Question 2 Report

Rohan Gore

~rmg9725

Files attached:

- midterm_q2.ipynb

Question 2.1: Customer Spending Analysis

Detailed Approach

The customer spending analysis aims to identify which customers spend the most money overall by computing the total amount spent by each customer, excluding cancelled orders. The approach involves several key steps:

1. **Data Loading**: First, we load the three JSON datasets (customers, orders, and order_items) using Spark's JSON reader.
2. **Order Total Calculation**: For each order, we calculate the total amount by multiplying the quantity by the unit price for each item in the order and then summing these values.
3. **Filtering Valid Orders**: We filter out cancelled orders since they should not count toward customer spending.
4. **Customer Spending Aggregation**: We group the valid orders by customer_id and calculate the total spending for each customer.
5. **Joining with Customer Data**: We join the spending data with customer information to get additional details like customer name and tier.
6. **Rounding and Sorting**: We round the total spent to exactly two decimal places and sort the results in descending order to identify the highest spenders.

Real-Life Use Cases

- **Customer Loyalty Programs**: Businesses can identify their highest-spending customers and offer them special rewards or premium membership status, increasing retention and lifetime value of these valuable customers.
- **Targeted Marketing Campaigns**: By understanding which customers spend the most, businesses can create personalized marketing campaigns that cater to their preferences and purchasing habits, improving marketing ROI and customer engagement.

Question 2.2: Category Preference Analysis

Detailed Approach

The category preference analysis aims to identify the most popular product categories for each customer tier (Gold, Silver, Bronze). The approach involves:

1. **Data Loading**: Loading the three JSON datasets as in Question 2.1.
2. **Category Extraction**: Since categories are stored as an array in the orders data, we use the explode function to convert each category in the array to a separate row.
3. **Filtering Valid Orders**: We filter out cancelled orders to ensure we're only analyzing categories from completed or processing orders.
4. **Joining with Customer Data**: We join the category data with customer information to get the tier information.
5. **Category Counting by Tier**: We count the occurrences of each category within each tier to determine popularity.
6. **Ranking Categories**: We use window functions to rank categories within each tier based on their count.
7. **Filtering Top Categories**: We filter to get only the most popular category (rank = 1) for each tier.
8. **Separate Display by Tier**: We filter the results to display separate tables for Gold, Silver, and Bronze tiers.

Real-Life Use Cases

- **Personalized Product Recommendations**: Businesses can use tier-based category preferences to recommend relevant products to customers in each tier, increasing cross-selling opportunities and customer satisfaction.
- **Inventory Management**: Understanding which categories are most popular among different customer tiers helps businesses optimize inventory planning and ensure adequate stock for high-demand categories.

Question 2.3: Price Range Preferences

Detailed Approach

This analysis provides insights into whether higher-tier customers (e.g., Gold) purchase more premium products compared to lower-tier customers (e.g., Silver or Bronze). The approach involves:

1. **Price Classification**: Implementing the provided `classify_price` function as a User-Defined Function (UDF) to categorize products into three price ranges:
    - Budget: Products priced below $50.00
    - Mid-range: Products priced between $50.00 and $199.99
    - Premium: Products priced at $200.00 or above
2. **Item Total Calculation**: For each order item, calculating the total amount by multiplying the quantity by the unit price.
3. **Filtering Valid Orders**: Filtering out cancelled orders since they should not be included in the spending analysis.
4. **Data Integration**: Joining the classified items with valid orders and customer information to associate each purchase with the customer's tier.
5. **Spending and Product Count Aggregation**: Grouping the data by customer tier and price category to calculate the total spending and product count in each segment.
6. **Percentage Calculation**: For each tier, calculating what percentage of their total spending goes to each price category.
7. **Rounding and Formatting**: Rounding all monetary values and percentages to exactly two decimal places for consistency and readability.
8. **Separate Display by Tier**: Filtering the results to display separate tables for Gold, Silver, and Bronze tiers.

Real-Life Use Cases

- **Tiered Pricing Strategy Optimization**: Businesses can optimize their tiered pricing strategy by understanding how different customer segments respond to products across price ranges, potentially increasing revenue by aligning pricing with customer preferences.
- **Targeted Product Development**: Companies can develop products that align with the price preferences of each customer tier, focusing resources on creating offerings that resonate with their most valuable customer segments.

Question 2.4: Top Customer Identification

Detailed Approach

The top customer identification analysis aims to identify the top 2 highest-spending customers within each tier-price category combination. This provides a granular view of who the most valuable customers are within each specific segment. The approach involves:

1. **Data Integration**: Using the classified items DataFrame from Question 2.3, we join it with valid orders and customer information to associate each purchase with the customer's information and tier.
2. **Spending Aggregation by Customer, Tier, and Price Category**: We group the data by customer ID, name, tier, and price category to calculate the total spending for each customer within each tier-price category combination.
3. **Ranking Customers**: We use window functions to rank customers within each tier-price category combination based on their total spending.
4. **Filtering Top Customers**: We filter to get only the top 2 customers (rank <= 2) for each tier-price category combination.
5. **Rounding and Formatting**: We round all monetary values to exactly two decimal places for consistency and readability.
6. **Separate Display by Tier**: We filter the results to display separate tables for Gold, Silver, and Bronze tiers.

Real-Life Use Cases

- **VIP Customer Programs**: Businesses can identify their most valuable customers within each segment and offer them special VIP treatment, exclusive access to new products, or personalized services, enhancing loyalty among high-value customers.
- **Personalized Marketing**: Companies can create highly targeted marketing campaigns for top customers in each segment, tailoring messages and offers to their specific preferences and spending patterns, resulting in higher engagement and conversion rates.