Instructions & Dataset Overview for final exam

This dataset consists of 7,000 observations, each representing a unique customer. The variables represent different characteristics of the customers and their interactions with a website, as well as various demographic and behavioral factors. This dataset should be used for regression, classification, and clustering analyses, which will help derive insights into customer behavior, marketing strategies, and predictive modeling.

Here is the explanation to each variable used in the dataset

CustomerID: Unique identifier for each customer.

Age: Customer's age

AnnualIncome: Customer's annual income

WebsiteVisits: Number of times a customer visits the website.

Clicks: Number of times the customer clicks on items or links.

Conversions: Number of successful conversions (e.g., purchases, signups).

Impressions: Number of times the customer sees ads on the website.

MarketingChannel: The marketing channel used to reach the customer (e.g., Email Marketing, Social Media).

BounceRate: Percentage of visitors who leave the site after viewing one page.

ScrollRate: Percentage of the page that the customer scrolls through.

LoyaltyPoints: Points earned through loyalty programs.

SuccessfulConversion: Whether the customer successfully converted (1 = Yes, 0 = No).

TimeSpentOnPage: The time the customer spends on the website (in minutes).

CustomerAgeGroup: The age group the customer belongs to (e.g., 18-24, 25-34).

DeviceType: The device used by the customer (e.g., Mobile, Desktop, Tablet).

AdViewFrequency: How often the customer sees ads.

CustomerLoyaltyStatus: The loyalty status of the customer (e.g., New, Returning, Loyal).

SeasonalityFactor: The season during which the data was collected (e.g., Winter, Summer).

ReferralSource: How the customer found the website (e.g., Organic Search, Paid Ads).

WishlistAddition: The number of items added to the customer's wishlist.

CartlistAddition: The number of items added to the cart.

ActualLeadQuality: The quality of the lead based on purchasing behavior (High, Medium, Low).

NPS_Score: Net Promoter Score, which measures customer loyalty and satisfaction (1 to 10).

CES_Score: Customer Effort Score, measuring how easy it is for the customer to interact (1 to 7).

CSAT_Score: Customer Satisfaction Score (1 to 5).

Has_Pet: Whether the customer has a pet (0 = No, 1 = Yes).

Sun_Sign: The customer's astrological sun sign (e.g., Aries, Taurus).

DailyCoffeeCups: Number of cups of coffee the customer drinks daily (0 to 5).

ProductCategory: Category of products the customer interacts with or buys (e.g., Electronics, Clothing).

SubscriptionType: Type of subscription the customer has (e.g., Basic, Premium, None).

CLTV: Customer Lifetime Value, the total revenue generated by the customer over time.

Number of Purchases: The number of purchases the customer has made.

Churn Risk: The likelihood that the customer will stop using the service (e.g., High, Medium, Low).

LoyaltyScore: A score representing the customer's loyalty, based on factors like engagement and purchases.

Your group will be using the provided dataset to perform a range of analyses, including regression, classification, and clustering. Each of these methods will help derive insights and predictions based on the customer data.

Report Structure:

- **Introduction:**
  - The **Introduction** section should clearly outline all chosen variables and provide a detailed justification for their inclusion in the analysis. Explain the relevance of each variable in the context of the problem you're trying to solve or the insight you're trying to gain.
  - In this section, outline the **workflow** of your analysis, explaining how the regression, classification, and clustering methodologies will be applied. This serves as the foundation for the entire report and helps to connect all the sections cohesively.
- **Section 1: Regression Analysis**
  - In this section, you will perform **at least two multiple regression analyses**. Each regression model should include one **dependent variable** and three **independent variables**.
  - For each regression, provide a detailed explanation of the results
  - Interpret the **output** from the regression analysis (using Python), including metrics such as **R-squared**, **Betas**, **P-values**, **T-statistics. Please make sure to paste the outputs in the report for better clarity.**
  - **Interpretation of the regression output**: Each coefficient (Beta) should be explained, i.e., what each one represents in relation to the dependent variable using the multile linear regression equation.
  - Discuss whether the model is performing well or underperforming. If the model isn't performing as expected, suggest ways to improve its effectiveness, such as adding variables, transforming features, or using regularization techniques.
  - You may also include **charts and graphs** to help visualize the results and support your explanation
- **Section 2: Prediction using Classification Algorithms**
  - In this section, you will perform at least two classification algorithms for prediction. You can choose from a variety of variables for predicting, but it is important to wisely select the variable your group wants to focus on. Ensure that the chosen variable is relevant and provides meaningful insights for the analysis.
  - Remember to select multiple levels for the target variable. Discuss and arrive at a consensus on which level offers the best accuracy for your chosen model.

- Paste the output of all the models at different levels using multi-card formats or tables, showing key evaluation metrics such as Precision, Recall, F1 Score.
- At the end of the section, explain which prediction metric has the best accuracy.
- Additionally, you may include charts and graphs to help visualize the results and provide a deeper understanding of the model performance

Section 3: Clustering Algorithms

- In this section, you will perform at least two clustering predictions using two different variables from the dataset.
- Identify which variables from the dataset are suitable for clustering. Choose variables that will help effectively segment the data into meaningful clusters.
- Apply clustering algorithms to identify the clusters based on the selected variables.
- Utilize charts and graphs (e.g., scatter plots) to visualize the resulting clusters.
- Provide appropriate explanations of the clustering results, highlighting key observations and insights. Also explain suitable measures on the clusters those which are not performing best. Give recommendations on how make those clusters perform better.

Instructions for Report Writing

- The report should be written in **English** and submitted before the **deadline** of **11th January 2026** 23:59h on **e-campus**.
- **Only one person** from the group needs to submit the report on behalf of the entire group.
- There is **no ideal page limit** for the report, as it will include graphs, charts, and algorithm outputs. A report of up to **30 pages** is acceptable.
- Please ensure you **adhere to the format specified before** for consistency.
- The report will be checked for **AI-generated content** and **plagiarism**, so make sure to include **relevant in-text citations** wherever needed. If you use in-text citations, remember to include a **reference list** at the end of the report.
- **Formatting requirements**: **Font**: Times New Roman, **font Size**: 12, **Spacing**: 1.5

- The **file name** of the report should be: **“Group _No_Final_Report”**. Ensure consistency in naming.
- ***Note****: After the deadline,* ***no reports will be accepted via email***.

Instructions for Presentations

- The **presentations** are scheduled for **15th January 2026** from **15:30 to 18:30**.
- Each group will have a **20-minute slot**, with the presentation order being selected randomly (your slot will be conveyed beforehand). The order of the group members will also be determined randomly.
- Each group will have **15 minutes** for the presentation and **5 minutes** for the Q&A session.
    - **Each group member** must answer at least **one question** during the Q&A to earn points for this section.
- The presentation should be created using **Microsoft PowerPoint** and must be **submitted on e-campus** before the deadline of **15th January 2026 at 10:00**.
- The **file name** of the PowerPoint presentation should be: **“Group _No”**. Please ensure consistency in naming.
- ***Note****: After the deadline,* ***no presentations will be accepted via email***.