Final Project

Instructions

1. Solve the problem using **pyspark** library.
2. Using Pandas library or other APIs to solve the problem **will not be accepted** as a solution.
3. All steps should be explained in detail using comments or markdown text.
4. Write the queries based on the questions given in the questions section.
5. Each step should have a heading and detailed explanations in plain text.
6. **Databricks** platform has to be used for this project.
7. Upload the data file to the location: ***/FileStore/tables/project/orders.csv and /FileStore/tables/project/customers.json***
8. All steps should be executed correctly
9. Use **PEP8** standards for coding (Refer to link).
10. Submit the notebook as **.dbc** file.
11. The notebook should have the Names and IDs of all members of your group at the beginning.
12. Please do not copy code from your friends. Plagiarism cases will be penalized. Direct copying will result in zero marks for both the groups involved.
13. Notebooks that are not properly documented or commented will be penalized

---

Dataset

- **About Dataset**
  - The "Online Sales in the USA" dataset contains the transactional data pertaining to various products, encompassing diverse merchandise and electronic categories across multiple states. Given the substantial shift of the internet-accessible population towards e-commerce, prominent retail entities are actively pursuing strategies to augment their profitability. Sales analysis stands as a pivotal methodology adopted by these retailers, aiming to optimize sales through a comprehensive understanding of customer purchasing behaviors and trends.
- Download the dataset from the following link:
  - **Link to Data**
- The data consists of 2 files:
  - **Orders.csv**
    - order_id, order_date, status, item_id, qty_ordered, price, value, discount_amount, total, category, payment_method, cust_id
  - **Customers.json**
    - City, County, Customer Since, E Mail, Gender, Place Name, Region, State, Zip, Age, cust_id, full_name

Problem Statement

Design a Data Lake and its zones based on the above datasets ingestion and develop queries to find key ***insights about market, products and customers.***

Problems to be Solved:

Landing Zone:

1. Read both the datasets (CSV and JSON files), apply the necessary schema and ingest the datasets into the landing zone.

Staging Zone:

Store the datasets in staging zone using appropriate schema, formats and partitions as specified below.

1. Join both datasets and store the dataset in columnar format.
2. The above-joined file should be stored as two separate destination files/tables.
    a. First one without any partitions
    b. Second one with partitions based on year and month.

Curated Zone:

Run the below queries on the above files/tables. Please choose to run the queries on the appropriate staging table (partitioned or non-partitioned). Also, create the charts or dashboards if applicable to any of the below queries.

1. Find **revenue** generated by different **categories** for the month of **11/2020**.
2. Which **top 5 categories** have a maximum number of **refunds** in the year **2020**?
3. Find a **total number of orders** by **each category for each month and year** in the dataset?
4. Segment customers by age: **0-20 as young, 20-35 adults, 35-55 middle-ages and >55 Old**. Find the **total spend** (in percentage of total spend of categories) by

customers by different age segments by different categories.

- E.g., Categories Total Spend Young Adults Middle-Age Old
- E.g., In appliances category total spend is X amount of which a,b,c,d are percentages contributed from Young, Adults, Middle-Age and Old respectively. (a+b+c+d=100)

5. Spend by **gender** across different **categories** in terms of **percentages**.
6. Find **top 5 customers** for each **month**. The results should be as follows:
   - E.g., Year, Month, Custid, Name, country, Gender, Total Spend, Rank
7. Calculate the **RFM** values for **each customer** (by customer id).
   - *RFM Represents R (Recency) F (Frequency) M (Monetary Value) (Click on the link to read more about RFM)*
   - For Recency calculation, use 31/10/2021 as last date. So, the recency for any customer should be how many weeks before he or she has made the last purchase from the date of 01/10/2021.
8. Find **top 10 customers** based on **frequency** and **monetary value**. Sort them based on first frequency and then monetary value.