Reflection Paper – K-Means Clustering Project

What I Implemented

In this project, I implemented **K-Means clustering** on a customer dataset with features *Income* and *Spending Score*. My workflow followed these steps:

1. **Data Preparation**: Loaded the dataset and selected the two numerical features (Income_$ and SpendingScore). Missing values were filled with the median of the column.
2. **Scaling**: Applied StandardScaler to normalize the data so that income and spending score were on comparable scales.
3. **Choosing K (Elbow Method)**: Ran a loop from K=1 to 10 and printed the Sum of Squared Errors (SSE). This helped visualize the elbow point.
4. **Clustering & Labeling**: Ran K-Means with the chosen number of clusters, assigned labels to each row, and added them back to the dataset.
5. **Metrics**: Evaluated results using **Silhouette Score** and **Davies–Bouldin Index (DBI)**.
6. **Cluster Centers**: Transformed cluster centers back into original units (income and spending score).
7. **Output**: Exported the labeled dataset to spending_labeled_clusters.csv.

Choosing K

Based on the printed results:

- **SSE** dropped sharply until around **K=4–5**, after which the improvement slowed down.
- The **Silhouette Score** was **0.369**, showing moderate cluster separation.
- The **DBI** was **0.991**, which is reasonably low (closer to 0 means better).

From these indicators, I chose **K=5** because:

- The elbow was visible around K=4–5.
- Adding more clusters (e.g., K=9–10) gave smaller SSE but risked overfitting.
- K=5 balanced interpretability and performance.

Cluster Interpretation

Based on the cluster centers (Income vs. Spending Score), here is how I interpreted them in plain language:

1. **Cluster 0 (Middle Income, Average Spending)**
    - Customers have moderate income (~$64k) and balanced spending.
    - **Business Action**: Target with general promotions and seasonal offers.
2. **Cluster 1 (Low Income, Low Spending)**
    - Income is around ~$33k and spending is low.
    - **Business Action**: Offer budget-friendly products or discounts to encourage more purchases.
3. **Cluster 2 (Low Income, High Spending)**
    - Lower income (~$25k) but high spending scores.
    - **Business Action**: Provide loyalty rewards or membership cards to retain these enthusiastic customers.
4. **Cluster 3 (High Income, High Spending)**
    - Wealthier customers (~$94k+) who also spend a lot.
    - **Business Action**: Upsell premium products and offer exclusive luxury services.
5. **Cluster 4 (Very Low Income, Very High Spending)**
    - Very low income (~$23k) but extremely high spending scores.
    - **Business Action**: Monitor carefully – may be aspirational buyers. Consider installment payment options or credit-based offers.

Limitations & Next Steps

Limitations:

- Only two features (*Income* and *Spending Score*) were used, which may oversimplify customer behavior.
- Other variables like **Age, Gender, Region, Visit Frequency, or Online Purchases** could improve the segmentation.
- Clusters may change if scaling or initialization changes (since K-Means is sensitive to starting points).

Next Steps:

- **Add a third feature (e.g., Age)** to see if it provides more meaningful clusters.
- **Try DBSCAN or Hierarchical Clustering** to compare with K-Means.
- **Visualize clusters** in 2D and 3D plots for better interpretability.
- Re-run metrics to confirm whether the new segmentation improves business insights.