Orders Data Quality Report

Detected Data Issues found by LLM

- Redundant columns: 'order_id' (string) and 'ord_id' (integer) both seem to represent the same order identifier. One should be dropped or investigated for discrepancies.

- Redundant columns: 'customer_id' (integer) and 'cust_id' (string) both seem to represent the same customer identifier. One should be dropped or investigated for discrepancies.

- Redundant columns: 'quantity' and 'qty'. It's unclear if these represent the same quantity or different types of quantities.

- Inconsistent date formats in 'order_date': Some values are in 'M/D/YYYY' format, while others are in 'YYYY-MM-DD' format. This column needs to be standardized.

- Missing values in 'order_date': The 'order_date' column has 9.0% null values.

- Missing values in 'order_datetime': The 'order_datetime' column has 19.3% null values.

- Missing values in 'status': The 'status' column has 15.1% null values.

- Missing values in 'order_status': The 'order_status' column has 13.3% null values.

- Missing values in 'notes': The 'notes' column has 78.6% null values. Consider if this column is necessary or if the missing values should be addressed.

- Missing values in 'tracking_number': The 'tracking_number' column has 37.6% null values.

- Inconsistent casing in 'status' and 'order_status': Values like 'delivered', 'processing', 'CANCELLED', 'pending', 'SHIPPED' show mixed casing. This should be standardized (e.g., all lowercase or title case).

- The 'notes' column has only one unique value, 'Special delivery instructions'. This suggests it might not be useful or is not being populated correctly. It could be a candidate for removal or further investigation.

- Potential data type mismatch: 'order_id' is string, while 'ord_id' is integer. One may need to be cast to match the other.

- Potential data type mismatch: 'cust_id' is string, while 'customer_id' is integer. One may need to be cast to match the other.

- The sample records show discrepancies between unit price and price with quantity. e.g. order_id ORD_00160, unit_price is 300.3 and quantity is 10, so price should be around 3000, but price is 55.56. Check price calculation.

- shipping_address column only contains Oak Ave and Pine Rd. This may be an artificially generated dataset or requires further investigation.

Column-wise Summary generated using utility

Column   : order_id
Types   : str
Sample Values : ['ORD_00001', 'ORD_00002', 'ORD_00003', 'ORD_00004', 'ORD_00005']
Unique Count : 1000
Null %   : 0.0%
Notes   : None

Column   : ord_id

Types   : int

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

Unique Count : 1000

Null %   : 0.0%

Notes   : None

Column   : customer_id

Types   : int

Sample Values : [180, 337, 384, 22, 168]

Unique Count : 432

Null %   : 0.0%

Notes   : None

Column   : cust_id

Types   : str

Sample Values : ['CUST_0180', 'CUST_0337', 'CUST_0384', 'CUST_0022', 'CUST_0168']

Unique Count : 432

Null %   : 0.0%

Notes   : None

Column   : order_date

Types   : str

Sample Values : ['5/9/2023', '2023-11-30', '2023-01-17', '2023-08-06', '1/13/2023']

Unique Count : 480

Null %   : 9.0%

Notes   : None

Column   : order_datetime

Types   : str

Sample Values : ['2023-01-03T17:48:00.000Z', '2023-01-20T08:04:00.000Z', '2023-02-20T13:37:00.000Z',
'2023-09-02T02:34:00.000Z', '2023-02-15T22:46:00.000Z']

Unique Count : 806

Null %   : 19.3%

Notes   : None

Column   : product_id

Types   : str

Sample Values : ['PROD_198', 'PROD_044', 'PROD_001', 'PROD_129', 'PROD_074']

Unique Count : 199

Null % : 0.0%

Notes : None

Column : item_id

Types : int

Sample Values : [146, 41, 100, 153, 157]

Unique Count : 200

Null % : 0.0%

Notes : None

Column : quantity

Types : int

Sample Values : [8, 1, 9, 2, 7]

Unique Count : 10

Null % : 0.0%

Notes : None

Column : qty

Types : int

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

Unique Count : 5

Null % : 0.0%

Notes : None

Column : unit_price

Types : float

Sample Values : [130.62, 36.21, 146.71, 109.64, 413.82]

Unique Count : 990

Null % : 0.0%

Notes : None

Column : price

Types : float

Sample Values : [404.94, 324.83, 145.07, 232.03, 69.9]

Unique Count : 986

Null % : 0.0%

Notes : None

Column   : total_amount

Types   : float

Sample Values : [140.51, 887.54, 352.45, 674.66, 813.13]

Unique Count : 994

Null %   : 0.0%

Notes   : None

Column   : order_total

Types   : float

Sample Values : [570.3, 700.84, 834.33, 203.85, 178.07]

Unique Count : 993

Null %   : 0.0%

Notes   : None

Column   : shipping_cost

Types   : float

Sample Values : [0.0, 12.26, 17.65, 7.0, 13.77]

Unique Count : 591

Null %   : 0.0%

Notes   : None

Column   : tax

Types   : float

Sample Values : [9.98, 13.2, 14.0, 10.51, 50.25]

Unique Count : 894

Null %   : 0.0%

Notes   : None

Column   : discount

Types   : float

Sample Values : [0.0, 18.36, 30.96, 14.69, 38.16]

Unique Count : 279

Null %   : 0.0%

Notes   : None

Column   : status

Types   : str

Sample Values : ['delivered', 'processing', 'CANCELLED', 'pending', 'SHIPPED']

Unique Count : 6

Null % : 15.1%

Notes : None

Column : order_status

Types : str

Sample Values : ['SHIPPED', 'CANCELLED', 'processing', 'pending', 'delivered']

Unique Count : 6

Null % : 13.3%

Notes : None

Column : payment_method

Types : str

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

Unique Count : 5

Null % : 0.0%

Notes : None

Column : shipping_address

Types : str

Sample Values : ['8984 Oak Ave', '6820 Oak Ave', '751 Oak Ave', '9781 Oak Ave', '6747 Pine Rd']

Unique Count : 984

Null % : 0.0%

Notes : None

Column : notes

Types : str

Sample Values : ['Special delivery instructions']

Unique Count : 1

Null % : 78.6%

Notes : None

Column : tracking_number

Types : str

Sample Values : ['TRK510443', 'TRK118744', 'TRK745174', 'TRK919582', 'TRK429287']

Unique Count : 624

Null % : 37.6%

Notes   : None