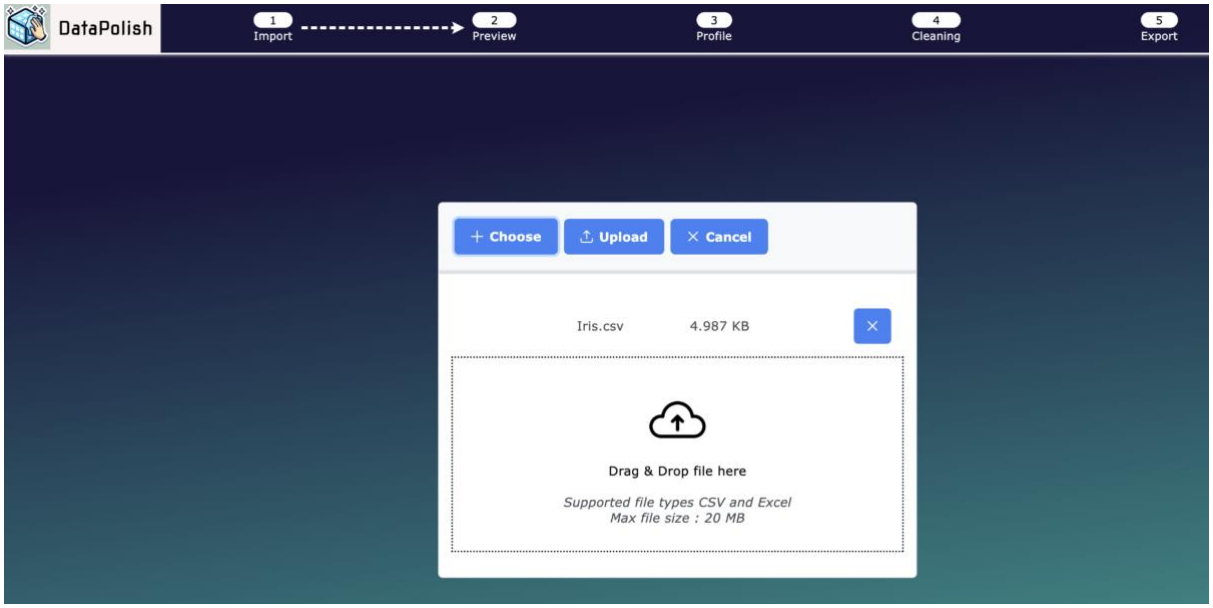
Table of Contents

1 Data Upload

1.1 Starting Your Data Journey

Welcome to DataPolish. The initial step in your data processing journey begins with the data upload interface, a straightforward and user-friendly environment where you can begin by uploading your dataset.



1.2 Uploading Your Dataset:

Supported Formats: Ensure your dataset is in either CSV or Excel format, adhering to the 20MB file size limit for optimal performance.

Upload Methods: You have two convenient options to upload your data:

Choose File: Click the 'Choose' button to select a file from your local directories.

Drag and Drop: Alternatively, you can drag your file and drop it into the designated area on the page.

Confirmation: Once your file appears in the upload area, confirm by clicking the 'Upload' button. Your file is now ready for the next phase.

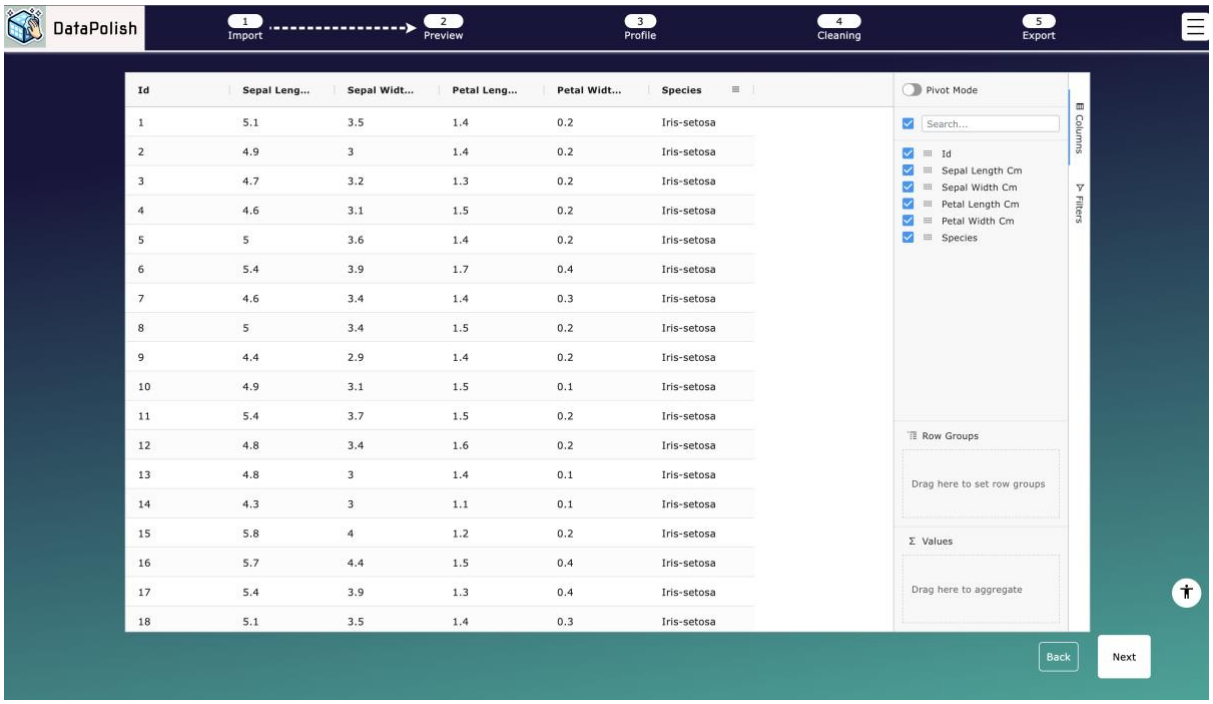
1.2.1 Navigation:

After the upload is successful, click the 'Next' button to transition to the Preview page, where you can review your uploaded data before moving on to cleaning and analysis.

2 Data Preview

2.1 Data Inspection and Customization:

Once your data is uploaded, the Preview page offers a snapshot view of the first 200 rows, allowing you to inspect and interact with your dataset. This step is crucial for verifying the correct upload and initial data quality.



2.2 Interactive Data Preview Features:

Column Expansion: Click to expand any column and view its full content, ensuring you see all data without truncation.

Autosizing Columns: Use the 'Autosize This Column' or 'Autosize All Columns' options for optimal column width and data visibility.

Column Adjustment: Feel free to rearrange the column order, pin columns for stationary reference, or reset columns to their original state post-filtering.

Data Sorting: Sort data within columns in ascending or descending order to identify patterns or outliers quickly.

Filtering for Focus: Employ the 'Filters' feature to drill down into your data. This tool allows for selective viewing, searching, and homing in on key data points.

Pivot Table Analysis: Activate the Pivot Mode to transform your data into a pivot table format for more complex, multidimensional analysis.

Row Grouping and Aggregation: Implement row groupings to categorize your data, and apply various aggregations such as average, count, or sum to glean insights.

Group Management: Manage your grouped data with ease using functionalities like 'Un-Group All', 'Expand All Row Groups', and 'Collapse All Row Groups' to adapt the view to your analysis needs.

2.2.1 Navigating the Data Journey:

After reviewing your data in the preview, click the 'Next' button to move forward to the data profiling step.

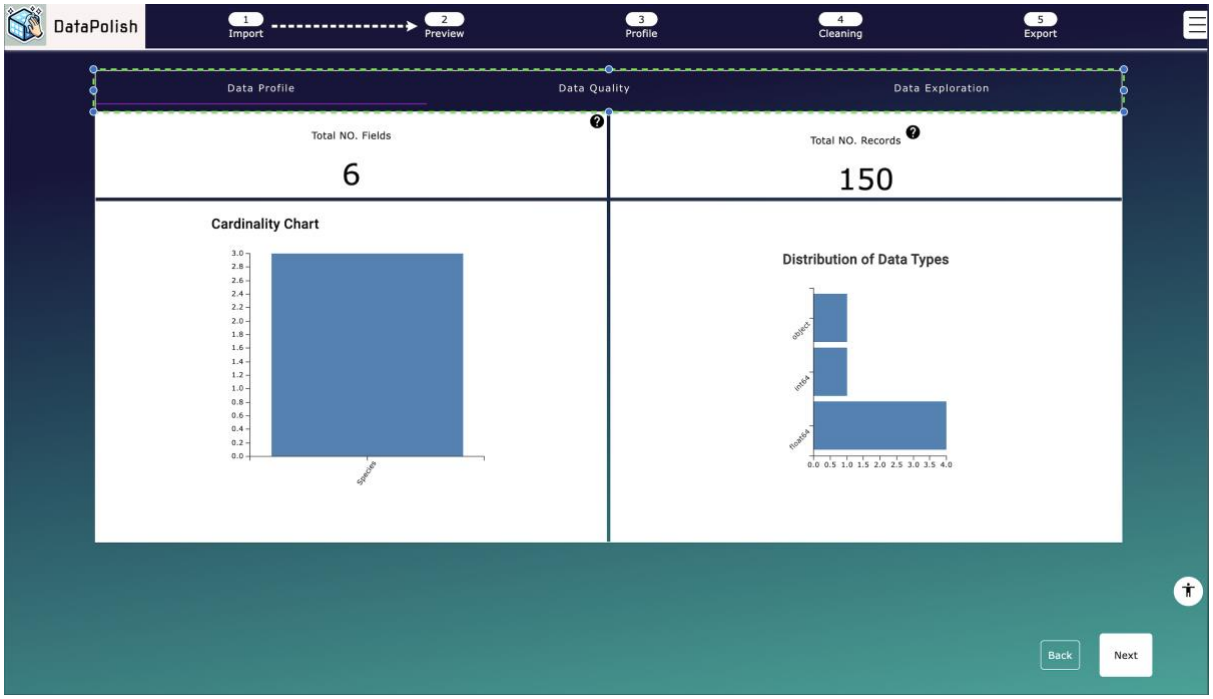
3 Data Polish Dashboard

This section is dedicated to the Data Polish Dashboard, an intuitive interface for data management and exploration. The dashboard is organized into three distinct sections:

Data Profile: Here, you can understand the structure and composition of your data.
Data Quality: This section is dedicated to evaluating the quality of your data, helping ensure accuracy and reliability.
Data Exploration: Explore and interact with your data in various ways to gain deeper insights.

3.1 Navigating the Dashboard:

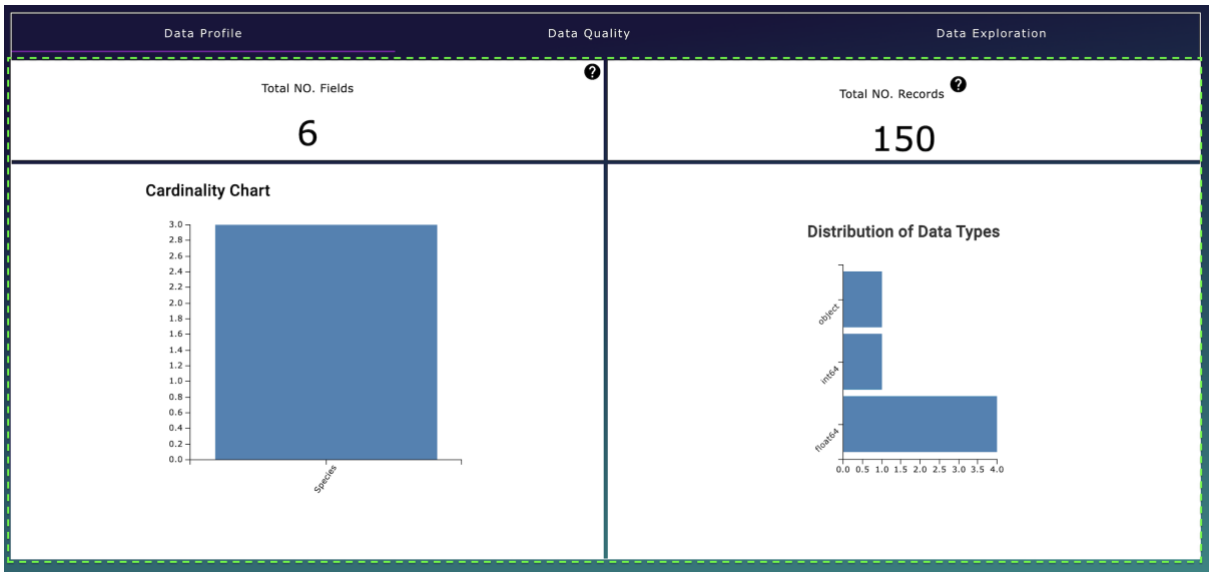
Easily navigate between these sections using the three tabs located at the top of the dashboard.
The tab you are currently viewing will be distinctly marked with a purple underline for easy identification.
Use these tabs to effortlessly switch between different aspects of your data, enhancing your data management experience.



3.2 Data Profile Tab

3.2.1 Overview:

The Data Profile helps you understand the fundamental structure of your data. It's designed to provide an overview of your data.



3.2.2

3.2.3 Key Features:

Record Count: View the total number of records in your dataset.

Field Count: View the total number of fields, providing a glimpse into the dataset's complexity.

Cardinality of Text Fields: Understand the uniqueness of text data by exploring the cardinality, indicating the diversity in text fields.

Data Type Distribution: Examine the distribution of data types within your dataset. This feature helps in understanding the proportion of different data types like numeric, text, dates, etc.

3.2.4 Visual Insights:

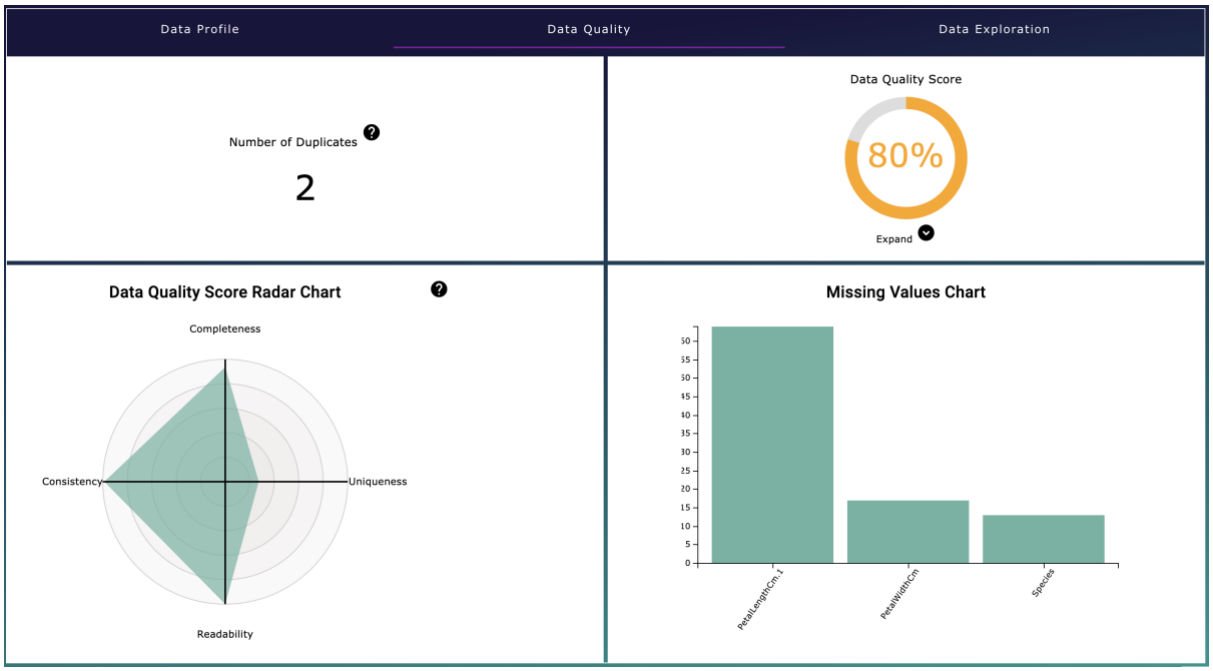
Below, you'll find a series of charts that visually represent these data attributes. These charts are designed to offer a clear and immediate understanding of your data's structure and composition.

Use the Data Profile tab to gain a deep and immediate understanding of the structure and nature of your data, setting the foundation for further data quality and exploration activities.

3.3 Data Quality tab

3.3.1 Overview

The Data Quality tab is designed to provide you with an understanding of the quality of your data. This tab is a powerful tool for identifying and guiding you to clean your data.



3.3.2 Key Features:

Duplicate Record Identification: Discover the number of duplicate records in your dataset.

Data Quality Score: Gives you a summarized view of your data's quality through a comprehensive score. This score is a quick indicator of the overall health of your data.

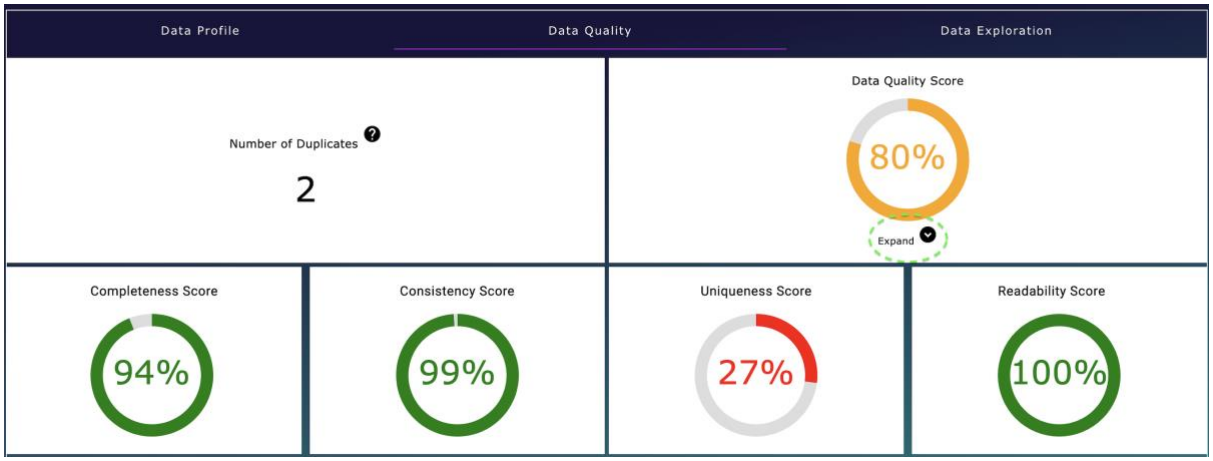
Missing Values Chart: Easily identify fields with missing values and quantify the extent of these gaps in your data.

3.3.3 Data Quality Score Attributes:

The Data Quality Score is broken down into four key attributes, each representing an aspect of data quality:

Completeness: Assessing whether all necessary data fields are filled.
Uniqueness: Evaluating the exclusivity of data entries.
Consistency: Ensuring that data formats and values are uniform across the dataset.
Readability: Determining the ease with which data can be read and understood by users and processing systems:

View these attributes in detail using the radar chart, providing a visual representation of your data's strengths and areas for improvement.
Alternatively, click 'Expand' under the Data Quality score metric for a more detailed breakdown, as illustrated below.



3.4 Data Exploration tab

3.4.1 Overview

The Data Exploration tab is an interactive and insightful component of your dashboard, designed to enable you to delve deep into your data and uncover insights. This tab is instrumental in visualizing and comprehending the complexities and relationships within your dataset.

3.4.2 Key Features:

Numerical Data Distribution: Examine the spread and behavior of your numerical data. This feature helps you understand the range, median, and common patterns in your numerical data.

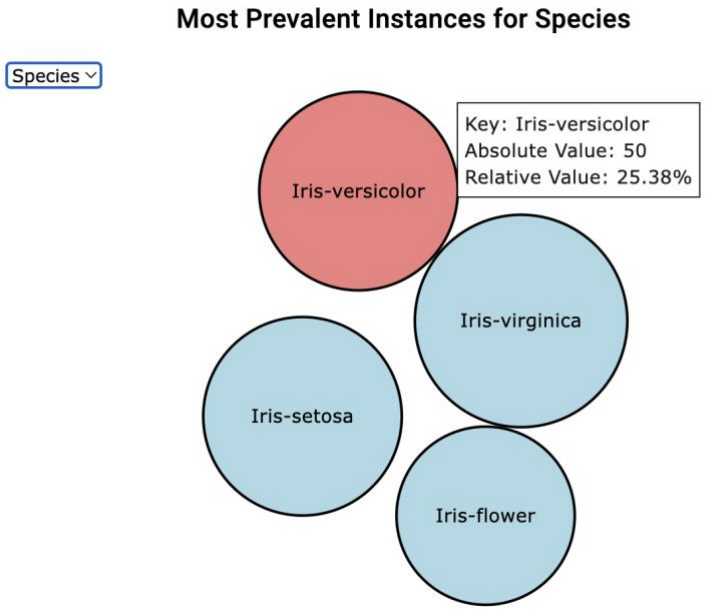
Correlation Discovery: Discover the relationships between various numerical variables. This is crucial for identifying trends and dependencies within your data.
Outlier Scatter Plot Visualization: Utilize scatter plots to visually identify and analyze outliers in your data. This tool is key in recognizing anomalies and unusual patterns.
Bubble Chart for Categorical Data: Dive deeper into your categorical data with the bubble chart. This interactive tool enables you to identify and understand the most prevalent values in a selected column, providing a clear view of categorical data distributions.

3.4.3 Interactivity and Insight:

The Data Exploration tab is not just about viewing data; it's about interacting with it. Each feature is designed to provide a hands-on experience, allowing you to engage with your data in a dynamic way.

Through these tools, you can effectively visualize complex data relationships, making it easier to extract meaningful insights and make informed decisions about your data cleaning.

3.5 Quality of Life Features in the Dashboard



3.5.1 Consistent Color Scheme:

Color-Blind Safe Palette: We have carefully selected colors that are accessible to individuals with color vision deficiencies. This ensures that all users can effectively interpret and enjoy our visualizations.

Unified Color Consistency: All visual elements adhere to a consistent color theme, resonating with the dashboard's blue hue. This consistency not only enhances aesthetic appeal but also aids in intuitive navigation.

3.5.2 Interactive Tooltips:

Enhanced Information: Non-metric visualizations are equipped with informative tooltips. These tooltips provide additional context and details, enriching your understanding of the data.

Selection Highlighting: Interactive elements respond to your cursor by highlighting selections. This feature allows for easy identification and navigation of choices within the dashboard.

3.5.3 Helpful Text:

Guidance and Context: Each visualization comes with explanatory help text. This text addresses two key questions: What is the visualization showing, and why is it important? This feature aids in comprehending the relevance and application of the data presented.

3.5.4

3.5.5 Smooth Transitions:

Feedback Through Animation: The dashboard includes subtle transitions, signaling that your data is being processed and analyzed. These animations provide a responsive and engaging user experience.

3.5.6 Dynamic Titles:

Contextual Awareness: Certain visualizations feature dynamic titles that adapt based on the fields being analyzed. This helps you quickly grasp the focus and context of each visualization.

3.5.7 Zoom Functionality in Bubble Chart:

Enhanced Navigation: The bubble chart is equipped with zoom in and zoom out capabilities. This functionality is especially useful when dealing with dense data, allowing you to adjust the view for a clearer understanding of all data points.

These quality-of-life features are designed to enhance your interaction with the dashboard, ensuring a user-friendly, accessible, and informative experience.

3.5.8 Navigating the Data Journey:

After getting a getting a deeper understanding of your data using the data polish dashboard click next on the bottom right to get to the next step, cleaning your data.

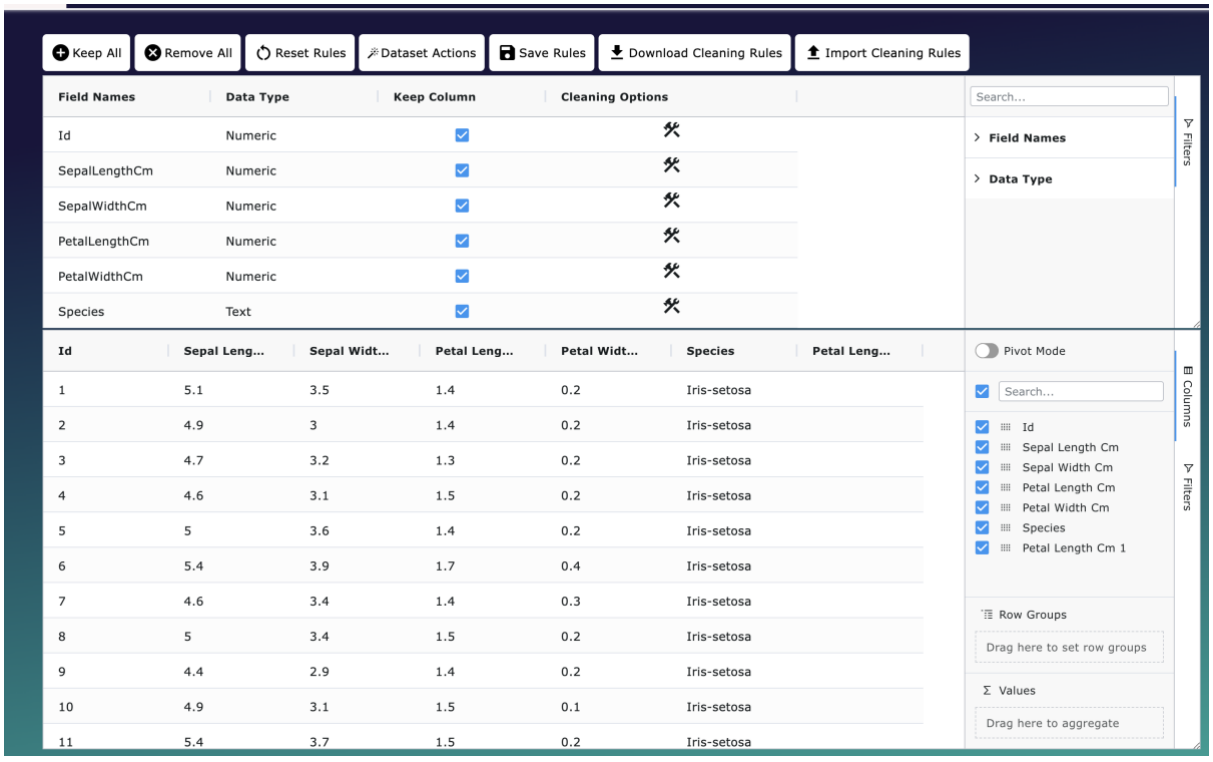
4 Data Cleaning

4.1 Overview

Welcome to the Data Cleaning Page, the fourth step in your data cleansing journey. This part of the manual provides a concise walkthrough of the page's features, from the top-left corner to the central cleaning functions.

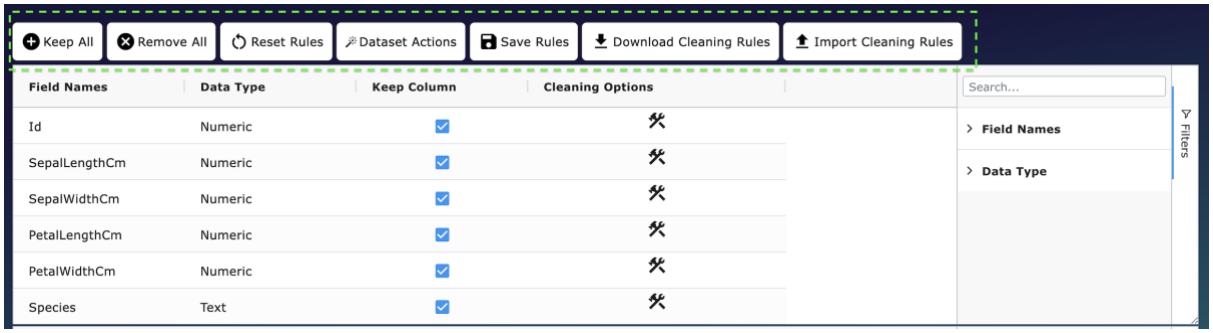
As you navigate this page, you'll find the layout familiar, especially the bottom section, which mirrors the Preview Page from step two. This design consistency ensures ease of use, allowing you to seamlessly continue your data cleaning process with familiar tools and previews at your disposal.

In the following sections, we'll guide you through each feature and function, ensuring you have all the necessary tools for efficient data cleaning and optimization.



4.2 Top section buttons:

The top section of the Data Cleaning Page features a series of buttons, each designed to streamline your data cleaning process.



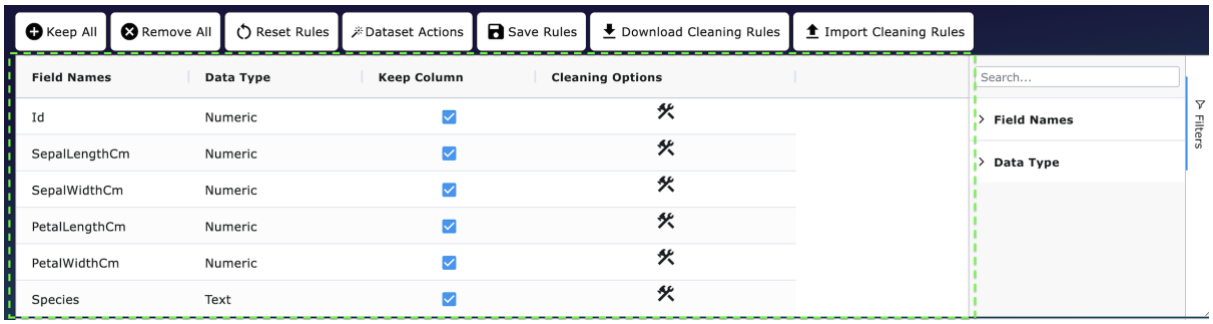
Here's a brief guide, moving from left to right:

1. **Keep All:** This button selects all checkboxes in the 'Keep' column, indicating that you wish to retain all fields.
2. **Remove All:** Clicking this will deselect all columns in the data cleaning checkboxes, effectively removing them from the dataset.
3. **Reset Rules**: Use this to revert all the changes you've made on the page, resetting your data cleaning actions to their original state.
4. **Dataset Actions:** This opens a menu for dataset-wide cleaning actions like 'Remove Duplicates.' While currently limited to this action, future updates will expand these options. Your feedback is welcome to help us improve this feature.
5. **Save Rules:** It's essential to save the rules you create for data cleaning, ensuring the actions are correctly applied. This button does just that.
6. **Download Cleaning Rules:** This feature enables you to download your data cleaning configurations. It's useful for reviewing changes on one page or reapplying the same rules later, ensuring consistency in data cleaning.
7. **Import Cleaning Rules:** If you have previously configured rules, use this button to upload them. It saves time and maintains consistent data cleaning across sessions.

Each of these buttons is an integral part of the data cleaning toolkit, designed to make your experience as efficient and user-friendly as possible.

4.3 Column Cleaning

Situated in the central area of the Data Cleaning Page is the Column Cleaning section. Here, you'll find a vertically arranged table listing each column of your dataset that can be cleaned.

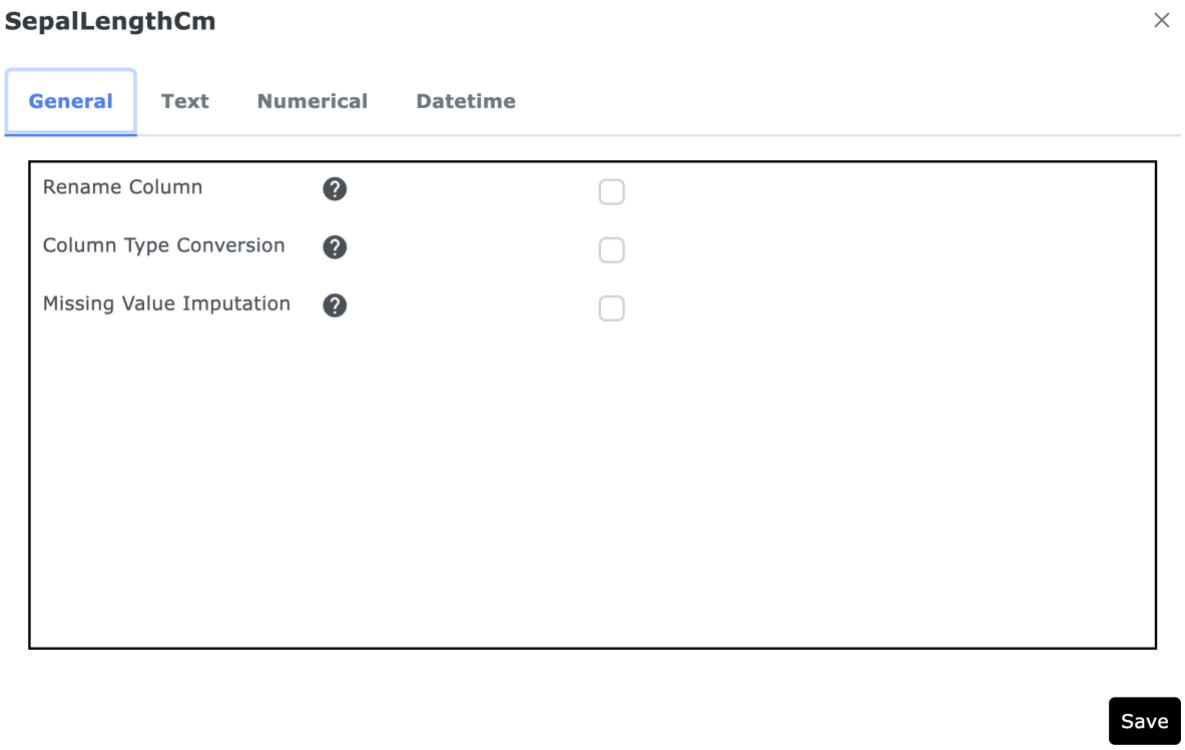


The table is organized as follows, from left to right:

1. **Field Name:** This column displays the name of the field in your dataset that you wish to clean.
2. **Data Type:** Here, the data type of each field is listed. This information is crucial as it dictates the cleaning options applicable to that field.
3. **Keep Column:** Next to each field name, there are checkboxes. Tick these to include the field in your final dataset or leave them unticked to exclude the field.
4. **Cleaning Options:** By clicking the cleaning icon in this column, you can access a variety of cleaning actions specifically tailored for your data.

Column Cleaning Actions Overview:

Clicking on the cleaning options icon triggers a menu, as depicted below. This menu categorizes the cleaning actions into four key sections: General, Text, Numerical, and Datetime.



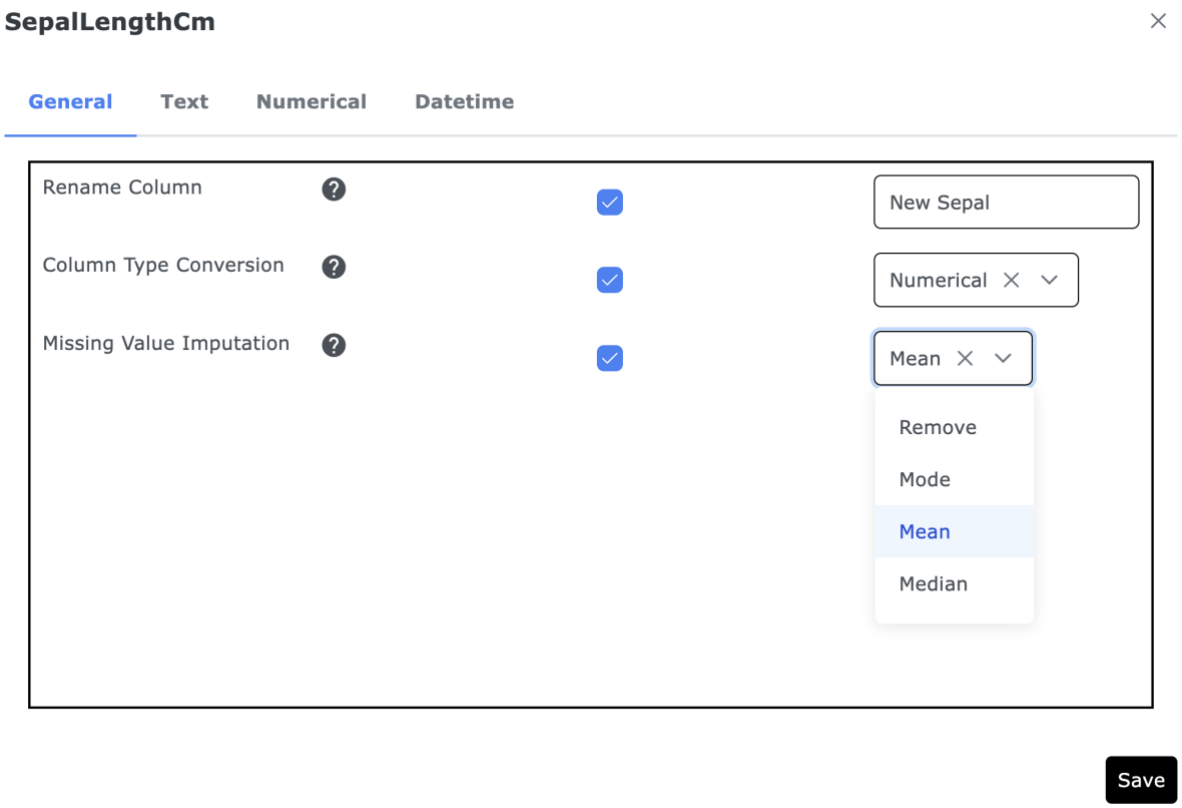
Each section contains relevant actions tailored to the data type of the column:

1. General Actions: Here, you can perform basic operations such as renaming the column or converting its data type.
2. Text Actions: This tab includes text-specific cleaning functions, allowing you to manipulate and refine text data.
3. Numerical Actions: For numerical data, you can access operations that are essential for number-based analytics.
4. Datetime Actions: This section provides options for adjusting date and time data for better chronological analysis.

Once you have selected and applied the desired cleaning actions, you can save your configurations using the 'Save' button located at the bottom right of the menu.
Each action is designed to be intuitive, ensuring a smooth and efficient data cleaning process.

4.3.1 General Actions in Column Cleaning

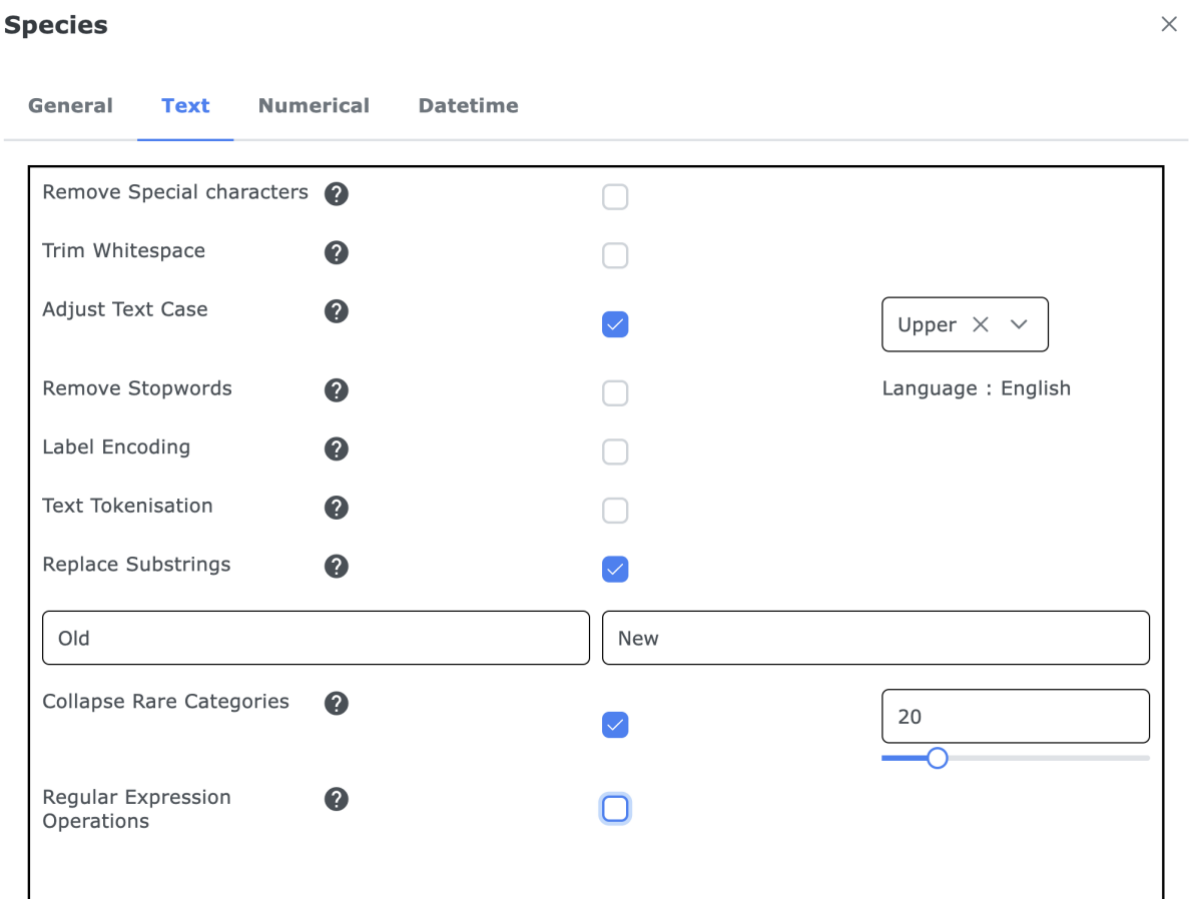
The General tab within the Column Cleaning menu offers three essential actions applicable to all column types, enhancing your data's integrity and usability:



1. **Rename Column:** When you select 'Rename Column,' a textbox will appear, allowing you to enter a new name for your column, providing the flexibility to align the column names with your data analysis requirements.
2. **Column Type Conversion:** Upon choosing 'Column Type Conversion,' a dropdown menu presents you with data type options, such as numeric or text. This functionality enables you to transform the column to better suit the data it represents.
3. **Missing Value Imputation:** By selecting 'Missing Value Imputation,' another dropdown menu offers you strategies for handling missing data, including options to fill in missing values with the mode, median, or to remove them entirely.

4.3.2 Text Actions in Column Cleaning

Within the Text tab of the Column Cleaning menu, users can access specialized actions to refine text data, ensuring enhanced readability and data quality for analysis. Use only with text type columns.



1. **Remove Special Characters:** Cleanses your text columns by eliminating any special characters, retaining only alphanumeric characters and spaces. This action improves clarity by removing extraneous symbols that could hinder readability or analysis.

2. **Trim Whitespace**: Automatically trims leading and trailing spaces from your text data. This essential step reduces noise, prevents processing errors, and maintains data quality.

3. **Adjust Text Case:** Allows you to standardize text to upper case, lower case, or title case, ensuring uniformity across your dataset and facilitating consistent data formatting.

4. **Remove Stopwords:** Strips common stopwords from text data, a crucial preprocessing step in NLP that streamlines data by eliminating frequently occurring, but often irrelevant words.

5. **Text Tokenization:** Splits text into individual words or tokens, preparing your text for further natural language processing or machine learning tasks.

6. **Replace Substrings:** Offers the ability to substitute specified segments of text with new ones, aiding in the standardization and consistency of textual data.

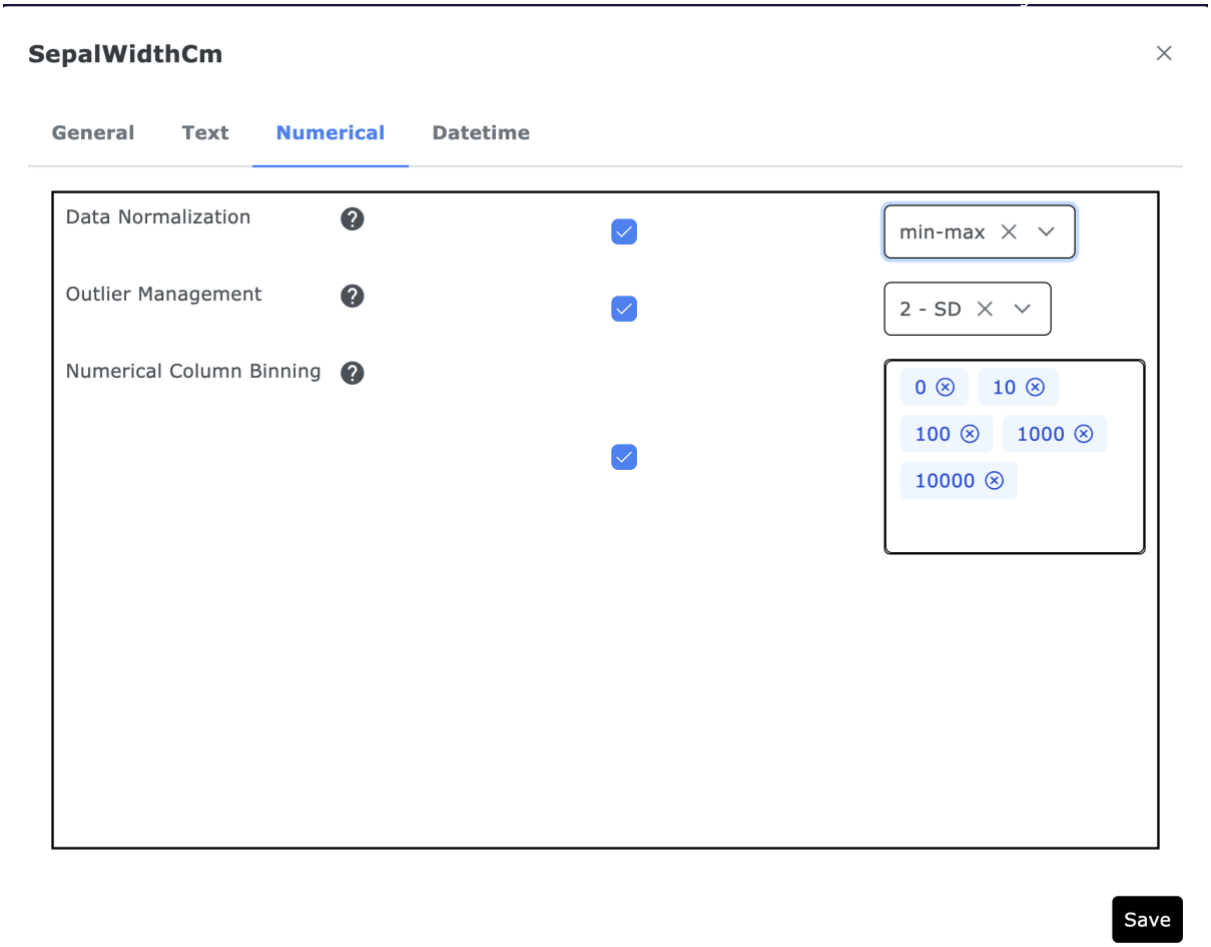
7. **Collapse Rare Categories:** Groups infrequent categories within your data into a single 'Other' category, simplifying your dataset and potentially enhancing analysis and machine learning model performance.

8. **Regular Expression Operations:** Applies complex regex patterns to replace or extract specific strings, allowing for sophisticated text transformations.

9. **Label Encoding:** Transforms categorical text into numerical labels, making the data suitable for machine learning models that require numerical input.

4.3.3 Numerical Actions in Column Cleaning

The Numerical tab in the Column Cleaning menu presents a suite of actions specifically tailored for numerical data manipulation, enhancing data suitability for analysis and machine learning algorithms. Use only with numeric type columns.



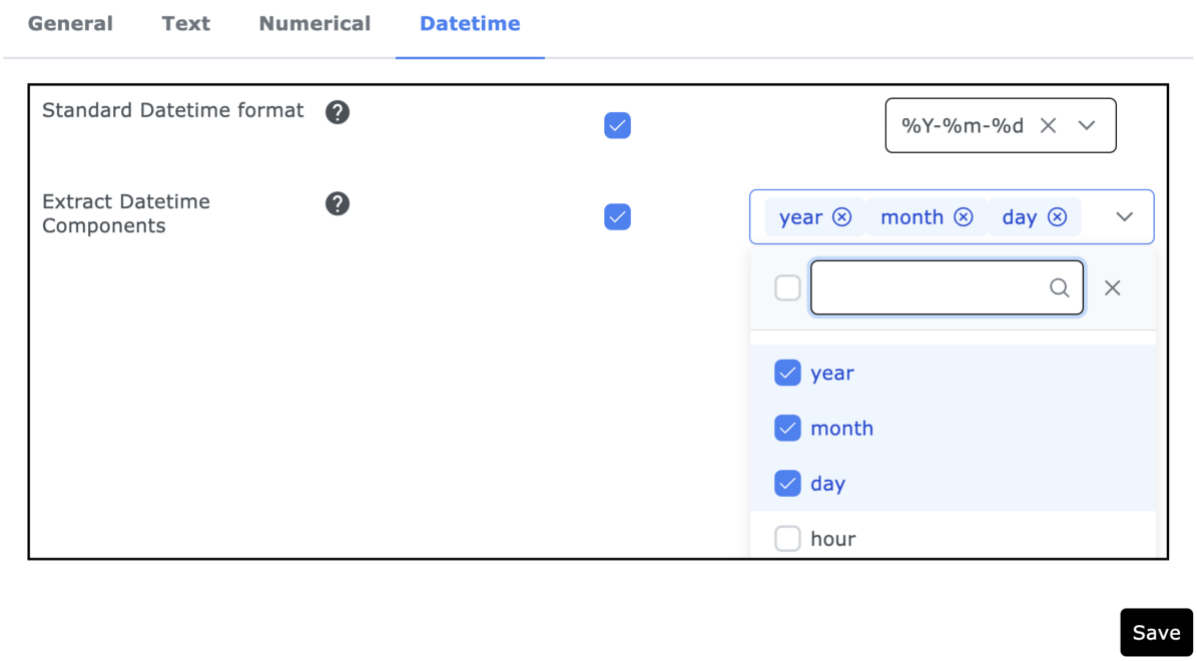
1. **Data Normalization:** This function allows you to standardize data within a column through min-max scaling or z-score normalization. Normalization is vital for harmonizing the scale of numerical data, which can significantly improve the performance and accuracy of various machine learning models.

2. **Outlier Management:** Outliers can skew the results of data analysis and model training. This action lets you mitigate their impact by removing them based on standard deviation criteria. It's particularly useful for algorithms that are sensitive to data distribution, ensuring a more representative dataset.

3. **Numerical Column Binning**: Transform numeric data into categorical by binning values into predefined ranges. Binning helps to reveal non-linear relationships within the data and is especially beneficial for linear models that presume linear feature-to-target relationships.

4.3.4 Datetime cleaning actions

In the Datetime tab of the Column Cleaning menu, users can find actions tailored to manage and optimize date and time data for enhanced analysis and modeling. Use only with datetime type columns.



1. **Standard Datetime Format:** This feature enables the standardization of date and time values by parsing and converting them into a uniform datetime format. This standardization is crucial for models that rely on temporal data, such as time series forecasting models, ensuring that the date and time information is precise and consistently formatted.

2. **Extract Datetime Components:** With this action, you can decompose a datetime object into its constituent parts—such as year, month, and day—and add them as separate new columns to your dataset. This breakdown facilitates a more granular level of analysis, allowing for the examination of patterns and trends related to specific time components.

4.4 Quality of Life Features on the Data Cleaning Page

Logical Layout: The data cleaning page is laid out in a logical flow, with actions organized from top to bottom, reflecting the user's cleaning journey for straightforward navigation.

Interactive Tooltips:

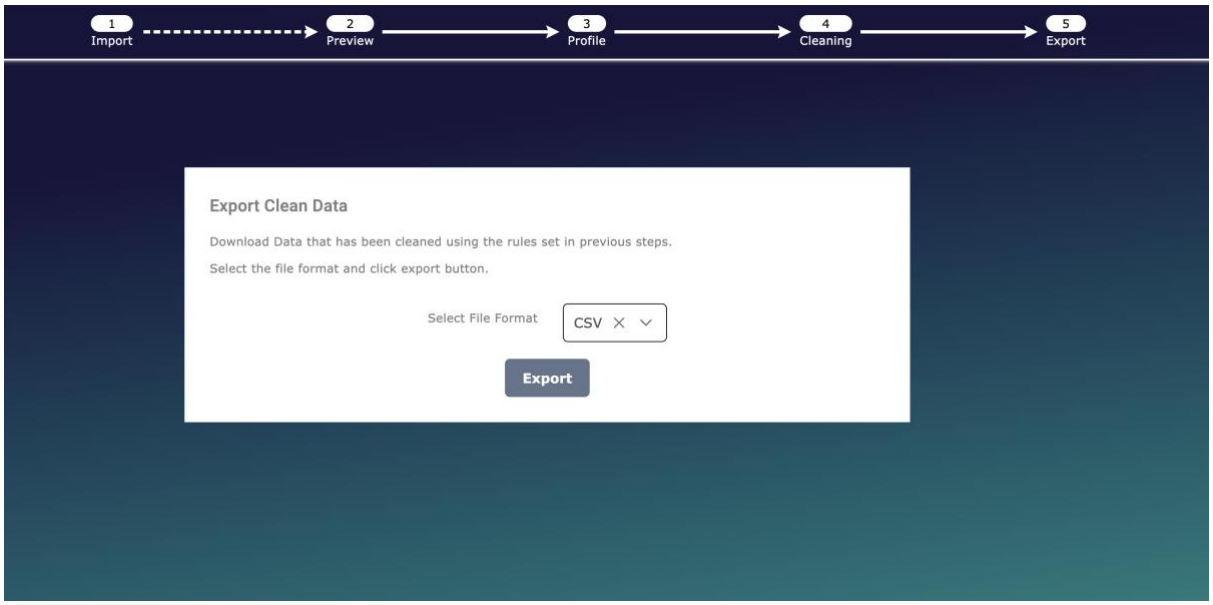
Informative Help Text: Every function on the page, from field selection to rule application, is equipped with interactive tooltips that emerge upon hover, offering explanations and helpful tips to enhance the user's understanding and efficiency.

Bulk Editing Tools: 'Keep All' and 'Remove All' buttons enable quick, comprehensive edits to the dataset columns, facilitating rapid data manipulation.

Rule Management: Users can easily save, download, or import cleaning rules, which streamlines repetitive tasks and ensures consistency across different datasets or projects.

5 Downloading your Clean Data

Congratulations you have now reached the final step of your data cleaning journey with DataPolish.



5.1 Data Export:

Select File Format: Determine the format for your cleaned data by choosing from the available options in the dropdown menu—CSV or Excel—to best suit your project's needs or personal preference.

Downloading: Finalize your data cleaning process by clicking the 'Export' button. Your data, now polished according to the rules set in the previous steps, will be prepared for download.

Your download will begin immediately, providing you with a clean dataset that is ready for analysis, sharing, or integration into your next workflow stage.