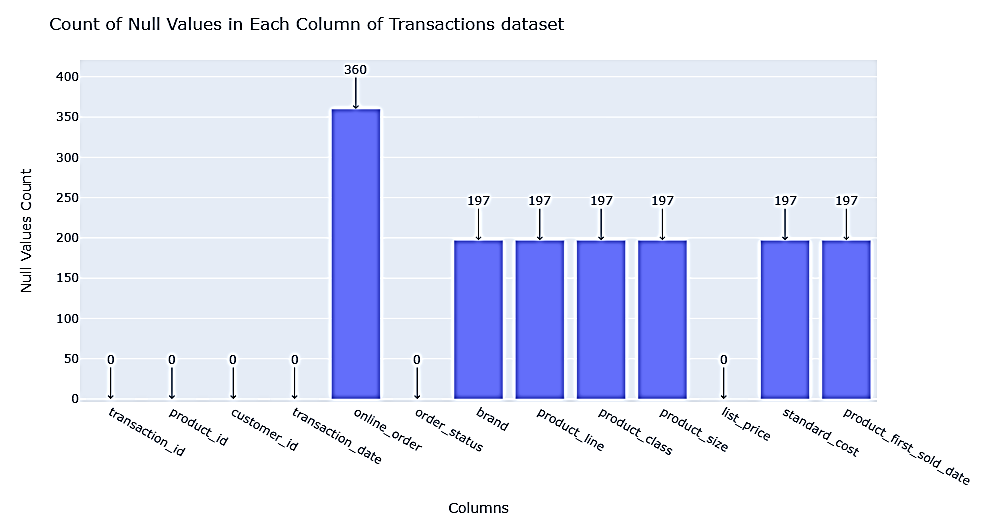
To Manager,

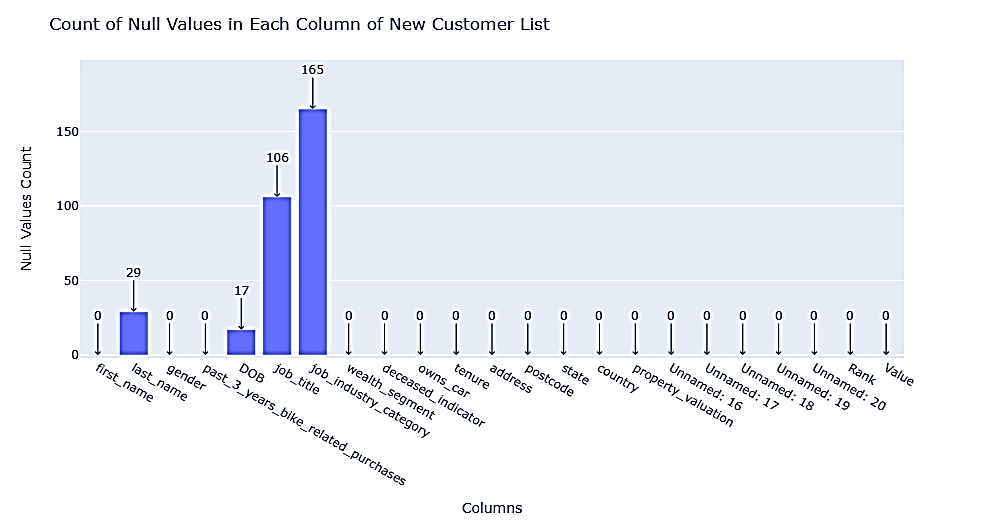
KPMG Data Analytics Team

As per Data Quality Assessment task performed by me on the provided datasets, I am here with the insights from it.



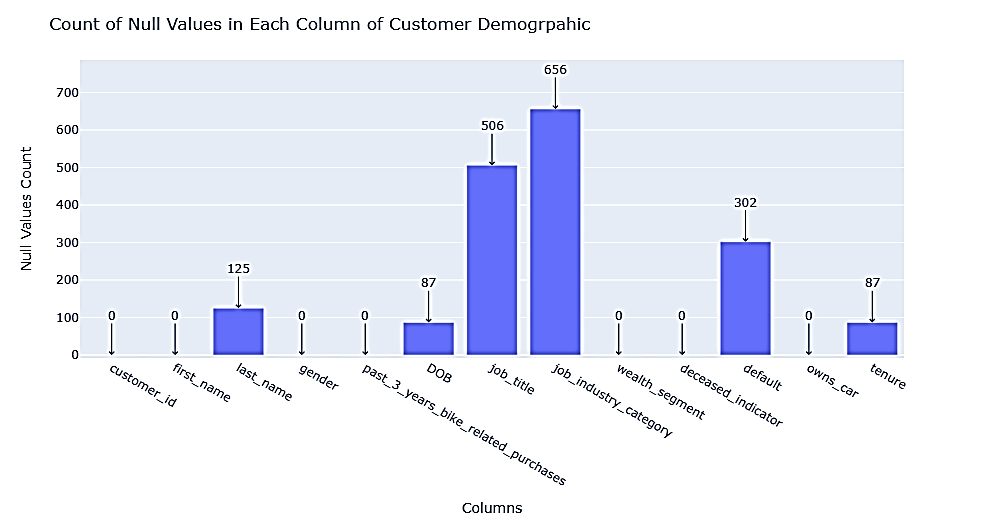
In the Transactions Dataset: -

* **Columns of Transactions Data that are unnecessary or they contain the wrong data from the analysis perspective are *product\_first\_sold\_date -> This column must contain data in datetime format but it has data in float type.***
* **There are 360 null values in the online\_order column of Transactions dataset.**
* **There are 197 null values in the brand column of Transactions dataset.**
* **There are 197 null values in the product\_line column of Transactions dataset.**
* **There are 197 null values in the product\_class column of Transactions dataset.**
* **There are 197 null values in the product\_size column of Transactions dataset.**
* **There are 197 null values in the standard\_cost column of Transactions dataset.**
* **There are 0 duplicate rows in the Transactions dataset.**
* **The date range is** 2017-01-01 00:00:00 **to** 2017-12-30 00:00:00
* **There are 20000 rows and 13 columns in the Transactions dataset.**



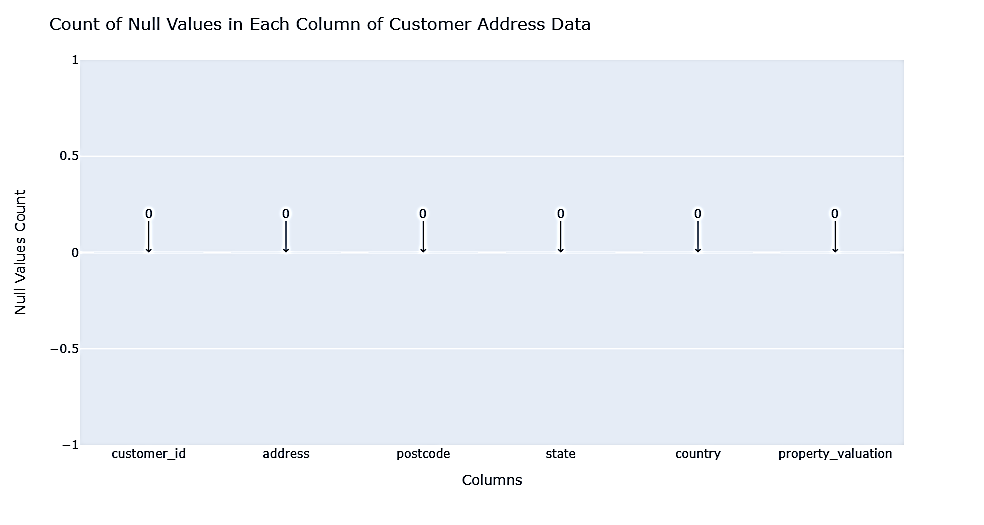
In the New Customer List Dataset: -

* **Columns of New Customer List Data that are unnecessary or they contain the wrong data from the analysis perspective are *Unnamed: 16, Unnamed: 17, Unnamed: 18, Unnamed: 19 and Unnamed: 20***
* **There are 29 Null values in last\_name column of the New Customer List Dataset.**
* **There are 17 Null values in DOB column of the New Customer List Dataset.**
* **There are 106 Null values in job\_title column of the New Customer List Dataset.**
* **There are 165 Null values in job\_industry\_category column of the New Customer List Dataset.**
* **There are 0 duplicate rows in the New Customer List dataset.**
* **The date range is** 1938-06-08 00:00:00 **to** 2002-02-27 00:00:00
* **There are 1000 rows and 23 columns in the New Customer List Dataset.**

******

In the Customer Demographic Dataset: -

* **There are 4000 rows and 13 columns**
* **The date range is** 1843-12-21 00:00:00 **to** 2002-03-11 00:00:00
* **Columns of Customer Demographic Data that are unnecessary or they contain the wrong data from the analysis perspective are *Default -> This column has nothing to do with analysis part. It’s just some random column containing some random values from the server***
* **There are 125 null values in the last\_name column of the Customer Demographic Dataset.**
* **There are 87 null values in the DOB column of the Customer Demographic Dataset.**
* **There are 506 null values in the job\_title column of the Customer Demographic Dataset.**
* **There are 656 null values in the job\_industry\_category column of the Customer Demographic Dataset.**
* **There are 302 null values in the default column of the Customer Demographic Dataset.**
* **There are 87 null values in the tenure column of the Customer Demographic Dataset.**
* **There are 0 duplicate rows in the Customer Demographic Dataset.**



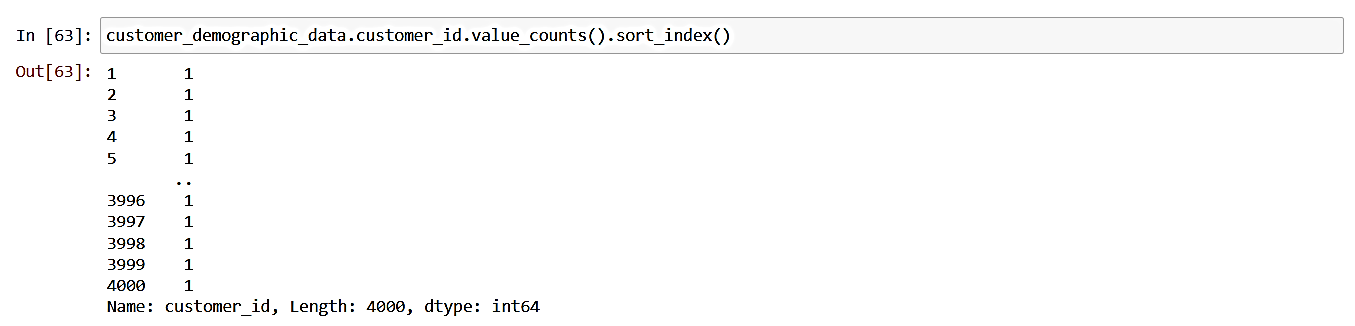
In the Customer Address Data: -

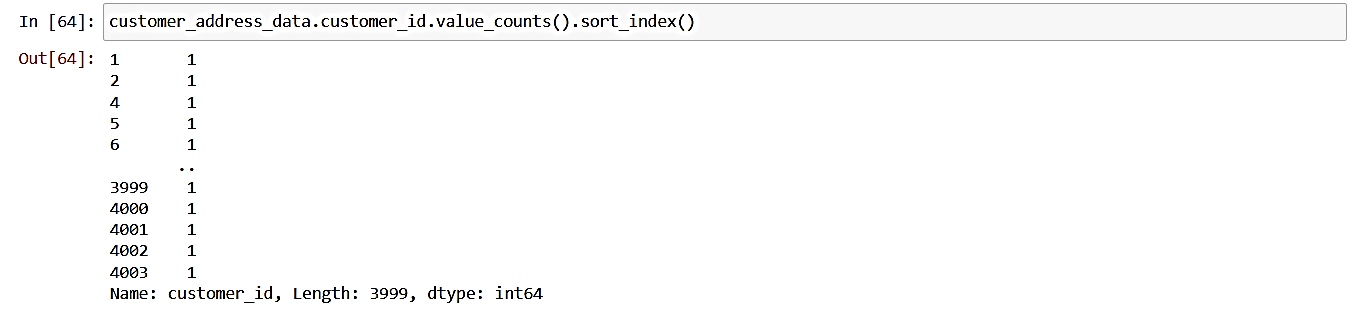
* **There are 4000 rows and 6 columns.**
* **There are 0 duplicate rows in Customer Address Dataset.**

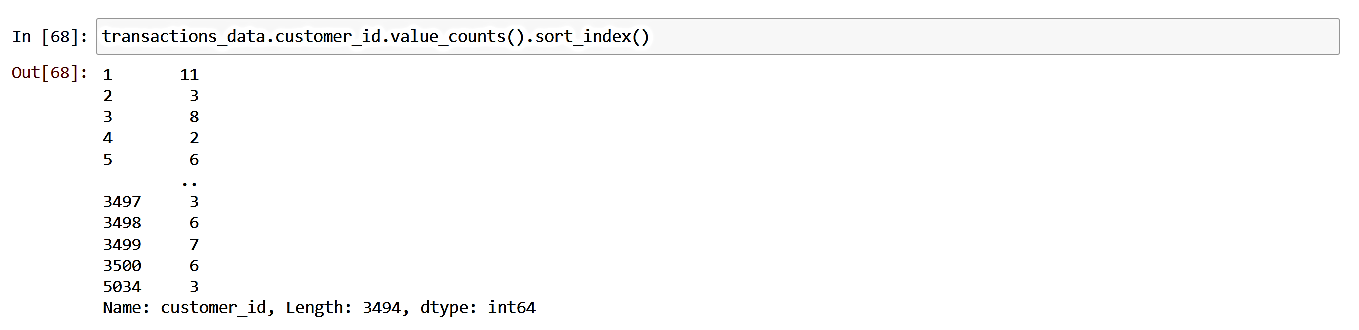
**The column named ‘customer\_id’ is the main column in all 3 datasets namely ->**

**Transactions Dataset, Customer Demographic Dataset and Customer Address Dataset**

**The following the customer\_id’s present in each of 3 datasets: -**

****

****

****