive business growth. Specifically, the focus will be on predicting customers' responses to marketing campaigns and optimizing promotional strategies to maximize their effectiveness and return on investment (ROI).

Note: Use the **marketing_campaign.csv** dataset for performing the below tasks

Excel Tasks:

1. **Data Exploration**:

   Task 1: Create a statistical summary for numerical features

   Task 2: Create a line chart for the number of enrolments by year

   Task 3: Give a cross-tabulated count for response values against education

   Task 4: Make a boxplot on income and write your observations

   Task 5: Calculate the age of customers and make a histogram of that

   Task 6: Visualize the response against Marital_Status

SQL Tasks:

2. **Data Loading:**
   - Create a schema named "retail_data"
   - Set "retail_data" as the default schema
   - Create tables to store the retail transaction data
   - Set ct_customer as the datetime field while loading the data and apply the appropriate date format

3. **Data Preprocessing:**
    - Calculate the total number of customer encounters in the marketing campaign dataset
    - Identify the top 10 most purchased products in the dataset, such as Wines, Meat Products, etc.
    - Find the count of response values
    - Determine the distribution of customers based on their education level and marital status
    - Identify the average income of customers who participated in the marketing campaign
    - Calculate the total number of promotions accepted by customers in each campaign
    - Identify the distribution of customers' responses to the last campaign
    - Calculate the average number of children and teenagers in customers' households
    - Create an Age column by subtracting year_birth from the current year
    - Create Age_group columns based on the below condition:
      WHEN Age BETWEEN 18 AND 25 THEN '18-25'
      WHEN Age BETWEEN 26 AND 35 THEN '26-35'
      WHEN Age BETWEEN 36 AND 45 THEN '36-45'
      WHEN Age BETWEEN 46 AND 55 THEN '46-55'
      ELSE '56+'
    - Determine the average number of visits per month for customers in each age group

Python Tasks:

4. **Exploratory Data Analysis**

    Task 1: General Data Overview:

    a. Check the first few rows of the dataset to understand its structure

    b. Check the data types of each column

    c. Check for any missing values in the dataset

    Task 2: Descriptive Statistics:

a. Compute summary statistics for numerical columns (mean, median, min, max, and standard deviation)
b. Explore the distribution of numerical variables using histograms or box plots

Task 3: Univariate Analysis:

a. Explore the distribution of each numerical variable using histograms or kernel density plots
b. Explore the distribution of each categorical variable using bar plots or pie charts
c. Identify outliers in numerical variables using box plots or scatter plots

Task 4: Bivariate Analysis:

a. Explore the relationship between numerical variables and the target variable (**Response**) using scatter plots or correlation matrices
b. Explore the relationship between categorical variables and the target variable using bar plots or chi-square tests
c. Explore the relationship between numerical and categorical variables using box plots or violin plots

Observations: Write an analysis report on performing exploratory data analysis (EDA) using Python in the context of building a fraud detection system for Retail Analytics

PowerBI Tasks:

5. **Interactive Dashboard Design in PowerBI:**
   - Visualize customer income distribution grouped by year of registration.
     - Open Power BI Desktop and load the **marketing_campaign.csv** file using the **Get Data > Xlsx** option.
     - Navigate to the **Data view** and verify data types of required columns.
     - Go to Modelling, click **New column**:
       Year Registered = YEAR('marketing_campaign'[Dt_Customer])
     - Drag **Number Diagnoses** and **Number Emergency** to the **X-axis** field.
     - In Fields pane**, right-click Income > New group > Select Bin, set size = 5000**, click OK, rename to **Income Bin**
     - Insert a **Stacked Column Chart**.
     - Drag **Income Bin** to **X-axis**.
     - Drag **Count of ID** to **Y-axis.**
     - Drag **Year Registered** to **Legend**
     - Optionally, enable **Data Labels** under the Format pane for better readability.

   - Create Examine education levels and marital status using Clustered bar chart.

- Create a **Clustered bar chart**.
- Drag **Education** to **Y-axis**
- Drag **Count of ID** to **X-axis**
- Drag **Marital_Status** to **Legend**.
- Optionally, enable **Data Labels** under the Format pane for better readability.

- Explore relationship between income and wine spending
  - Create a new column by clicking on Modelling
  - Use this DAX format : **Normalized Wine Spending = 'marketing_campaign'[MntWines] / 'marketing_campaign'[Income]**
  - Drag **Sum of Income** to **X-axis**
  - Drag **Sum of Normalized Wine Spending** to **Y-axis**
  - Optional, Drag Sum of **MntWines** to **Size**

- Analyze frequency of purchases across categories
  - Insert a **Stacked Column Chart** on new page
  - Drag **ID** to **X-axis**
  - Drag **MntWines, MntFruits, MntMeatProducts** to **Y-axis**
  - Go to the **Filters pane** on the right.
  - Drag **ID** into the **Filters on this visual** area.
  - Change filter type to **Top N**.
  - Choose: **Show Top:** 20 (or 50, up to you)
  - **By Value:** use TotalSpend measure created earlier
  - Click **Apply Filter**

- Create a dashboard with all the visualizations
  - Click on the dashboard icon at the bottom
  - Copy the Visuals from all the pages and paste them on the new page dashboard.