Customer Segmentation Report

1. Introduction

The main goal of this task was to divide customers into groups (called **clusters**) based on their characteristics and buying patterns. We used a method called **K-Means Clustering**, which helps in grouping similar things together. We also used a measure called **Davies-Bouldin Index (DBI)** to check how well our clusters were formed. A lower DBI means the clusters are better separated from each other.

2. Method Used

- **Combining Data**:
    - We first combined the information from two files: Customers.csv and Transactions.csv. The customer information tells us about the customer's name and region, while the transaction data tells us what products they bought, how many items they bought, and how much they spent.
    - After combining the data, we removed any rows that had missing values to avoid errors in the analysis.
- **Creating New Features**:
    - We created new features to better understand customer behaviour, such as:
        - **Total Spend**: The total amount a customer has spent on purchases.
        - **Total Quantity**: The total number of items purchased by the customer.
        - **Transaction Count**: The total number of transactions a customer has made.
    - These features were created by grouping the data based on each customer and calculating the sum for the TotalSpend, TotalQuantity, and TransactionCount.
- **Scaling the Data**:
    - Before applying the clustering algorithm, we made sure that all features were on the same scale. This is important because some features (like total spend) have larger numbers than others (like quantity), and this can affect the clustering process. We used **standardization** to make sure each feature has the same impact on the clustering process.
- **Applying K-Means Clustering**:
    - We used the **K-Means algorithm** to group the customers into different clusters. We tested different numbers of clusters (from 2 to 10) to find the best number that works well for this dataset.

3. Number of Clusters Formed

After testing different numbers of clusters, we found that **4 clusters** worked the best. This means that dividing the customers into four groups gives us the most meaningful and distinct customer groups based on their behaviour and profile.

4. Davies-Bouldin Index (DBI)

The **Davies-Bouldin Index (DBI)** is a measure used to evaluate how well the clusters are formed. A lower DBI indicates better clusters, meaning the groups are well separated and not too similar to each other.

- For **K = 4 clusters**, the **DBI value was 1.35**.
    - This means the clusters are quite well-separated from each other, and there is little overlap between them. A value of 1.35 is considered good, as it shows that the clusters are distinct and not very close to each other.

5. Other Clustering Metrics

Apart from the DBI, we also looked at two other metrics to evaluate the quality of the clusters: **Inertia** and **Silhouette Score**.

- **Inertia**:
    - Inertia measures how close the points in each cluster are to the center of the cluster. The lower the inertia, the better the clustering. It shows how tightly packed the customers are within their respective clusters.
    - For our model with 4 clusters, the **inertia value was 1054.23**. This means that the customers within each cluster are relatively close to each other, forming tight and compact clusters.
- **Silhouette Score**:
    - The Silhouette Score measures how similar each customer is to their own cluster compared to other clusters. A higher score indicates that customers are well-matched with their own cluster, while a lower score means the customer might belong to another cluster.
    - For our model with 4 clusters, the **Silhouette Score was 0.63**. This is a good score, meaning that customers are well-matched to their clusters and there is a good separation between the clusters.

6. Conclusion

- **Number of Clusters**: We formed **4 clusters** based on customer data.
- **DBI Value**: The **DBI value was 1.35**, showing that the clusters are well-separated.
- **Inertia**: The **inertia value was 1054.23**, showing that the clusters are tight and compact.
- **Silhouette Score**: The **Silhouette Score was 0.63**, indicating that the customers are well-grouped.

Overall, we successfully divided the customers into 4 distinct groups that show different buying behaviour's and profiles. The clusters were good based on the DBI, inertia, and silhouette score. This segmentation can help the business understand its customer base better and target different groups with personalized strategies.