Customers Data Quality Report

Detected Data Issues found by LLM

- Redundant columns: 'customer_id' (int) and 'cust_id' (str) both seem to identify the customer, but 'cust_id' has a specific format ('CUST_0001').
- Redundant columns: 'email' and 'email_address' both store email information, but they have different null percentages and potentially different values.
- Redundant columns: 'phone' and 'phone_number' both store phone number information, but they have different null percentages and potentially different formats.
- Redundant columns: 'zip_code' (float) and 'postal_code' (str) both store postal code information, but have different data types and null percentages.
- Redundant columns: 'registration_date' and 'reg_date' both store registration date information, but have different formats and null percentages.
- Inconsistent naming: 'customer_name' contains values that seem like usernames (e.g., 'henry.davis123', 'jane_doe', 'alice.johnson@email.com') while 'full_name' contains actual names. The sample records confirm name swapping between columns.
- Mixed data types: 'zip_code' is a float, which is unusual for zip codes. It should be a string to accommodate leading zeros and potentially non-numeric characters.
- Missing values: High percentage of null values in 'phone' (58.4%), 'zip_code' (28.2%), 'registration_date' (21.6%), 'preferred_payment' (22.4%), 'age' (36.4%), 'birth_date' (58.2%) and 'segment' (21.6%).
- Inconsistent formatting/casing: 'city' has inconsistent casing (e.g., 'Houston', 'Phoenix', 'New York', 'Chicago', 'NYC'). It also contains values like 'new_york' and 'la'.
- Inconsistent formatting/casing: 'state' has both full names (e.g., 'California', 'New York') and abbreviations (e.g., 'AZ', 'IL', 'PA').
- Inconsistent formatting: 'phone_number' has different formats, some with parentheses and spaces (e.g., '(555) 376-9467') and some are empty strings.
- Inconsistent formatting: 'registration_date' has different date formats (e.g., '12/11/2023', '6/12/2022', '2022-04-09').
- Inconsistent formatting: 'status' and 'customer_status' have inconsistent casing (e.g., 'suspended', 'ACTIVE', 'INACTIVE', 'active'). They also contain empty strings.
- Invalid entries: 'email_address' contains empty strings, which are invalid email addresses.
- Invalid entries: 'gender' contains empty strings and 'Other', which might need further clarification or standardization.
- Potential data inconsistency: Discrepancies between 'status' and 'customer_status'. For example, a customer can have 'status' as 'ACTIVE' and 'customer_status' as 'INACTIVE'.
- Potential data inconsistency: Customer names are being populated incorrectly. 'customer_name' should be the username and 'full_name' should be the actual name, but the sample data shows them populated in reverse.
- Missing values: 'status' and 'customer_status' columns contain NULL values.
- Mixed data types: The 'total_spent' column is of type float, but the sample data contains the values as strings. This needs to be converted to float.

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
----------------------------------------
```