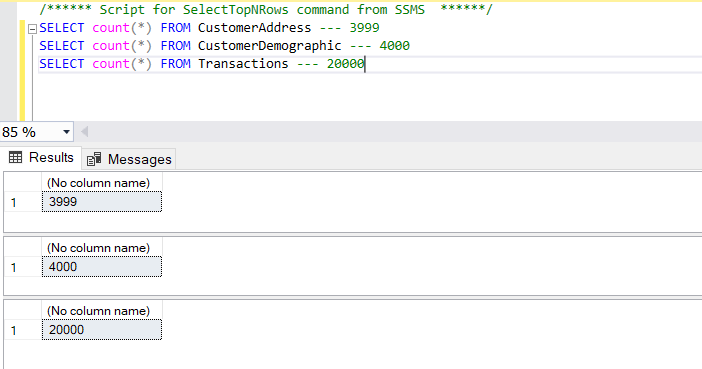
**Sprocket Central Data Profiling**

The data is loaded into SQL Server using a Talend Open Studio job

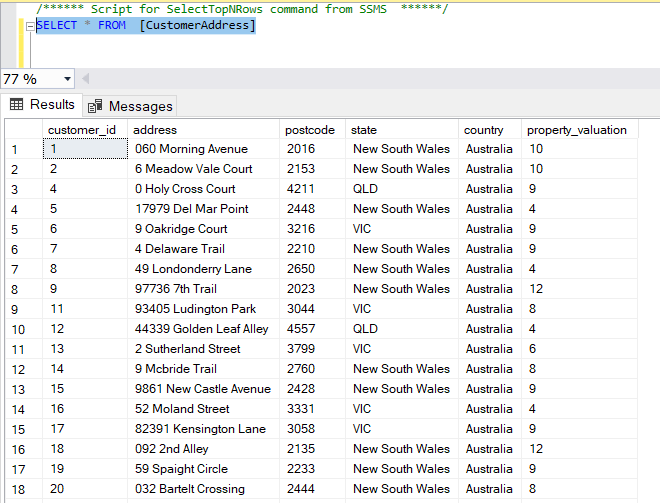
Since we are interested in the Data Quality Check lets consider it as a Staging Area, for now please ignore any referential integrity on the tables as it is a straight dump from the Excel Sheet.

**Total Counts of the Tables:**

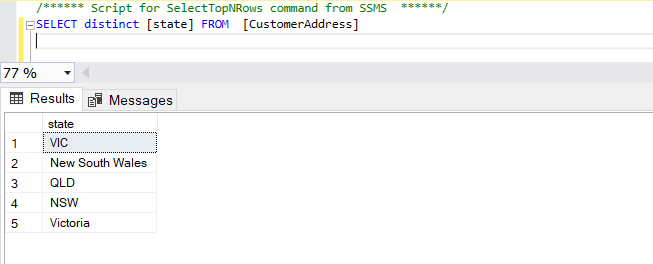
The following are the row counts



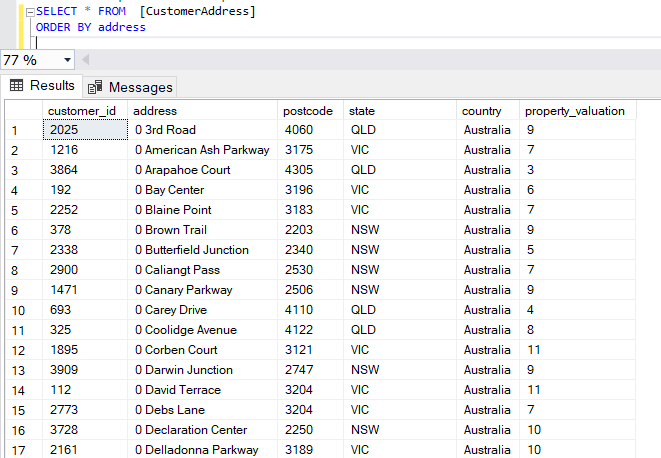
1. **CustomerAddress Table Profiling:**

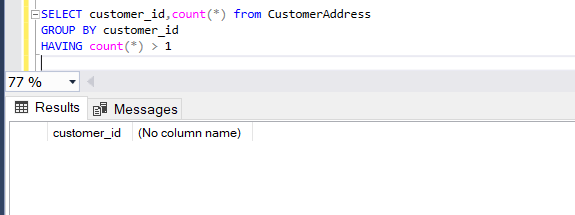
CustomerAddress first look**:** looks like the datatypes are consistent in the table

State column values are inconsistent:

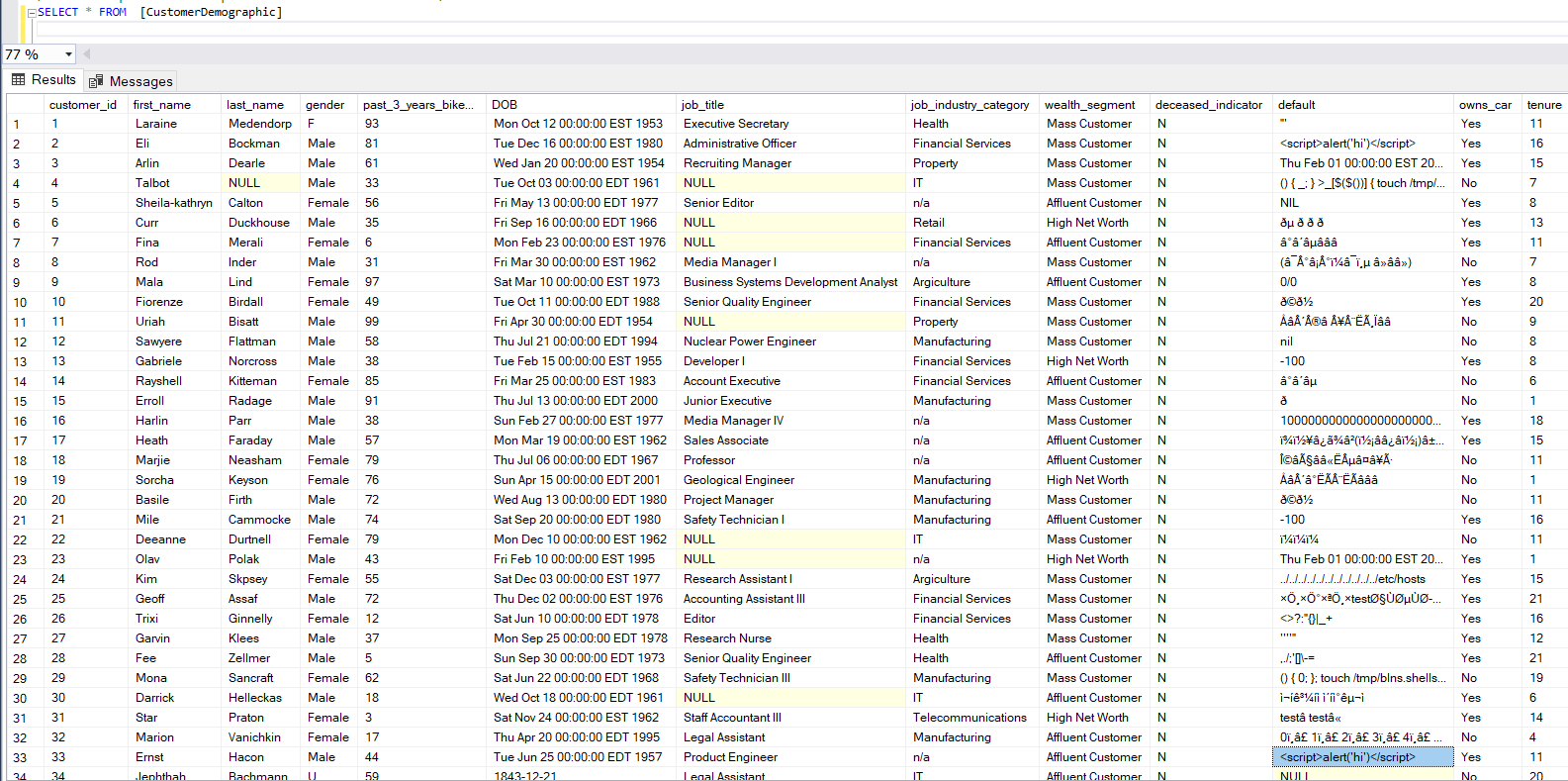


Looks like we need to trim the leading zeros from address column after:

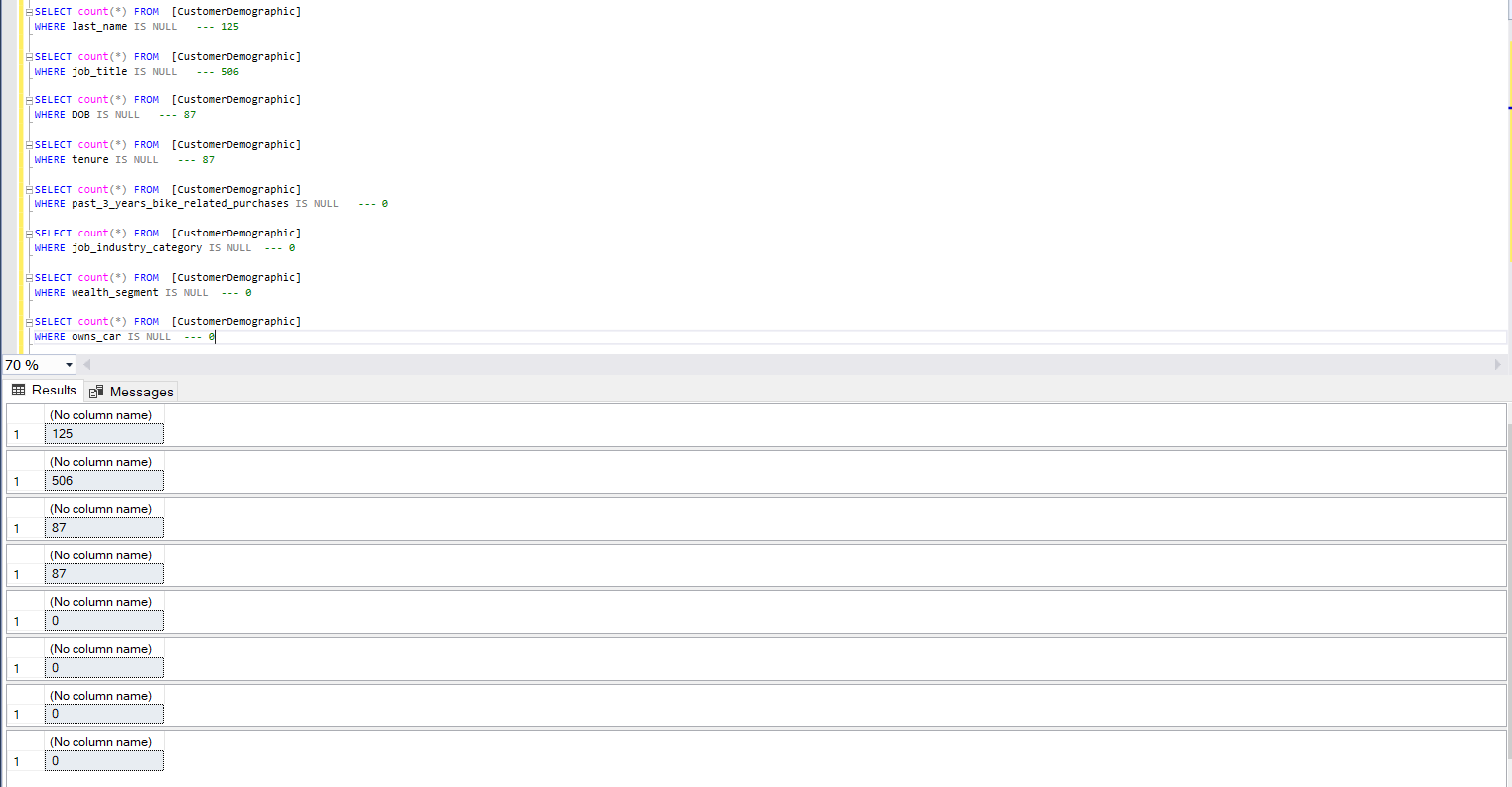


No duplicate customers are present:

1. **CustomerDemographic Table Profiling**:

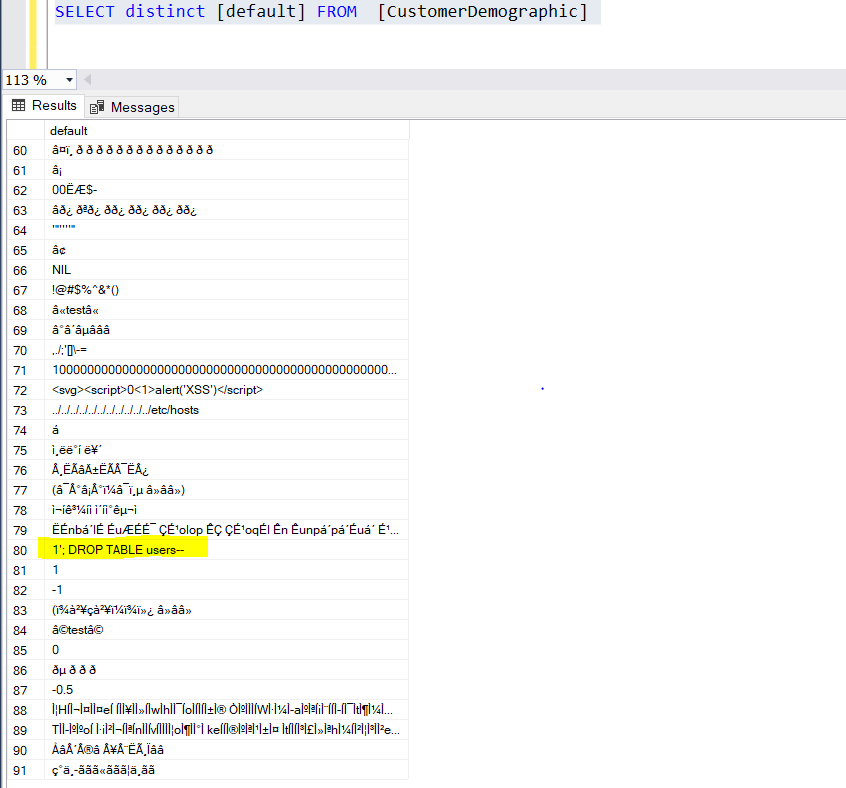
One can see there few NULL values present

Let’s check if there are any NULL values present for important columns and their count:

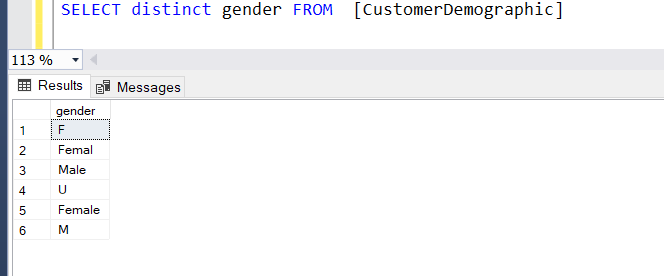


Looks like few columns are having NULL values which must be substituted or remove via correct measures.

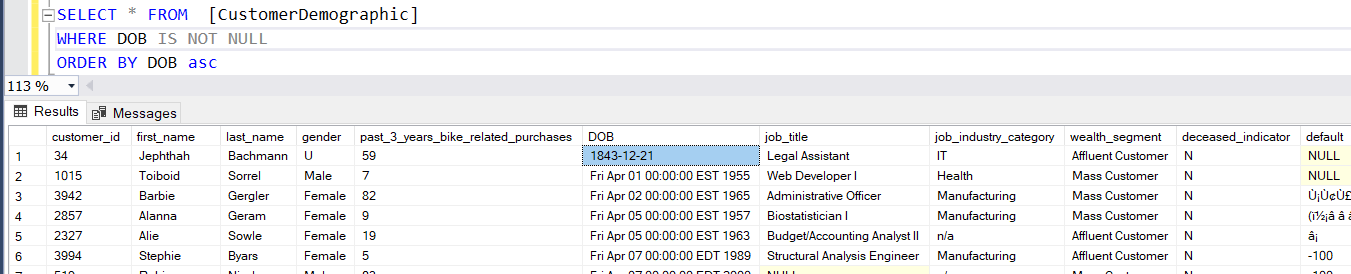
After looking at the data values one will doubt if we need this column anymore.

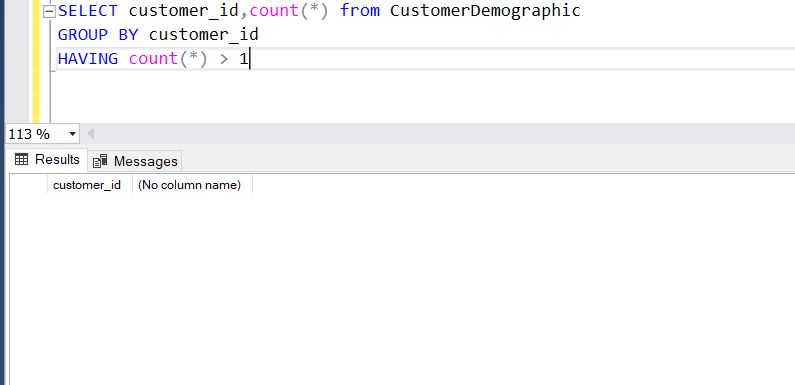
Also was someone trying DROP a table. Looks like a SQL injection, we should validate the front end properly or remove this column if not required:

Gender column has inconsistent data:



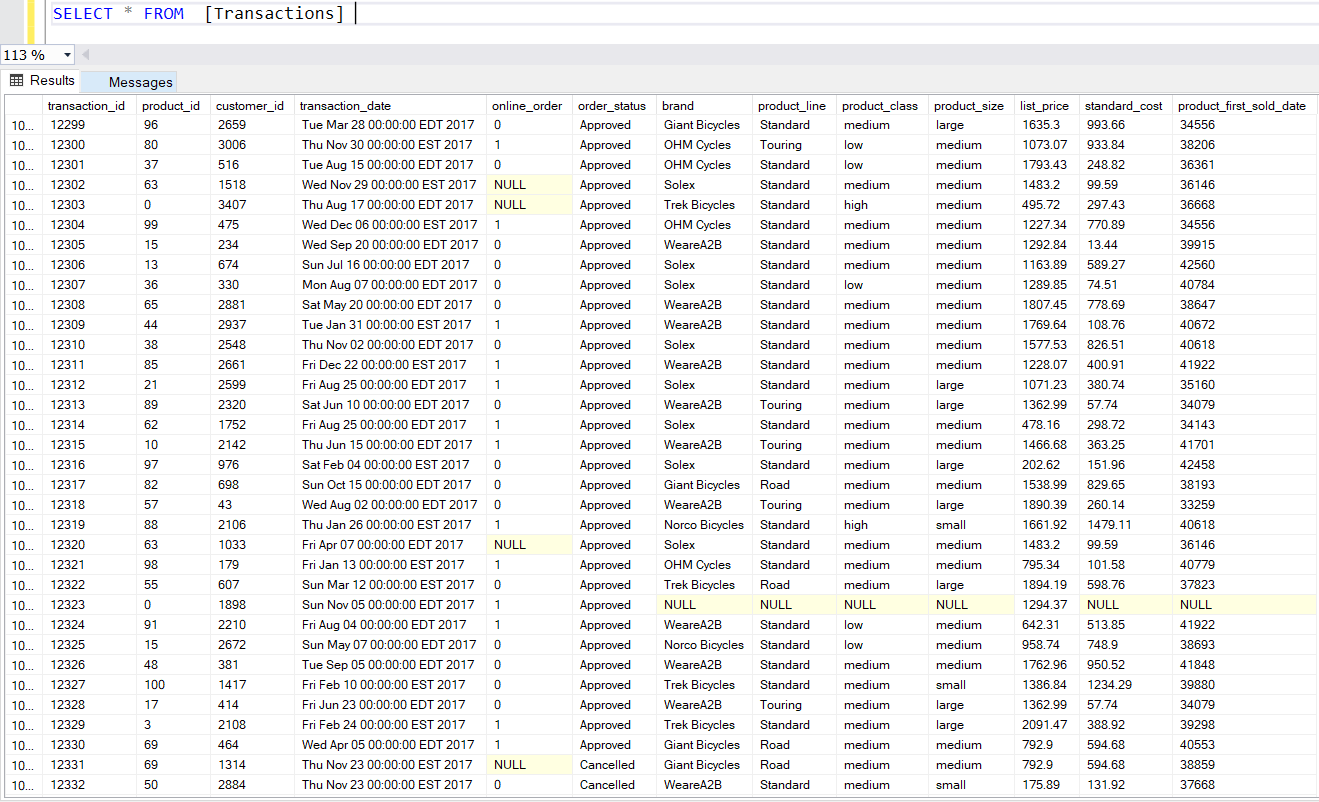
DOB format issue or typo:



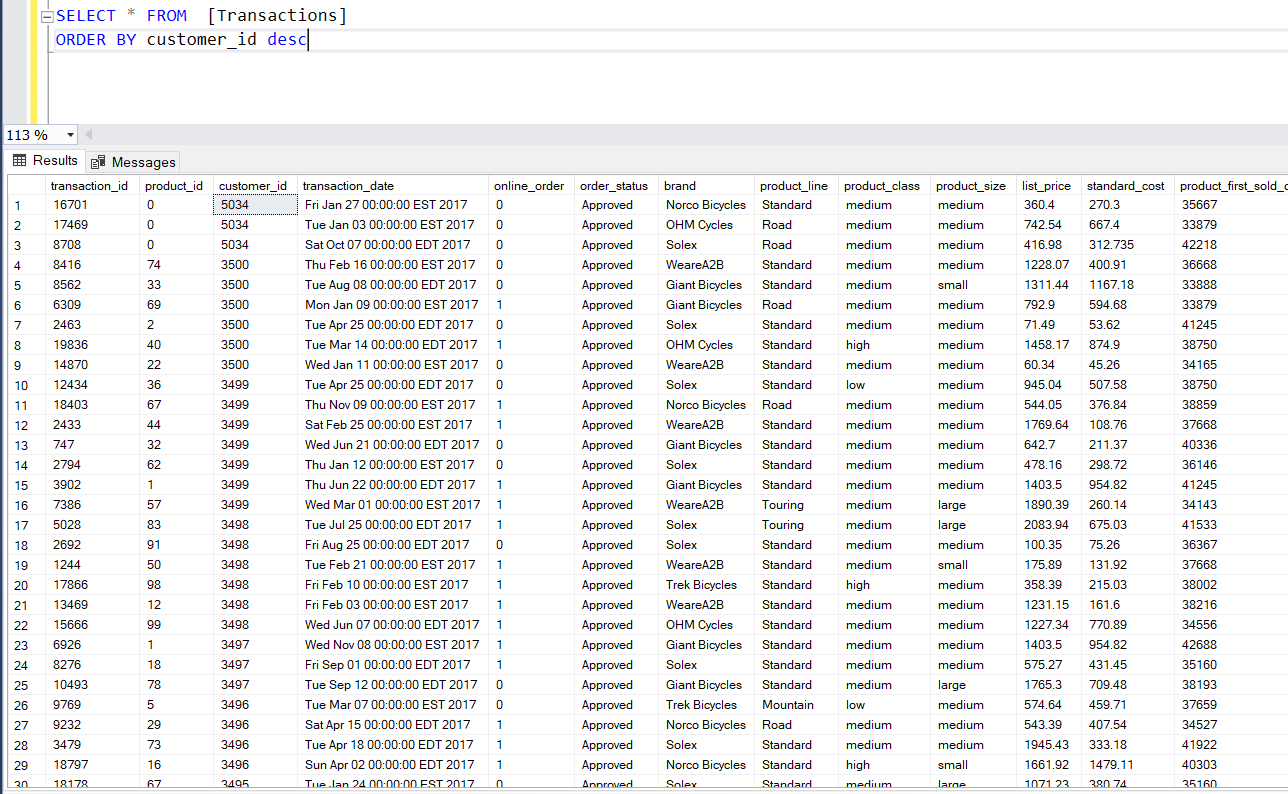
No duplicate customers in CustomerDemographic table:

1. **Transactions Table Profiling:**

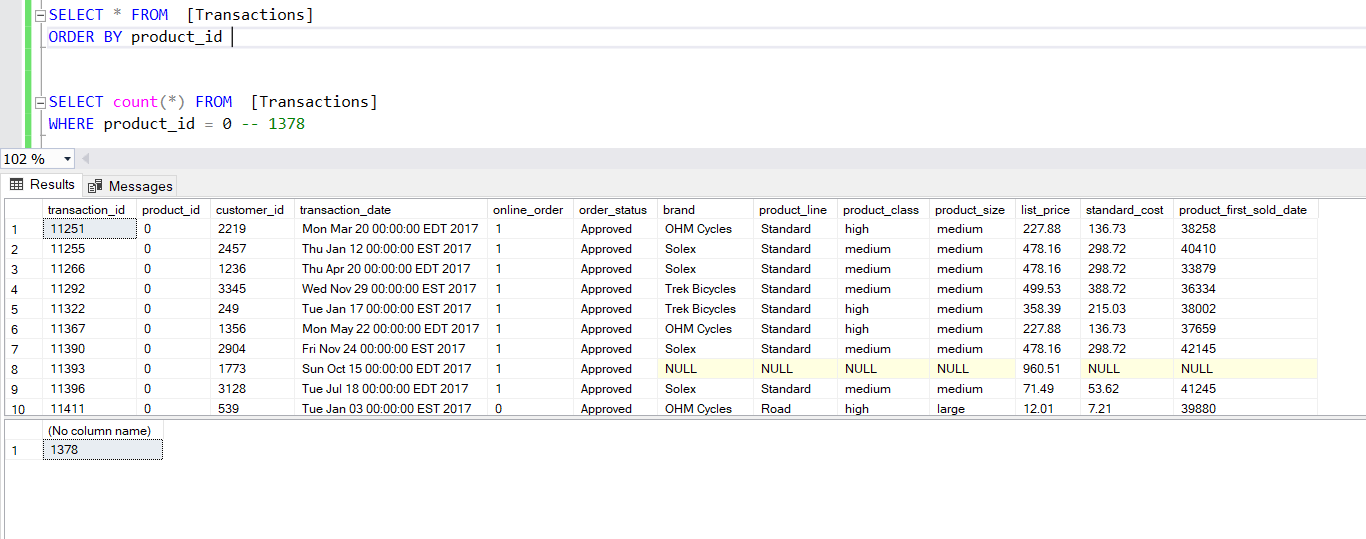
Looks like few columns are having missing values and columns: brand, product\_line, product\_class, standard\_cost, product\_first\_sold\_date are inter-related:



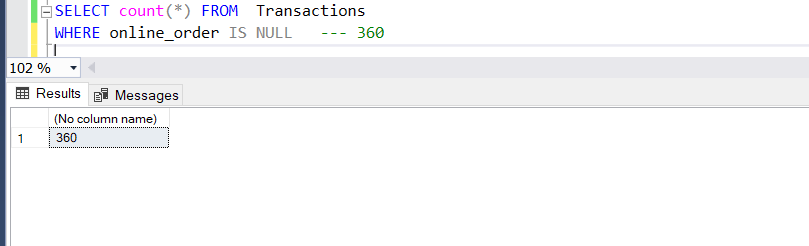
Looks like there are few outliers in the tables with respect to customer\_id and invalid product\_id:



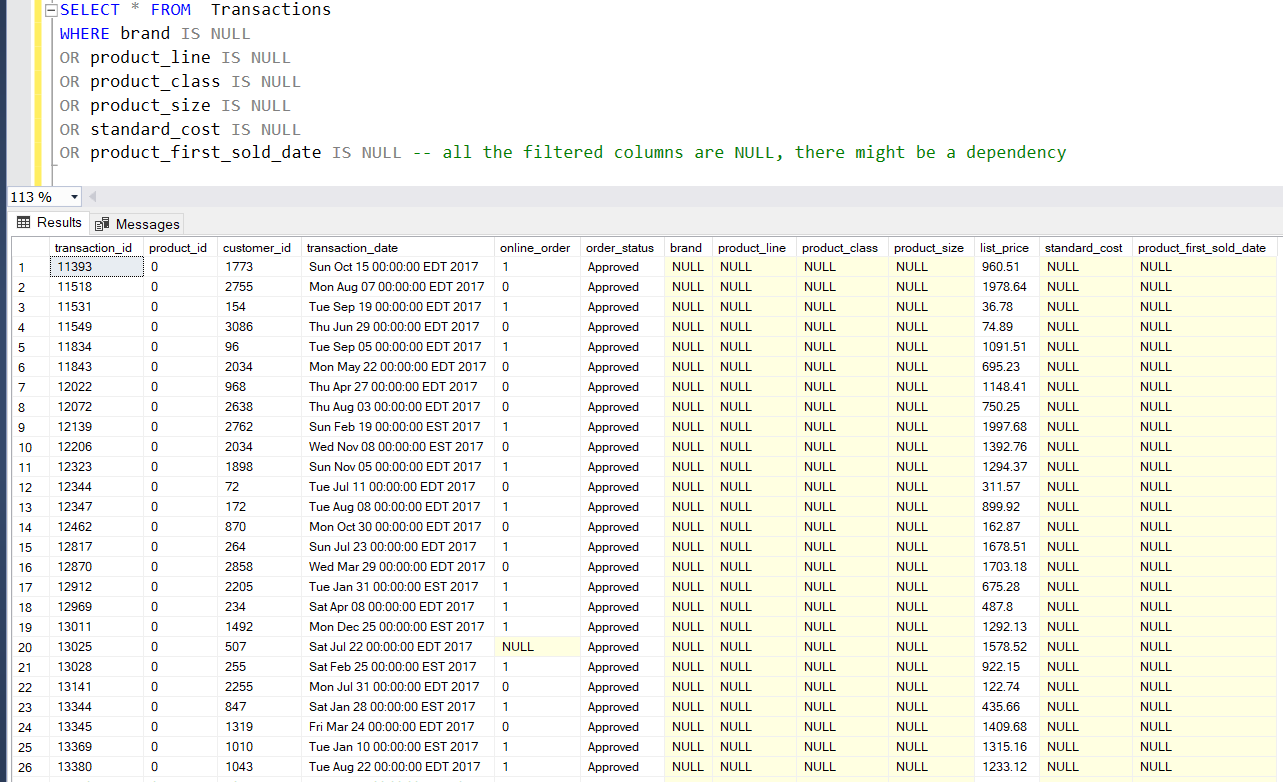
There are 1378 rows with product\_id = 0:



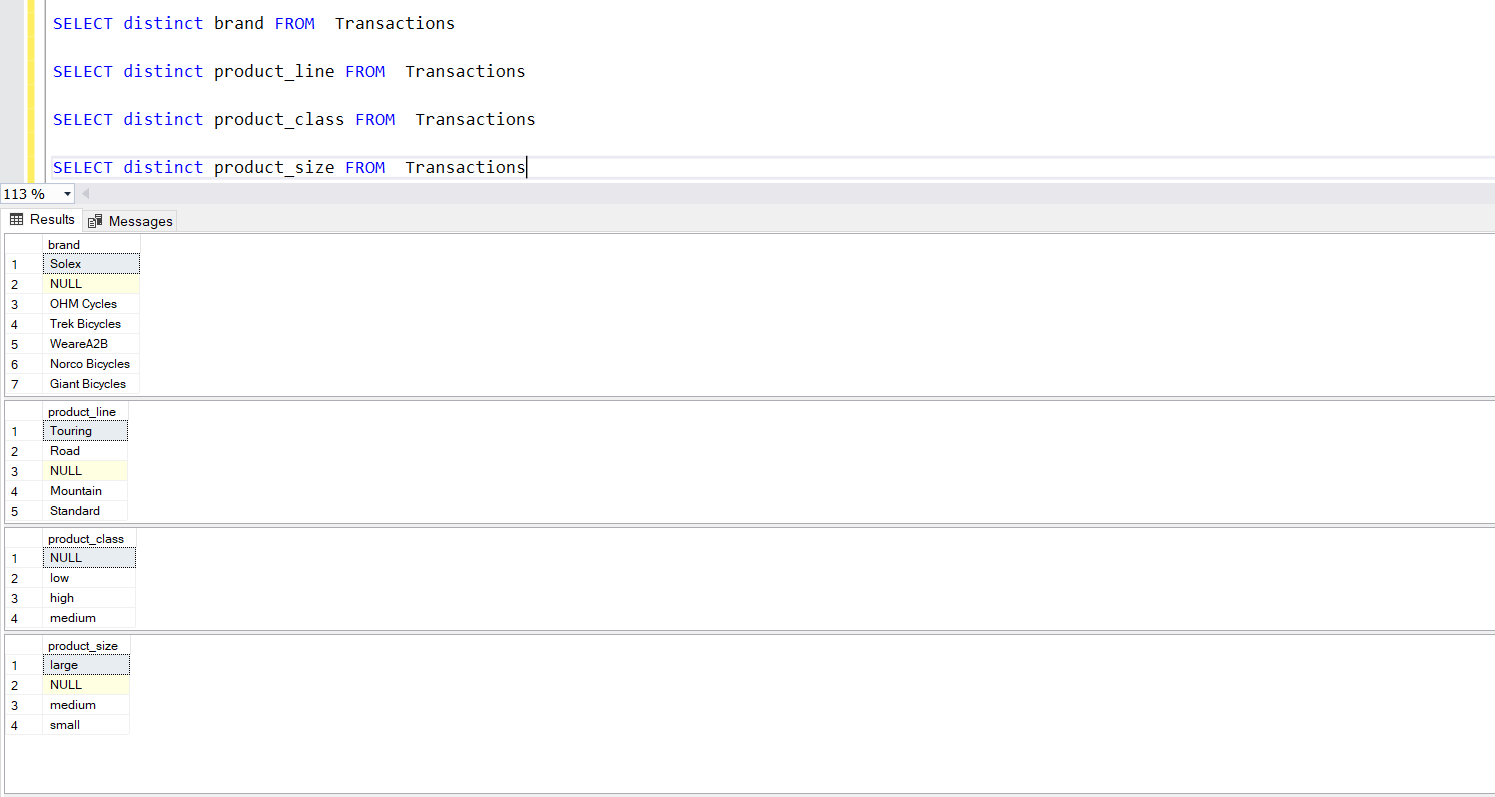
There 360 rows with NULL online\_order:



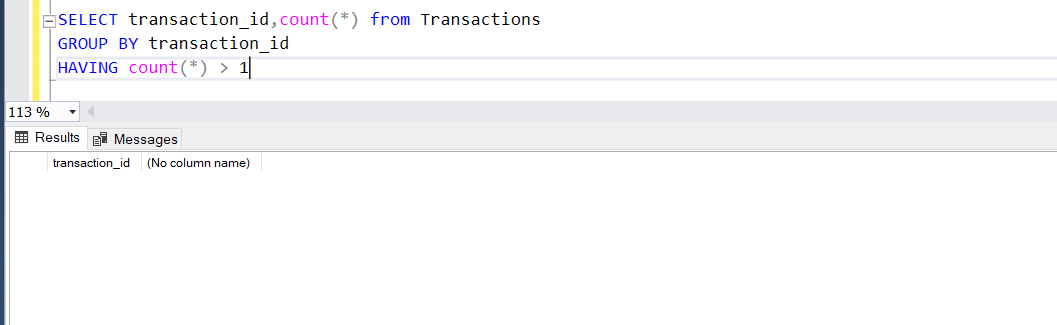
All the values are NULL for the following columns:



Looks like the distinct values of the following columns are consistent except for NULL values:

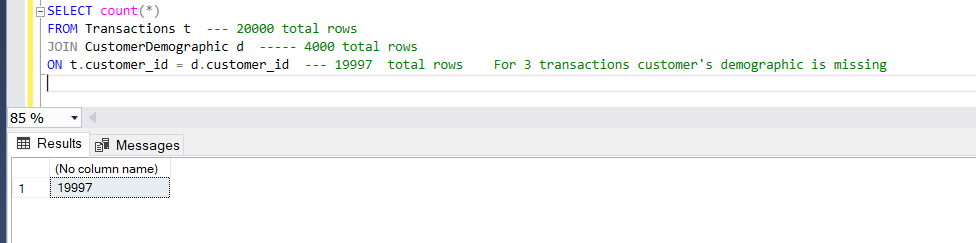


No duplicate customers in the Transactions table:

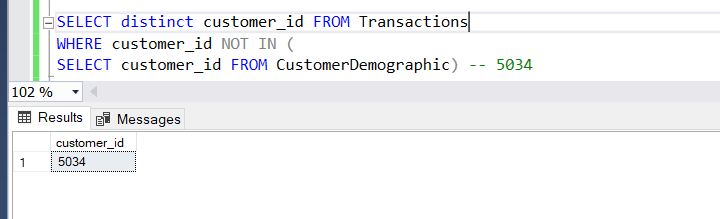


1. **JOINING the tables and validating the data:**

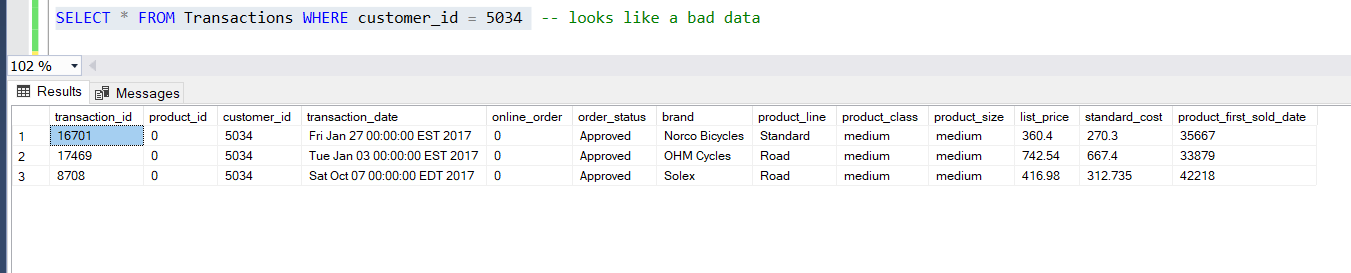
Joining Transactions with CustomerDemographic: 3 rows are not joined



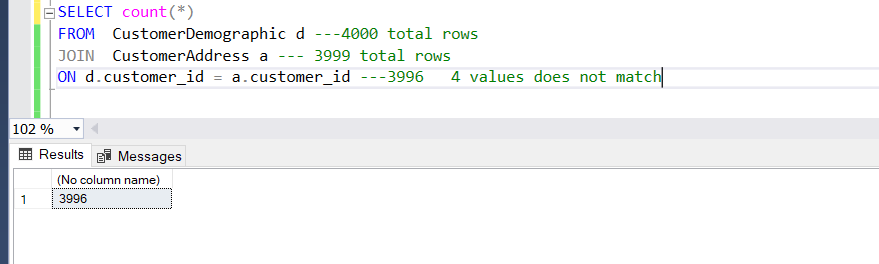
Customer number 5034 is missing:



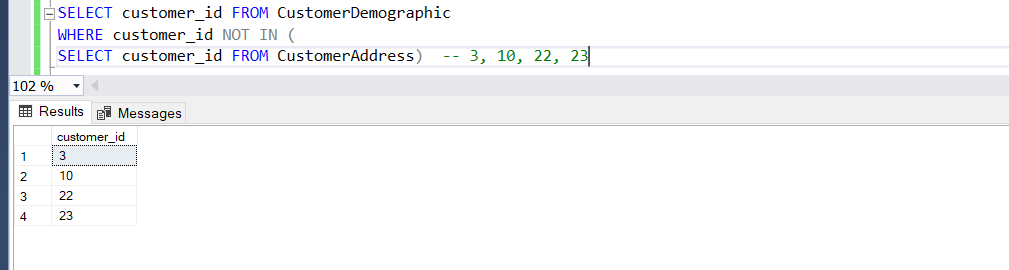
Let’s find out those 3 missing records: looks like bad data with product\_id as 0



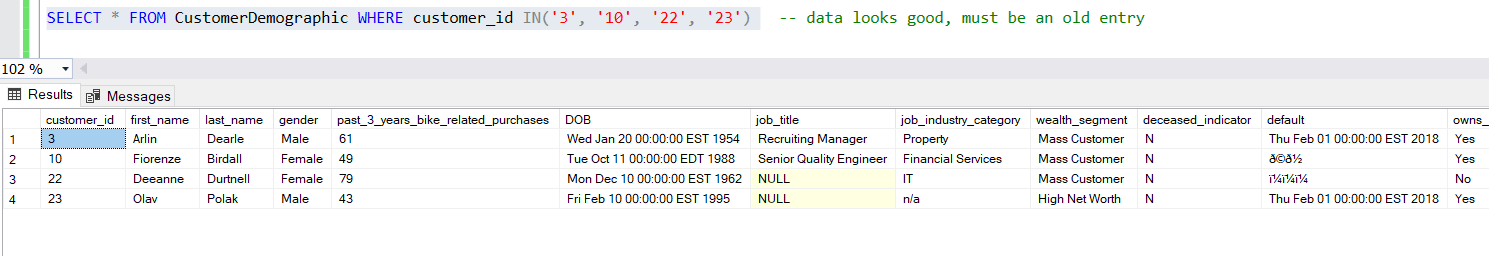
Joining Demographics with Address: 4 values are missing



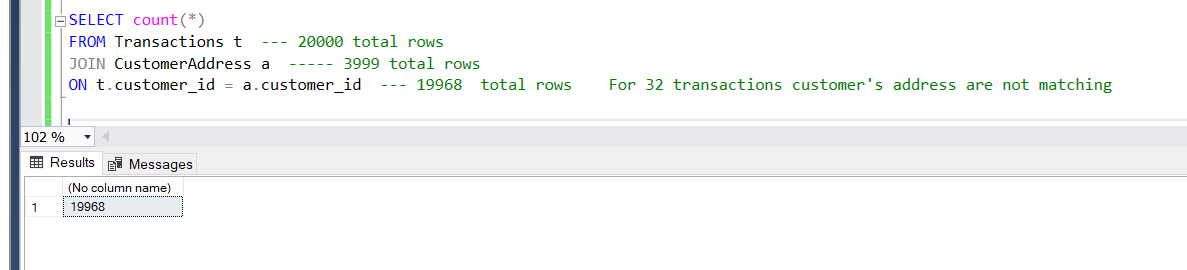
The following are the 4-missing customer\_ids:



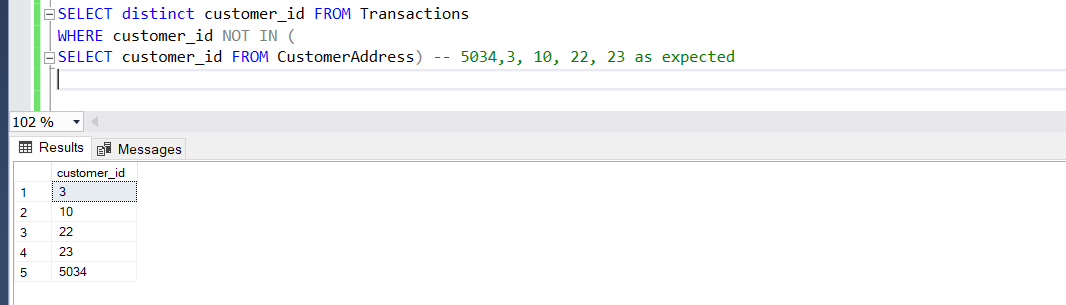
Let’s look at the above missing customers:



Joining Transactions with Address:



As expected, 5 customer ids are missing:



1. **CREATED a View for Analysis:**

CREATE VIEW Sprocket\_Analysis AS

SELECT /\*\* all the columns \*\*/

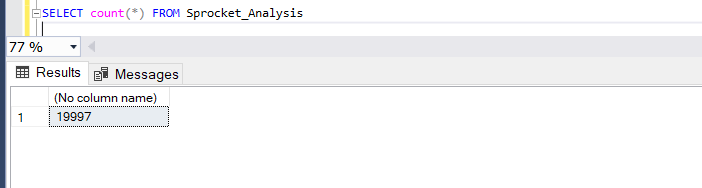
FROM Transactions t --- 20000 total rows

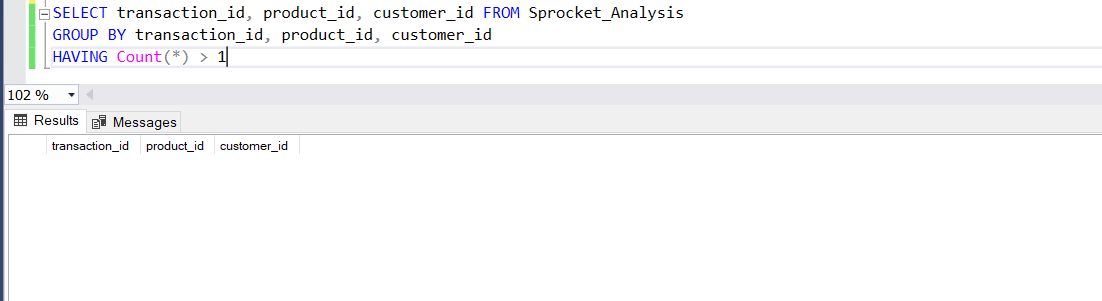
JOIN CustomerDemographic d ----- 4000 total rows

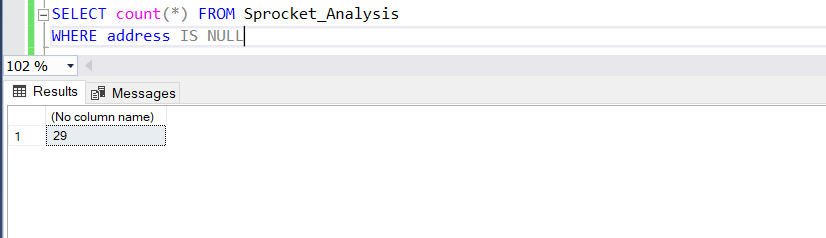
ON t.customer\_id = d.customer\_id

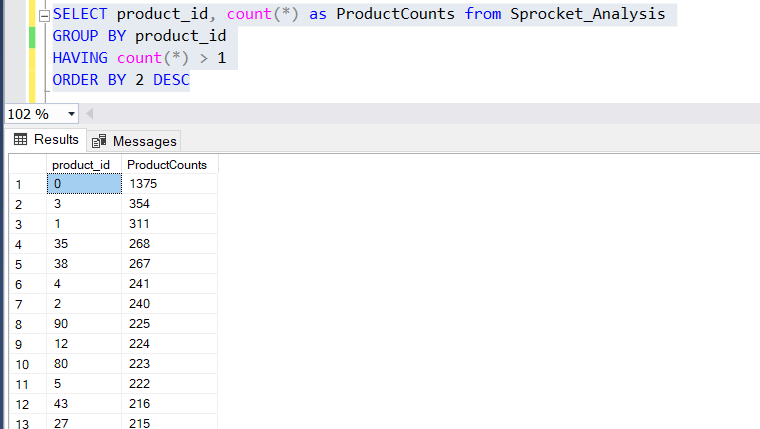
LEFT JOIN CustomerAddress a --- 3999 total rows

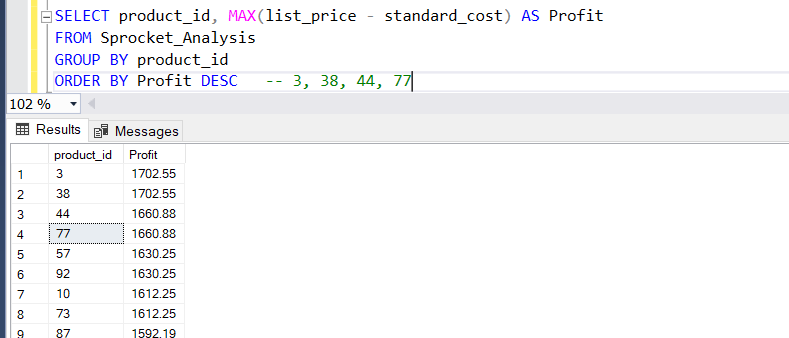
ON d.customer\_id = a.customer\_id ------------ 19997

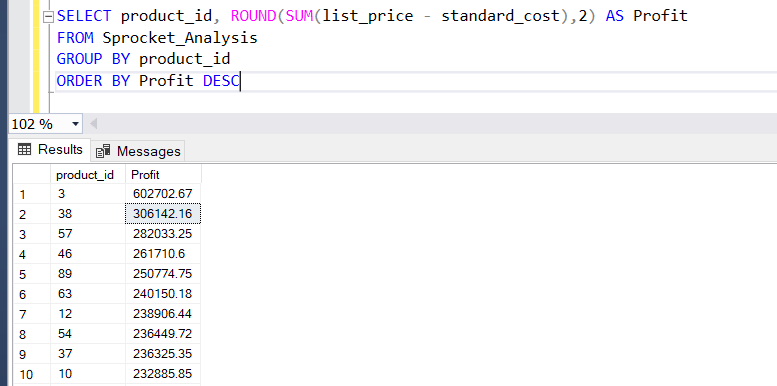
Total Count: 19997 (since we have excluded the customer\_id 5034 which made no sense with product\_id as 0)

Checking duplicates:

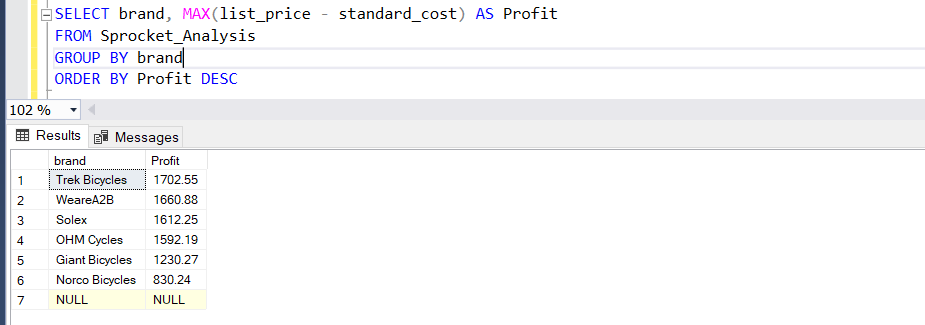
After joining all three tables there are 29 customers with missing Address information:

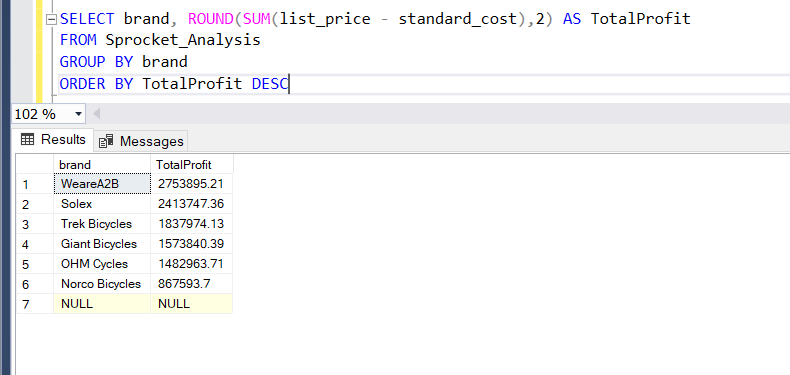
Transactions by product counts: top 3 products are 3, 1, 35

Products by MAX profit gained:

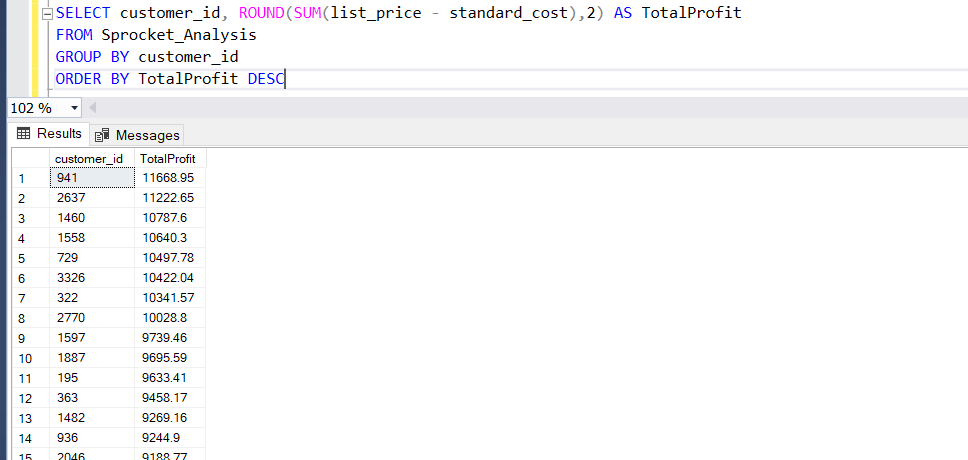
Products by Total Profit gained:

One cannot understand what this product is, lets group by brand and check the profitable brand: Brand by Max Profit:



Brand by Total Profit:

Customer by Total Profit:



CONCLUSION:

We have some data quality issue with the dataset, few values are inconsistent, there are missing values across the dataset which needs to be handle and there are few irrelevant data.

The following is the Excel containing column wise detail quality assessment and possible solutions:

