Personal Development School Analytics Challenge

Welcome to the Personal Development School Analytics Challenge! This assessment evaluates your SQL skills, data pipeline design, and ability to create insightful visualizations and summaries. We look forward to seeing your approach and insights.

Problem Statement: Cohort Subscription Retention Analysis

Background:

You are working for an edtech company that offers various subscription plans. The goal is to produce a dynamic **Waterfall Retention Report** that tracks how many users from each subscription cohort remain active over time, specifically month by month, since the beginning of their subscription. You are provided with data on subscription start dates, cancellation dates (if applicable), and subscription plans, which have varying durations.

Objective

1. **Cohort Definition**:
    - Define cohorts based on the **subscription_started year and month**.
2. **Month 1 to Month N Retention**:
    - For each cohort, calculate how many users remain active in each subsequent month after their subscription starts.
    - Users should be counted as "active" only during the months they are subscribed:
        - If a user cancels, they should be counted as active only until the month of cancellation.
        - If a user is still subscribed (i.e., no cancellation), they should continue to be counted as active through the current month.
3. **Handling Subscription Plan Types**:
    - **Monthly**: Renews every month.
    - **Quarterly**: Renews every 3 months.
    - **Semi-Annual**: Renews every 6 months.
    - **Annual**: Renews every 12 months.
    - **Single Course**: Treated as a 24-month subscription.
    - **Lifetime**: Also treated as a 24-month subscription.
    - **Attachment-Bootcamp**:- Renews every month.

For example:

- If a user starts their subscription in **October 2019** and their subscription lasts for 3 months, they should be counted in **Month 1 (October)**, **Month 2 (November)**, and **Month 3 (December)**.
- If a user started in **June 2024** and has not canceled their subscription yet, they should be counted in **Month 1 (June 2024)**, **Month 2 (July 2024)**, and subsequent months until the current month.

Requirements

1. **SQL Query**:
    - Write a SQL query to dynamically generate a report for each cohort's retention from Month 1 through Month N, determined by the maximum subscription duration.
    - The query output should include:
        - A column for **subscription_started** (formatted as "YYYY/MM").
        - Columns for each **Month_1, Month_2, Month_3**, etc., showing active user counts for each month.
2. **Dataset Preparation**:
    - **Google Sheets**: Import the dataset into a Google Sheet and sort it by the creation date.
3. **Data Pipeline**:
    - **Ingestion and Storage**:
        - Design a data pipeline to pull data from Google Sheets, process it, and push it to a data warehouse (BigQuery or Snowflake).
        - Store the raw data file in CSV format within a data lake (GCP Cloud Storage or AWS S3).
    - **Scheduling and Orchestration**:
        - The pipeline should trigger every 8 hours.
        - Use **Apache Airflow** for pipeline orchestration.
    - **Data Pipeline Implementation**:
        - Implement the pipeline using **Python**.
4. **Visualization and Reporting**:
    - **View Creation**: Write a view in the data warehouse based on the SQL query for the retention analysis.
    - **Dashboard**: Create a visualization dashboard in **Looker Studio, Power BI, or Tableau** to present insights on churn and retention.
5. **Deliverables**:

- Submit:
  - **Airflow DAG files** for the data pipeline.
  - **SQL query** for the retention and churn analysis view.
  - **Link to the visualization dashboard** showcasing insights.
  - A **Google Doc** summarizing key insights and trends observed.

Example Output:

subscription_started	Month_1	Month_2	Month_3	Month_4	...
2019/10	100	80	70	50	...
2020/01	150	130	100	90	...
2024/06	200	190	180	170	...

Constraints:

- The data must be dynamic and adapt to varying subscription durations across cohorts.
- Users who have not canceled their subscription should continue to be counted in the current month.
- Use efficient querying techniques, as the data may involve a large number of subscriptions and long timeframes.

Additional Notes:

- Consider edge cases such as subscriptions canceled on the same day, long-term subscribers (e.g., lifetime plans), and dynamic updates to the number of months based on actual data.

Evaluation Criteria:

- Clarity and efficiency of the SQL query.
- Design, accuracy, and reliability of the data pipeline.
- Quality and insightfulness of the visualization and summary insights.