Data Science Assignment: eCommerce Transactions Dataset

Overview:

You are provided with an eCommerce Transactions dataset consisting of three files: `Customers.csv`, `Products.csv`, and `Transactions.csv`. Your task is to perform exploratory data analysis (EDA), build predictive models, and derive actionable insights. This assignment will test your data analysis, machine learning, and business insight generation skills.

Customers.csv:
https://drive.google.com/file/d/1bu_--mo79VdUG9oin4ybfFGRUSXAe-WE/view?usp=sharing
Products.csv :
https://drive.google.com/file/d/1IKuDizVapw-hyktwfpoAoaGtHtTNHfd0/view?usp=sharing
Transactions.csv :
https://drive.google.com/file/d/1saEqdbBB-vuk2hxoAf4TzDEsykdKlzbF/view?usp=sharing

---

Files Description:

1. **Customers.csv**
   - `CustomerID`: Unique identifier for each customer.
   - `CustomerName`: Name of the customer.
   - `Region`: Continent where the customer resides.
   - `SignupDate`: Date when the customer signed up.
2. **Products.csv**
   - `ProductID`: Unique identifier for each product.
   - `ProductName`: Name of the product.
   - `Category`: Product category.
   - `Price`: Product price in USD.
3. **Transactions.csv**
   - `TransactionID`: Unique identifier for each transaction.
   - `CustomerID`: ID of the customer who made the transaction.
   - `ProductID`: ID of the product sold.
   - `TransactionDate`: Date of the transaction.

- ○ `Quantity`: Quantity of the product purchased.
- ○ `TotalValue`: Total value of the transaction.
- ○ `Price`: Price of the product sold.

Assignment Tasks:

Task 1: Exploratory Data Analysis (EDA) and Business Insights

1. Perform EDA on the provided dataset.
2. Derive at least **5 business insights** from the EDA.
   - ○ Write these insights in **short point-wise sentences** (maximum 100 words per insight).

Deliverables:

- A Jupyter Notebook/Python script containing your EDA code.
- A PDF report with business insights (maximum 500 words).

Task 2: Lookalike Model

Build a **Lookalike Model** that takes a user's information as input and recommends **3 similar customers** based on their profile and transaction history. The model should:

- Use both **customer** and **product** information.
- Assign a **similarity score** to each recommended customer.

Deliverables:

- Give the top 3 lookalikes with there similarity scores for the first 20 customers (CustomerID: C0001 - C0020) in Customers.csv. Form an "`Lookalike.csv`" which has just one map: Map<cust_id, List<cust_id, score>>
- A Jupyter Notebook/Python script explaining your model development.

Evaluation Criteria:

- Model accuracy and logic.
- Quality of recommendations and similarity scores.

Task 3: Customer Segmentation / Clustering

Perform **customer segmentation** using clustering techniques. Use both **profile information** (from `Customers.csv`) and **transaction information** (from `Transactions.csv`).

- You have the flexibility to choose any clustering algorithm and any number of clusters in between(2 and 10)
- Calculate clustering metrics, including the **DB Index(Evaluation will be done on this)**.
- Visualise your clusters using relevant plots.

Deliverables:

- A report on your clustering results, including:
  - The number of clusters formed.
  - DB Index value.
  - Other relevant clustering metrics.
- A Jupyter Notebook/Python script containing your clustering code.

Evaluation Criteria:

- Clustering logic and metrics.
- Visual representation of clusters.

Submission Instructions:

1. **GitHub Link**
   - Upload all the PDF and code files in a public GitHub repository.
2. **File Naming Convention:**
   - Use the following naming convention for all your files:
     - `FirstName_LastName_EDA.pdf`
     - `FirstName_LastName_EDA.ipynb`
     - `FirstName_LastName_Lookalike.csv`
     - `FirstName_LastName_Lookalike.ipynb`
     - `FirstName_LastName_Clustering.pdf`
     - `FirstName_LastName_Clustering.ipynb`

Evaluation Process:

Your submissions will be evaluated based on the following criteria:

Task	Weightage
Exploratory Data Analysis	25%
Business Insights	15%
Lookalike Model	30%
Customer Segmentation	30%

Given the large number of submissions, the evaluation will be automated as much as possible. Ensure your file formats and naming conventions are accurate to avoid disqualification.

Final Note:

This comprehensive assignment requires critical thinking and practical application of data science concepts. Focus on creating clean, efficient code and providing meaningful insights that can help the company improve its business strategy.

Good luck!