**1. Introduction**

This report summarises the findings from an extensive data analysis of Champo Carpets' sales records. The dataset comprises 13,135 entries detailing carpet orders across various countries, item categories, and quality names. The analysis included exploratory data analysis (EDA) and clustering exercises to uncover customer distribution, spending patterns, and item popularity.

**1.1 Key Insights Derived (EDA):**

* **Customer Distribution:** The U.S. leads in orders (9,196), followed by the U.K. (1,491) and Italy (551). Smaller markets like China, Israel, and UAE present expansion opportunities.
* **Order Size Variability:** The average order area is 44.73 sq. units, ranging widely (0.04 to 1,024 sq. units), reflecting the diverse space needs of customers.
* **Spending Patterns:** Average order spending is $2,392.04, varying greatly up to $599,719.68, indicating a wide range of customer spending behaviours.
* **Item Popularity:** 'Hand Tufted' rugs are the most popular (4,670 orders), influencing inventory strategies. 'Durries' and 'Double Back' rugs also show high demand.
* **Bulk Order Trends:** A slight negative correlation between item quantity and total area suggests bulk ordering of smaller items.
* **Quality-Driven Revenue:** High-quality products like 'Tufted 52C Wool+Visc' generate considerable revenue, highlighting the demand for premium items.

**A graph with different colored bars

Description automatically generatedA graph with red line

Description automatically generated**

**A graph with different colored bars

Description automatically generatedA graph with red dots

Description automatically generated**

**1.2 Segmenting Customers by Spending and Engagement (Stage 1 and Stage 2)**

Customers were categorised based on their order frequency and total spending as part of the segmentation process. Two main features were engineered:

* **Frequency**: The number of unique orders per customer.
* **Monetary Value**: The total spend per customer.

**Comparative Analysis of Clustering Algorithms  
K-Means Clustering:** Applied to customer frequency and monetary value, identifying three clusters:

* **Cluster 0:** Highly engaged, high spending (average frequency 290.83, monetary value $1,496,781.65).
* **Cluster 1:** Moderately engaged, moderate spending (average frequency 16.61, monetary value $292,044.87).
* **Cluster 2:** Lower engagement, extremely high spending (average frequency 36.00, monetary value $11,341,052.51). A high Silhouette Score (0.7745) indicates effective customer segmentation.

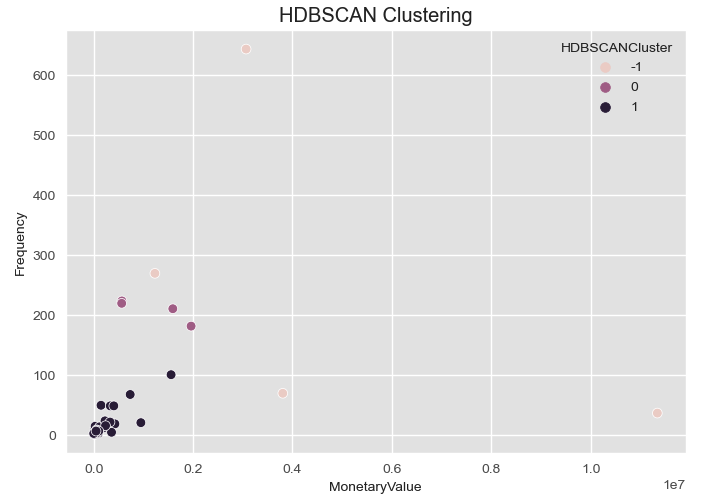
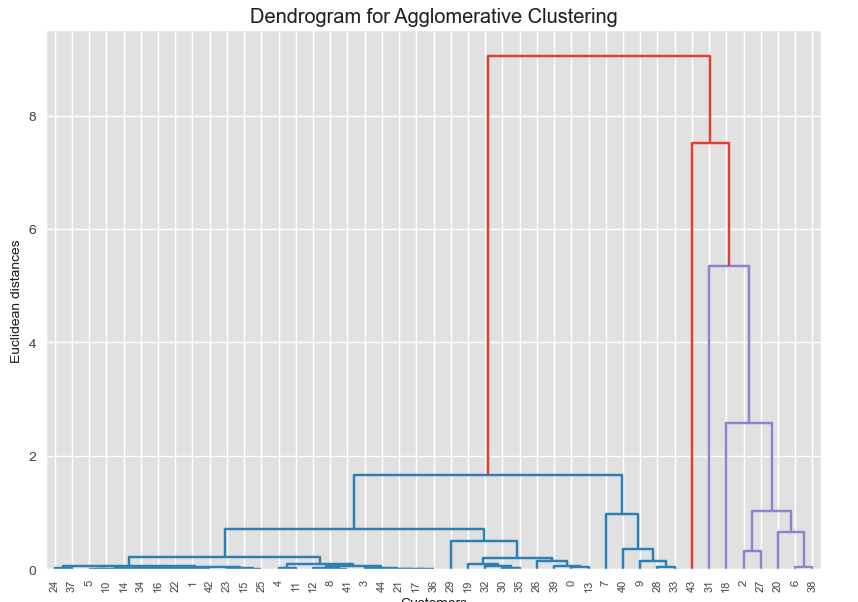
**Agglomerative Clustering:** Revealed similar customer segments with a high Silhouette Score (0.7718), confirming the effectiveness of the segmentation.

**HDBSCAN Clustering:**

* **Cluster -1 (Outliers):** Varied frequency, high monetary value (average $4,860,987.58).
* **Cluster 0:** High frequency and monetary value.
* **Cluster 1:** Lower frequency, moderate monetary value. Supported by Calinski-Harabasz and Davies-Bouldin indexes, indicating well-defined clusters.

**A graph with a number of dots

Description automatically generatedA graph with numbers and lines

Description automatically generated**

**1.3 Extended Feature Set Clustering (Stage 3)  
K-Means Clustering with Extended Features:**

* **Cluster 0**: High engagement and spending with diverse product interaction.
* **Cluster 1**: Moderate engagement and spending, average product diversity.
* **Cluster 2**: Extremely high spending, lower product diversity.
* **Cluster 3**: Low engagement, high spending on a limited product range. These clusters offer a detailed view of customer behavior and preferences.

A graph with red dots

Description automatically generatedA graph with blue and purple dots

Description automatically generated

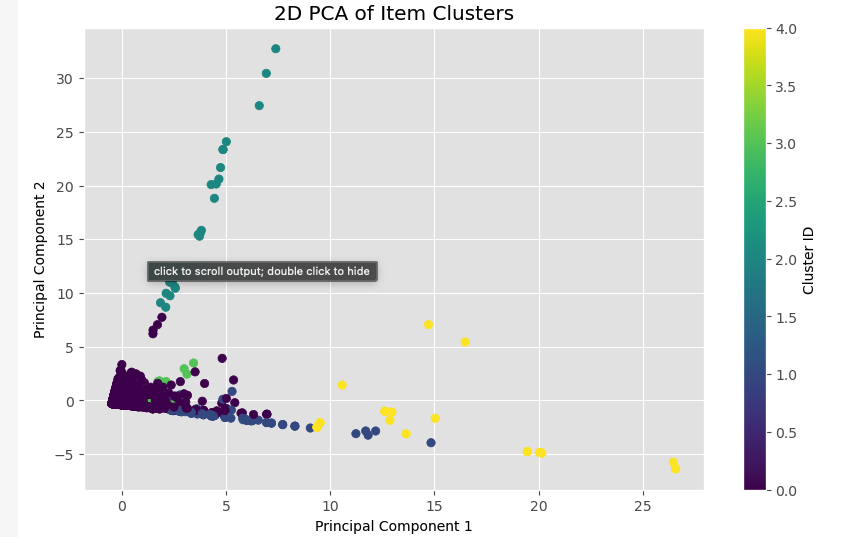
**1.4 Product Segmentation**

**General Observations:**

* The clustering predominantly identifies 'General' segments, indicating a majority of products with moderate quantities ordered and revenue generated.
* 'High Revenue' and 'High Quantity' segments emerge as distinct categories with specific characteristics.

**Key Insights:**

1. **General Segments** (Clusters 0, 1, 3):
   * Represent the bulk of products (5637, 81, 105 items respectively).
   * Characterized by average order quantities and revenues.
   * Suggests these are regularly purchased items, essential for maintaining consistent inventory.
2. **High Revenue Segment** (Cluster 2):
   * Includes 27 items with significantly high total revenue ($339,862.13 on average).
   * Low order quantity but very high revenue per unit.
   * Indicates these are premium or high-value products, potentially with high margins.
3. **High Quantity Segment** (Cluster 4):
   * Comprises 25 items with extremely high quantities ordered (average 5111.16).
   * Large volume sales, but lower revenue per unit.
   * Likely bulk items or essentials, suitable for volume-based marketing strategies.



**Strategic Recommendations:**

* **Focus on High Revenue Items**: Prioritize marketing and stock for premium products in the High Revenue segment.
* **Volume Discounts for High Quantity Items**: Implement bulk sale strategies for the High Quantity segment to maximize volume turnover.
* **Consistent Supply of General Items**: Ensure regular availability of General segment products to maintain steady sales.

**Silhouette Score Analysis:**

* A high silhouette score of 0.8771 suggests well-defined and distinct segments, indicating the effectiveness of the clustering approach in differentiating product categories based on ordering patterns and revenue.

**1.5 Product Cluster Based on Purchase Pattern**  
Interpreting some of these clusters based on the items they contain:

**Largest Clusters**

* **Cluster 107**: This cluster is quite large and contains a diverse range of products (from 2308 to 37036 and many in between). This diversity suggests that Cluster 107 might represent a general category of products that are commonly purchased in various combinations. It could include everyday items or products that are frequently bought together due to their complementary nature or broad appeal.

**Smallest Clusters**

* **Clusters 15, 64, and 88**: These clusters contain very few items (one or two items in each cluster). These might represent unique or niche products that are seldom bought or are only bought in very specific circumstances. The small size of these clusters indicates that these items don't commonly co-occur with others in your dataset.
* **Cluster 77**: A smaller cluster with a few items, possibly indicating a set of products that are often bought together, but not as commonly as those in larger clusters. This could be indicative of a more specialized shopping behavior.
* **Cluster 224**: Similar to Cluster 223, this cluster has a moderate number of items and may represent a specific product category or a common set of items purchased together.

**Three Key Lessons Learned:**

1. **Importance of Data-Driven Decision Making**: This exercise underscores the critical role of data analysis in uncovering hidden patterns and insights, which can drive strategic business decisions. Leveraging data effectively can lead to more informed and impactful strategies in marketing, sales, and operations.
2. **Algorithm Selection and Application Nuances**: The varied results from different clustering algorithms (K-Means, Agglomerative, HDBSCAN) highlight the importance of choosing the right tool for the specific nature of the data. It also demonstrates the need to understand the strengths and limitations of each algorithm to extract the most meaningful insights.
3. **Adaptability and Continuous Learning**: The initial challenges faced, such as handling sparse data and interpreting complex results, emphasize the need for adaptability and continuous learning in the field of business analytics. The ability to pivot strategies based on new findings and to continually refine analytical techniques is crucial for staying ahead in a rapidly evolving business landscape.