Loan data summary:

Step-by-Step Summary of Data Cleaning Process

Step 1: Organize Data Quality Reports

First, I arranged the data quality reports by generating two new spreadsheets to separate categorical and continuous variables. This allowed me to evaluate the data more effectively.

Categorical Data Quality Report: This report included metrics such as count, unique values, top (most frequent category), and frequency for each categorical feature.

Continuous Data Quality Report: This report included metrics such as count, mean, standard deviation, min, 25th percentile, median, 75th percentile, max, mode, median, cardinality, and outliers for each continuous feature.

Step 2: Identify Outliers and Bad Data

Using the data quality reports, I employed various elimination procedures to find outliers and incorrect data.

Missing Values: I identified features with a significant percentage of missing values. Features with more than 90 missing values were considered for deletion.

Constants: I looked for features with constant values (only one unique value) because they don't provide valuable information. For this, I used cardinality from the analysis.

Duplicates: To ensure data quality, I identified and eliminated duplicate entries from the dataset.

Outliers in Age: I found that the "Age" feature had a lot of null values and some extreme outlier values (e.g., -2 for age seemed like an error).

Step 3: Clean the Data

Based on the analysis, I selected features that needed to be deleted and described the reasons for doing so:

Age: Despite being an essential demographic factor, there was a large number of missing data (513 recorded observations out of 614). This rendered it untrustworthy, so it was removed.

Other Features: Features with a high number of missing values, constant values, and minimal variability were also eliminated.

Step 4: Documenting the Steps for Dropping Features

I documented each step of the feature elimination process:

High Percentage of Missing Values: Features with more than 90 missing values were dropped.

Constant Features: Features with only one unique value were dropped.

Duplicate Rows: Duplicate rows were identified and removed from the dataset.

Step 5: Final Cleaned Data

After cleaning the data, I kept the features that were deemed valuable based on their data quality metrics and significance for analysis:

Retained Features: These included crucial identifiers (Loan_ID, Customer), demographic information (Gender, Married, Dependents), financial information (ApplicantIncome, CoapplicantIncome), loan specifics (LoanAmount, Loan_Amount_Term), and the target variable (Loan_Status).

Columns Removed and Reasons:

High Cardinality:

Loan_ID: High cardinality, unique identifier.

Application_Number: High cardinality, unique identifier.

Customer: High cardinality, unique identifier.

Constant:

Application_Type: Constant feature with a single unique value ("Loan").

Duplicate Features:

Amount_Request and LoanAmount: Both features represent the loan amount. I eliminated Amount_Request.

Application_Number and Loan_ID: Both features are unique identifiers. I eliminated Application_Number.

Reasoning:

Here's how I use cardinality to evaluate data quality:

High Cardinality:

For continuous features, high cardinality is often preferable since it shows a large range of values, providing more information and flexibility.

Very high cardinality for categorical features might be problematic at times, especially if it leads to sparse data. This can make modeling more challenging and necessitate more complex strategies for handling.

Low Cardinality:

Low cardinality in continuous features can indicate a lack of variability, implying that the feature isn't particularly informative.

Low cardinality in categorical features is generally preferred since it simplifies the model and makes it easier to interpret.

Constant Values:

If a feature has a cardinality of one, it indicates that all values are the same. These features give no valuable information and should be removed from the dataset.

Additional Considerations:

I did not drop the Gender feature because, despite its low cardinality, it is an important demographic factor and did not have a significant number of null values like the "Age" feature.