Dallas Dataset Statistics

The study that is offered offers a thorough examination of the Dallas dataset that includes information pertinent to the food industry. The report includes a wide range of factors designed to address any discrepancy found in the dataset. With 114 variables spread across 78,400 observations, the dataset offers a sizable amount of data for analysis.

**Data Quality:** The existence of 64,543,27 missing cells, or 72.2% of the dataset, is a serious problem for data quality. Given the significant percentage of missing data, urgent action is needed. 42 duplicate rows are also found in the dataset, however they only make up 0.1% of the total.

**Memory Usage:** The dataset has an average record size of 912.0 B and takes up 68.2 MiB in memory. Understanding the memory needs for processing and analysis depends on this information.  
**Possible Consequences:** The significant proportion of missing data makes statistical analysis difficult and could necessitate the adoption of imputation techniques or a review of data gathering strategies. The dataset appears to be rather huge based on the memory utilisation, and working with it properly may need the use of appropriate data handling techniques.  
Our recommendation is to look into the reasons behind the missing data and utilise suitable imputation techniques in order to enhance the dataset's integrity. To make sure that every observation is distinct, the duplicates should be looked at and, if needed, eliminated.

**ETL:**

**Conclusion:** There are many potential and difficulties with the dataset. Before conducting a trustworthy analysis, it is imperative to address the alarming amount of missing data. The amount of text variables in the dataset is very small, but it can require specific text analysis methods due to its high percentage. The dataset has great insights if it is cleaned and prepared properly.