# **Data Quality Assessment- Report Sunil Goroji**

This is the detailed report of all the data quality issues identified in Transactions, NewCustomerList, CustomerDemographic, CustomerAddress data sheets of the excel workbook we are working with. The issues with these sheets are explained below accordingly.

**Transactions**:

* We have identified missing values in multiple columns, in which the online\_order column was the column with the most missing values. This can lead to biased or incomplete results and impact our overall quality of analysis.
* For the product\_first\_sold\_date column, we recommend converting the data to the appropriate date format, to ensure that we obtain accurate analysis results.

**NewCustomerList:**

* The dataset 'NewCustomerList' has missing data and is missing information in the job\_title and job\_industry\_category columns. About 11% and 17% of data, respectively are missing. This could impact the analysis and should be addressed by either imputing missing data or dropping rows with missing data, depending on the data and business context.
* We also found some inconsistencies in the dataset, for example, we identified 5 unnamed columns with random data points. This makes it difficult to ensure the accuracy of the data, as it is unclear what this data represents. To address this, we recommend dropping these unnamed columns or investigating their origin to confirm their validity.

**CustomerDemographic:**

* Corrupt data in the "default" column: The "default" column has corrupted data, meaning that the values in the column may be incorrect, incomplete, or inconsistent. It is unclear what specifically is wrong with the data in this column, but it could potentially include values that are not valid or do not conform to the expected format.
* Inconsistent names in the "gender" column: The "gender" column has inconsistent names, which means that the values in the column do not conform to a standard format or set of values. For example, some records may use "M" to represent male, while others use "Male". This inconsistency could lead to errors in analysis or make it difficult to combine the data with other datasets that use a different naming convention.
* And again, the common problem which is incomplete or missing data in multiple columns.

**CustomerAddress:**

* This final dataset was completely clean except for some consistency issues in the column ‘state’ column.

"In conclusion, our analysis has revealed a number of data quality issues that could impact the accuracy and completeness of the project. These include missing data, inconsistencies, and inaccuracies in certain data fields, which could lead to inaccurate insights and decisions.

To address these issues, we recommend implementing a number of data cleaning and validation procedures, including filling in missing data, standardizing data fields, and verifying the accuracy of the data against external sources. In addition, we recommend establishing clear data governance processes to ensure ongoing data quality.

The benefits of these actions include increased accuracy and reliability of the data, improved decision-making, and a better understanding of the underlying trends and patterns in the data. We believe that by addressing these data quality issues, we can ensure that the project is based on a solid foundation of high-quality data and will be more likely to achieve the desired outcomes.

We look forward to working with you to implement these recommendations and improve the project’s data quality."