# Module End Assignment 5

**Total Marks: 25**

**Project Title:**

**Customer Segmentation and Purchase Prediction Using Machine Learning Techniques**

**Problem Statement:** You are a data Scientist working for an e-commerce company. The company has provided you with a dataset containing information about customer behavior. The goal is to understand patterns in customer data, predict whether a customer will purchase a product, and segment the customers for targeted marketing campaigns.

**Data set Link:**

**https://drive.google.com/file/d/1x5Ly7rQtdNyeeLKSB9jsVfzywxsYrPGw/view?usp=sharing**

## 1. Descriptive Statistics, Visualization, and Pre-processing (10 Marks)

Perform the following steps using Python:

**(a)** Calculate the **mean, median, and standard deviation** for the following columns: **(2 Marks)**

- `Age`, `Total Spend`, `Items Purchased`, `Average Rating`, and `Days Since Last Purchase`.

**(b)** Handle **categorical variables**: **(2 Marks)**

- Apply **label encoding** to `Gender`, `Membership Type`, and `Satisfaction Level`.
- Apply **one-hot encoding** to `City`.

**(c)** Apply **feature scaling**: **(2 Marks)**

- Normalize and Standardize `Total Spend` and `Items Purchased` columns. Display both results.

**(d)** Create a **boxplot** of `Total Spend`. Detect and treat**(4 Marks)**

---

## 2. Classification and Clustering Insights (15 Marks)

**(a)** Use the processed data to **predict customer satisfaction** (`Satisfaction Level`) as a classification problem: **(5 Marks)**

- Encode labels (e.g., Satisfied = 2, Neutral = 1, Unsatisfied = 0)
- Use **Logistic Regression** to build a model. Show accuracy and confusion matrix.

**(b)** Use **K-Means Clustering** to segment the customers: **(5 Marks)**

- Apply clustering on numeric features (after scaling).
- Use **Elbow Method** to determine optimal number of clusters.
- Visualize clusters with a scatter plot of any 2 key features.

**(c)** Identify **key features** influencing satisfaction: **(5 Marks)**

- Use **SelectKBest** and correlation matrix to find the most relevant features.