**PRCE-003 Mall-Customers**

**Project Overview:**

This project focused on analyzing mall customer data using PySpark, a powerful framework for data processing. The analysis aimed to uncover insights into customer behavior, spending patterns, and demographic influences on spending score.

**Key Steps:**

1. Environment Setup:

- Created a new Conda environment named "mallcustomer"

>>>conda create --name mallcustomer

>>>conda activate mallcustomer

- Installed necessary packages: PySpark (py4j is a necessary lib required), OpenJDK

>>>conda install -c conda-forge openjdk

>>>conda install -c conda-forge pyspark

>>>jupyter notebook

2. Data Loading and Initial Exploration:

- Utilized PySpark's SparkSession to create a DataFrame from the CSV file (since the data is in structured form.

- Explored the data structure using methods like printSchema(), show(), and describe().

3. Data Cleaning and Preprocessing:

- Renamed columns for clarity and consistency: corrected “Genre” to “gender”, and removed special characters (like k$, 1-100) from column names.

- Checked for and addressed invalid entries, null values, and duplicates (in customerID).

- Created age groups for more meaningful analysis.

4. Data Analysis:

- Conducted statistical analysis on spending scores

- Examined spending patterns by gender and age groups

- Calculated correlations between numerical variables: The correlation heatmap reveals minimal relationships between variables.

* Age weakly correlates negatively (-0.33) with spending score.
* Annual income shows negligible correlations with age (-0.012) and spending score (0.0099).
* Overall, these variables don't strongly predict each other.

- Categorized spending scores into Low, Medium, and High

5. Visualization:

- Created various plots using Seaborn and Matplotlib:

* Histograms for age, annual income, and spending score distributions
* Correlation heatmap for numerical variables
* Scatter plot of annual income vs. spending score
* Bar plot of average spending score by gender

**Key Inferences:**

1. The project revealed insights into the distribution of customer ages, annual incomes, and their relation to spending scores.

2. Gender-based analysis showed differences in spending patterns between male and female customers.

3. Correlations between age, annual income, and spending score were examined, potentially revealing minimal relationships.

4. Age group analysis provided insights into how different age brackets tend to spend.

5. The categorization of spending scores helped in understanding the overall distribution of customer spending habits.

**Challenges Faced:**

1. The project highlighted the need for more practice in PySpark SQL, as it combines elements of both Python and SQL.

2. Data Path Handling: An initial challenge was encountered with file path syntax, requiring the use of raw string literal (r”file path”) to properly read the CSV file. Also, Google Drive link to data file wasn’t readable in the notebook.

**Conclusion:**

This project provided valuable hands-on experience in using PySpark for data cleaning and analysis, with statistical analysis and visualization techniques. The challenges faced highlight areas for further practice in PySpark SQL.