**Documentation: User Segmentation Using K-Means Clustering**

**Overview of the SQL Query**

The SQL query retrieves user transaction data to analyze purchasing behaviors. It consists of the following steps:

1. **UserOrderStats Common Table Expression (CTE):**
   * Groups data by user\_id.
   * Aggregates metrics:
     + order\_count: Number of orders placed by a user.
     + total\_order\_amount: Total monetary value of the user's orders.
   * Data source: public.orders and public.groups\_carts tables.
2. **Final Selection:**
   * Retrieves the aggregated metrics (user\_id, order\_count, total\_order\_amount) for subsequent clustering analysis.

This query ensures a clean, user-centric dataset that serves as input for the clustering algorithm.

**Clustering Method: K-Means**

* **Algorithm Overview:**
  + K-Means is an iterative clustering algorithm that partitions data into k clusters.
  + Each cluster is represented by its centroid, and the algorithm minimizes the within-cluster sum of squares (WCSS).
* **Chosen Number of Clusters (k=3):**
  + Based on business requirements and exploratory analysis, three clusters were deemed sufficient to represent user groups:
    - High-value customers.
    - Moderate/occasional buyers.
    - Low-value/rare buyers.
* **Features Used for Clustering:**
  + order\_count: Indicates user activity frequency.
  + total\_order\_amount: Indicates monetary value contributed by the user.
* **Implementation:**
  + Scikit-learn’s KMeans was used with n\_clusters=3 and a fixed random state for reproducibility.

**Benefits of This Analysis**

1. **Customer Segmentation:**
   * Identifies distinct user groups based on their purchasing behaviors.
2. **Targeted Marketing:**
   * High-value customers: Prioritize for loyalty programs or premium offers.
   * Occasional buyers: Encourage more frequent purchases with targeted promotions.
   * Rare buyers: Understand barriers and re-engage with personalized strategies.
3. **Resource Allocation:**
   * Helps businesses allocate marketing resources efficiently by focusing on impactful customer segments.
4. **Dashboard Integration:**
   * Cluster results can be visualized on business dashboards to provide actionable insights to stakeholders.

**Number of Clusters**

* The analysis used **three clusters**:
  + Cluster 0: Represents users with moderate activity and spending.
  + Cluster 1: Represents high-value, frequent buyers.
  + Cluster 2: Represents low-value or infrequent buyers.

**Visualizations Included**

1. **Scatter Plot:**
   * Displays the clustering of users based on order\_count and total\_order\_amount.
2. **Histogram of Order Counts by Cluster:**
   * Shows the distribution of order frequencies within each cluster.
3. **Histogram of Total Order Amounts by Cluster:**
   * Highlights spending patterns of users across clusters.
4. **Cluster Centers Plot:**
   * Visualizes the centroids of clusters to understand their positioning in the feature space.

**Future Enhancements**

1. **Dynamic Clustering:**
   * Experiment with different values of k to validate the choice of clusters using metrics like the elbow method.
2. **Feature Expansion:**
   * Incorporate additional features, such as average order value, recency, and product categories purchased.
3. **Time-Based Analysis:**
   * Analyze trends over time to understand evolving user behaviors.
4. **Integration with Business Metrics:**
   * Align clusters with key performance indicators (KPIs) to track business outcomes effectively.

This analysis empowers businesses to derive actionable insights from their user data, enhancing decision-making and user engagement strategies.