Experiment -1

Handling Missing Values on a Dataset

(Dataset to be used: Cleveland Clinic Heart Disease Dataset

Available at: https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset/)

a) Apply filter "**Replace Missing Values"** to fill missing values with means and modes
b) Apply filter "**Replace Missing with User Constant"** to fill missing values with user-supplied values
c) Apply filter "**Remove"** to remove the attributes with missing values
d) Apply filter **"Replace with Missing Value"** to introduce missing values

---

Submitted By:

Name: Bharath K

USN: 23BCAR0252

Lab Batch: I

Load a Dataset and identify if missing values exist

Steps to Download and Upload the Dataset for Evaluation

1. **Access the Dataset Link**:
    - First, open the Google Classroom document where the link to the dataset is provided. Click on the link, which will redirect you to the Kaggle website.
2. **Sign Up/Login on Kaggle**:
    - If you do not already have a Kaggle account, sign up using your email address or through Google/Facebook. If you already have an account, simply log in.
3. **Download the Dataset**:
    - After logging in, you will be directed to the Kaggle page containing the dataset. Click the **"Download"** button to download the dataset in a zip folder format.
4. **Extract the Zip File**:
    - Once the download is complete, locate the zip file on your computer. Right-click on the file and select **"Extract"** to unzip it into a folder where you can easily access it.
5. **Open Wika Platform**:
    - Log in to the Wika platform where you will be working with the dataset. Once logged in, find the **"Open File"** button or option.
6. **Upload the Dataset**:
    - Click on **"Open File"**, and navigate to the folder where you extracted the zip file. Select the extracted file to upload it to the Wika platform.
7. **Verify Data Loading**:
    - After the file is successfully uploaded, confirm that the dataset has been loaded correctly. It should contain **14 attributes** and **303 instances**, as expected.

Handling Missing Values in the Dataset

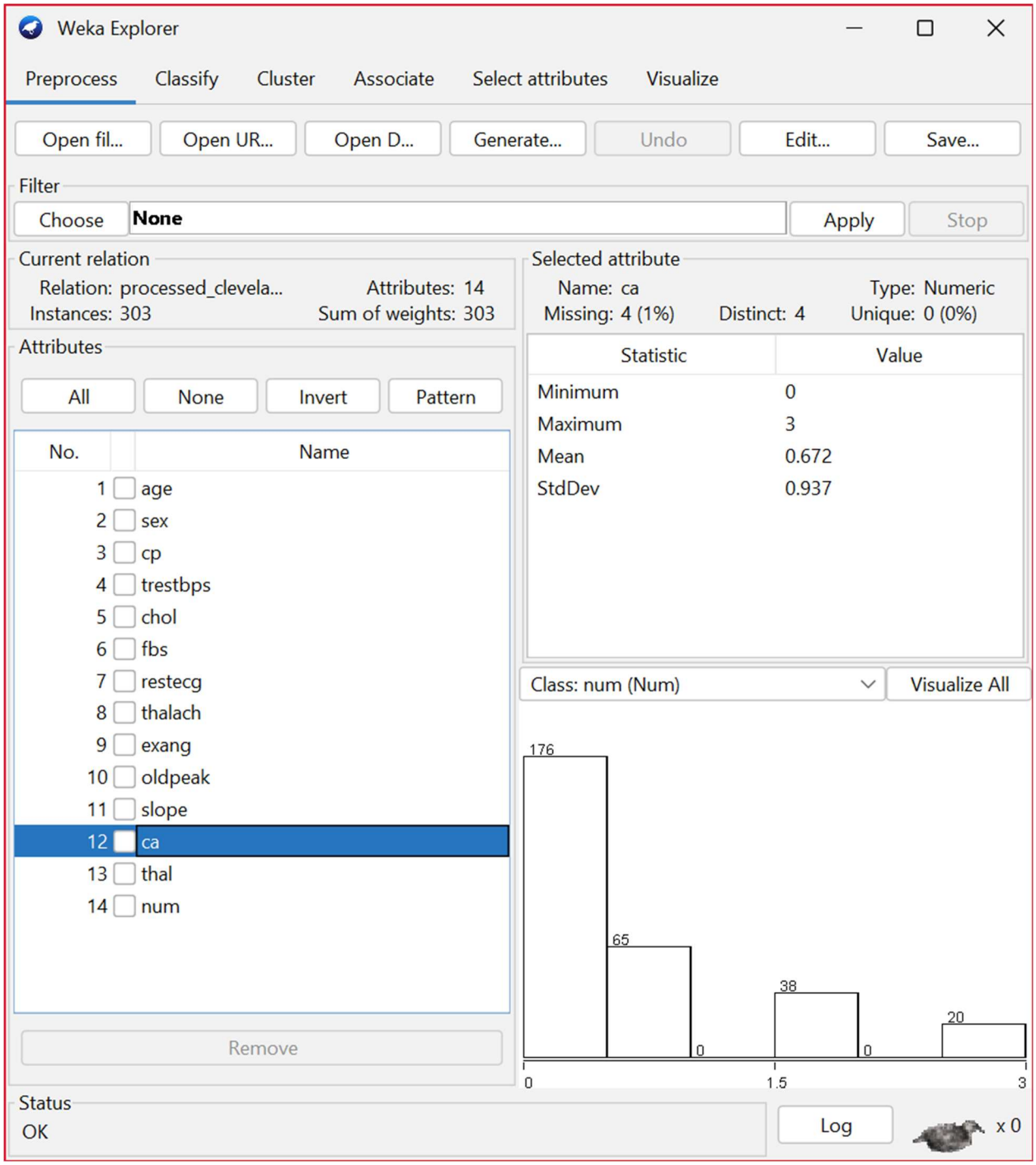
The dataset has missing values in two attributes: **Ca** and **Thal**. These missing values could impact the accuracy of our analysis and visualizations, so it's essential to handle them properly.

- **Missing Values**:
    - The **Ca** attribute has 4 missing values, which account for 1% of the total data.

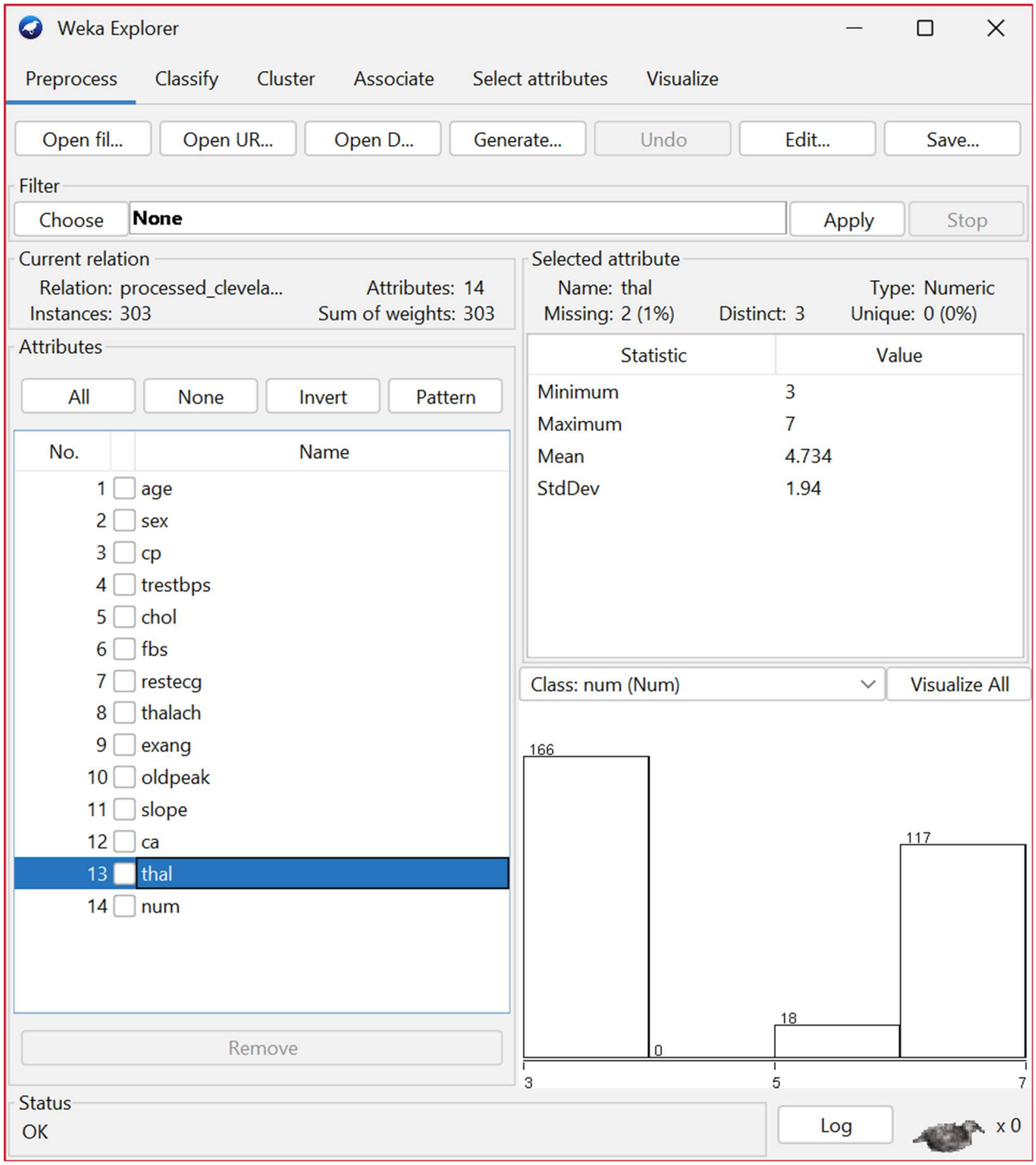
- The **Thal** attribute has 2 missing values, which also account for 1% of the total data.

Steps to Address Missing Values

1. **Identify Missing Values**:
   - First, check the dataset for missing values. The **Ca** and **Thal** attributes have missing values, as seen in the provided screenshots.
2. **Apply Filters to Handle Missing Data**:
   - To handle these missing values, I will apply filters to the dataset. These filters will help mitigate the impact of missing data, ensuring more accurate analysis and visualizations. Some options are:
     - **Imputation**: Replacing missing values with the mean, median, or mode for numerical attributes (like the **Ca** attribute).
     - **Deletion**: Removing rows with missing values if they don't significantly affect the dataset.
     - **Prediction**: Using machine learning algorithms to predict and fill in missing values based on other available data.
3. **Effect on Data Analysis**:
   - By addressing these missing values, I will improve the dataset's integrity, ensuring any subsequent analysis, visualizations, and models built on this data will be accurate and reliable.



The image showing missing values in the "Ca" attribute indicates that there are 4 missing values, which account for 1% of the total data.



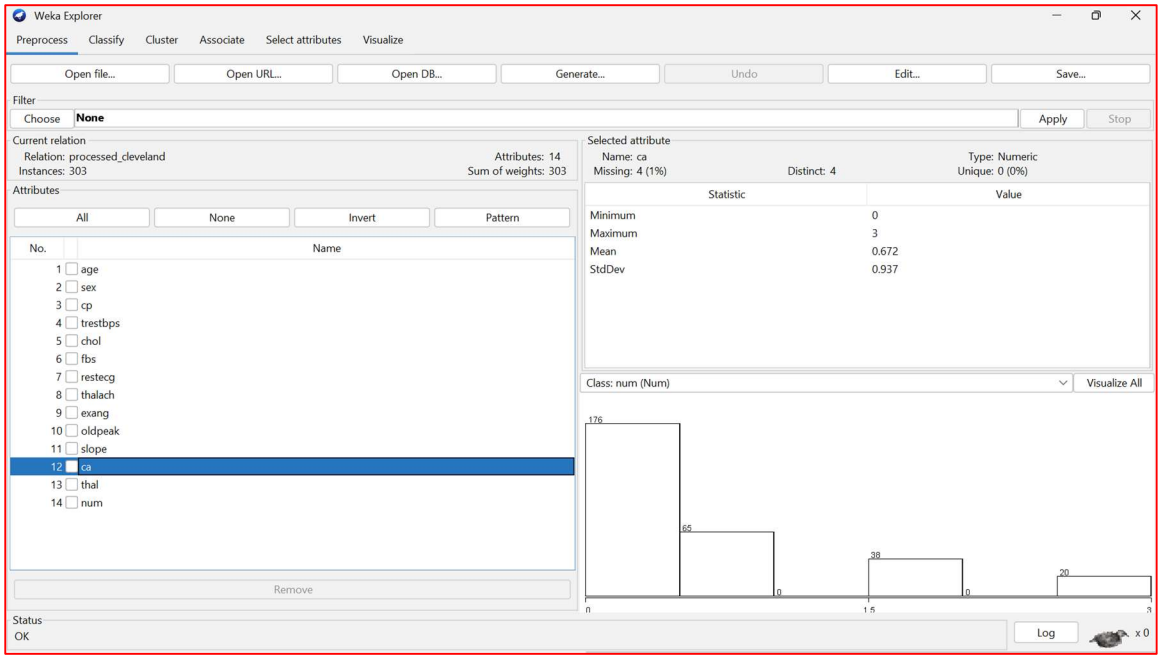
The image showing missing values in the "Thal" attribute indicates that there are 2 missing values, which account for 1% of the total data.

Applying the "Replace Missing Values" Filter

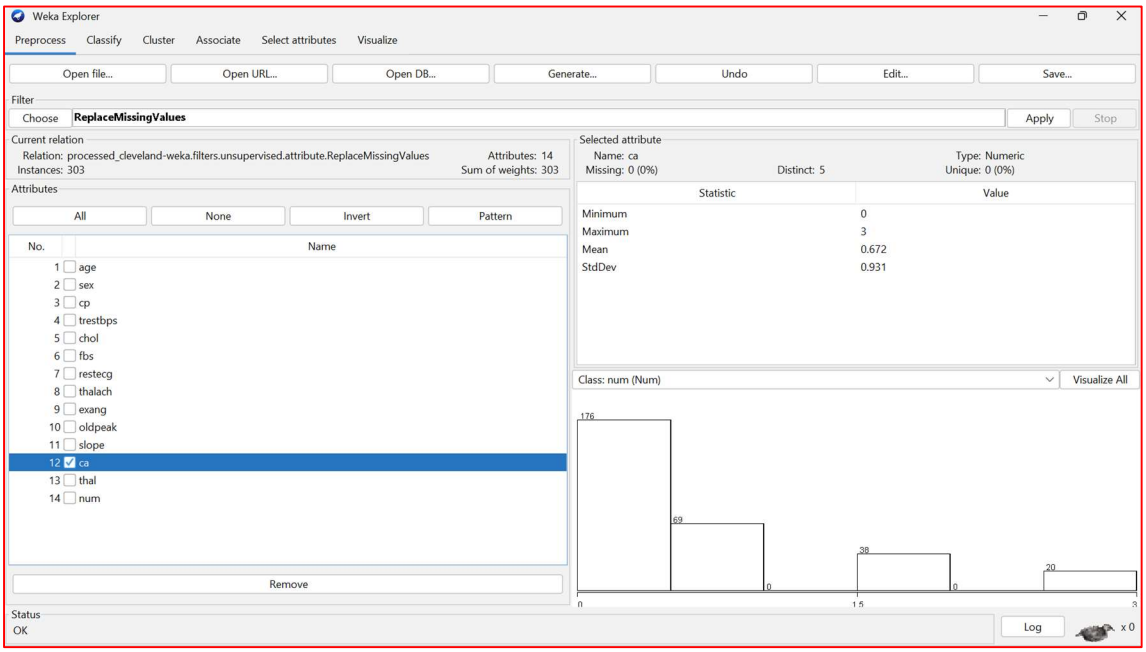
To handle the missing values in the Ca and Thal attributes, the first filter we will apply is the "Replace Missing Values" filter. This filter will help in reducing the missing values and improving the overall data integrity.

Steps to Apply the "Replace Missing Values" Filter:

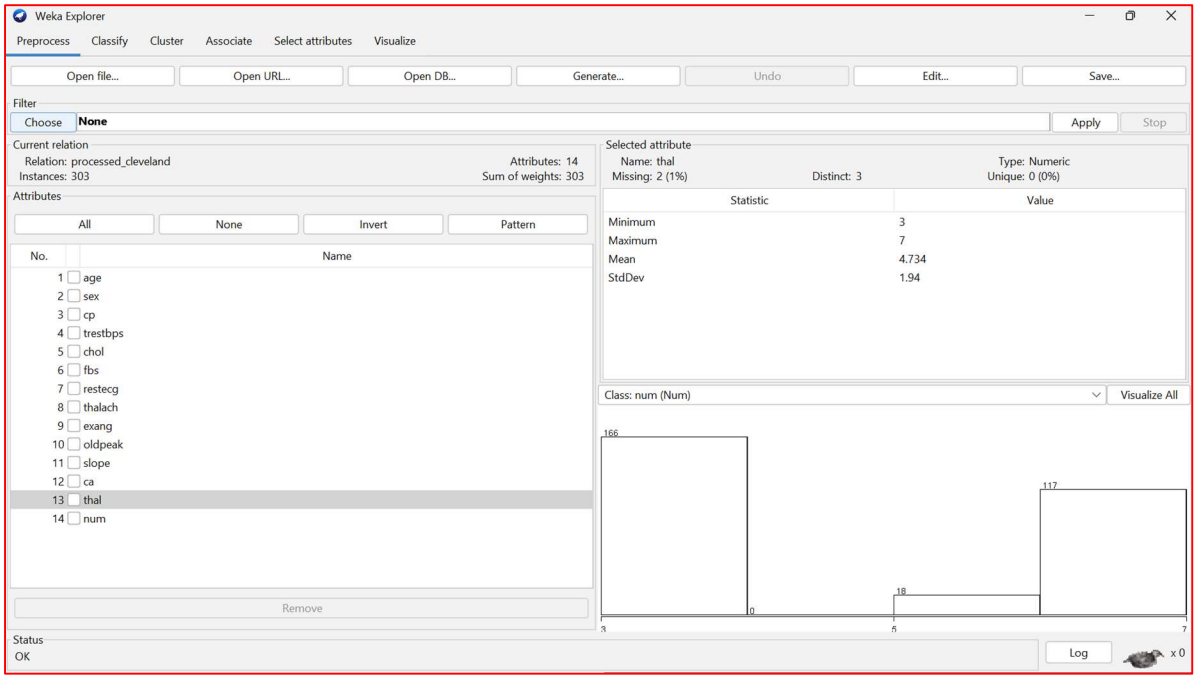
1. **Navigate to the Filter Option:**
    - In the Wika platform, click on the "Choose Filter" option.
    - Then, navigate to Filters and select Unsupervised Data.
2. **Select the "Replace Missing Values" Filter:**
    - Within the Unsupervised Data section, choose the "Replace Missing Values" filter.
3. **Effect on Missing Values:**
    - After selecting this filter, the missing values in both the Ca and Thal attributes will be automatically reduced to zero, as shown in the screenshots below.
        - Before applying the filter, there were 4 missing values for Ca and 2 missing values for Thal.
        - After applying the filter, both attributes will have zero missing values.
4. **How It Works:**
    - For numeric data, the filter uses the mean of the attribute to replace the missing values.
    - For non-numeric data, such as Thal, the filter uses the mode (most frequent value) to replace the missing values.
5. **Outcome:**
    - By applying this filter, the data integrity is increased, as the missing values are effectively filled, making the dataset ready for further analysis and visualization.



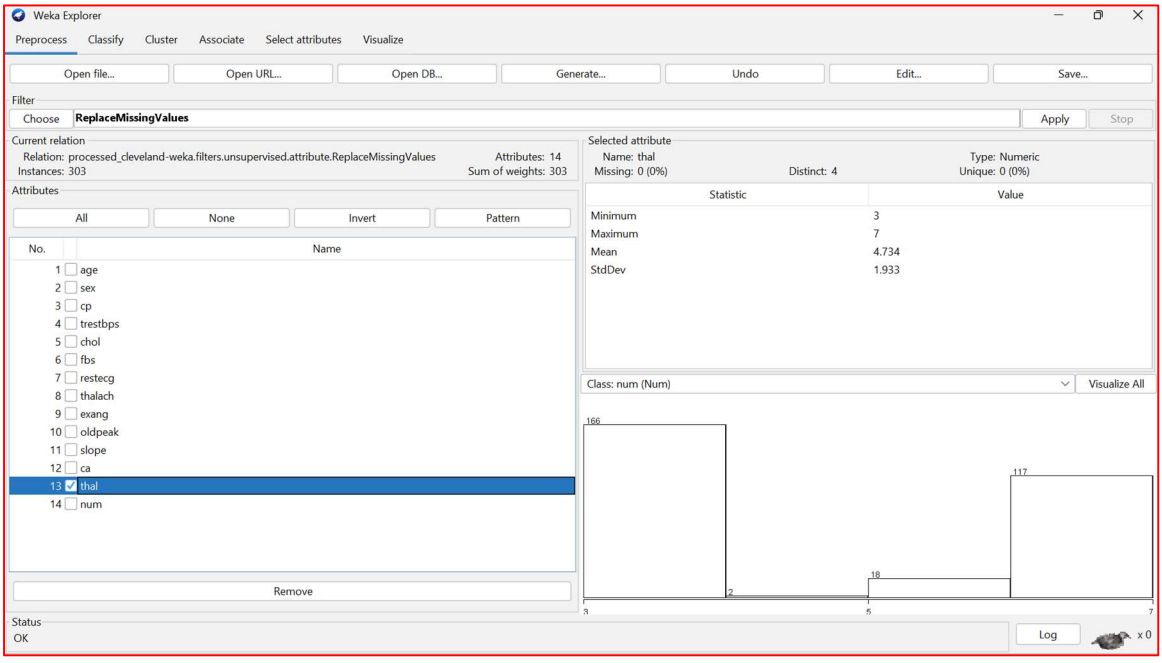
Before Image: *The "Ca" attribute contains 4 missing values (1%) in the dataset. This missing data can potentially affect the accuracy of analysis and visualizations if not addressed.*



After Image: *Once the appropriate filter is applied, the missing values in the "Ca" attribute are handled, resulting in no missing values. This ensures that the data integrity is restored and can be used for more accurate analysis.*



Before Image: *The "Thal" attribute contains 2 missing values (1%) in the dataset. These missing values could impact the quality of analysis and insights derived from the data if not handled properly.*



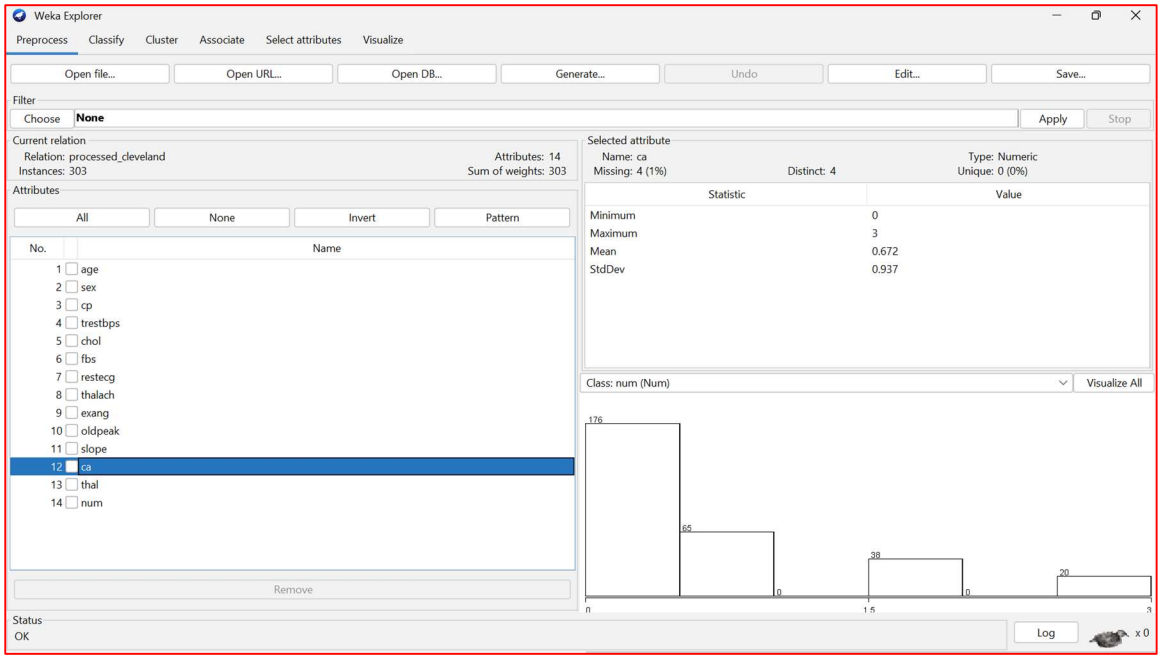
After Image: *After applying the relevant filter, the missing values in the "Thal" attribute are addressed, ensuring no missing data remains. This improves the dataset's integrity and prepares it for accurate analysis.*

Applying the "Replace Missing with User Constant" Filter

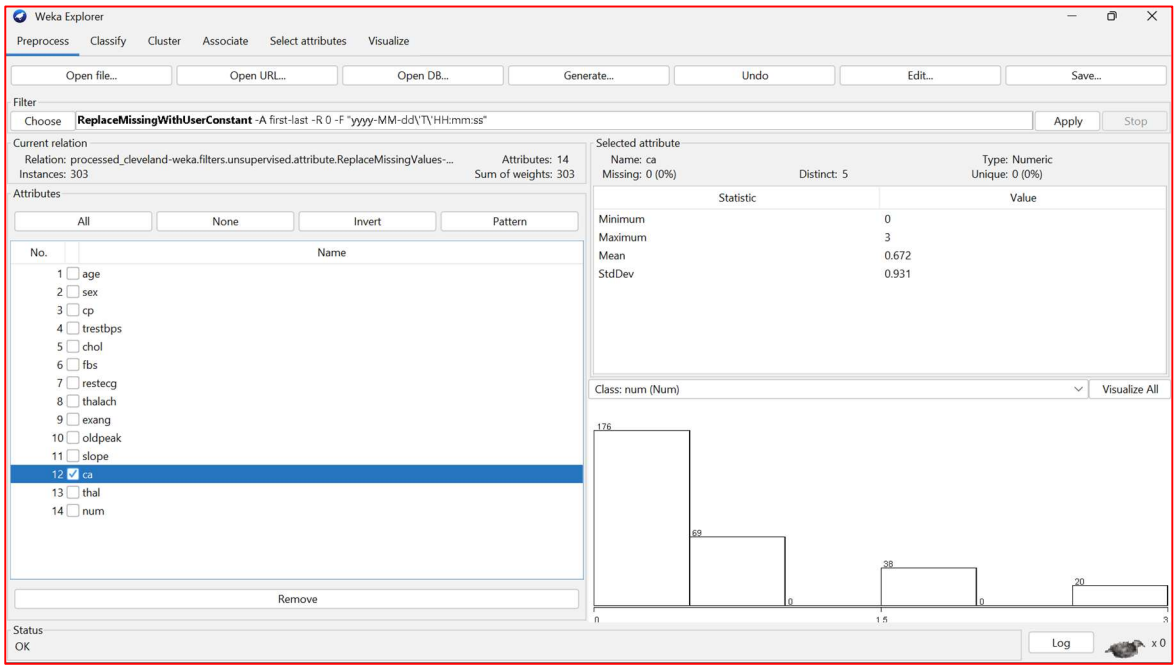
In addition to the "Replace Missing Values" filter, we will now apply the "Replace Missing with User Constant" filter to handle missing values in the dataset. This filter allows you to replace missing values with a custom constant value, giving you more control over the replacement.

Steps to Apply the "Replace Missing with User Constant" Filter:

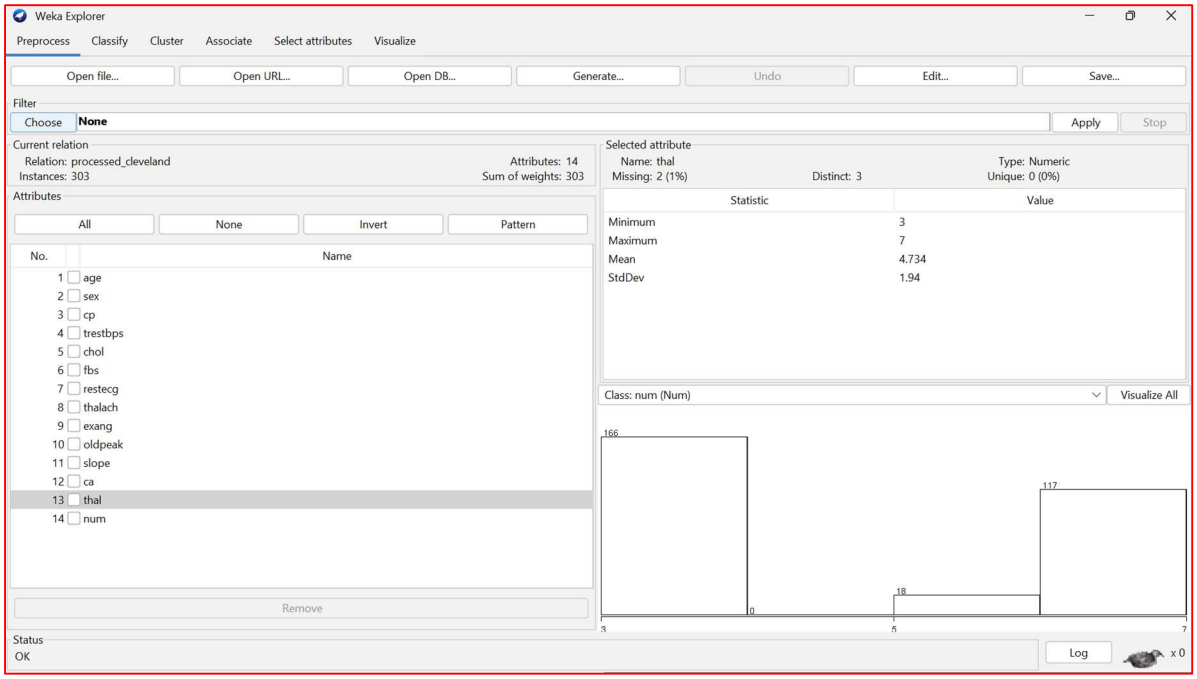
1. **Undo the Previous Filter:**
    - Before applying the "Replace Missing with User Constant" filter, you will need to undo the "Replace Missing Values" filter, as only one filter can be applied at a time.
2. **Navigate to the Filter Option:**
    - In the Wika platform, click on the "Choose Filter" option.
    - Navigate to Filters and select Unsupervised Data.
3. **Select the "Replace Missing with User Constant" Filter:**
    - Within the Unsupervised Data section, choose the "Replace Missing with User Constant" filter.
4. **No Prompt for Default Value:**
    - This filter does not prompt you to enter a default value when applying it. Instead, the missing values are automatically replaced with a predefined constant.
    - For example, the missing values in Ca could be replaced with 0, or the missing values in Thal could be replaced with "Unknown". The exact constant depends on your choice during the filtering process.
5. **Effect on Missing Values:**
    - After applying this filter, the missing values in both Ca and Thal will be replaced with the user-defined constant (e.g., 0 for Ca or "Unknown" for Thal). This ensures that no missing values remain, as illustrated in the before-and-after screenshots.
6. **Outcome:**
    - By applying the "Replace Missing with User Constant" filter, you gain flexibility in customizing how missing data is handled. This is particularly useful for cases where you want to clearly mark missing values with a specific constant.
    - Be mindful of how this constant might affect subsequent analysis or visualizations, especially when dealing with categorical data like Thal.



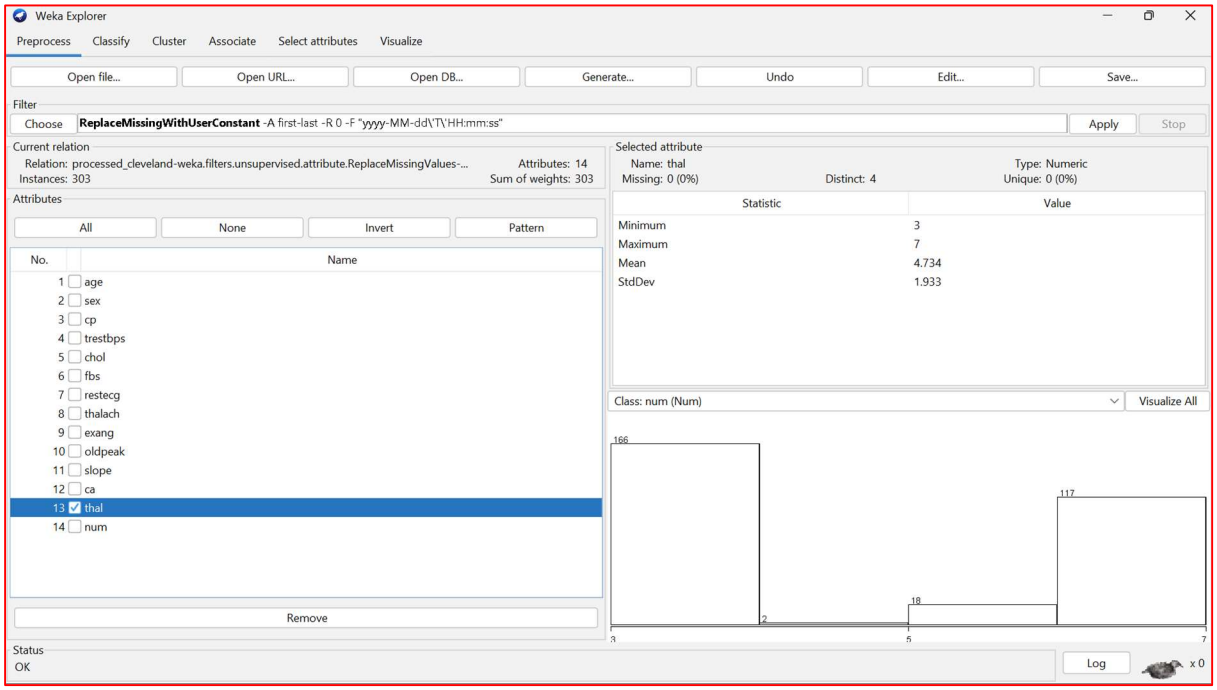
Before Image: *The "Ca" attribute contains 4 missing values (1%) in the dataset. This missing data can potentially affect the accuracy of analysis and visualizations if not addressed.*



After Image: *Once the appropriate filter is applied, the missing values in the "Ca" attribute are handled, resulting in no missing values. This ensures that the data integrity is restored and can be used for more accurate analysis.*



Before Image: *The "Thal" attribute contains 2 missing values (1%) in the dataset. These missing values could impact the quality of analysis and insights derived from the data if not handled properly.*



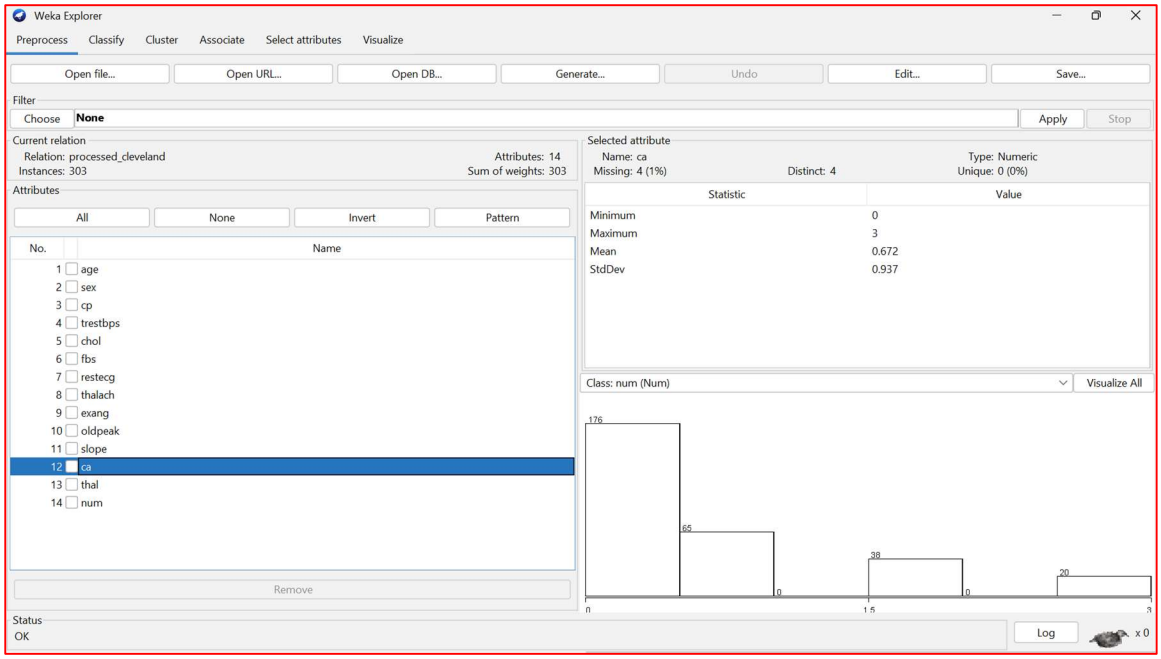
After Image: *After applying the relevant filter, the missing values in the "Thal" attribute are addressed, ensuring no missing data remains. This improves the dataset's integrity and prepares it for accurate analysis.*

Applying the "Remove" Filter

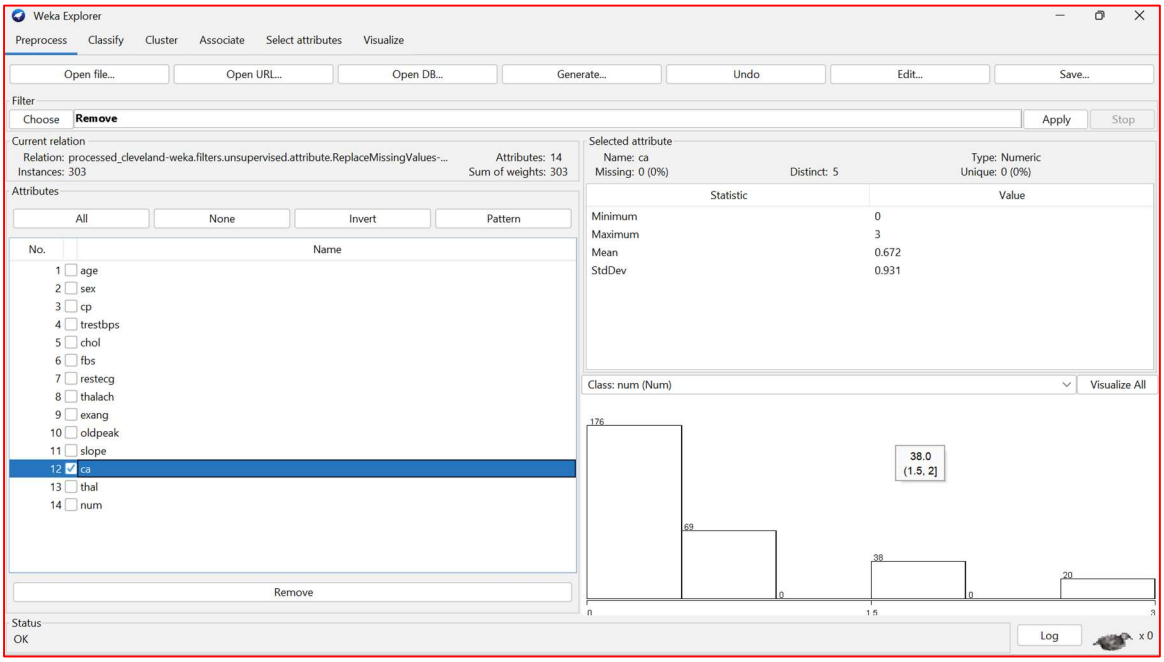
The **"Remove"** filter is used to eliminate rows with missing values, completely removing any data points that are incomplete. This filter can help simplify the dataset when dealing with missing values, but it may result in data loss if many rows are removed.

Steps to Apply the "Remove" Filter:

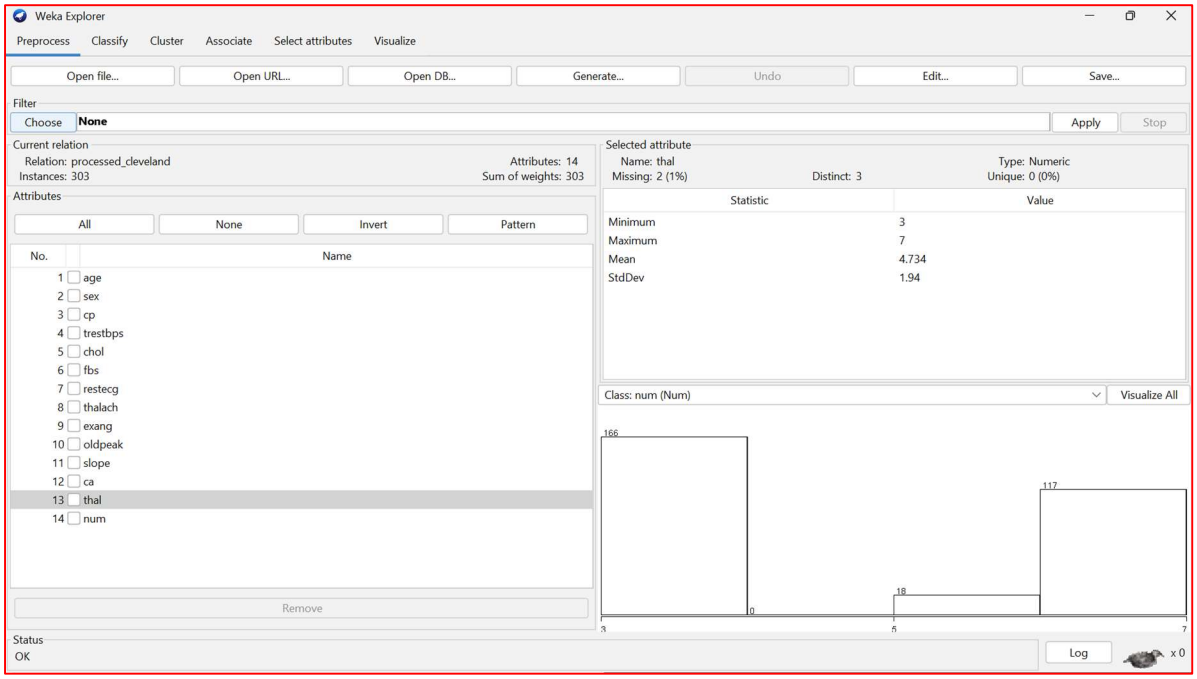
1. **Undo the Previous Filter**:
    - Before applying the **"Remove"** filter, you will need to undo the previously applied filters (such as the **"Replace Missing Values"** or **"Replace Missing with User Constant"** filter), as only one filter can be applied at a time.
    - To undo the previous filter, navigate to the filter settings and reset or remove the currently applied filter.
2. **Navigate to the Filter Option**:
    - In the Wika platform, click on the **"Choose Filter"** option.
    - Navigate to **Filters** and select **Unsupervised Data**.
3. **Select the "Remove" Filter**:
    - Within the **Unsupervised Data** section, choose the **"Remove"** filter.
4. **Effect on Missing Values**:
    - After applying this filter, the rows with missing values in **Ca** and **Thal** will be entirely removed from the dataset. This means that any instance where a missing value exists in either of these attributes will be excluded from the data.
    - This filter will reduce the number of instances (rows) in the dataset, but it will ensure that no missing values remain.
5. **Outcome**:
    - By applying the **"Remove"** filter, you eliminate rows with missing data, ensuring that your dataset is free of any incomplete instances. This can be useful when the missing values are minimal and removing the rows won't significantly impact the analysis.
    - However, it's important to keep in mind that this approach could lead to data loss, especially if there are many rows with missing values. In cases where a substantial amount of data is removed, it could affect the quality of your analysis.



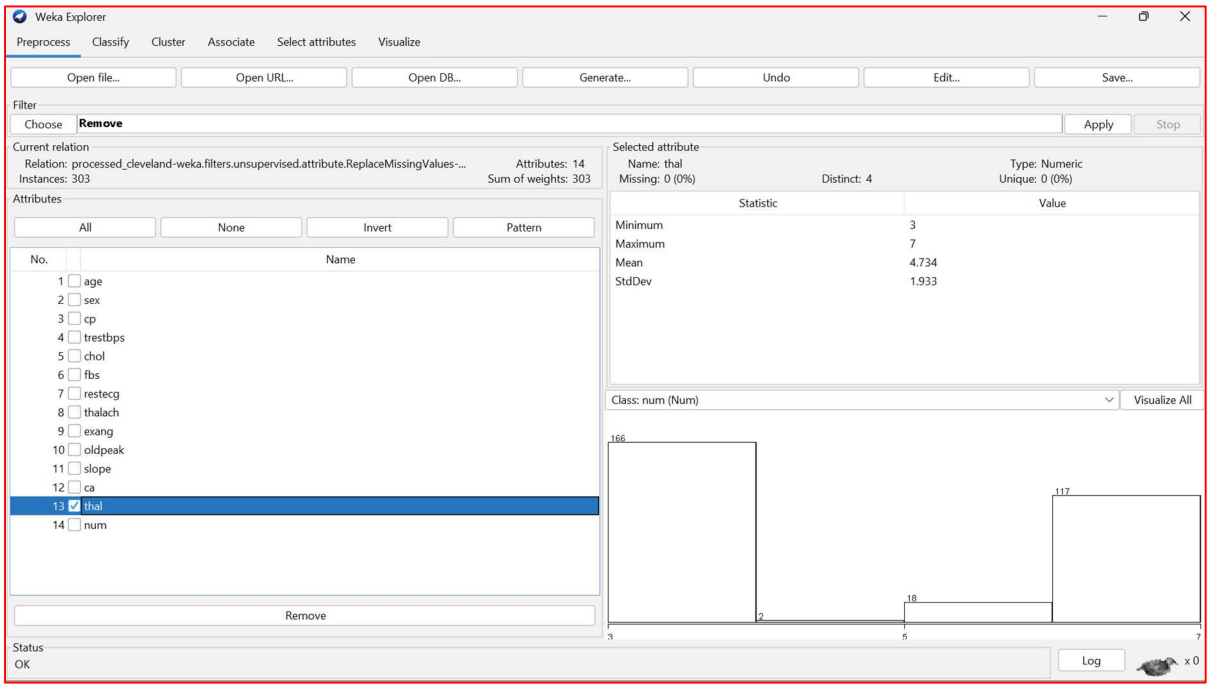
Before Image: *The "Ca" attribute contains 4 missing values (1%) in the dataset. This missing data can potentially affect the accuracy of analysis and visualizations if not addressed.*



After Image: *Once the appropriate filter is applied, the missing values in the "Ca" attribute are handled, resulting in no missing values. This ensures that the data integrity is restored and can be used for more accurate analysis.*



Before Image: *The "Thal" attribute contains 2 missing values (1%) in the dataset. These missing values could impact the quality of analysis and insights derived from the data if not handled properly.*



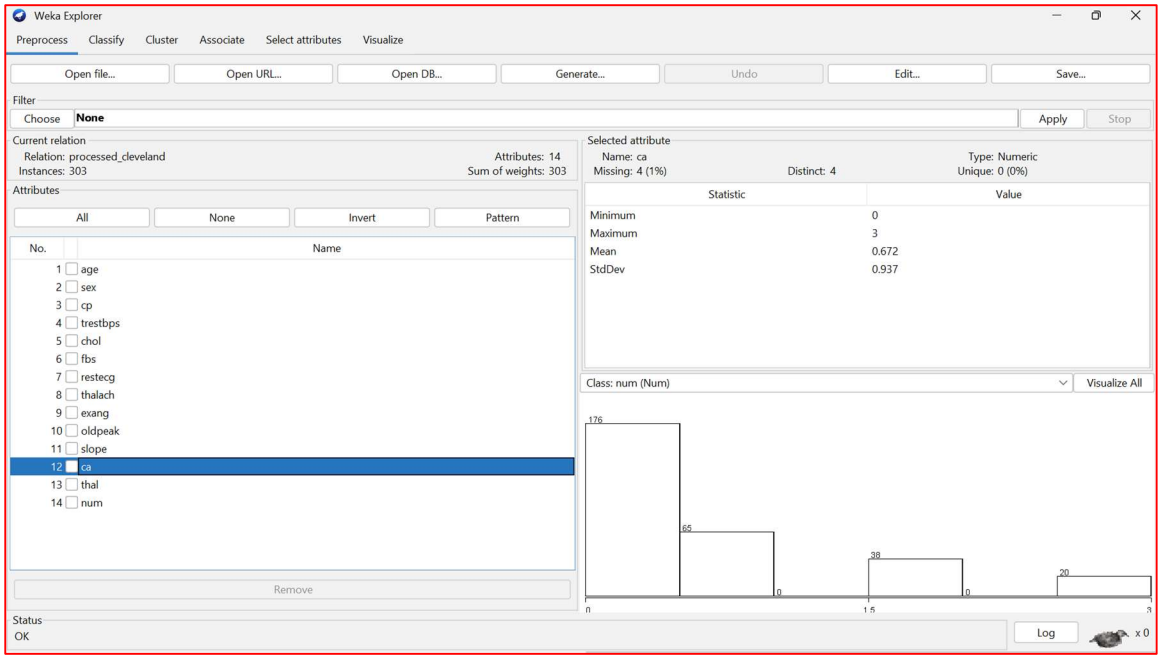
After Image: *After applying the relevant filter, the missing values in the "Thal" attribute are addressed, ensuring no missing data remains. This improves the dataset's integrity and prepares it for accurate analysis.*

Applying the "Replace with Missing Value" Filter

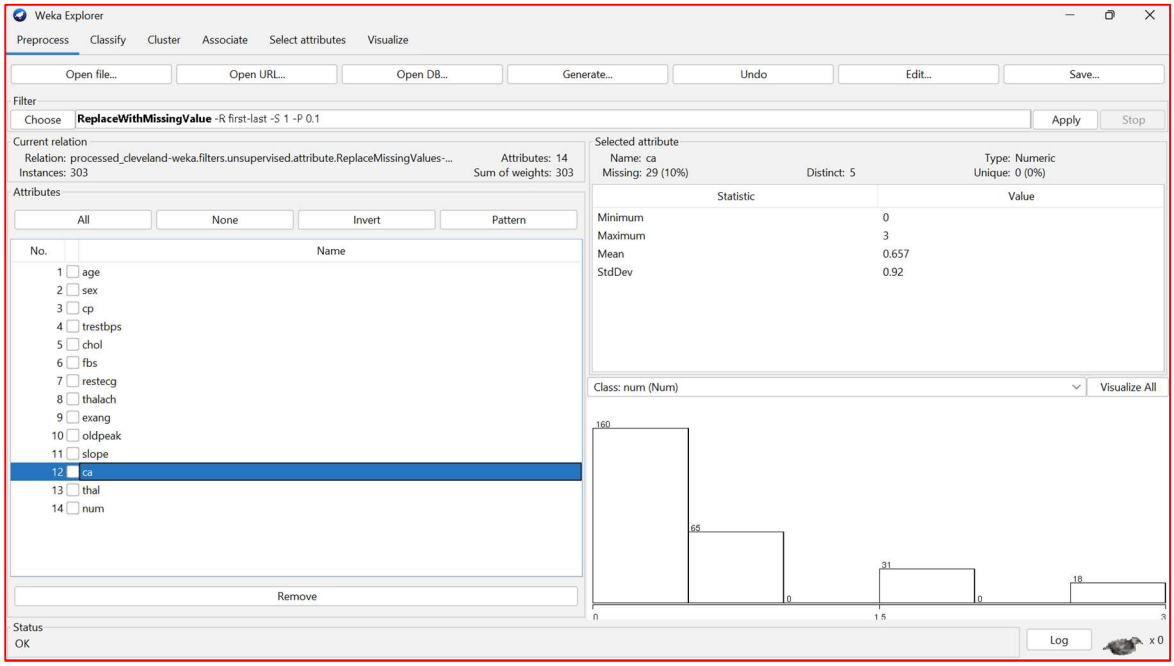
The "Replace with Missing Value" filter allows you to intentionally introduce missing values into the dataset. This can be useful for testing how the data processing pipeline handles missing data or for simulating missing values in a controlled environment.

Steps to Apply the "Replace with Missing Value" Filter:

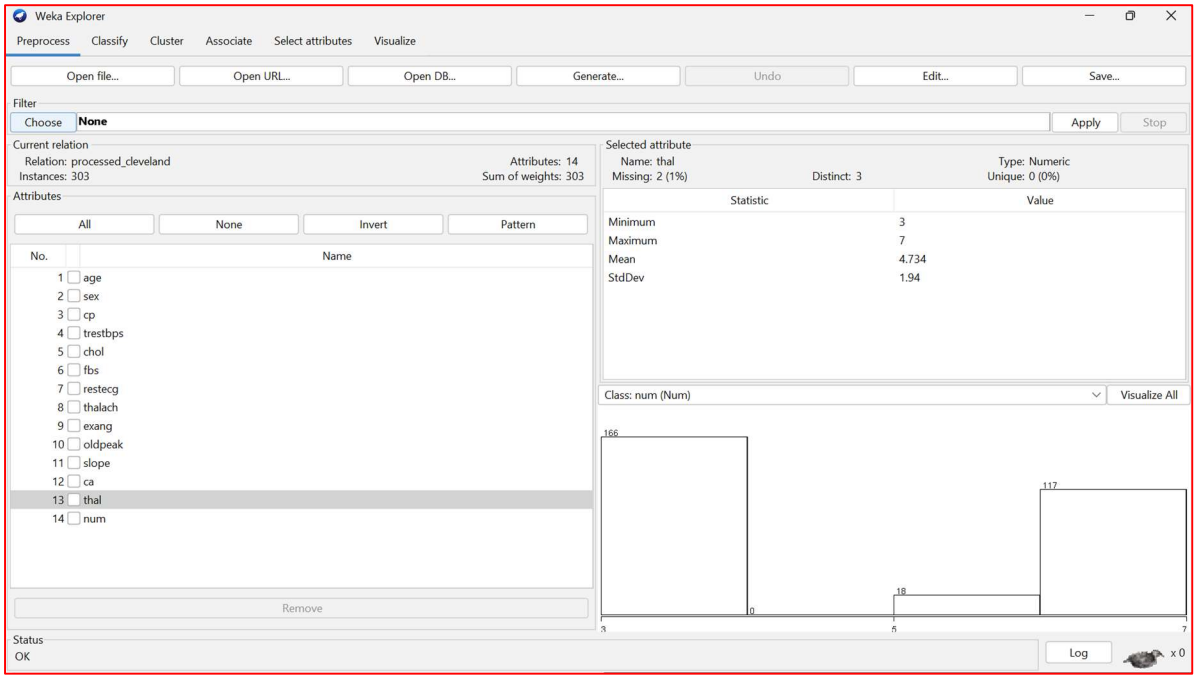
1. **Navigate to the Filter Option:**
    - In the Wika platform, click on the "Choose Filter" option.
    - Navigate to Filters and select Unsupervised Data.
2. **Select the "Replace with Missing Value" Filter:**
    - Within the Unsupervised Data section, choose the "Replace with Missing Value" filter.
3. **Configure the Filter:**
    - After selecting the filter, you will be able to configure which attributes (columns) should have their values replaced with missing values.
    - You can choose to replace specific attributes, such as Ca or Thal, with missing values, or you may choose to apply it to other attributes in the dataset.
4. **Effect on the Dataset:**
    - Once the filter is applied, the specified values in the selected attributes will be replaced with missing values (typically marked as NaN or null). This will simulate missing data for those attributes, allowing you to test how the dataset behaves when missing values are introduced.
    - For example, if you apply this filter to the Ca attribute, certain values in the Ca column will be replaced with missing values.
5. **Outcome:**
    - By applying the "Replace with Missing Value" filter, you introduce missing data into the dataset in a controlled manner. This can be helpful for simulating real-world scenarios where data might be incomplete or testing how different filters and models handle missing data.



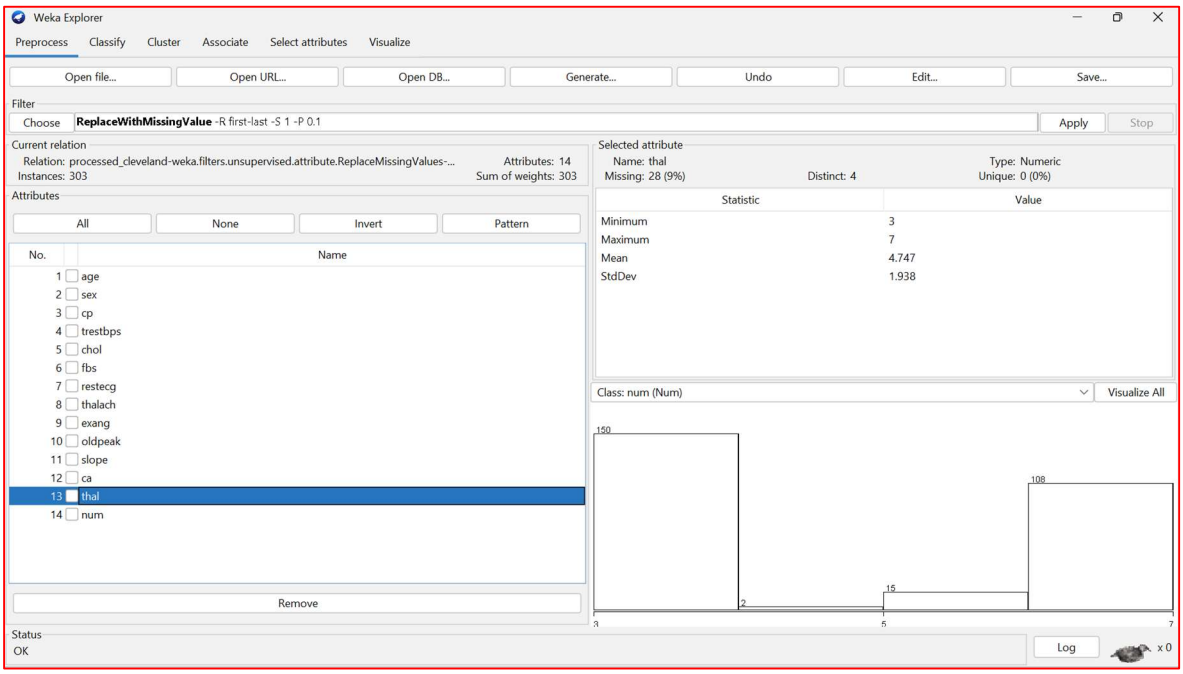
Before Image:*The Ca attribute originally contains 4 missing values (1%) in the dataset. At this point, there are no additional missing values introduced, and the dataset is being examined with its natural missing data.*



After Image: *After applying the "Replace with Missing Value" filter, the missing values in the Ca attribute have increased to 29 (10%). This filter intentionally replaces values in the Ca column with NaN or null values, simulating a higher level of missing data in the dataset. This change significantly impacts the completeness of the data.*



Before Image:*The **Thal** attribute originally had 2 missing values (1%) in the dataset. These missing values were part of the dataset's natural missing data and hadn't been altered yet.*



After Image: *After applying the "Replace with Missing Value" filter, the missing values in the Thal attribute have increased to 28 (9%). The filter replaces certain values in the Thal column with NaN or null, which results in a substantial increase in the missing data for this attribute.*

Dataset's Attribute Information:

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type
    - Value 1: typical angina
    - Value 2: atypical angina
    - Value 3: non-anginal pain
    - Value 4: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
    - Value 0: normal
    - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak: ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
    - Value 1: upsloping
    - Value 2: flat
    - Value 3: downsloping
12. ca: number of major vessels (0-3) colored by flourosopy (for calcification of vessels)
13. thal: results of nuclear stress test (3 = normal; 6 = fixed defect; 7 = reversable defect)
14. status: target variable representing diagnosis of heart disease (angiographic disease status) in any major vessel
    - Value 0: < 50% diameter narrowing
    - Value 1: > 50% diameter narrowing.