Customers Data Quality Report

Detected Data Issues found by LLM

- Duplicate Columns: customer_id and cust_id both seem to identify the customer, but one is int and the other is string.
- Duplicate Columns: customer_name and full_name both represent the customer's name, but have different formatting and content.
- Duplicate Columns: email and email_address both store email information, but have different values and null percentages.
- Duplicate Columns: phone and phone_number both store phone number information, but have different formats, and null percentages.
- Duplicate Columns: zip_code and postal_code both store zip code information, but zip_code has a high null percentage and is a float, while postal_code is a string.
- Duplicate Columns: registration_date and reg_date both capture registration date, but have varying formats and null percentages.
- Duplicate Columns: status and customer_status both capture customer status, but have different values and null percentages.
- Mixed Data Types: zip_code is a float, which is unusual for zip codes. It should likely be a string to preserve leading zeros.
- Missing Values: High percentage of missing values in email, phone, zip_code, registration_date, preferred_payment, age, birth_date and gender.
- Inconsistent Formatting: customer_name has mixed formats (e.g., 'frank-miller', 'Bob Wilson', 'alice.johnson@email.com').
- Inconsistent Formatting: city has mixed casing (e.g., 'Houston', 'New York', 'NYC', 'chicago', 'philadelphia').
- Inconsistent Formatting: state has abbreviations and full names (e.g., 'California', 'AZ', 'New York', 'IL', 'NY').
- Inconsistent Formatting: status and customer_status have mixed casing (e.g., 'ACTIVE', 'active', 'INACTIVE', 'inactive').
- Invalid Entries: phone_number contains empty strings.
- Invalid Entries: email_address contains empty strings.
- Data Type: total_spent is string instead of numeric
- Possible inconsistencies between state and city: Some cities appear to be in multiple states in the sample data (e.g. Chicago is in both IL and California)
- Age inconsistent with birth_date: The calculated age from the birth_date does not always match the age column. Missing values in either column makes verification difficult.
- Inconsistent Date Formats: registration_date has different date formats, e.g. '12/11/2023' and '2023-02-16'
- Empty Values: Several string columns contain empty strings ('') which may need to be treated as nulls depending on the context.

Column-wise Summary generated by utility

Column : customer_id

Types : int

Sample Values : [1, 2, 3, 4, 5]

Unique Count : 500

Null % : 0.0%

Notes : None

Column : cust_id

Types : str

Sample Values : ['CUST_0001', 'CUST_0002', 'CUST_0003', 'CUST_0004', 'CUST_0005']

Unique Count : 500

Null % : 0.0%

Notes : None

Column : customer_name

Types : str

Sample Values : ['henry.davis123', 'jane_doe', 'Grace Lee', 'John Smith', 'alice.johnson@email.com']

Unique Count : 10

Null % : 0.0%

Notes : None

Column : full_name

Types : str

Sample Values : ['John Smith', 'jane_doe', 'Charlie Brown', 'diana.prince', 'EVE WHITE']

Unique Count : 10

Null % : 0.0%

Notes : None

Column : email

Types : str

Sample Values : ['customer1@example.com', 'customer3@example.com', 'customer4@example.com', 'customer5@example.com', 'customer6@example.com']

Unique Count : 401

Null % : 19.8%

Notes : None

Column : email_address

Types : str

Sample Values : ['user1@domain.com', 'user2@domain.com', '', 'user5@domain.com', 'user6@domain.com']

Unique Count : 357

Null % : 0.0%

Notes : None

Column : phone

Types : str

Sample Values : ['555-3757', '555-2711', '555-3525', '555-5864', '555-6391']

Unique Count : 208

Null % : 58.4%

Notes : None

Column : phone_number

Types : str

Sample Values : ['', '(555) 376-9467', '(555) 998-4735', '(555) 930-4810', '(555) 612-1072']

Unique Count : 257

Null % : 0.0%

Notes : None

Column : address

Types : str

Sample Values : ['2356 Pine Rd', '9075 Oak Ave', '1522 Second Ave', '9885 Second Ave', '4240 Second Ave']

Unique Count : 494

Null % : 0.0%

Notes : None

Column : city

Types : str

Sample Values : ['Houston', 'Phoenix', 'New York', 'Chicago', 'NYC']

Unique Count : 10

Null % : 0.0%

Notes : None

Column : state

Types : str

Sample Values : ['California', 'AZ', 'New York', 'IL', 'PA']

Unique Count : 8

Null % : 0.0%

Notes : None

Column : zip_code

Types : float

Sample Values : [13375.0, 47793.0, 95669.0, 32649.0, 75004.0]

Unique Count : 358

Null % : 28.2%

Notes : None

Column : postal_code

Types : str

Sample Values : ['81012-7131', '43335-5087', '59904-1550', '60808-9661', '97458-9695']

Unique Count : 305

Null % : 0.0%

Notes : None

Column : registration_date

Types : str

Sample Values : ['12/11/2023', '6/12/2022', '11/19/2023', '1/28/2022', '8/16/2023']

Unique Count : 370

Null % : 21.6%

Notes : None

Column : reg_date

Types : str

Sample Values : ['', '2021-12-28', '2021-01-11', '2021-01-03', '2020-07-25']

Unique Count : 298

Null % : 0.0%

Notes : None

Column : status

Types : str

Sample Values : ['suspended', 'ACTIVE', 'INACTIVE', 'active', '']

Unique Count : 7

Null % : 11.6%

Notes : None

Column : customer_status

Types : str

Sample Values : ['active', '', 'pending', 'inactive', 'INACTIVE']

Unique Count : 7

Null % : 13.8%

Notes : None

Column : total_orders

Types : int

Sample Values : [43, 10, 7, 6, 20]

Unique Count : 50

Null % : 0.0%

Notes : None

Column : total_spent

Types : float

Sample Values : [2527.99, 1611.75, 156.96, 4229.64, 2685.28]

Unique Count : 500

Null % : 0.0%

Notes : None

Column : loyalty_points

Types : int

Sample Values : [690, 513, 461, 959, 79]

Unique Count : 394

Null % : 0.0%

Notes : None

Column : preferred_payment

Types : str

Sample Values : ['debit_card', 'paypal', 'credit_card', 'cash']

Unique Count : 4

Null % : 22.4%

Notes : None

Column : age

Types : float

Sample Values : [53.0, 34.0, 77.0, 20.0, 24.0]

Unique Count : 60

Null % : 36.4%

Notes : None

Column : birth_date

Types : str

Sample Values : ['1978-02-12', '1970-02-24', '1998-02-17', '1994-01-06', '1990-02-05']

Unique Count : 208

Null % : 58.2%

Notes : None

Column : gender

Types : str

Sample Values : ['F', '', 'Other', 'Male', 'Female']

Unique Count : 6

Null % : 11.8%

Notes : None

Column : segment

Types : str

Sample Values : ['regular', 'premium', 'new', 'vip']

Unique Count : 4

Null % : 21.6%

Notes : None
