Data Pipeline Project: Problem Description

Problem Description:
XYZ Company is facing challenges in aggregating, cleaning, and visualizing customer data collected from various Google Forms and Survey Monkey surveys. The goal is to create a data pipeline that efficiently collects, cleans, and presents this data in a unified dashboard.

Data Understanding
Type of Data: Qualitative and quantitative
The data collected includes responses from multiple Google Forms and Survey Monkey surveys. Each response includes personal information, feedback, preferences, product usage details, and demographic information.

Data Issues:

Missing Values (NA): Some survey responses contain missing values in certain fields.
Outliers: The data might contain outliers, especially in numerical fields like age and product usage frequency.
Skewed Data: Distribution of certain responses, such as product satisfaction ratings, might be skewed.

Data Analysis
Data Types:
The data includes both structured data (e.g., ratings, multiple-choice) and unstructured data (e.g., open-ended text responses).

Data Cleaning Challenges:

Missing Values: Missing values can affect the quality of analysis and visualization.
Outliers: Outliers can distort the overall trends and insights.
Skewed Data: Skewed data distribution might lead to misinterpretation of results.

Approaches to Address Data Issues
Handling Missing Values:

Use imputation techniques such as mean, median, or mode imputation for fields with missing values.
For unstructured text responses, consider using natural language processing (NLP) methods for imputation.
Outlier Treatment:

Identify outliers using statistical methods like z-scores or IQR (Interquartile Range).
Decide whether to remove extreme outliers or transform them to bring them within a reasonable range.
Addressing Skewed Data:

For skewed data, consider transforming the data using techniques like logarithmic or Box-Cox transformations.
Utilize appropriate visualization methods that can handle skewed data, such as violin plots.