Data Science Exercise for Master Students

# Objective

Perform customer segmentation based on purchasing behavior to tailor marketing strategies.

# Dataset Overview

Features: order\_number, order\_date, customer\_number, type, month, item\_number (SKU), quantity, category, revenue, customer\_source, order\_source

# Exercise Description

## Title: Customer Segmentation and Analysis

Goal: Segment customers to understand purchasing behavior and suggest targeted marketing strategies.

Note: This exercise is not about completing all tasks, but rather about demonstrating your skills and experiences in data science.

# Steps to Complete the Exercise

## Day 1: Data Preparation and Exploration

1. Data Loading and Cleaning:

- Load the dataset, handle missing values, and ensure correct data types.

2. Exploratory Data Analysis (EDA):

- Generate summary statistics and visualizations.

- Examine feature relationships.

3. Feature Engineering:

- Aggregate data by customer\_number to create features like total revenue and average order value.

- Calculate recency using order\_date.

## Day 2: Customer Segmentation and Analysis

4. Feature Scaling:

- Normalize features for consistency.

5. Clustering:

- Propose and apply a suitable clustering algorithm (e.g., K-means) with rationale.

- Determine optimal clusters using methods like the Elbow Method.

6. Cluster Analysis:

- Analyze and visualize cluster characteristics.

7. Marketing Strategy Suggestions:

- Propose marketing strategies based on cluster analysis.

# Analysis and Predictions to Address

1. Average Order Value (AOV):

- Calculate and analyze average order value.

2. SKU Demand Forecasting:

- Predict SKU quantities needed for the next three months.

3. Top Performing SKU Identification:

- Identify top-performing SKUs by revenue and quantity sold.

4. Customer Segmentation:

- Segment customers by order\_source, customer\_source, and customer\_type.

# Expected Deliverables

1. Jupyter Notebook: Documenting all processes and analysis.

2. Presentation: Summarizing findings and marketing suggestions.

3. Report: Detailed methodology, results, and implications.

# Tools and Libraries

Python: For data analysis.

Pandas: Data manipulation.

Matplotlib/Seaborn: Visualization.

Scikit-learn: Machine learning.

# Evaluation Criteria

1. Data Preparation and Cleaning: Thoroughness and accuracy.

2. EDA and Visualizations: Clarity and insights.

3. Feature Engineering: Effectiveness.

4. Clustering Implementation: Justification and accuracy.

5. Cluster Analysis and Insights: Practicality.

6. Presentation and Reporting: Clarity and professionalism.

# Final Demo

Duration: 45 minutes

Content: Jupyter Notebook demonstration with Q&A.

Objective: Showcase process, findings, and recommendations.

Note: Complete and submit all work in a single Jupyter Notebook.

This exercise provides hands-on experience with real-world data, covering essential steps from preparation to actionable insights. It focuses on demonstrating your data science skills and experiences.