**Title: Mall Customer Segmentation & Spending Score Analysis**

**Student Name**: *PTID-CDE-MAR-25-61*  
**Project ID**: PRCE-003  
**Tool Used**: PySpark  
**Dataset**: Mall\_Customers.csv

**1. Objective**

The goal of this project is to perform customer segmentation and analyze customer behavior based on their annual income and spending score using Apache PySpark. This helps understand purchasing patterns for targeted marketing strategies.

**2. Steps Performed**

**a. Data Loading**

* Loaded Mall\_Customers.csv using PySpark’s read.csv() function.
* Verified schema using df.printSchema() and previewed the dataset using df.show().

**b. Schema Summary**

* CustomerID - Integer
* Genre - String
* Age - Integer
* Annual Income (k$) - Integer
* Spending Score (1-100) - Integer

**c. Group-wise Analysis**

Grouped customers by Genre and calculated:

* **Average Annual Income**
* **Average Spending Score**

**Results**:

| **Gender** | **Avg Income** | **Avg Spending Score** |
| --- | --- | --- |
| Female | 62.27 k$ | 51.50 |
| Male | 62.27 k$ | 48.61 |

*Interpretation*: Female customers have a slightly higher average spending score despite equal income.

**Who spends more?**

* **Female customers** have a **higher average spending score (51.50)** than male customers (48.61), **even though both genders have similar average incomes (~62.27k$)**.
* This suggests **females may be more engaged or responsive to spending opportunities** at the mall.

**🔹 Gender trends:**

* Spending score differences indicate a **behavioral trend** — females tend to **spend more freely or frequently**, while males may be **more conservative** with spending.
* Useful for marketing campaigns targeting loyalty programs, promotions, or store placements.

**🔹 Income relation:**

* Despite both genders having the **same average income**, their spending patterns **do not align proportionally**.
* This highlights that **income is not the sole factor influencing spending score** — customer behavior varies even at equal income levels.

**3. Challenges Faced**

* pyspark module not found — fixed by reactivating the virtual environment and installing it using pip.
* CSV path issues — corrected the working directory and placed the file in the same folder.
* Column names with spaces ("Spending Score (1-100)") caused errors — resolved using backticks or renaming.
* Mistyped column name ("Gender" vs "Genre") — fixed by checking schema with df.printSchema().

4. **Conclusion**

This PySpark project successfully demonstrated customer segmentation using grouped analysis. Future improvements can include:

* Applying KMeans clustering for deeper segmentation
* Visualizing patterns using Pandas/Matplotlib
* Exporting results to dashboards or reports