PART 1 : Exploratory Data Analysis

The dataset comprises data from 45 B2B customers, each with multiple orders and items, varying in value. Through exploratory data analysis (EDA), we've examined key insights, focusing on the distribution and relationships of order quantities, areas, amounts, and specific customer and item details.

```
Total no. of customers= 45
Total no. of items= 5875
no of countries= 14
no of unique qualities= 184
```

Stage 1: Exploring these features and then segment the customers based on the analysis. The exploratory data analysis has provided the following key insights depicted through charts:

Insight 1 - Top customer by order frequency

Link for Graph - Order Frequency by customer.png

Customer M-1 has the highest order frequency, with a total of 2380 orders.The top five customers (M-1, P-5, A-9, JL, and C-1) have significantly higher order frequencies compared to the rest of the customers in the list.There is a gradual decrease in order frequency as we move down the list, with the lower-ranked customers having relatively lower order counts.The distribution of order frequencies provides insights into the importance and contribution of each customer to the overall business. Analyzing and understanding the top customers can help in developing targeted strategies for customer retention, satisfaction, and potentially identifying areas for business growth.

Insight 2 - Top products by quantity required

The code calculates and prints the total quantity required for each product ('ITEM_NAME') in descending order, indicating product popularity by demand. Key insights include identifying best-selling products, understanding demand patterns, and informing inventory management and marketing strategies based on customer preferences.

Link for Graph - Top 5 item Names by quantity ordered.png

The output reveals the top 10 products by order frequency, offering insights into the popularity hierarchy within a dataset. 'DURRY' emerges as the most popular product with a substantial demand of 319,635, indicating strong customer preference. The descending order of demand underscores the varying popularity of product types, with certain items being more sought after than others. Products like 'GUN TUFTED' and 'INDO-TIBBETAN' exhibit lower demand, reflecting diversity in customer preferences. The gradual decrease in order frequency emphasizes a wide range of products in the dataset. Analyzing order frequency provides valuable insights for inventory management, marketing strategies, and overall business planning, enabling data-driven decision-making for sustainable growth.

Insight 3 - Top 5 countries by sales

Here we analyze and visualize the top 5 countries based on total sales revenue.
Link for Graph - Top 5 country by sales.png

The graph shows the USA leading sales revenue at $27,082,870, guiding strategic focus in marketing and supply chain management. It reveals market size disparities and potential in countries like Brazil and China, with the average order size chart providing insights for tailored business strategies. These visualizations support data-driven decision-making for global market engagement and growth.

`Sales distribution across different countries: - CountryName`

We can see the dominance of the United States (USA) in sales revenue, significantly surpassing other countries. The descending order of sales amounts reveals varying contributions, indicating diverse market sizes and international sales diversification. Notably, the USA's substantial sales of $27,082,870 underscores its significant market impact. The detailed sales figures for countries like the UK to Brazil inform strategic business decisions, resource allocation, and indicate growth markets, enabling companies to focus on high-impact areas for sustainable expansion.

Insight 4 - Average order size by country

Link for Graph - Average Order size by country.png

The bar chart presents a clear and concise overview of the mean quantity required for each country. The title 'Average Order Size by Country' indicates the focus on understanding and comparing average order sizes across geographic regions. The chart's design, including a figure size of (10, 6) and rotated x-axis labels, enhances visual clarity. This information is crucial for businesses, as it allows them to identify patterns in customer behaviour related to order quantities. The insights gained from this visualization can inform strategic decisions in inventory management, marketing, and overall business planning, ensuring that businesses are aligned with the specific ordering preferences of customers in different countries.

Insight 5 - Number of repeat orders per customers

Link for Graph - Number of repeat order per customer.png

The bar chart effectively communicates the distribution of customers based on the count of repeat orders, with each bar corresponding to a customer code and its associated number of repeat orders. The title 'Number of Repeat Orders per Customer' indicates the focus on understanding customer loyalty or retention through the lens of repeat purchasing behaviour. The chart's design, featuring a figure size of (12, 8) and rotated x-axis labels, enhances visual clarity and facilitates easy interpretation. This information is valuable for businesses as it enables the identification of loyal customers who consistently place repeat orders, allowing for targeted customer engagement strategies and the enhancement of overall customer satisfaction and loyalty.

PART 2: Customer Segmentation

Stage 1 - Key insights for K- Means clustering:

Link for Graph - Elbow Method for Optimal number of cluster.png

In the above chart we can see the WCSS values for different cluster numbers. the "elbow" point signifies the optimal number of clusters, striking a balance between reducing WCSS and avoiding excessive cluster complexity. The elbow point is where adding more clusters yields diminishing returns in terms of improving model performance. The optimal number of clusters is chosen at this elbow, representing a trade-off between maximizing cluster separation and minimizing the number

of clusters. This visual aid helps in determining the most suitable cluster count for effective data segmentation.

Link for Graph - Customer Segements based on Frequency and Monetary Value.png

The output displays clusters with their respective average transaction frequency (Avg_Frequency) and average monetary value (Avg_Monetary).

- **Cluster 3** stands out with a high frequency of 643 transactions and a substantial average monetary value of approximately **$3,066,518** indicating a high-value segment.
- **Cluster 2** shows moderate frequency but an exceptionally high average monetary value of around $**11,341,050**.
- **Clusters 0 and 1** exhibit lower frequencies and monetary values, representing segments with fewer transactions and lower spending. The clustering helps identify and understand distinct customer segments based on their purchasing behaviour and value contributions.
- Targeted campaigns or loyalty programs can be designed based on the characteristics of each cluster.
- The **silhouette score** of **0.77** indicates a good separation between clusters in K-Means clustering.

Stage 2 - Key Insights for Agglomerative Clustering -

We are using Agglomerative Hierarchical Clustering with Ward linkage to group customers based on similarity. The dendrogram visually represents the merging of clusters, showing hierarchical relationships. This is for understanding how customers are organized into clusters based on scaled data similarity.

Link for Graph - Dendogram for Agglomerative Clustering.png

The Agglomerative Hierarchical Clustering with 4 clusters achieved a silhouette score of 0.77. This score indicates good separation and cohesion within clusters, suggesting effective grouping of similar customers. A higher silhouette score implies well-defined and distinct clusters. Agglomerative Clustering successfully organized the customer data, highlighting meaningful patterns in their behaviour. This information is valuable for businesses seeking to understand and categorize customers based on their shared characteristics, aiding in targeted marketing and personalized service strategies.

Key Insights for HDB Scan Clustering -

Link for Graph - HDB Scan Clusting.png

The output represents clusters with their average transaction frequency (Average_Frequency), mean monetary value (MeanMonetaryValue), and the count of customers in each cluster (Count). Overall, the clustering provides insights into different customer segments based on their transaction behaviour and monetary contributions Cluster 0 (Label: 2):This cluster has low-frequency customers but with an extremely high average monetary value, representing a segment of high-value, infrequent customers.

- Cluster 1 (Label: 3):Interpretation: This cluster comprises customers with very high frequency and substantial average monetary value, indicating a high-value, high-frequency segment.

- Cluster 2 (Label: 0): This cluster represents customers with moderate frequency and a substantial average monetary value, signifying a segment of valuable, moderately frequent customers.
- Cluster 3 (Label: 1): This cluster includes low-frequency customers with lower average monetary values, constituting a larger segment of less frequent and lower-spending customers.

Stage 3 - Key Insights - Analyzing the enhanced clusters

We have added two features 1.High Value v/s Low value items 2.Diversity of items

	CustomerCode	Frequency	Monetary	Cluster	Unique_Items	High_Value_Count
0	A-11	11	1.854041e+05	-1	23	13.0
1	A-6	2	6.247460e+03	3	25	1.0
2	A-9	210	1.592080e+06	-1	504	621.0
3	B-2	8	1.481116e+04	6	44	4.0
4	B-3	11	5.862686e+04	5	72	40.0

the enhanced customer data now includes new features:

1. Unique_Items: The number of unique items each customer has purchased, indicating the diversity of their orders.
2. High_Value_Count: The count of 'High Value' items purchased by each customer.

With these new features, we have reapplied clustering & used Agglomerative Clustering for exploring complex behvoiurs. It's a hierarchical clustering method that can capture more nuanced relationships between data points.