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 and the Continuous Data Quality Report

Step 2: Identify Outliers and Bad Data

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

Missing Values: I found characteristics that had a significant percentage of missing values. Features over 90 missing values were considered for deletion.
Constants: I looked for characteristics with constant values (meaning only one unique value) because they don't provide any valuable information. For this I looked at the cardinality from the analysis.
Duplicates: To assure data quality, I identified and eliminated duplicate entries from the dataset.
Age had a lot of null values and some values for age had extreme outlier values.
e.g. -2 for age seemed like an error and would have affected the data quality.

Step 3: Clean the Data

Based on the analysis, I selected characteristics 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, thus it was removed.

Other Features: Elimination criteria included features with a high number of missing data, stable values, and minimal variability.
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 attributes that were deemed valuable based on their data quality metrics and significance for analysis:

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

The eliminated features also included Application type:
Constant feature with single unique value: Loan

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 characteristics 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.

In my analysis of the loan dataset, I found that attributes such as "Loan_ID" and "Customer" had high cardinality, indicating that each loan application and client is unique. This is expected and required for these identifiers.

However, for features such as "Application_Type" and "Gender", low cardinality is preferable because there are only a few possible values (e.g., Loan/Not Loan, Male/Female), making the data easier to work with and analyze.

I didn’t drop gender out because I wasn’t 100% confident on letting it out just based on cardinality, also because I saw this as an important feature to reflect on a predictive model and also because this feature didn’t have a significant number of Null values like age.