Email Marketing Analytics Process

Objective

Measure converters for the launched email campaign and perform quality assurance (QA) on the campaign data.

Datasets :
https://drive.google.com/drive/folders/1RcKyXrBiWwAxQl2dQhNbsAMS7NF2eoR1?usp=sharing

QA Checks

Perform the following QA checks on campaign data, ensuring only records matching the Master Data are processed. Records failing any check will be logged in a separate file (qa_failed_records.csv).

1. **Master Data Match**: Only process campaign data records that match the Master Data on the MD5 hash of the email.
2. **Duplicate Check**: Ensure no duplicate records exist at the MD5 level.
3. **Non-Empty MD5**: No records should have an empty or NULL MD5 field.
4. **Date Validation**: Verify the following date hierarchy:
    - **Click Date** ≥ **Open Date** ≥ **Delivery Date**.
    - Definitions:
        - **Delivery Date (Del Date)**: Date when the campaign email was sent.
        - **Open Date**: Date when the email was opened.
        - **Click Date**: Date when a link in the email was clicked.
        - **Unsubscribed Date (Unsub)**: Date when the recipient unsubscribed.

Output for Failed Records

- Records failing any QA check will be stored in a separate file with the following columns:
    - MD5, Delivery Date, Open Date, Click Date, Unsub Date

Unsubscribed Data Handling

- Extract all unsubscribed records that match the Master Data on MD5.
- Store these records in a separate file (unsub_records.csv) with the following columns:
    - MD5, Delivery Date, Unsub Date.

Email Marketing Analytics Process

Converter Measurement

For records passing all QA checks, identify converters by matching with Customer Data provided by the client. A record is considered a converter if it matches on **either** of the following levels:

1. **Email MD5**: Match on the MD5 hash of the email.
2. **Postal Level**: Match on postal fields. (ADDREPLUS SUITE+ ZIP)

Converter Criteria

- The difference between the **Delivery Date** and the **Customer Date** (date of purchase or conversion) must satisfy:
    - **0 ≤ (Customer Date - Delivery Date) ≤ 30 days** (the window period).
- Output converter records in separate files, grouped by **Delivery Date**, as requested by the client. Suggested file format:
    - converters_del_date_YYYYMMDD.csv converters_del_date_2025-04-21.csv

Top 10 Customers Analysis

Identify the top 10 customers based on **CUST_NO** (Customer Number) and their associated **ORDER_NO** (Order Number). Each customer may have multiple orders.

Steps

1. Aggregate data by **CUST_NO** to count distinct **ORDER_NO** values.
2. Sort customers by the number of orders in descending order.
3. Select the top 10 customers.
4. For each of the top 10 customers, list all associated **ORDER_NO** values.

Output

- Store results in a file (top_10_customers.csv) with the following columns:
    - CUST_NO, ORDER_NO List (comma-separated list of order numbers).

Implementation Notes

- Use a robust data pipeline to handle data matching, deduplication, and date validations.
- Ensure all date fields are parsed in a consistent format (e.g., YYYY-MM-DD).
- Log all QA failures with clear reasons to facilitate debugging and reporting.
- Validate Customer Data for completeness (e.g., non-null MD5, postal code, and Customer Date) before matching.