

CUSTOMER SEGMENTATION ANALYSIS FOR E-COMMERCE OPTIMIZATION

BY

Venkata Sai Dhanush MIRIYALA

Project Description

This project explores customer segmentation for an online retail platform by analyzing purchase behaviors to enhance marketing strategies and boost customer engagement. By leveraging data-driven insights, we aim to help the e-commerce business target its audience effectively. Completed on July 26, 2025, this analysis combines real-world transactional data with advanced tools to deliver practical recommendations.

Dataset

- **Source:** The data is sourced from the Online Retail Dataset on Kaggle.
- **Link:** <https://www.kaggle.com/datasets/carrie1/ecommerce-data>
- **Explanation:** This dataset contains transactional records from a UK-based online retailer from 2010 to 2011, totaling over 541,877 entries. It offers a detailed view of customer purchasing patterns, making it ideal for segmentation to understand buying frequency, recency, and spending.
- **Features:**
 - InvoiceNo: Unique transaction identifier (string).
 - StockCode: Product code (string).
 - Description: Product description (string).
 - Quantity: Number of items per transaction (integer).
 - InvoiceDate: Transaction date and time (timestamp).
 - UnitPrice: Price per unit (float).
 - CustomerID: Unique customer identifier (string).
 - Country: Country of transaction (string).
- **Purpose:** These features enable the calculation of Recency, Frequency, and Monetary (RFM) metrics, forming the basis for customer segmentation.

Tools Used

- **Snowflake:** A cloud-based platform for efficient data storage and processing.
- **Python (Jupyter Notebook):** Used for RFM analysis and clustering with Snowpark, scikit-learn, and scipy.
- **Power BI:** Utilized to create interactive dashboards for visualizing results.
- **Git:** Employed for version control to manage project files.

Methodology

Our process involved transforming raw data into actionable insights through:

1. **Data Ingestion:** Uploaded the dataset into Snowflake for scalable management.
2. **RFM Analysis:** Computed Recency, Frequency, and Monetary values using Snowpark.
3. **Clustering:** Applied K-Means clustering with a Silhouette Score validation.
4. **Visualization:** Developed a Power BI dashboard to present findings.
5. **Validation:** Ensured data reliability with statistical tests.

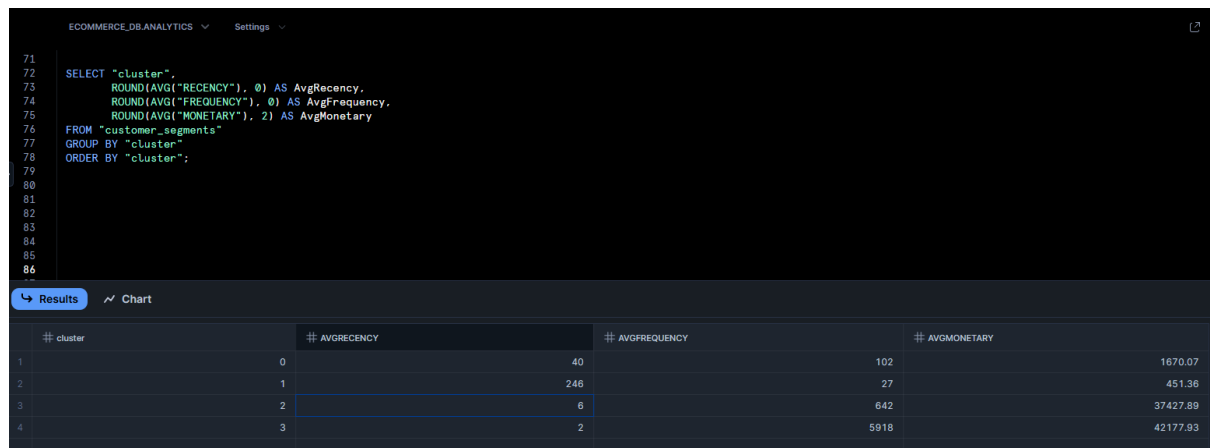
Pipelines

- **Data Pipeline:**
 - Loaded data.csv into Snowflake's ecommerce_stage.
 - Created customer_data table and derived rfm_data with RFM metrics.
 - Generated customer_segments table post-clustering.
- **Analysis Pipeline:**
 - Executed SQL queries for summaries.
 - Used Python for clustering and validation.
 - Visualized outcomes in Power BI.

Tables Generated for Power BI

- **customer_segments:**

- Columns: CustomerID, Recency, Frequency, Monetary, cluster
- Purpose: Stores clustered customer data for dashboard use.

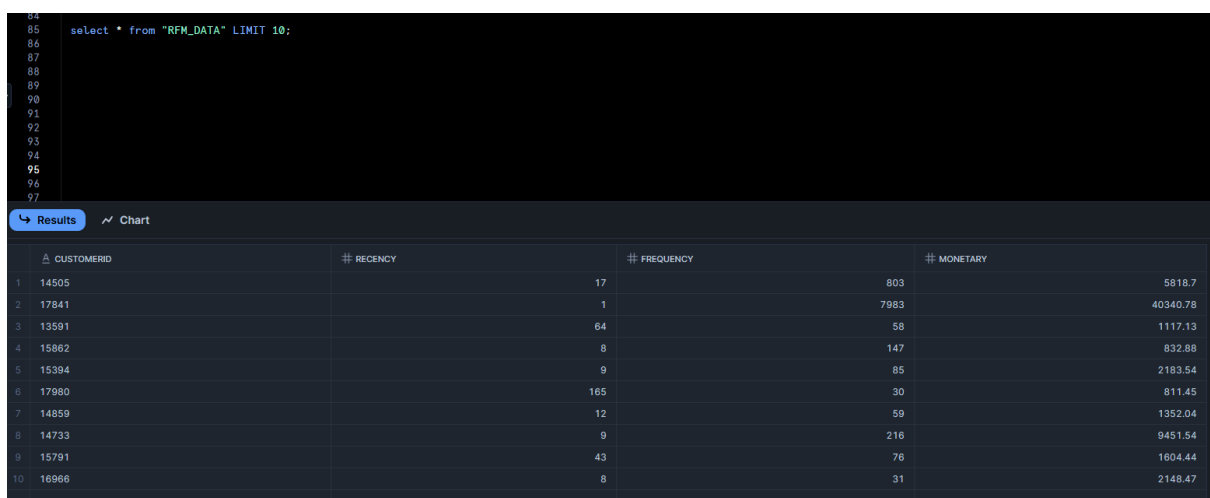


```
71 SELECT "cluster",
72        ROUND(AVG("REGENCY"), 0) AS AvgRecency,
73        ROUND(AVG("FREQUENCY"), 0) AS AvgFrequency,
74        ROUND(AVG("MONETARY"), 2) AS AvgMonetary
75 FROM "customer_segments"
76 GROUP BY "cluster"
77 ORDER BY "cluster";
```

# cluster	# AVGREGENCY	# AVGFREQUENCY	# AVGMONETARY
0	40	102	1670.07
1	246	27	451.36
2	6	642	37427.89
3	2	5918	42177.93

- **RFM_DATA:**

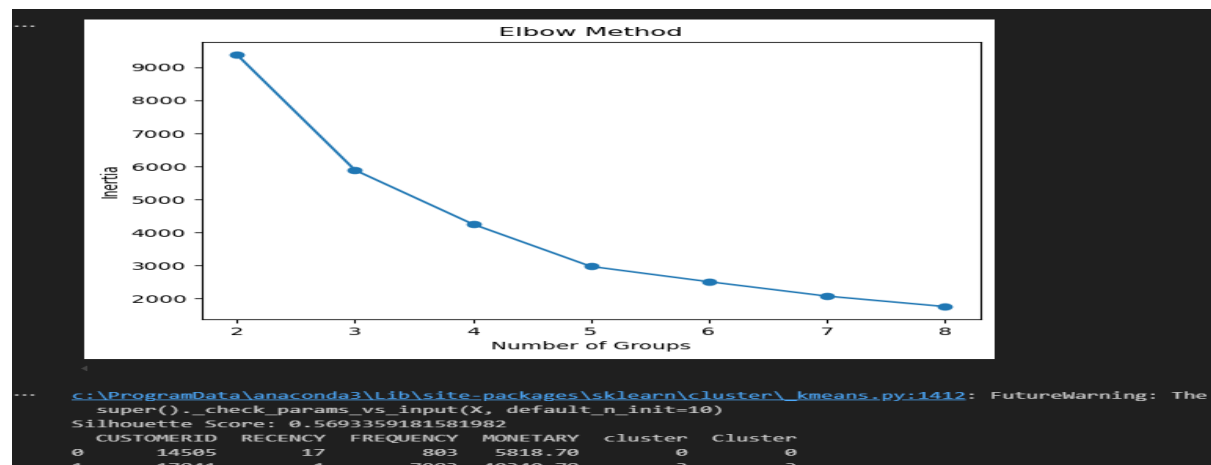
- Columns: CustomerID, Recency, Frequency, Monetary
- Purpose: Holds intermediate RFM metrics before clustering.



```
84 select * from "RFM_DATA" LIMIT 10;
```

# CUSTOMERID	# REGENCY	# FREQUENCY	# MONETARY
14505	17	803	5818.7
17841	1	7983	40340.78
13591	64	58	1117.13
15862	8	147	832.88
15394	9	85	2183.54
17980	165	30	811.45
14859	12	59	1352.04
14733	9	216	9451.54
15791	43	76	1604.44
16966	8	31	2148.47

In this we use Elbow method and Silhouette score.



Why We Chose 4 Clusters (from the Elbow Graph)

We use the **Elbow Method graph** to find the **optimal number of clusters**.

In the graph, the “**elbow point**” appears at **k = 4**, where the drop in inertia (error) starts to slow down.

This means:

Using 4 clusters gives a good balance between model accuracy and simplicity.

Choosing more than 4 adds complexity without significant improvement.

What is Silhouette Score?

The **Silhouette Score** measures **how well each point fits within its cluster** — and how clearly it’s separated from other clusters.

How it works:

- Score ranges from **-1 to +1**:
 - **+1** → Perfect clustering (well-separated, tight clusters)
 - **0** → Overlapping clusters
 - **-1** → Wrong clustering (points assigned to the wrong cluster)

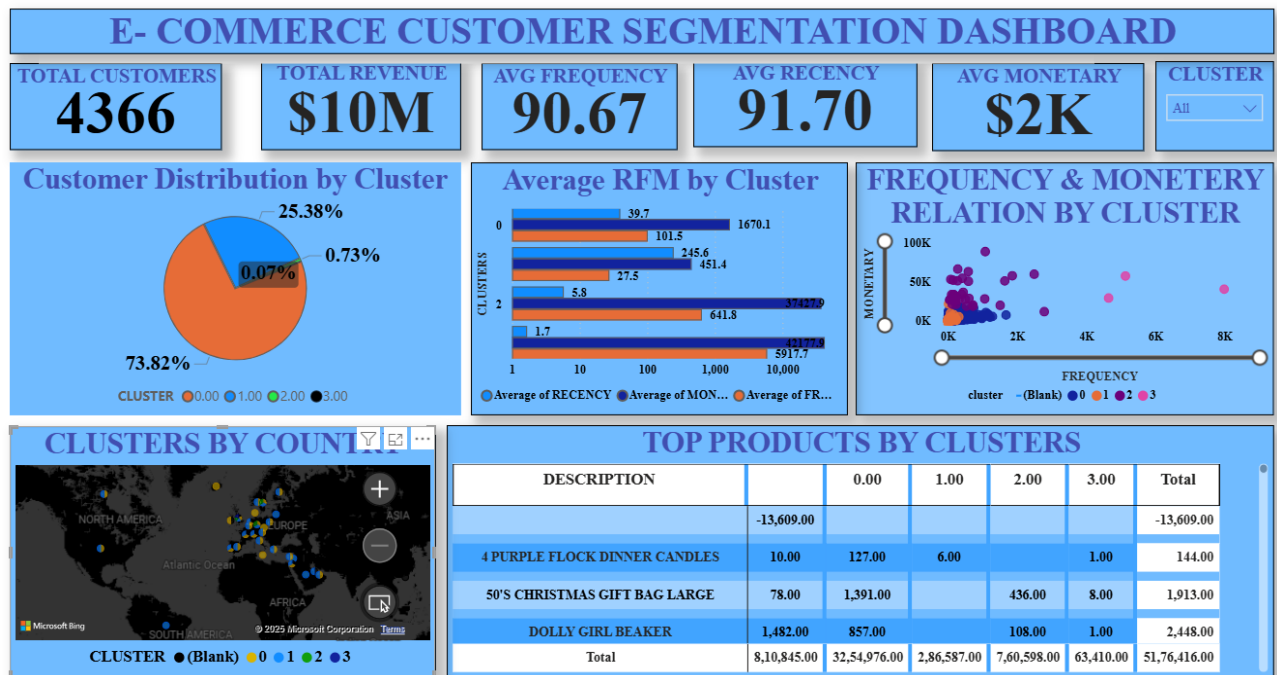
In our case the Silhouette Score: 0.569

This score means:

- Your 4 clusters are **well-formed** and **distinct**.
- The segmentation is **reliable** and meaningful for analysis or marketing actions.

We used the **Elbow Method graph** to select 4 clusters, avoiding unnecessary complexity, and validated the clustering quality with a **Silhouette Score of 0.569**, which indicates well-separated and meaningful customer segments.

Power BI Dashboard Reporting



We added KPIs below:

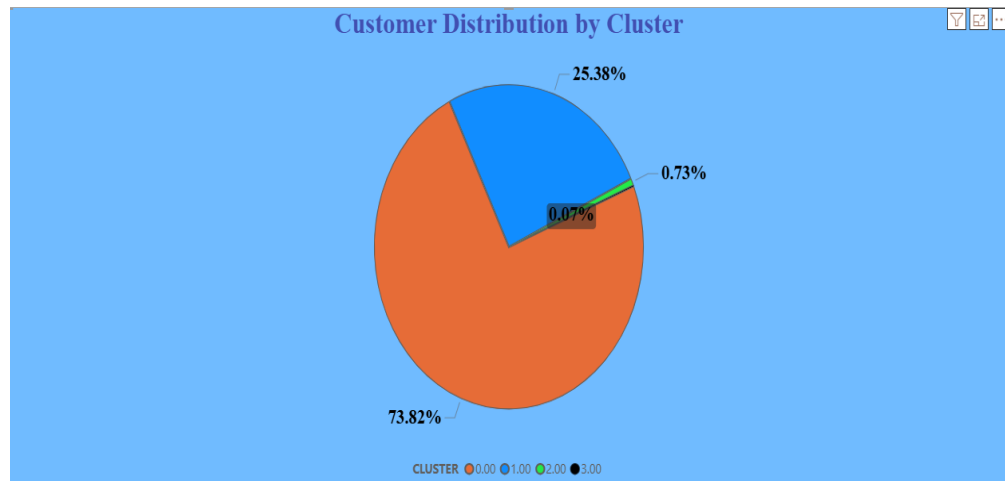
- Total Customers
- Total revenue
- Average Frequency
- Average Recency
- Average Monetary

• Slicer with Clusters:

- **Reporting:** The slicer offers a dropdown to filter data by clusters (0, 1, 2, 3), providing a user-friendly way to focus on specific segments. It enhances analysis by isolating each group's performance.
- **View:** A clean dropdown interface, allowing quick switches between clusters for detailed exploration.

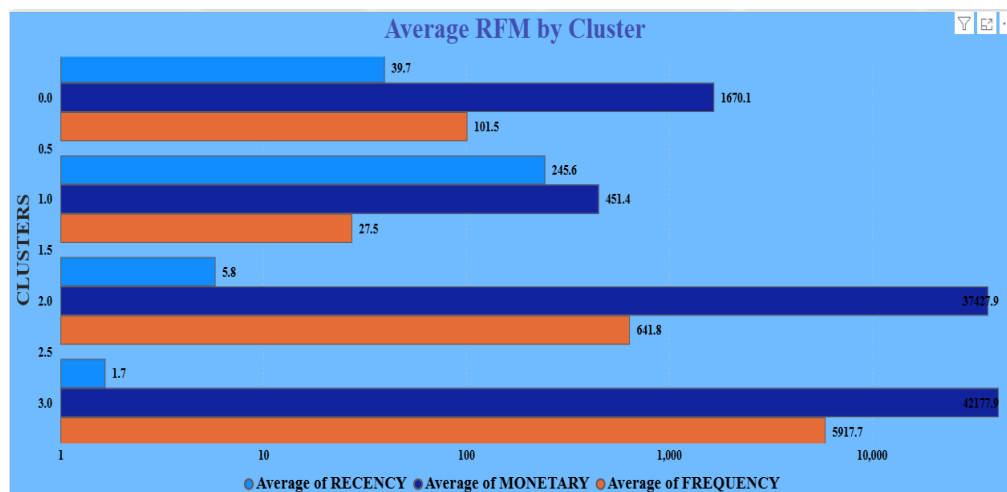
- **Distribution of Customers by Cluster (Pie Chart):**

- **Reporting:** This pie chart shows the proportion of customers across the four clusters. Given the SQL averages, Cluster 0 likely dominates due to its moderate activity, while Cluster 3, with its high frequency, is a small but significant slice.
- **View:** A colorful pie with segments varying in size, highlighting the largest group (likely Cluster 0) and the smallest (likely Cluster 3).



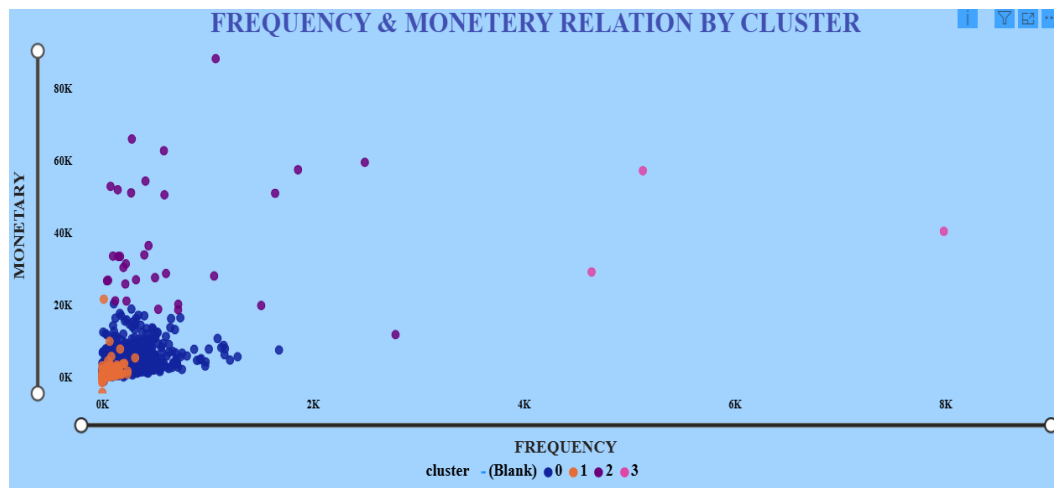
- **Average RFM by Cluster (Clustered Bar Chart):**

- **Reporting:** The bar chart compares average Recency, Frequency, and Monetary values. Cluster 3 stands out with a 2-day Recency, 5,918 purchases, and \$42,177.93 spend, indicating a premium segment, while Cluster 1 lags with 246 days and \$451.36.
- **View:** Clustered bars show Cluster 3 towering in Frequency and Monetary, with Cluster 1 lower across all metrics.



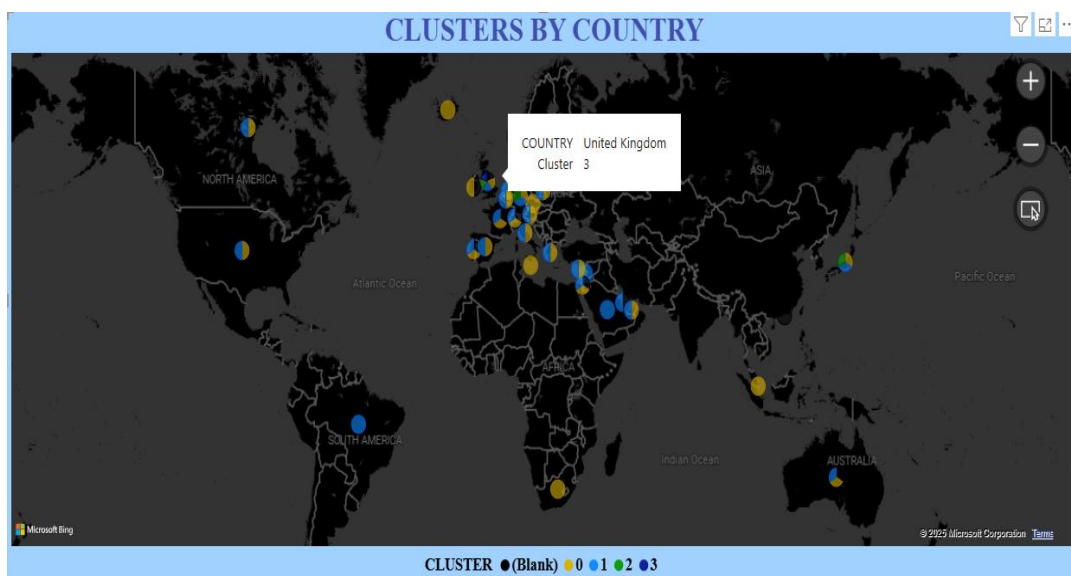
• Scatter Plot of Frequency vs. Monetary by Cluster:

- **Reporting:** This scatter plot maps Frequency against Monetary, color-coded by cluster. Cluster 3 appears in the upper-right with ~5,918 purchases and ~\$42,177.93, marking it as a high-value group, while Cluster 1 clusters lower left.
- **View:** A scatter with distinct clusters, Cluster 3 as a standout point, and others spread across lower values.



• Maps of Country by Cluster:

- **Reporting:** The map displays customer distribution by country per cluster. Given the dataset's UK focus, most clusters likely concentrate there, with Cluster 3 possibly showing a tighter regional pattern due to its small size.
- **View:** A map with shaded regions, predominantly the UK, with varying intensity by cluster.



• Total Products by Cluster:

- **Reporting:** This chart or table lists unique products per cluster. Cluster 3, with its high activity, may lead in product diversity, suggesting broad purchasing, while Cluster 1 might show fewer due to lower engagement.
- **View:** A table or bar chart with Cluster 3 potentially at the top for product count.

TOP PRODUCTS BY CLUSTERS						
DESCRIPTION		0.00	1.00	2.00	3.00	Total
AFGHAN SLIPPER SOCK PAIR	8.00	218.00	279.00		12.00	517.00
AGED GLASS SILVER T-LIGHT HOLDER	507.00	8,428.00	907.00	596.00	124.00	10,562.00
AIRLINE BAG VINTAGE JET SET BROWN	129.00	157.00	18.00	45.00	5.00	354.00
AIRLINE BAG VINTAGE JET SET RED	151.00	307.00	38.00	46.00	9.00	551.00
AIRLINE BAG VINTAGE JET SET WHITE	78.00	123.00	26.00	20.00	8.00	255.00
AIRLINE BAG VINTAGE TOKYO 78	317.00	520.00	130.00	77.00	42.00	1,086.00
AIRLINE BAG VINTAGE WORLD CHAMPION	69.00	172.00	20.00	16.00	4.00	281.00
Total	8,10,845.00	32,54,976.00	2,86,587.00	7,60,598.00	63,410.00	51,76,416.00

Findings and Business Impact

Cluster Insights:

- **Cluster 0:** With 40-day recency, 102 purchases, and \$1,670.07 spend, this large group represents a steady, engaged customer base, ideal for broad marketing.
- **Cluster 1:** At 246-day recency, 27 purchases, and \$451.36 spend, this segment is less active, signaling a need for re-engagement efforts.
- **Cluster 2:** With 6-day recency, 642 purchases, and \$37,427.89 spend, this group is highly loyal and valuable, warranting retention focus.
- **Cluster 3:** With 2-day recency, 5,918 purchases, and \$42,177.93 spend, this tiny segment is a powerhouse, likely including VIPs or bulk buyers.

Business Help

- These insights allow the e-commerce platform to:
 - Prioritize Cluster 3 for exclusive offers to maintain their \$42,177.93 average.
 - Enhance Cluster 2's loyalty with tailored incentives for their \$37,427.89 contribution.
 - Re-engage Cluster 1 with targeted promotions.
 - Leverage Cluster 0's size for scalable campaigns.

Suggestions for the Business

- **Marketing Strategies:**
 - Offer a premium membership for Cluster 3 with personalized deals.
 - Introduce loyalty rewards for Cluster 2 to sustain their high activity.
 - Send reactivation emails with discounts to Cluster 1.
 - Launch seasonal promotions for Cluster 0 to boost their \$1,670.07 average.
- **Inventory Management:**
 - Stock a wide range of products for Cluster 3 based on their diverse purchases.
 - Focus on UK-popular items as indicated by the map.
- **Geographic Focus:**
 - Concentrate marketing efforts in the UK, with plans to explore other regions.
- **Technology Use:**
 - Use Power BI dashboards for ongoing segment monitoring.

Conclusion

This project segmented customers into four distinct groups, with Cluster 3 emerging as the highest-value segment with a \$42,177.93 average spend. Leveraging Snowflake, Python, and Power BI, we've delivered a dashboard that equips the e-commerce business with data-driven insights to enhance engagement and revenue. Validated by a 0.569 Silhouette Score, this analysis lays a strong foundation for future growth and optimization.