Sales Data Analysis and Data Quality Check Report

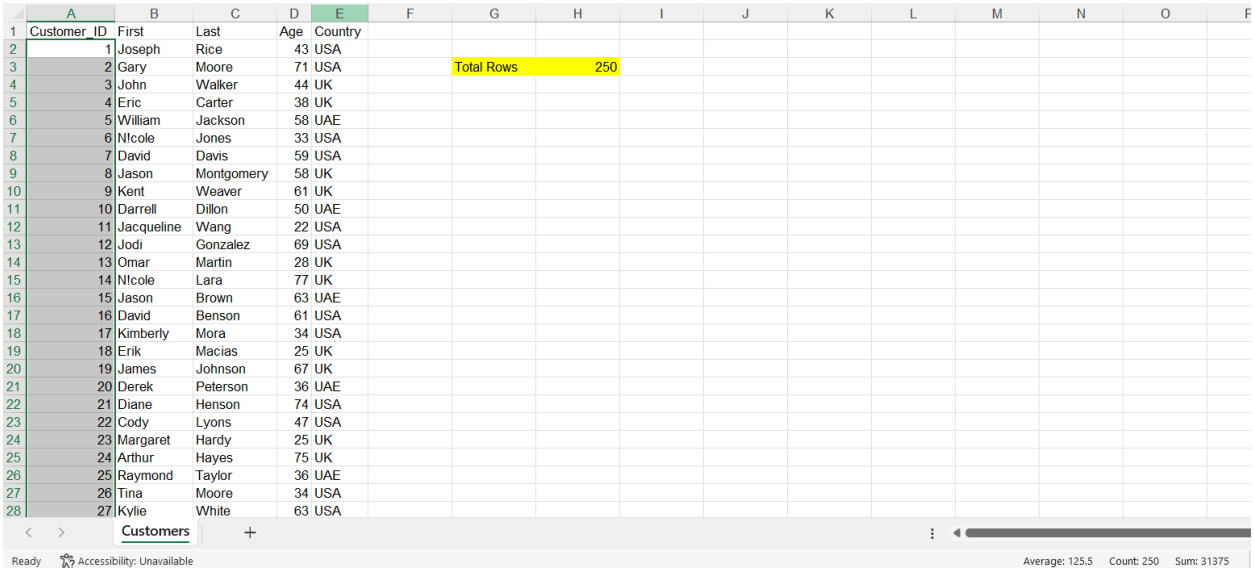
Author - Shantanu Zodey

Date – 30/05/2025

Source Data Verification.

Customer.xlsx –

- Consist of total 250 records.
- 5 Columns – (Customer_ID, First, Last, Age, Country)



Customer_ID	First	Last	Age	Country
1	Joseph	Rice	43	USA
2	Gary	Moore	71	USA
3	John	Walker	44	UK
4	Eric	Carter	38	UK
5	William	Jackson	58	UAE
6	N!cole	Jones	33	USA
7	David	Davis	59	USA
8	Jason	Montgomery	58	UK
9	Kent	Weaver	61	UK
10	Darrell	Dillon	50	UAE
11	Jacqueline	Wang	22	USA
12	Jodi	Gonzalez	69	USA
13	Omar	Martin	28	UK
14	N!cole	Lara	77	UK
15	Jason	Brown	63	UAE
16	David	Benson	61	USA
17	Kimberly	Mora	34	USA
18	Erik	Macias	25	UK
19	James	Johnson	67	UK
20	Derek	Peterson	36	UAE
21	Diane	Henson	74	USA
22	Cody	Lyons	47	USA
23	Margaret	Hardy	25	UK
24	Arthur	Hayes	75	UK
25	Raymond	Taylor	36	UAE
26	Tina	Moore	34	USA
27	Kylie	White	63	USA

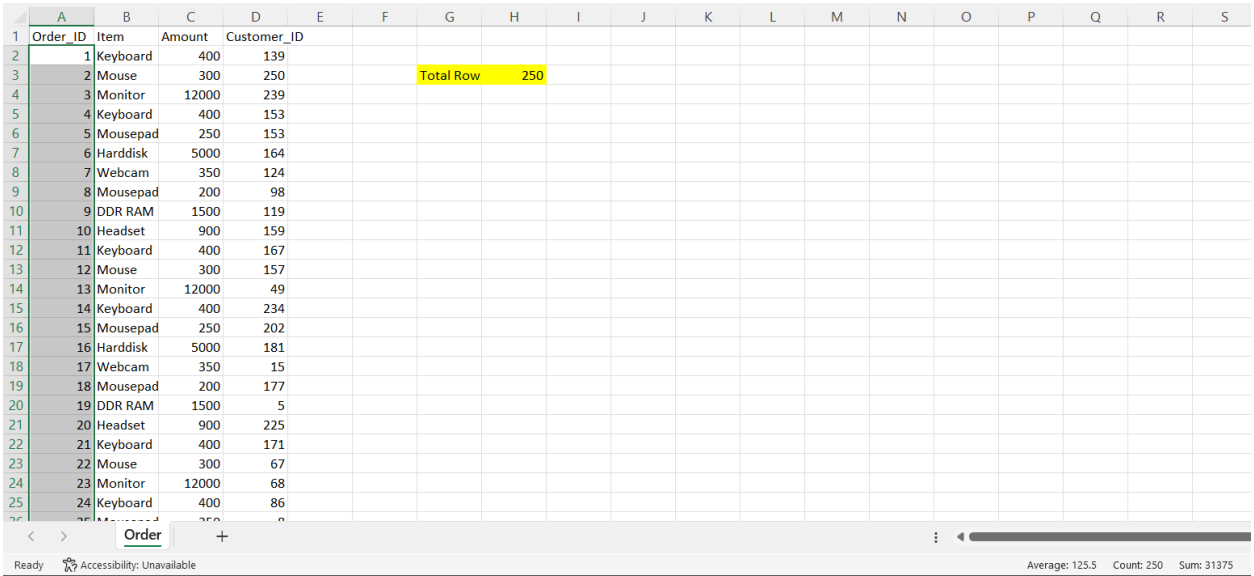
Total Rows 250

Customers

Ready Accessibility: Unavailable Average: 125.5 Count: 250 Sum: 31375

Orders.csv –

- Consist of total 250 records.
- 5 Columns – (Order_ID, Item, Amount, Customer_ID)



Order_ID	Item	Amount	Customer_ID
1	Keyboard	400	139
2	Mouse	300	250
3	Monitor	12000	239
4	Keyboard	400	153
5	Mousepad	250	153
6	Harddisk	5000	164
7	Webcam	350	124
8	Mousepad	200	98
9	DDR RAM	1500	119
10	Headset	900	159
11	Keyboard	400	167
12	Mouse	300	157
13	Monitor	12000	49
14	Keyboard	400	234
15	Mousepad	250	202
16	Harddisk	5000	181
17	Webcam	350	15
18	Mousepad	200	177
19	DDR RAM	1500	5
20	Headset	900	225
21	Keyboard	400	171
22	Mouse	300	67
23	Monitor	12000	68
24	Keyboard	400	86

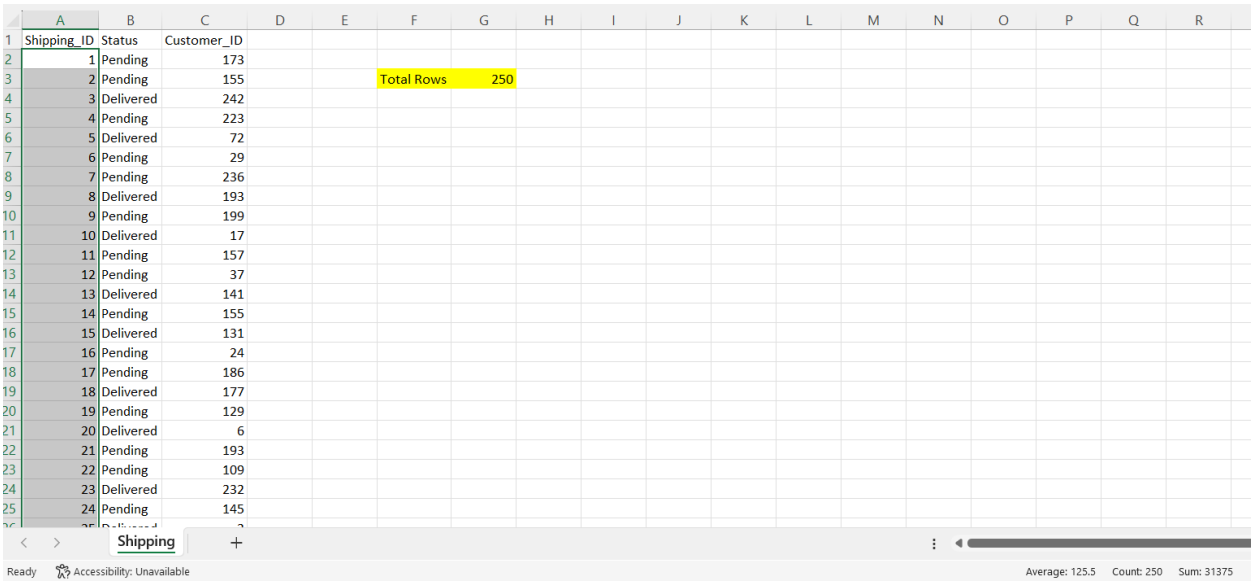
Total Row 250

Order

Ready Accessibility: Unavailable Average: 125.5 Count: 250 Sum: 31375

Shipping.json -> Shipping.csv –

- Consist of total 250 records.
- 5 Columns – (Shipping_ID, Status, Customer_ID)



Shipping_ID	Status	Customer_ID
1	Pending	173
2	Pending	155
3	Delivered	242
4	Pending	223
5	Delivered	72
6	Pending	29
7	Pending	236
8	Delivered	193
9	Pending	199
10	Delivered	17
11	Pending	157
12	Pending	37
13	Delivered	141
14	Pending	155
15	Delivered	131
16	Pending	24
17	Pending	186
18	Delivered	177
19	Pending	129
20	Delivered	6
21	Pending	193
22	Pending	109
23	Delivered	232
24	Pending	145

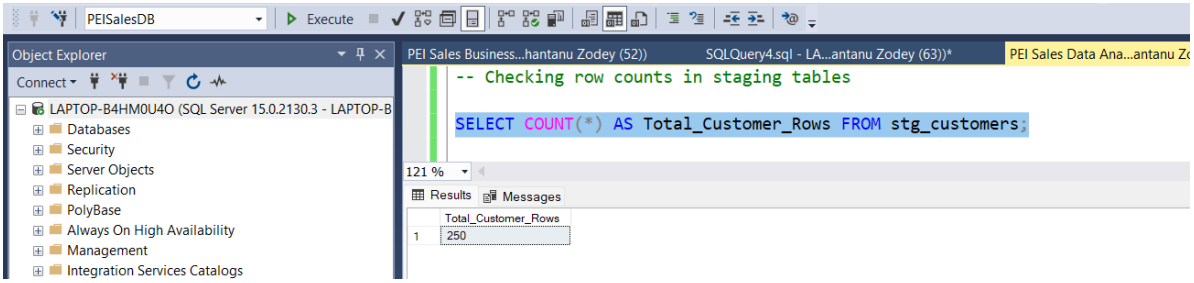
Total Rows 250

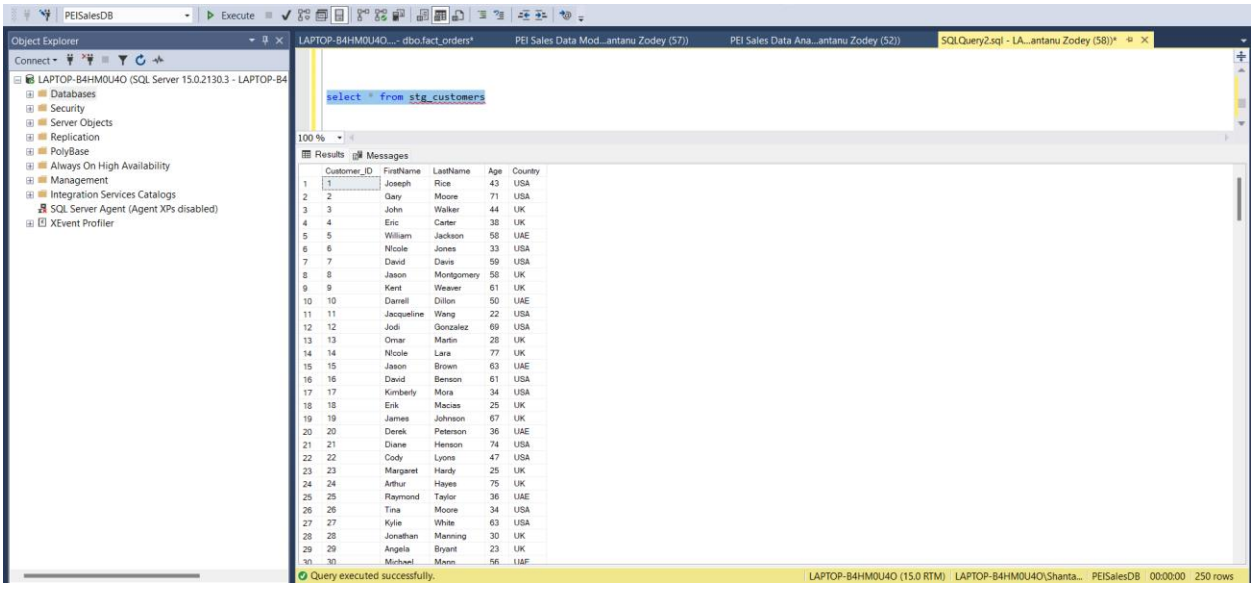
Shipping

Ready Accessibility: Unavailable Average: 125.5 Count: 250 Sum: 31375

Customer Data Quality

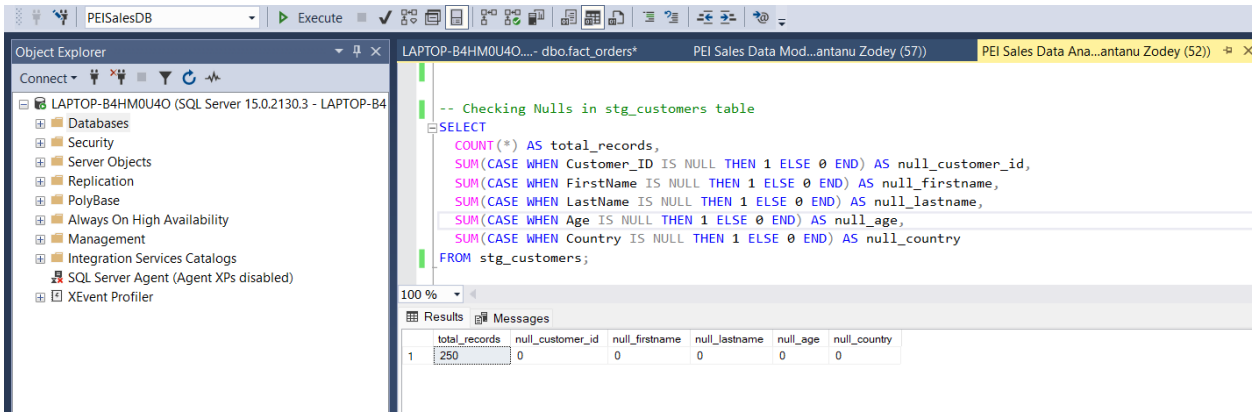
Validating row count and the records that are loaded into staging table from Customer.xlsx.





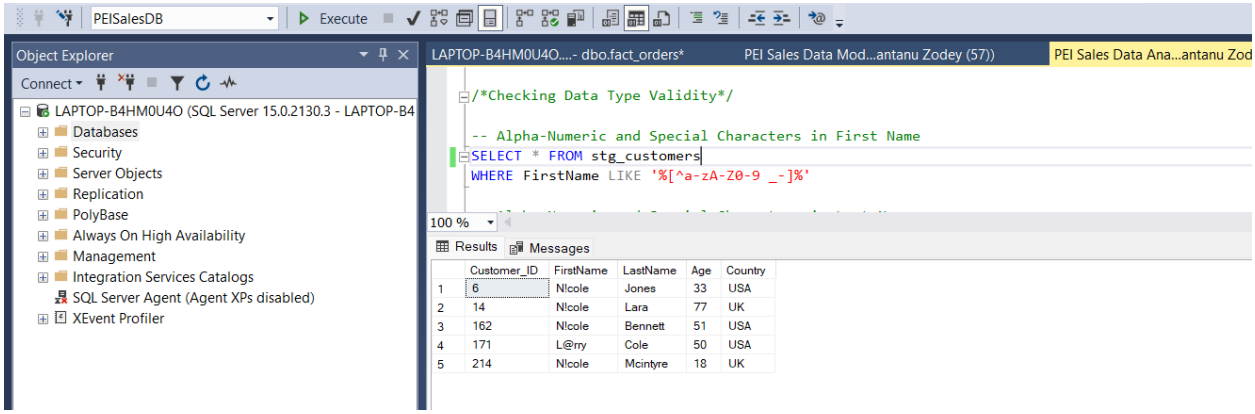
Completeness Checks

Checking Nulls in stg_customer table



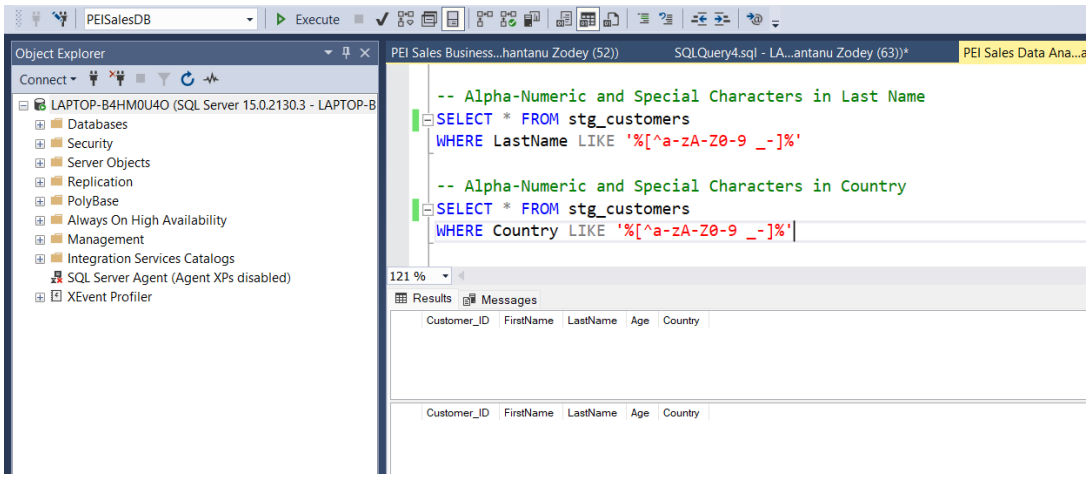
Accuracy Checks

Checking Alpha-Numeric and Special characters in First Name

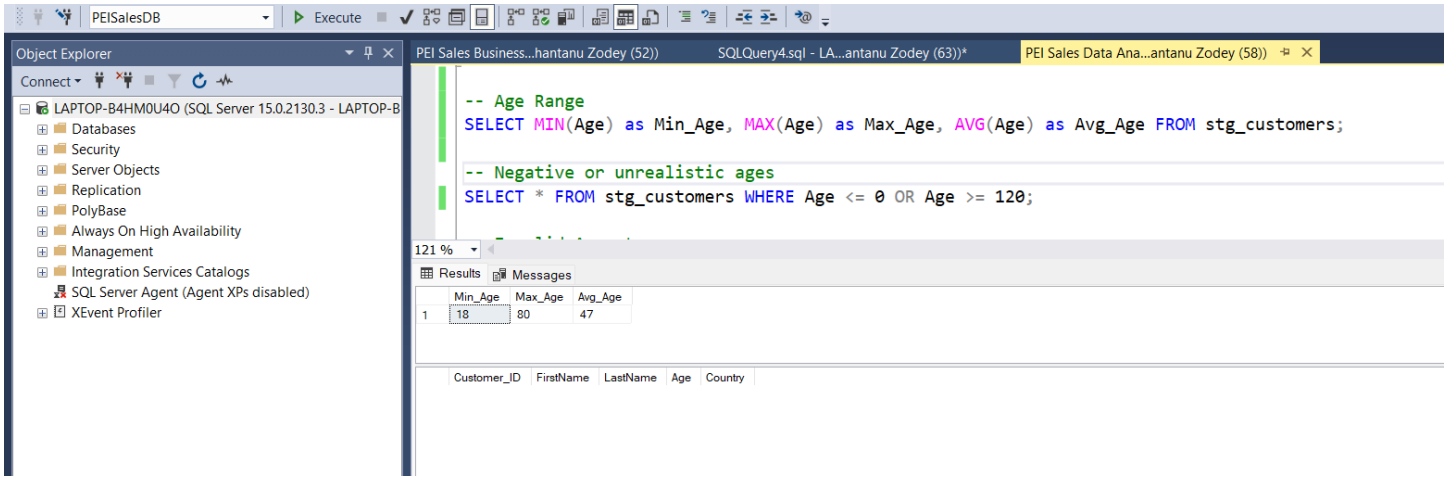


In the above screenshot, there are few records in First Name that contains special characters. This column needs transformation in order to get the clean and accurate data.

Checking Alpha-Numeric and Special characters in Last Name and Country

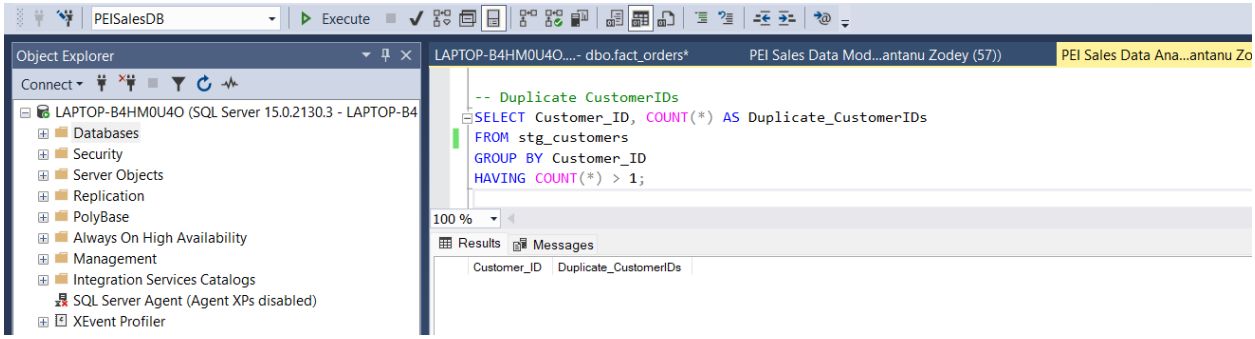


Checking Negative or Unrealistic Ages.



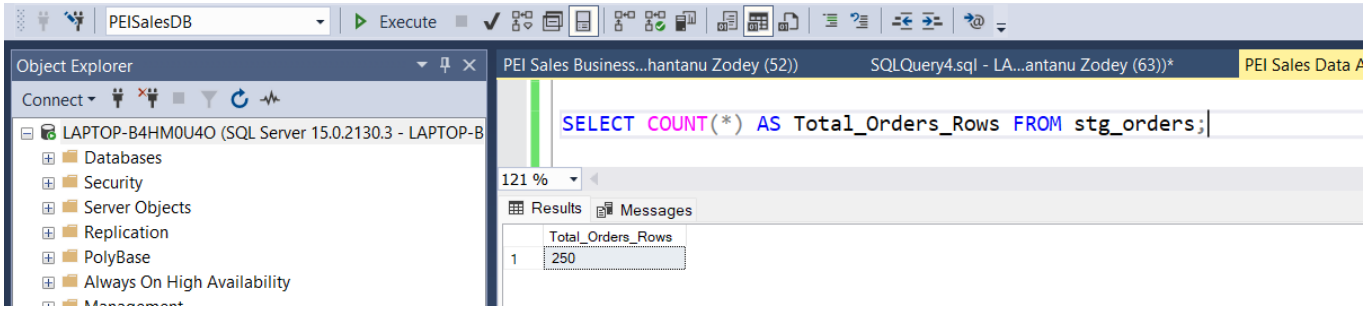
Reliability Checks

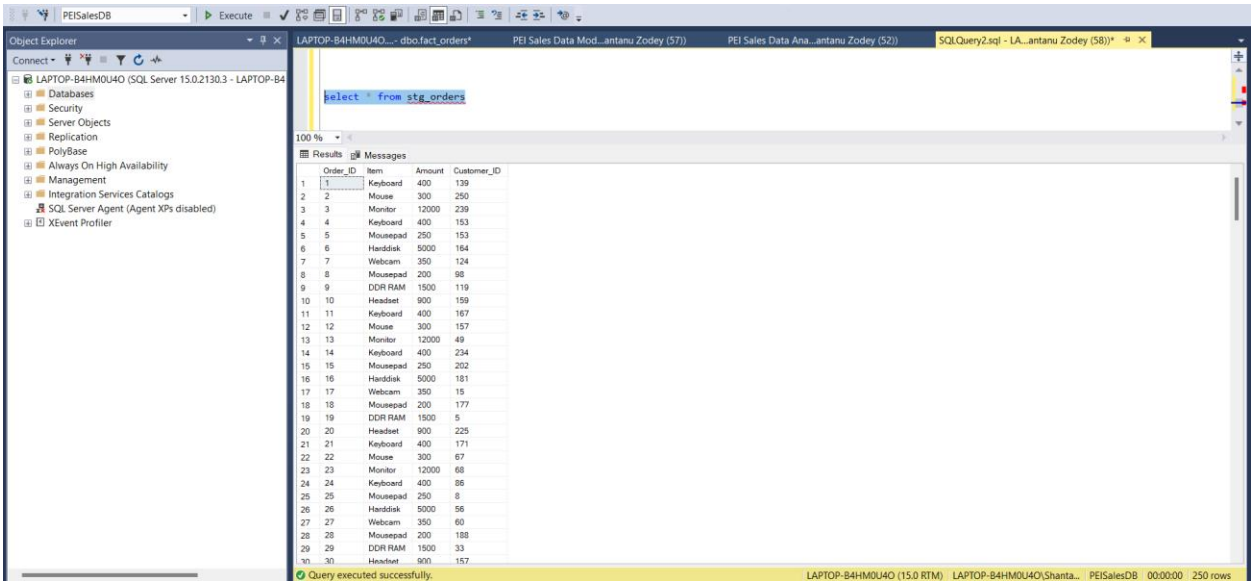
Checking Duplicate Customer IDs in stg_customer table



Orders Data Quality

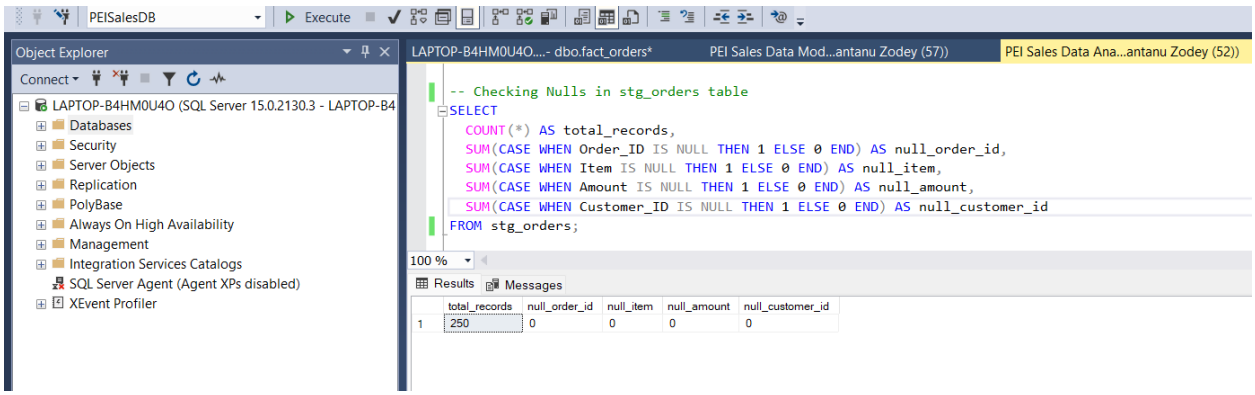
Validating row count and the records that are loaded into staging table from Customer.xlsx.



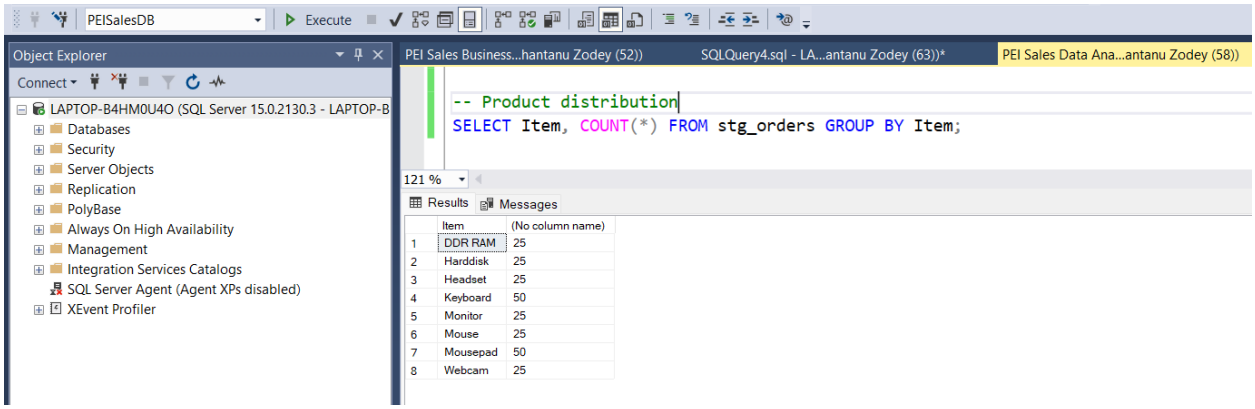


Completeness Checks

Checking Nulls in stg_orders table

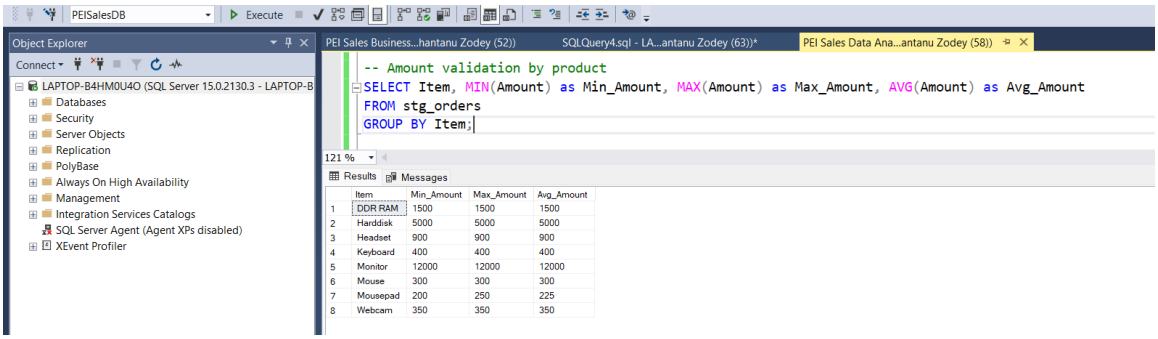


Checking Product Distribution

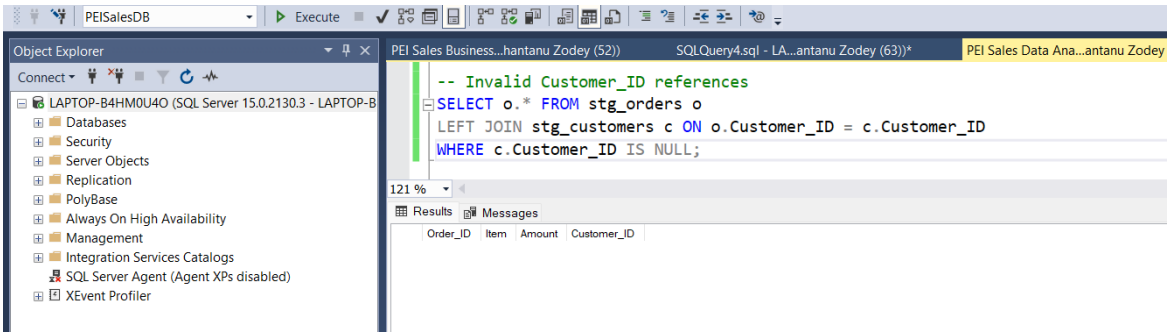


Accuracy Checks

Amount Validation

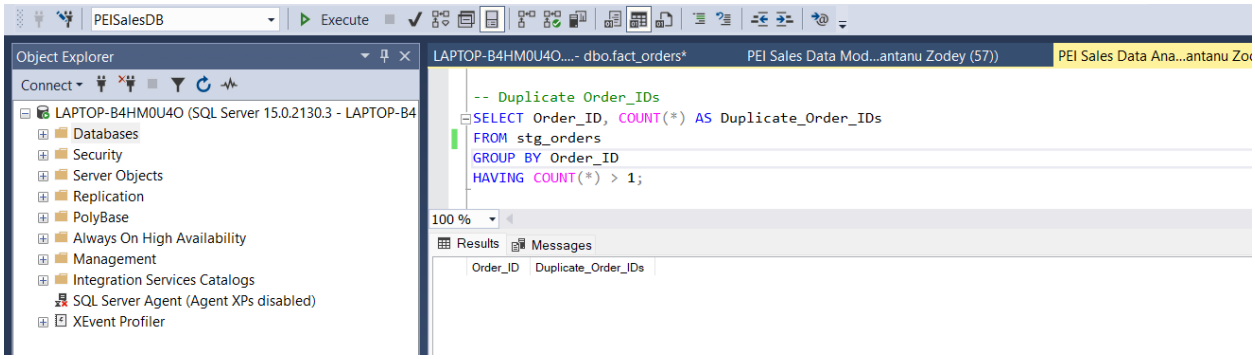


Invalid Customer_ID references



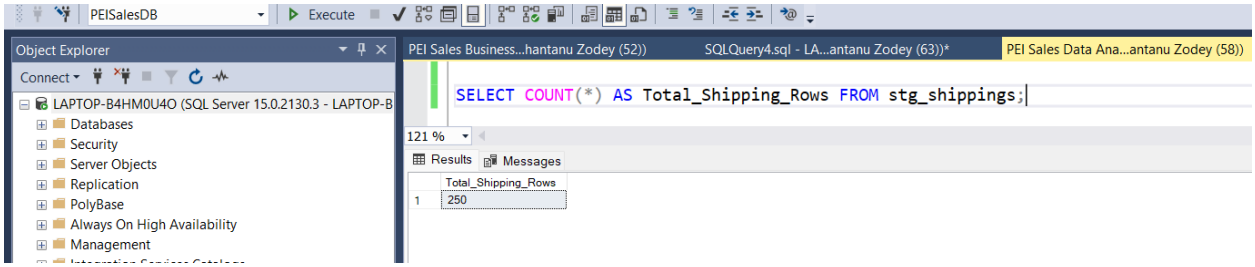
Reliability Checks

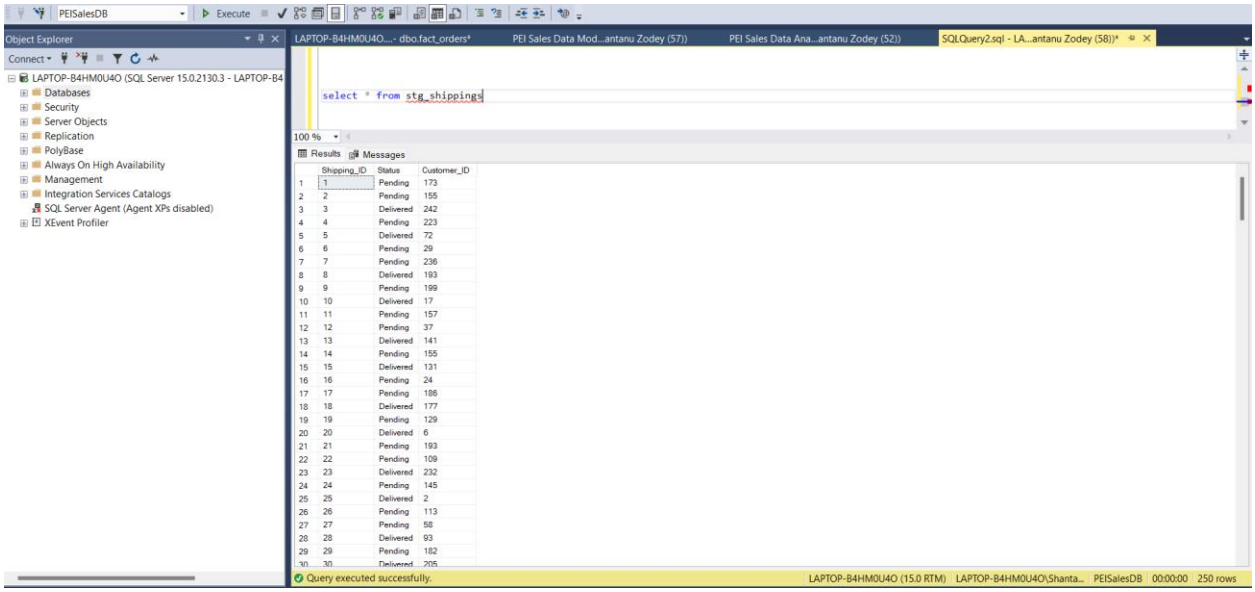
Checking Duplicate Order_IDs in stg_order table



Shipping Data Quality

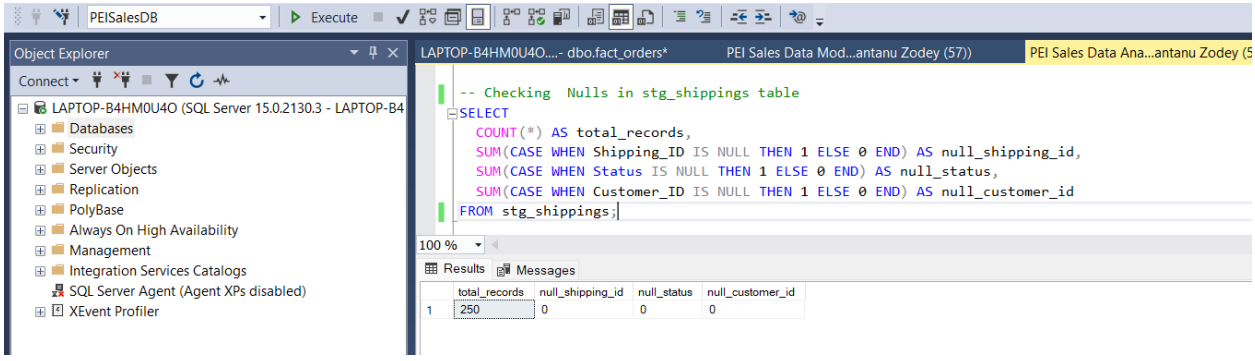
Validating row count and the records that are loaded into staging table from Customer.xlsx.



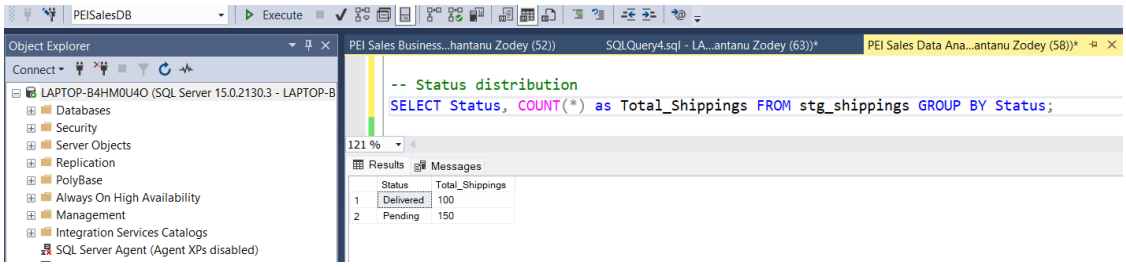


Completeness Checks

Checking Nulls in stg_shipping table

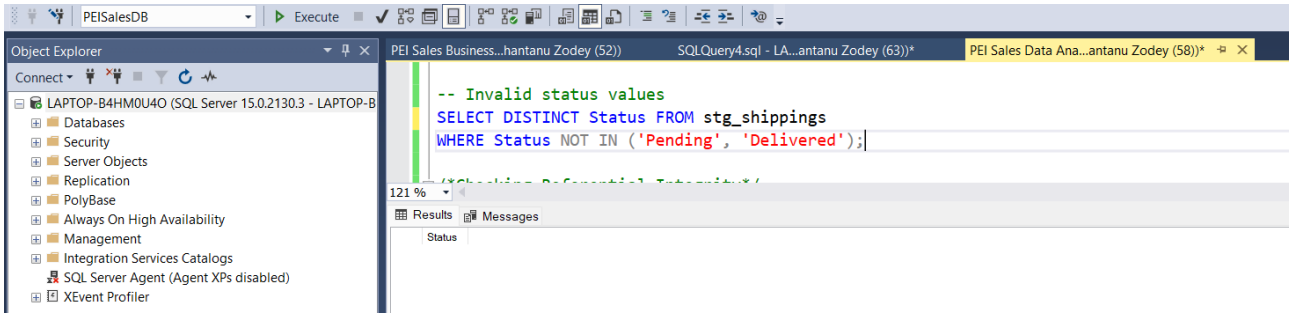


Shipping by Status Distribution



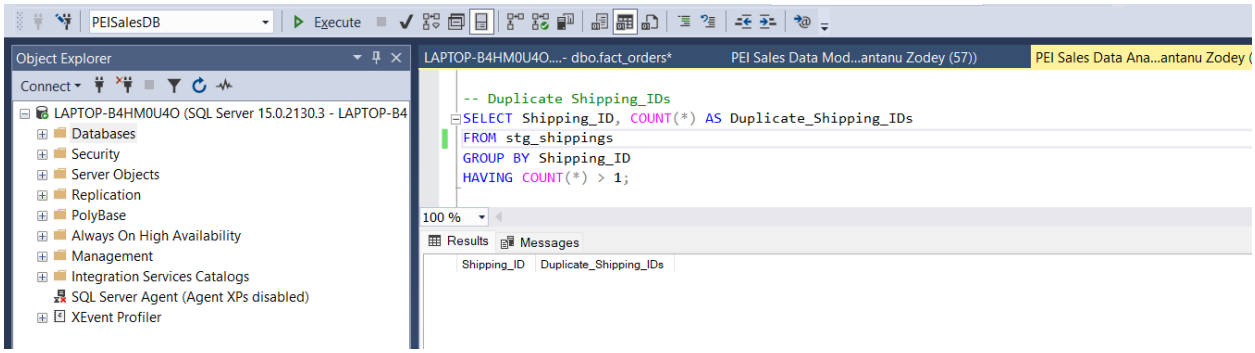
Accuracy Checks

Checking other status apart from Pending and Delivered

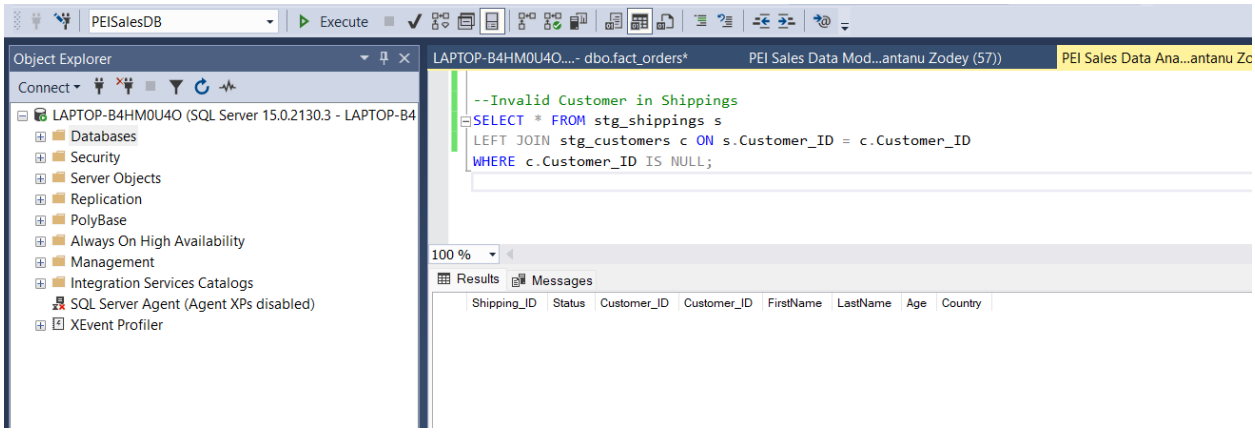


Reliability Checks

Checking Duplicate Order_IDs in stg_shipping table



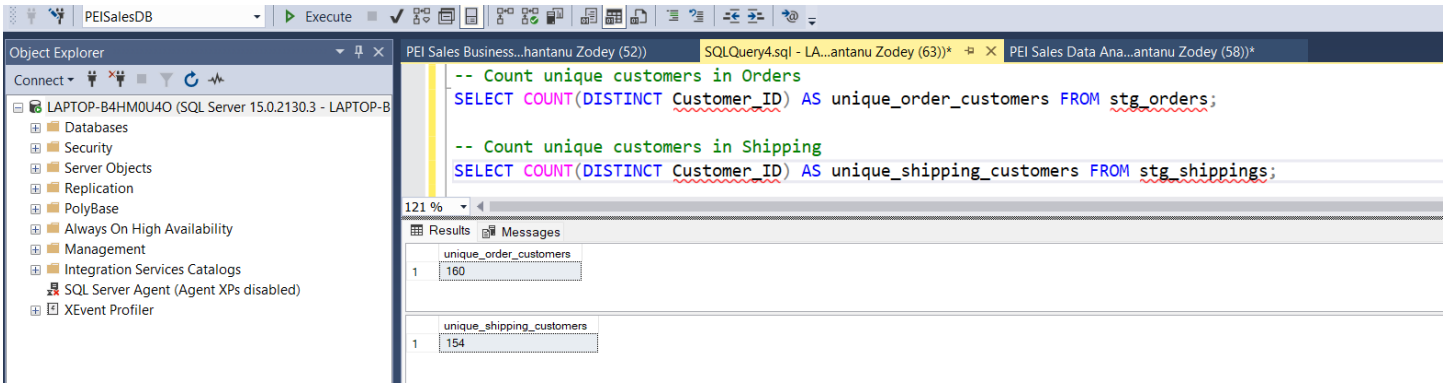
Invalid customers present in stg_shippings



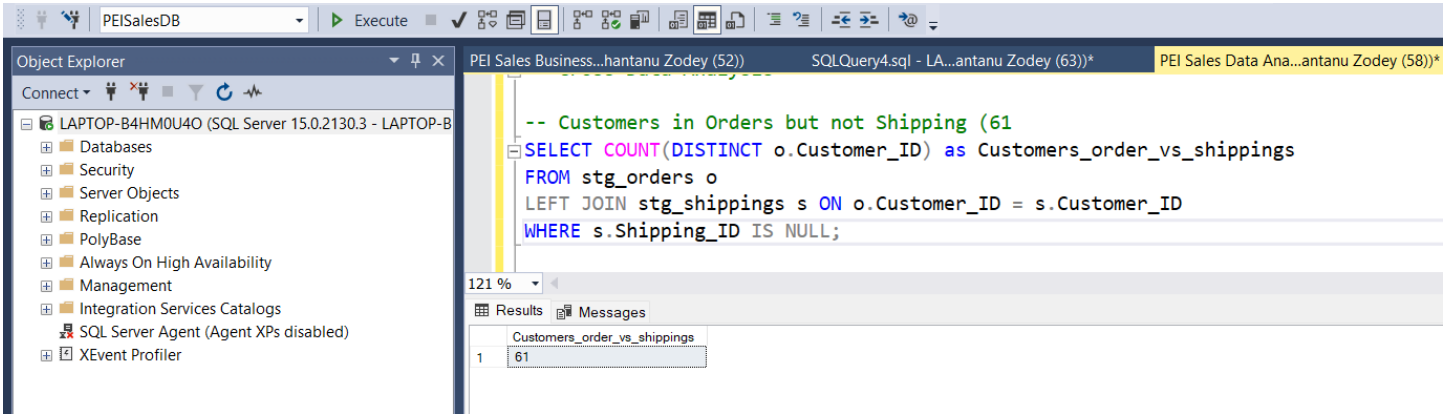
Cross-Dataset Integrity Analysis

Orders vs Shipping

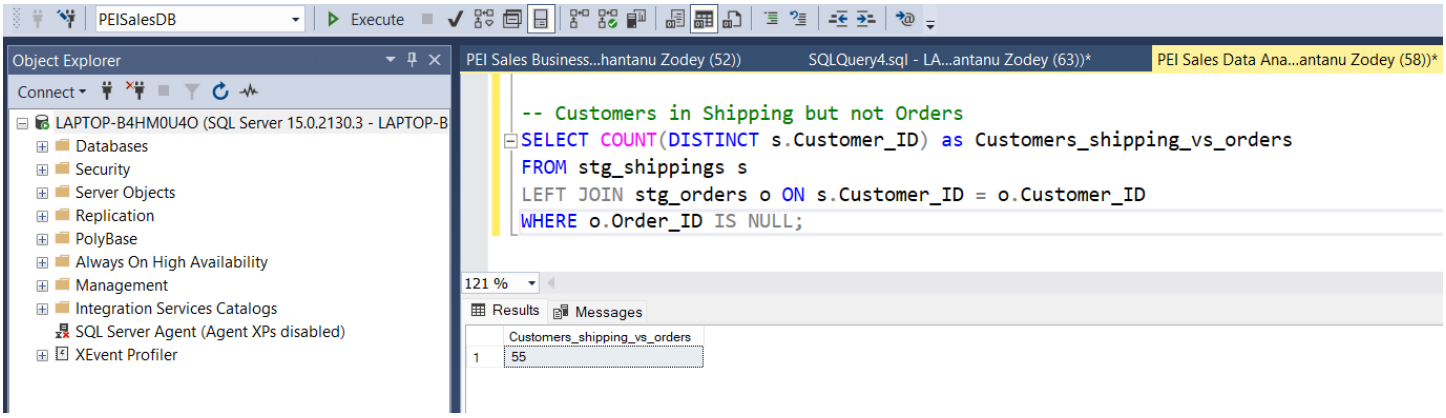
Unique CustomerIDs in Orders and Shippings table



Customer_IDs that are present in Orders but not in Shipping tables



Customer_IDs that are present in Shipping but not in Order tables

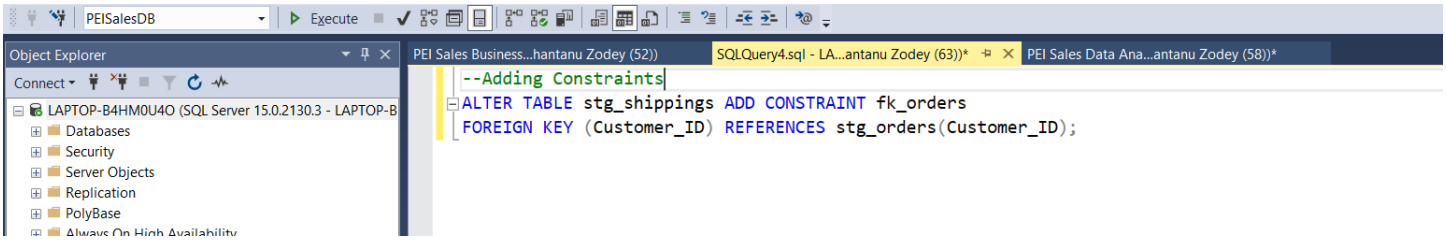


Key Findings

1. **Orders Analysis**:
    - Found 160 unique Customer_ID values in stg_orders
    - When compared with stg_shippings, 61 of these didn't exist in shipping records
    - These represent orders that were placed but never shipped
2. **Shipping Analysis**:
    - Found 154 unique Customer_ID values in stg_shippings
    - When compared with stg_orders, 55 had shipping records but no orders
    - These are problematic as shipping should always reference an order

Recommendations

Preventive Measure to tackle this issue:



Implementing FOREIGN KEY constraint between Orders and Shipping tables will help prevent invalid records.