USER MANUAL

DataCleaningTool

Masters Thesis Student: Devosmita Chatterjee [1]
Department of Mathematical Sciences, Chalmers University of Technology

Industrial Supervisor: Sven Ahlinder [2]
Powertrain Department, Volvo Group Trucks Technology

Academic Supervisor: Anton Johansson [3]
Department of Mathematical Sciences, Chalmers University of Technology

[1]E.mail: chatterjeedevosmita267@gmail.com
[2]E.mail: svenahlinder@gmail.com
[3]E.mail: johaant@chalmers.se

Contents

1 Overview **2**

2 App Installation **4**

3 Getting Started **5**
3.1 Import Data with Features in Columns Button . . . 7

4 Data Cleaning Widgets **8**
4.1 Current Data Widget . . . 8
4.2 Data Properties Widget . . . 9
4.2.1 Id Button . . . 10
4.2.2 Feature Names Button . . . 13
4.2.3 Change Case Button . . . 16
4.2.4 Remove Extra Space Button . . . 20
4.2.5 Delete Rows Button . . . 28
4.2.6 Sort Features Button . . . 31
4.2.7 Delete Feature Button . . . 33
4.3 Numerical Features Widget . . . 36
4.3.1 Numerical Feature Cell Selection Button . . . 36
4.3.2 Remove Observations Button . . . 38
4.3.3 Delete Rows Button . . . 41
4.4 Datetime Features Widget . . . 44
4.4.1 Datetime Feature Cell Selection Button . . . 44
4.4.2 Convert To Excel DATEVALUE Button . . . 46
4.4.3 Change Format Button . . . 46
4.4.4 Remove Observations Button . . . 49
4.4.5 Delete Rows Button . . . 49
4.5 Text Features Widget . . . 52
4.5.1 Select Similar Categories Button . . . 53
4.5.2 Text Feature Cell Selection Button . . . 56
4.5.3 Label Encoding Button . . . 58
4.5.4 One Hot Encoding Button . . . 61
4.5.5 Remove Observations Button . . . 65
4.5.6 Delete Rows Button . . . 65
4.6 Imputation Widget . . . 66
4.6.1 Delete Feature Button . . . 66
4.6.2 Impute Button . . . 69
4.7 Data Transformation Widget . . . 71
4.7.1 Transform Button . . . 72
4.8 Save Data . . . 75
4.8.1 Save Button . . . 75
4.9 Results . . . 78
4.9.1 Generate Report Button . . . 78

5 Other Attributes **81**
5.1 Resize Button . . . 81
5.2 Undo Button . . . 81
5.3 Help Button . . . 81

Chapter 1

Overview

Presently, large amount of data generated by organizations drives its business decisions. The data is usually inconsistent, inaccurate and incomplete. Poor data quality may lead to incorrect decisions for the organizations and hence, negatively affect organizations. Thus, high quality data is of utmost priority to use the data effectively. Data cleaning is the ultimate way to solve the data quality issues. But, data cleaning is really a time consuming task. Thus, tools which can help with the task are needed. This demands data cleaning tools for systematically examining data for errors and automatically cleaning them using algorithms. These data cleaning tools help organizations save time and increase their efficiency.

DataCleaningTool is a user friendly, free and open source data cleaning standalone application developed to achieve the task of data cleaning in a cooperative way. This application is able to identify the potential data problems and report results and recommendations such that users can clean data effectively with its assistance. The major data problems encountered by DataCleaningTool and the possible approaches to fix them are as follows.

Incorrect data type

- Example: Numerical instead of string entries.
- Possible Approach: Set data type constraint.

Inconsistent feature names or columns

- Example: Feature names or columns have inconsistent capitalizations.
- Possible Approach: Use uppercase or lowercase characters.

Typographical errors

- Example: Extra white spaces.
- Possible Approach: Remove extra white spaces.

Nonsensical data

- Example: Age = -1.
- Possible Approach: Set range constraint to variable - Age ≥ 0.

Extrapolation errors

- Example: A model of glacial retreat: $V = 100 - 2t$ where V = volume of ice, t = time variable, and $t = 0$ AD. If we extrapolate to earlier than $t = 0$, then ice volume becomes bigger. Mathematically, we can extrapolate back in time but then the ice volume of the glacier would exceed the total volume of the earth which is absurd.
- Possible Approach: Set range constraint to variable - $t \geq 0$.

Truncation error (Volvo)

- Example: Difference between the actual value (2.99792458×10^8) and the truncated value up to two decimals (2.99×10^8).
- Possible Approach: Use long format [1].

Time stamp errors (Volvo)

- Example: The first failure time can show time prior to when the electric vehicles were produced if the vehicle clock has not been correctly set.
- Possible Approach: Set cross-field validation constraint to variable - first failure time of a vehicle $>$ time when the vehicle was produced.

Fault code count (Volvo)

- Example: Fault codes stored by the on-board computer diagnostic system notify about a problem found in the car. Sometimes although an issue is notified, failure count $= 0$.
- Possible Approach: Set range constraint to variable - Failure count > 0.

Missing data

- Example: NaN or ‘ ’.
- Possible Approach: Imputation using MissForest method. [2].

Outliers

- Example: Fraudulent credit card transactions.
- Possible Approach: Z-score [3].

Chapter 2

App Installation

DataCleaningTool is a standalone application that can run on Windows platform. DataCleaningTool is a standalone application created from Matlab functions so that it can be used to run Matlab compiled program on computers that do not have Matlab installed. The Matlab Compiler Runtime enables to run standalone application compiled within Matlab.

DataCleaningTool has been developed and tested in MATLAB Version: R2018b and requires the following toolboxes: System Identification Toolbox, Statistics and Machine Learning Toolbox, Financial Toolbox, and MATLAB Report Generator. The DataCleaningTool app installation package can be downloaded from the github repository https://github.com/devosmitachatterjee2018/DataCleaningTool. The following steps show how to install the DataCleaningTool application from the repository.

Step 1. Download 'Standalone Desktop App/for_redistribution.zip' and unzip it to a preferred location.

Step 2. Run the executable file "DataCleaningTool.exe" and follow instructions. If not already present, the MATLAB Compiler Runtime (mcr) R2018b will be downloaded from the web and installed automatically.

Step 3. Once installed, the app is added to the Start Menu in Windows.

Step 4. Click the app icon to run the program.

Chapter 3

Getting Started

DataCleaningTool is a data cleaning application which consists of multiple widgets and buttons. DataCleaningTool is shown in figure 3.1. The properties of DataCleaningTool are

- DataCleaningTool always opens in a full screen mode. The application can be resized to a reduced size.
- Each widget provides specific statistical information about the data.
- Each button aims to clean data by resolving inconsistencies, smoothing noisy data, identifying outliers, removing outliers or filling in missing observations.
- All buttons are black in color. Pressing a button each time changes the button color from black to grey color and then again to black. The button remains grey in color until it completes its specific task and all widgets gets updated accordingly.
- Pressing any button turns the Undo button to blue color. The Undo button remains blue in color until last activity can be undone.
- Sliders and their corresponding edit boxes are interdependable.
- User can find help in using DataCleaningTool.

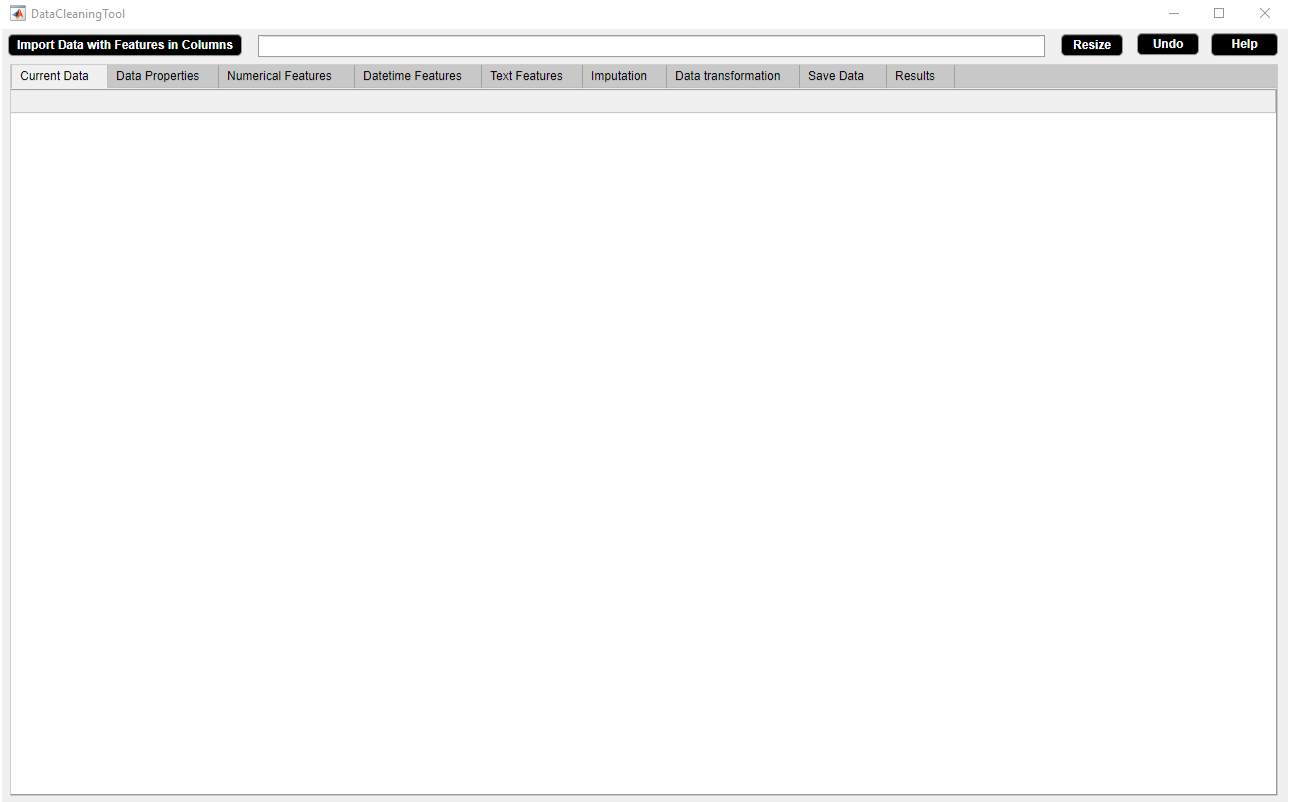


Figure 3.1: DataCleaningTool.

We demonstrate the DataCleaningTool using an example dataset 'demodata.csv'. The example dataset is obtained by tweaking the coronavirus dataset [4]. The example dataset is of dimension 127×12. The example dataset consists of the following features.

1. Serial_Number: Unique identifier to a country.
2. Country_Region: Name of the country.
3. Population_Size: Size of the population of the country.
4. tourism: Number of international arrivals in the country.
5. Date_FirstFatality: Date of the first fatality in the country.
6. Date_FirstConfirmedCase: Date of the first confirmed case in the country.
7. Latitude: Geographic coordinate of the country.
8. Longtitude: Geographic coordinate of the country.
9. mean_Age: Mean age of the population of the country.
10. Lockdown_Date: Date of the lockdown in the country.
11. Lockdown_Type: Level of the lockdown (full or partial) in the country.
12. Country_Code: Geographical code representing the country.

Using the example dataset, we will show the steps how to clean data using the DataCleaningTool. First we wish to understand our data by doing a descriptive statistics analysis of our dataset. In Descriptive Statistics, we are describing and summarizing our data, either through numerical calculations or graphs. Secondly we distinguish id feature 'Serial_Number' from other numerical features. Next we detect inconsistent capitalization of feature names such as 'Serial_Number', 'Country_Region', 'Population_Size', 'tourism', 'Date_FirstFatality', 'Date_FirstConfirmedCase', 'Latitude', 'Longtitude', 'mean_Age', 'Lockdown_Date', 'Lockdown_Type', 'Country_Code' and unify inconsistent capitalization of feature names. Then we wish to extract data for the countries whose 'Population_Size' is greater than 'Tourism'. So we set cross-field validation constraint to remove irrelevant observations. Then we wish to extract data for the countries whose maximum 'Mean_Age' is 45. So we set the range constraint to remove irrelevant observations. We delete feature 'Longitude' since it contains a large percentage of missing observations. We illustrate missing observations by missingness plot and impute missing observations using missForest method. Lastly, we log transform the numerical feature 'Population_Size' which makes the feature less skewed.

3.1 Import Data with Features in Columns Button

Loads data from comma-separated (.csv), Excel (.xlsx), tab-delimited (.txt), data (.dat) files and then reads the data into table.

Application

- Reduce truncation errors upto 15 decimal places using long decimal format.

Example

Step 1: Click **Import Data with Features in Columns** button.

Step 2: **Import Data with Features in Columns** button in use turns grey in color and an open dialog box appears. Browse for an input file.

Step 3: **Import Data with Features in Columns** button returns back to its original color once it completes its task. The full path of the selected file is displayed and the file is loaded.

We use **Import Data with Features in Columns** button to load the example data 'demodata.csv'. Figures 3.2-3.4 illustrate how to use **Import Data with Features in Columns** button.

Figure 3.2: Step 1. Import Data with Features in Columns Button

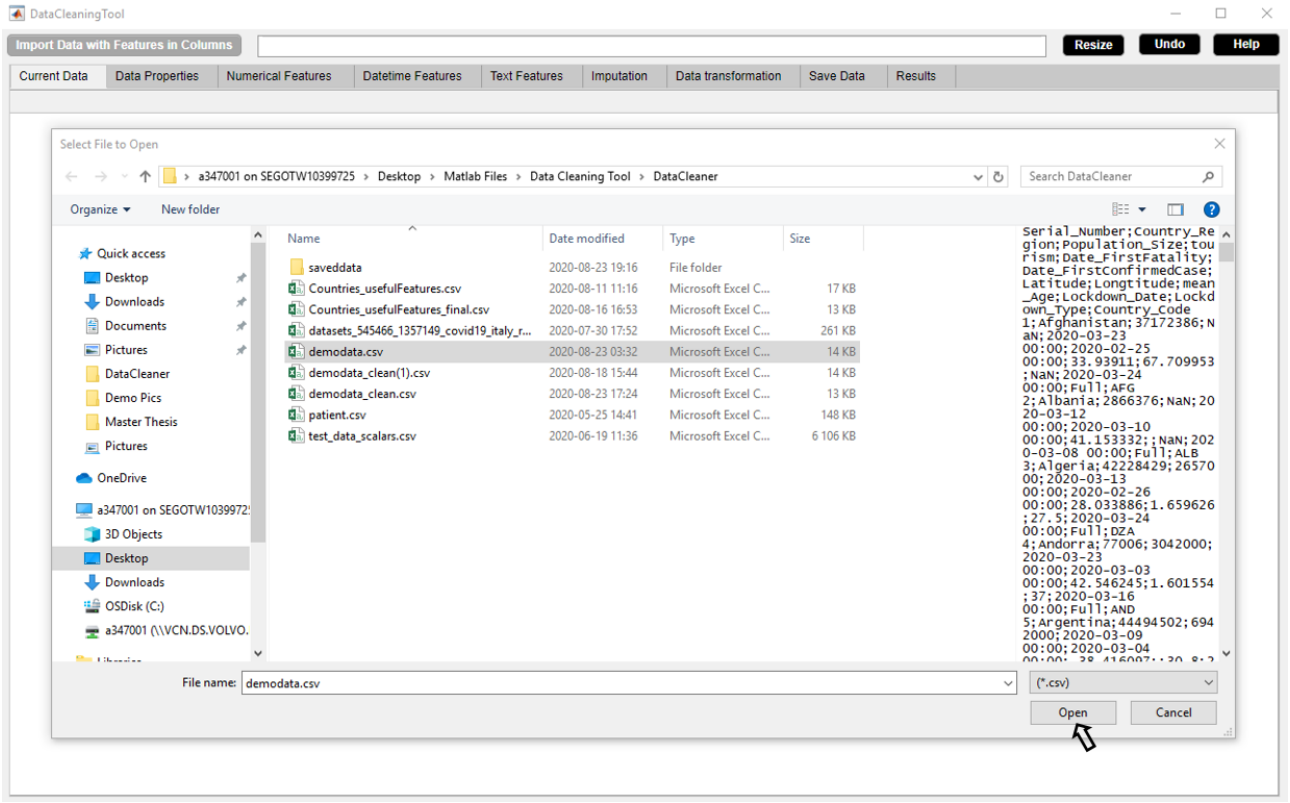


Figure 3.3: Step 2. Import Data with Features in Columns Button

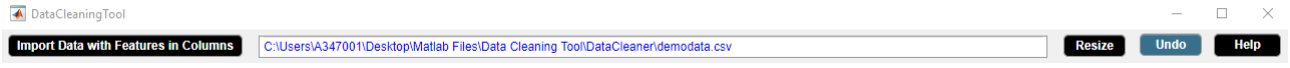


Figure 3.4: Step 3. Import Data with Features in Columns Button

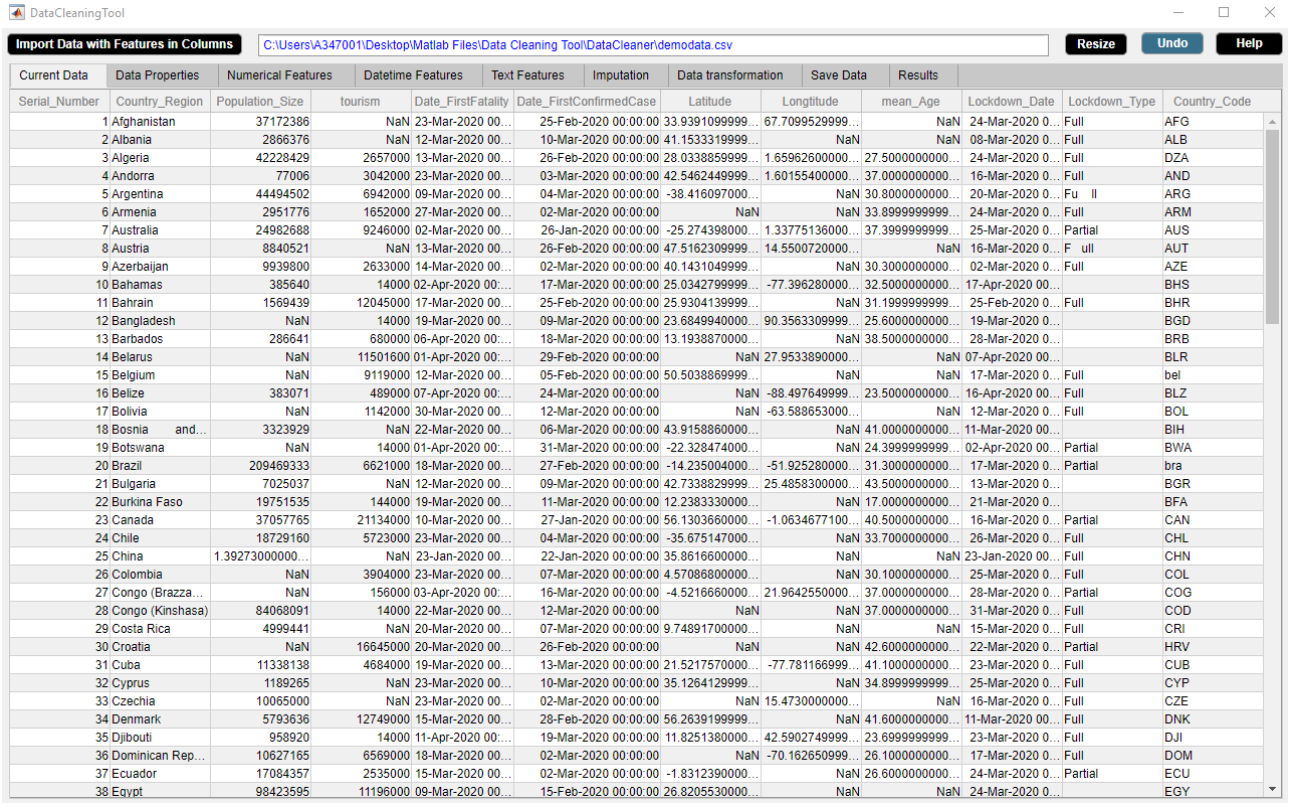
Chapter 4

Data Cleaning Widgets

4.1 Current Data Widget

The **Current Data** widget displays the input data in table format. The **Current Data** widget is shown in figure 4.1. The properties of the Current Data widget are as follows.

- The widget shows the presence of round off errors in numerical features.
- The widget shows the presence of inconsistent capitalization of feature names and features.
- The widget shows the existence of extra whitespaces in text features.
- Default datetime format is 'dd-MMM-yyyy HH:mm:ss' for datetime features.
- The widget shows the presence of missing numerical observations represented by NaNs.
- The widget shows the presence of missing datetime observations represented by NaTs.
- The widget shows the presence of missing text observations represented by empty strings.
- The updated table can be visualized after each activity since the widget gets updated accordingly.



Serial_Number	Country_Region	Population_Size	tourism	Date_FirstFatality	Date_FirstConfirmedCase	Latitude	Longtitude	mean_Age	Lockdown_Date	Lockdown_Type	Country_Code
1	Afghanistan	37172386	NaN	23-Mar-2020 00...	25-Feb-2020 00:00:00	33.9391099999...	67.7099529999...	NaN	24-Mar-2020 0...	Full	AFG
2	Albania	2866376	NaN	12-Mar-2020 00...	10-Mar-2020 00:00:00	41.1533319999...	NaN	NaN	08-Mar-2020 0...	Full	ALB
3	Algeria	42228429	2657000	13-Mar-2020 00...	26-Feb-2020 00:00:00	28.0338859999...	1.65962600000...	27.5000000000...	24-Mar-2020 0...	Full	DZA
4	Andorra	77006	3042000	23-Mar-2020 00...	03-Mar-2020 00:00:00	42.5462449999...	1.60155400000...	37.0000000000...	16-Mar-2020 0...	Full	AND
5	Argentina	44494502	6942000	09-Mar-2020 00...	04-Mar-2020 00:00:00	-38.416097000...	NaN	30.8000000000...	20-Mar-2020 0...	Fu ll	ARG
6	Armenia	2951776	1652000	27-Mar-2020 00...	02-Mar-2020 00:00:00	NaN	NaN	33.8999999999...	24-Mar-2020 0...	Full	ARM
7	Australia	24982688	9246000	02-Mar-2020 00...	26-Jan-2020 00:00:00	-25.274398000...	1.33775136000...	37.3999999999...	25-Mar-2020 0...	Partial	AUS
8	Austria	8840521	NaN	13-Mar-2020 00...	26-Feb-2020 00:00:00	47.5162309999...	14.5500720000...	NaN	16-Mar-2020 0...	F ull	AUT
9	Azerbaijan	9939800	2633000	14-Mar-2020 00...	02-Mar-2020 00:00:00	40.1431049999...	NaN	30.3000000000...	02-Mar-2020 0...	Full	AZE
10	Bahamas	385640	14000	02-Apr-2020 00:...	17-Mar-2020 00:00:00	25.0342799999...	-77.396280000...	32.5000000000...	17-Apr-2020 00...		BHS
11	Bahrain	1569439	12045000	17-Mar-2020 00...	25-Feb-2020 00:00:00	25.9304139999...	NaN	31.1999999999...	25-Feb-2020 0...	Full	BHR
12	Bangladesh	NaN	14000	19-Mar-2020 00...	09-Mar-2020 00:00:00	23.6849940000...	90.3563309999...	25.6000000000...	19-Mar-2020 0...		BGD
13	Barbados	286641	680000	06-Apr-2020 00:...	18-Mar-2020 00:00:00	13.1938870000...	NaN	38.5000000000...	28-Mar-2020 0...		BRB
14	Belarus	NaN	11501600	01-Apr-2020 00:...	29-Feb-2020 00:00:00	NaN	27.9533890000...	NaN	07-Apr-2020 00...		BLR
15	Belgium	NaN	9119000	12-Mar-2020 00...	05-Feb-2020 00:00:00	50.5038869999...	NaN	NaN	17-Mar-2020 0...	Full	bel
16	Belize	383071	489000	07-Apr-2020 00:...	24-Mar-2020 00:00:00	NaN	-88.497649999...	23.5000000000...	16-Apr-2020 00...	Full	BLZ
17	Bolivia	NaN	1142000	30-Mar-2020 00...	12-Mar-2020 00:00:00	NaN	-63.588653000...	NaN	12-Mar-2020 0...	Full	BOL
18	Bosnia and...	3323929	NaN	22-Mar-2020 00...	06-Mar-2020 00:00:00	43.9158860000...	NaN	41.0000000000...	11-Mar-2020 00...		BIH
19	Botswana	NaN	14000	01-Apr-2020 00:...	31-Mar-2020 00:00:00	-22.328474000...	NaN	24.3999999999...	02-Apr-2020 00...	Partial	BWA
20	Brazil	209469333	6621000	18-Mar-2020 00...	27-Feb-2020 00:00:00	-14.235004000...	-51.925280000...	31.3000000000...	17-Mar-2020 0...	Partial	bra
21	Bulgaria	7025037	NaN	12-Mar-2020 00...	09-Mar-2020 00:00:00	42.7338829999...	25.4858300000...	43.5000000000...	13-Mar-2020 0...		BGR
22	Burkina Faso	19751535	144000	19-Mar-2020 00...	11-Mar-2020 00:00:00	12.2383330000...	NaN	17.0000000000...	21-Mar-2020 0...		BFA
23	Canada	37057765	21134000	10-Mar-2020 00...	27-Jan-2020 00:00:00	56.1303660000...	-1.0634677100...	40.5000000000...	16-Mar-2020 0...	Partial	CAN
24	Chile	18729160	5723000	23-Mar-2020 00...	04-Mar-2020 00:00:00	-35.675147000...	NaN	33.7000000000...	26-Mar-2020 0...	Full	CHL
25	China	1.39273000000...	NaN	23-Jan-2020 00...	22-Jan-2020 00:00:00	35.8616600000...	NaN	NaN	23-Jan-2020 00...	Full	CHN
26	Colombia	NaN	3904000	23-Mar-2020 00...	07-Mar-2020 00:00:00	4.57086800000...	NaN	30.1000000000...	25-Mar-2020 0...	Full	COL
27	Congo (Brazza...	NaN	156000	03-Apr-2020 00:...	16-Mar-2020 00:00:00	-4.5216660000...	21.9642550000...	37.0000000000...	28-Mar-2020 0...	Partial	COG
28	Congo (Kinshasa)	84068091	14000	22-Mar-2020 00...	12-Mar-2020 00:00:00	NaN	NaN	37.0000000000...	31-Mar-2020 0...	Full	COD
29	Costa Rica	4999441	NaN	20-Mar-2020 00...	07-Mar-2020 00:00:00	9.74891700000...	NaN	NaN	15-Mar-2020 0...	Full	CRI
30	Croatia	NaN	16645000	20-Mar-2020 00...	26-Feb-2020 00:00:00	NaN	NaN	42.6000000000...	22-Mar-2020 0...	Partial	HRV
31	Cuba	11338138	4684000	19-Mar-2020 00...	13-Mar-2020 00:00:00	21.5217570000...	-77.781166999...	41.1000000000...	23-Mar-2020 0...	Full	CUB
32	Cyprus	1189265	NaN	23-Mar-2020 00...	10-Mar-2020 00:00:00	35.1264129999...	NaN	34.8999999999...	25-Mar-2020 0...	Full	CYP
33	Czechia	10065000	NaN	23-Mar-2020 00...	02-Mar-2020 00:00:00	NaN	15.4730000000...	NaN	16-Mar-2020 0...	Full	CZE
34	Denmark	5793636	12749000	15-Mar-2020 00...	28-Feb-2020 00:00:00	56.2639199999...	NaN	41.6000000000...	11-Mar-2020 00...	Full	DNK
35	Djibouti	958920	14000	11-Apr-2020 00:...	19-Mar-2020 00:00:00	11.8251380000...	42.5902749999...	23.6999999999...	23-Mar-2020 0...	Full	DJI
36	Dominican Rep...	10627165	6569000	18-Mar-2020 00...	02-Mar-2020 00:00:00	NaN	-70.162650999...	26.1000000000...	17-Mar-2020 0...	Full	DOM
37	Ecuador	17084357	2535000	15-Mar-2020 00...	02-Mar-2020 00:00:00	-1.8312390000...	NaN	26.6000000000...	24-Mar-2020 0...	Partial	ECU
38	Egypt	98423595	11196000	09-Mar-2020 00...	15-Feb-2020 00:00:00	26.8205530000...	NaN	NaN	24-Mar-2020 0...		EGY

Figure 4.1: Current Data Widget.

4.2 Data Properties Widget

The Data Properties widget displays several statistical aspects of the data. The Data Properties widget is shown in figure 4.2. The properties of the Data Properties widget are as follows.

- The widget automatically discovers the datatypes of features of the input data set and shows the numerical features, the datetime features and the text features separately.
- The widget summarizes the characteristics of a data set such as file size in megabytes, number of rows and columns, number of id, numerical, datetime and text features, number of duplicate rows and columns, and number of deleted rows and columns.
- The widget shows the percentage of missing observations in the data set and the percentage of missing observations in each feature. The widget presents two visual methods for missing data - the missingness plot and the missing observations percentage plot. The missingness plot indicates the missing value occurence in the data. The missing observations percentage plot indicates the percentage of missing observations in each feature. This study of missing data helps to determine the missing data mechanism and hence choose strategies like listwise deletion, pairwise deletion, dropping features, imputation which can be applied to handle missing data so that they can be used for analysis and modelling.
- The information in the widget gets updated after each activity.

Figure 4.2: Data Properties Widget.

4.2.1 Id Button

Separates id features from numerical or datetime or text features. Here id feature represents a unique identifier field in the data.

Application

- Avoid overfitting problem which occurs due to a unique identifier among features.

Example

Step 1: Select a feature from **Numerical Feature** or **Datetime Feature** or **Text Feature** list box in the **Data Properties** widget.

Step 2: Click **Id** button.

Step 3: **Id** button in use turns grey in color.

Step 4: **Id** button returns back to its original color once it completes its task.

In the example data, Serial_Number represents unique identifier to a country. We use **Id** button to seperate id feature 'Serial_Number' from numerical features. Figures 4.3-4.6 illustrate how to use **Id** button.

Figure 4.3: Step 1. Id Button

Figure 4.4: Step 2. Id Button

Figure 4.5: Step 3. Id Button

Figure 4.6: Step 4. Id Button

4.2.2 Feature Names Button

Changes letter case of all feature names to one of the cases - lower case or upper case or capitalized case.

Application

- Fix structural errors such as unify inconsistent capitalization of feature names.

Example

Step 1: Check if there is any inconsistency in feature names capitalization.

Step 2: Select case from **Feature Names** dropdown menu.

Step 3: Click **Feature Names** button.

Step 4: **Feature Names** button in use turns grey in color.

Step 5: **Feature Names** button returns back to its original color once it completes its task.

In the example data, the feature names 'Serial_Number', 'Country_Region', 'Population_Size', 'tourism', 'Date_FirstFatality', 'Date_FirstConfirmedCase', 'Latitude', 'Longtitude', 'mean_Age', 'Lockdown_Date', 'Lockdown_Type', and 'Country_Code' have inconsistent capitalization. We use **Feature Names** button to capitalize first letter of each feature name so as to unify inconsistent capitalization of feature names. Figures 4.7-4.11 illustrate how to use **Feature Names** button.

Figure 4.7: Step 1. Feature Names Button

Figure 4.8: Step 2. Feature Names Button

Figure 4.9: Step 3. Feature Names Button

Figure 4.10: Step 4. Feature Names Button

Figure 4.11: Step 5. Feature Names Button

4.2.3 Change Case Button

Change letter case of a feature to one of the cases- lower case or upper case or capitalized case.

Application

- Fix structural errors such as unify inconsistent capitalization of a feature column.

Example

Step 1: Check if there is any inconsistency in feature capitalization in the **Current Data** widget.

Step 2: Select case from **Change Case** dropdown menu.

Step 3: Select the inconsistent feature from **Numerical Feature** or **Datetime Feature** or **Text Feature** list box in the **Data Properties** widget.

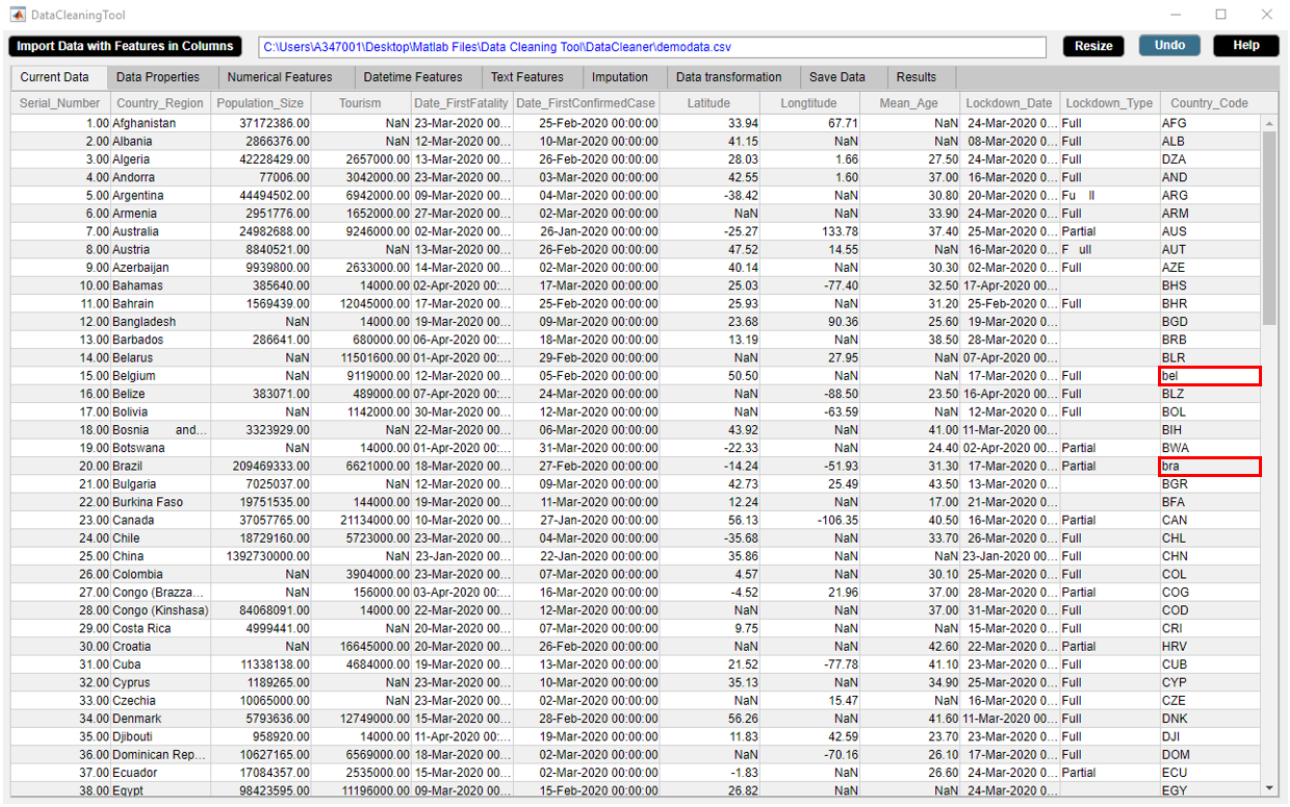
Step 4: Click **Change Case** button.

Step 5: **Change Case** button in use turns grey in color.

Step 6: **Change Case** button returns back to its original color once it completes its task.

Step 7: Check the change in **Current Data** widget.

In the example data, the feature column 'Country_Code' has inconsistent capitalization. The whole feature column 'Country_Code' is in upper case except fifteenth observation 'bel' and twenth observation 'bra'. We use **Change Case** button to change the whole column to upper case so as to unify inconsistent capitalization of the feature. Figures 4.12-4.18 illustrate how to use **Change Case** button.



Serial_Number	Country_Region	Population_Size	Tourism	Date_FirstFatality	Date_FirstConfirmedCase	Latitude	Longtitude	Mean_Age	Lockdown_Date	Lockdown_Type	Country_Code
1.00	Afghanistan	37172386.00	NaN	23-Mar-2020 00...	25-Feb-2020 00:00:00	33.94	67.71	NaN	24-Mar-2020 0...	Full	AFG
2.00	Albania	2866376.00	NaN	12-Mar-2020 00...	10-Mar-2020 00:00:00	41.15	NaN	NaN	08-Mar-2020 0...	Full	ALB
3.00	Algeria	42228429.00	2657000.00	13-Mar-2020 00...	26-Feb-2020 00:00:00	28.03	1.66	27.50	24-Mar-2020 0...	Full	DZA
4.00	Andorra	77006.00	3042000.00	23-Mar-2020 00...	03-Mar-2020 00:00:00	42.55	1.60	37.00	16-Mar-2020 0...	Full	AND
5.00	Argentina	44494502.00	6942000.00	09-Mar-2020 00...	04-Mar-2020 00:00:00	-38.42	NaN	30.80	20-Mar-2020 0...	Fu ll	ARG
6.00	Armenia	2951776.00	1652000.00	27-Mar-2020 00...	02-Mar-2020 00:00:00	NaN	NaN	33.90	24-Mar-2020 0...	Full	ARM
7.00	Australia	24982688.00	9246000.00	02-Mar-2020 00...	26-Jan-2020 00:00:00	-25.27	133.78	37.40	25-Mar-2020 0...	Partial	AUS
8.00	Austria	8840521.00	NaN	13-Mar-2020 00...	26-Feb-2020 00:00:00	47.52	14.55	NaN	16-Mar-2020 0...	F ull	AUT
9.00	Azerbaijan	9939800.00	2633000.00	14-Mar-2020 00...	02-Mar-2020 00:00:00	40.14	NaN	30.30	02-Mar-2020 0...	Full	AZE
10.00	Bahamas	385640.00	14000.00	02-Apr-2020 00:...	17-Mar-2020 00:00:00	25.03	-77.40	32.50	17-Apr-2020 00...		BHS
11.00	Bahrain	1569439.00	12045000.00	17-Mar-2020 00...	25-Feb-2020 00:00:00	25.93	NaN	31.20	25-Feb-2020 0...	Full	BHR
12.00	Bangladesh	NaN	14000.00	19-Mar-2020 00...	09-Mar-2020 00:00:00	23.68	90.36	25.60	19-Mar-2020 0...		BGD
13.00	Barbados	286641.00	680000.00	06-Apr-2020 00:...	18-Mar-2020 00:00:00	13.19	NaN	38.50	28-Mar-2020 0...		BRB
14.00	Belarus	NaN	11501600.00	01-Apr-2020 00:...	29-Feb-2020 00:00:00	NaN	27.95	NaN	07-Apr-2020 00...		BLR
15.00	Belgium	NaN	9119000.00	12-Mar-2020 00...	05-Feb-2020 00:00:00	50.50	NaN	NaN	17-Mar-2020 0...	Full	bel
16.00	Belize	383071.00	489000.00	07-Apr-2020 00:...	24-Mar-2020 00:00:00	NaN	-88.50	23.50	16-Apr-2020 00...	Full	BLZ
17.00	Bolivia	NaN	1142000.00	30-Mar-2020 00...	12-Mar-2020 00:00:00	NaN	-63.59	NaN	12-Mar-2020 0...	Full	BOL
18.00	Bosnia and...	3323929.00	NaN	22-Mar-2020 00...	06-Mar-2020 00:00:00	43.92	NaN	41.00	11-Mar-2020 00...		BIH
19.00	Botswana	NaN	14000.00	01-Apr-2020 00:...	31-Mar-2020 00:00:00	-22.33	NaN	24.40	02-Apr-2020 00...	Partial	BWA
20.00	Brazil	209469333.00	6621000.00	18-Mar-2020 00...	27-Feb-2020 00:00:00	-14.24	-51.93	31.30	17-Mar-2020 0...	Partial	bra
21.00	Bulgaria	7025037.00	NaN	12-Mar-2020 00...	09-Mar-2020 00:00:00	42.73	25.49	43.50	13-Mar-2020 0...		BGR
22.00	Burkina Faso	19751535.00	144000.00	19-Mar-2020 00...	11-Mar-2020 00:00:00	12.24	NaN	17.00	21-Mar-2020 0...		BFA
23.00	Canada	37057765.00	21134000.00	10-Mar-2020 00...	27-Jan-2020 00:00:00	56.13	-106.35	40.50	16-Mar-2020 0...	Partial	CAN
24.00	Chile	18729160.00	5723000.00	23-Mar-2020 00...	04-Mar-2020 00:00:00	-35.68	NaN	33.70	26-Mar-2020 0...	Full	CHL
25.00	China	1392730000.00	NaN	23-Jan-2020 00...	22-Jan-2020 00:00:00	35.86	NaN	NaN	23-Jan-2020 00...	Full	CHN
26.00	Colombia	NaN	3904000.00	23-Mar-2020 00...	07-Mar-2020 00:00:00	4.57	NaN	30.10	25-Mar-2020 0...	Full	COL
27.00	Congo (Brazza...	NaN	156000.00	03-Apr-2020 00:...	16-Mar-2020 00:00:00	-4.52	21.96	37.00	28-Mar-2020 0...	Partial	COG
28.00	Congo (Kinshasa)	84068091.00	14000.00	22-Mar-2020 00...	12-Mar-2020 00:00:00	NaN	NaN	37.00	31-Mar-2020 0...	Full	COD
29.00	Costa Rica	4999441.00	NaN	20-Mar-2020 00...	07-Mar-2020 00:00:00	9.75	NaN	NaN	15-Mar-2020 0...	Full	CRI
30.00	Croatia	NaN	16645000.00	20-Mar-2020 00...	26-Feb-2020 00:00:00	NaN	NaN	42.60	22-Mar-2020 0...	Partial	HRV
31.00	Cuba	11338138.00	4684000.00	19-Mar-2020 00...	13-Mar-2020 00:00:00	21.52	-77.78	41.10	23-Mar-2020 0...	Full	CUB
32.00	Cyprus	1189265.00	NaN	23-Mar-2020 00...	10-Mar-2020 00:00:00	35.13	NaN	34.90	25-Mar-2020 0...	Full	CYP
33.00	Czechia	10065000.00	NaN	23-Mar-2020 00...	02-Mar-2020 00:00:00	NaN	15.47	NaN	16-Mar-2020 0...	Full	CZE
34.00	Denmark	5793636.00	12749000.00	15-Mar-2020 00...	28-Feb-2020 00:00:00	56.26	NaN	41.60	11-Mar-2020 00...	Full	DNK
35.00	Djibouti	958920.00	14000.00	11-Apr-2020 00:...	19-Mar-2020 00:00:00	11.83	42.59	23.70	23-Mar-2020 0...	Full	DJI
36.00	Dominican Rep...	10627165.00	6569000.00	18-Mar-2020 00...	02-Mar-2020 00:00:00	NaN	-70.16	26.10	17-Mar-2020 0...	Full	DOM
37.00	Ecuador	17084357.00	2535000.00	15-Mar-2020 00...	02-Mar-2020 00:00:00	-1.83	NaN	26.60	24-Mar-2020 0...	Partial	ECU
38.00	Egypt	98423595.00	11196000.00	09-Mar-2020 00...	15-Feb-2020 00:00:00	26.82	NaN	NaN	24-Mar-2020 0...		EGY

Figure 4.12: Step 1. Change Case Button

Figure 4.13: Step 2. Change Case Button

Figure 4.14: Step 3. Change Case Button

Figure 4.15: Step 4. Change Case Button

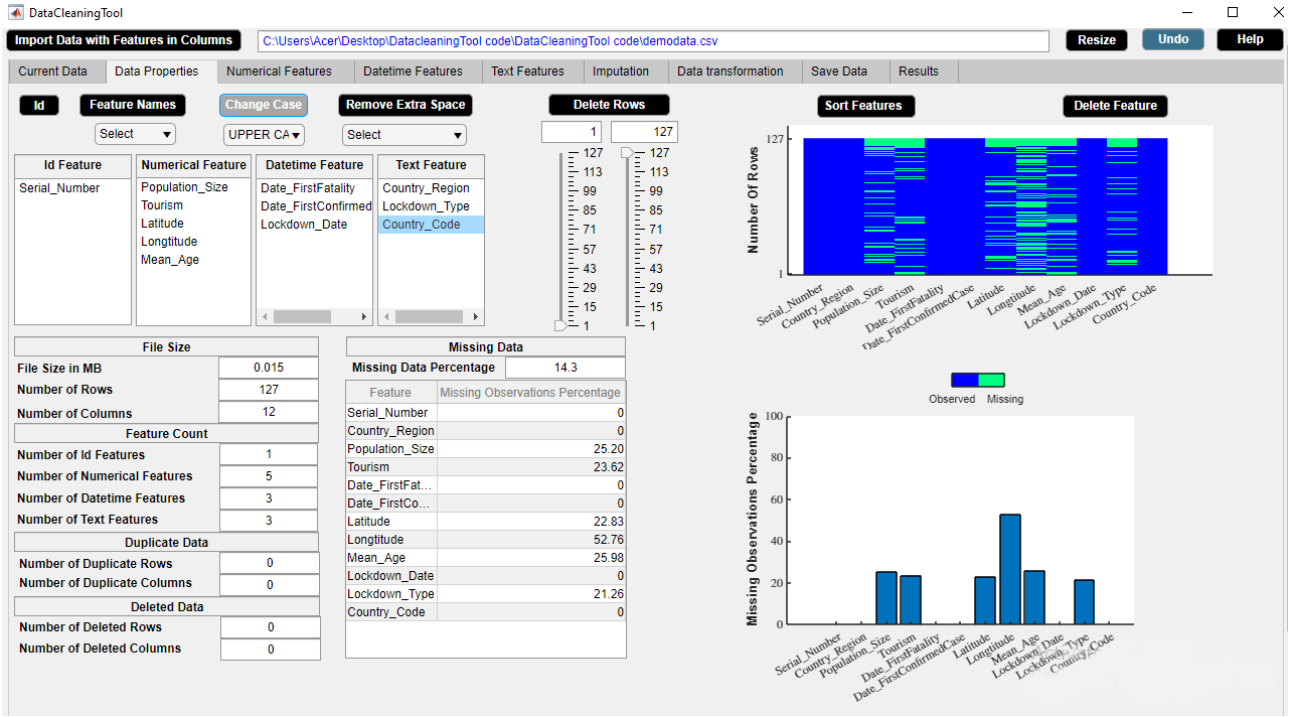


Figure 4.16: Step 5. Change Case Button

Figure 4.17: Step 6. Change Case Button

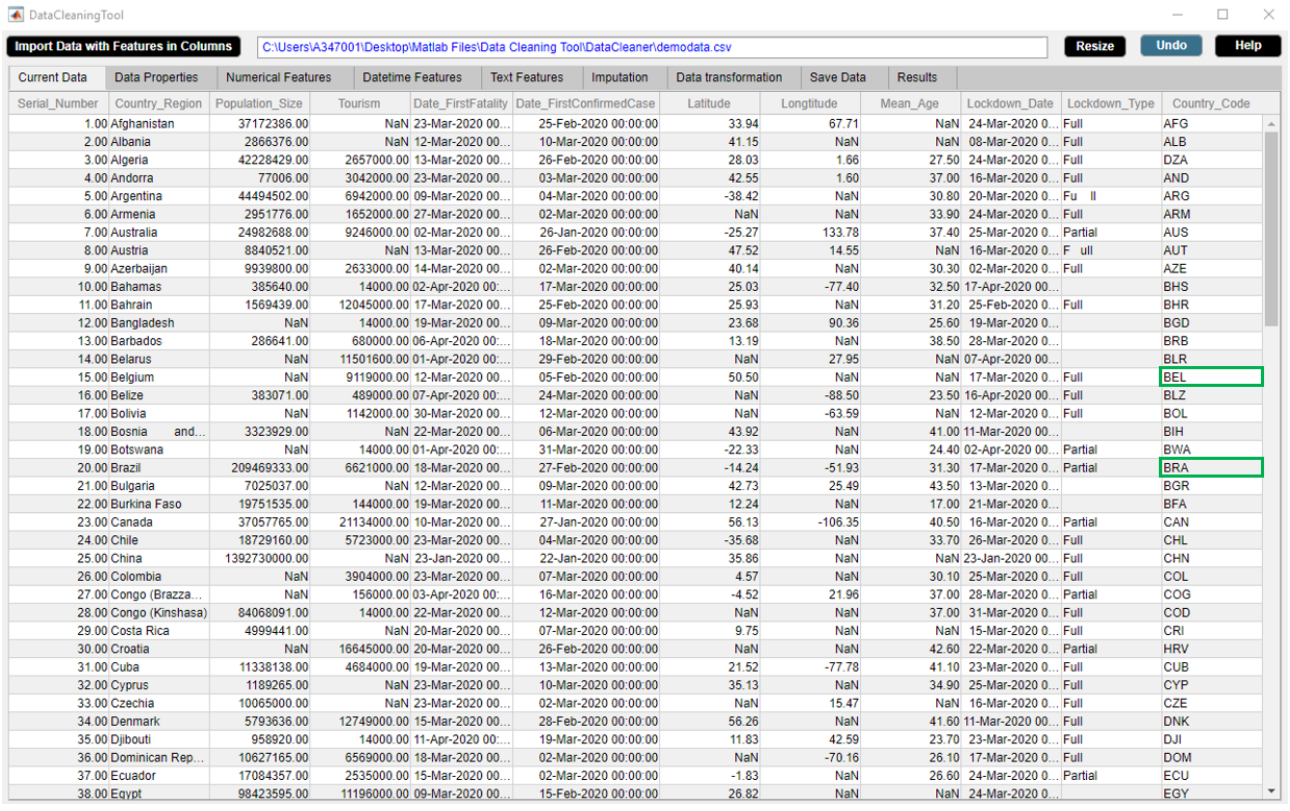


Figure 4.18: Step 7. Change Case Button

4.2.4 Remove Extra Space Button

Removes either all spaces or to only one whitespace in a string of a feature.

Application

- Fix structural errors such as typographical errors.

Example

Step 1: Check if there is any extra space in a feature in the **Current Data** widget.

Step 2: Select any one option from **Remove Extra Space** dropdown menu.

Step 3: Select the feature from **Numerical Features** or **Datetime Features** or **Text Features** list box in the **Data Properties** widget.

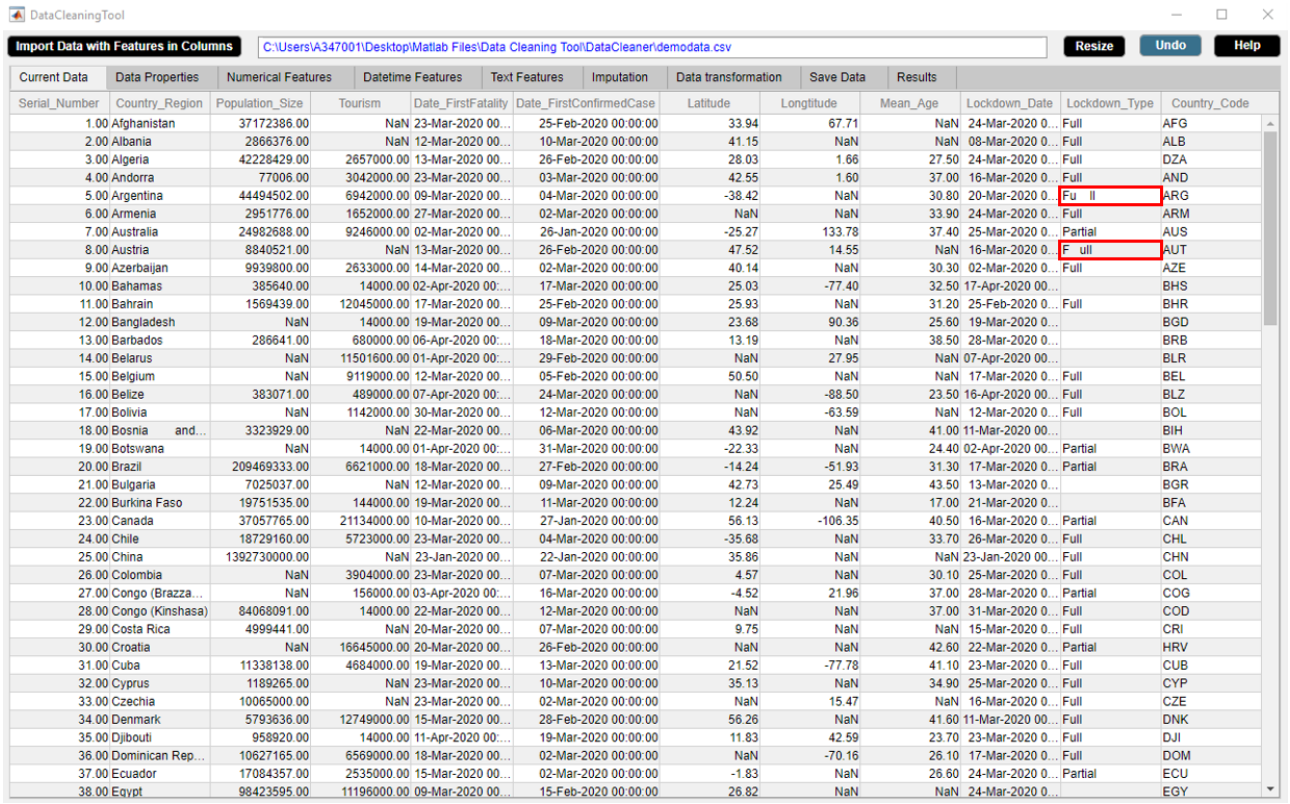
Step 4: Click **Remove Extra Space** button.

Step 5: **Remove Extra Space** button in use turns grey in color.

Step 6: **Remove Extra Space** button returns back to its original color once it completes its task.

Step 7: Check the change in **Current Data** widget.

In the example data, the feature 'Lockdown_type' is either 'Full' or 'Partial'. The fifth and eighth observations in feature column 'Country_Code' are 'Fu ll' and 'F ull'. We use **Remove Extra Space** button to remove all spaces in the whole column. Figures 4.19-4.25 illustrate how to use **Remove Extra Space** button.



Serial_Number	Country_Region	Population_Size	Tourism	Date_FirstFatality	Date_FirstConfirmedCase	Latitude	Longtitude	Mean_Age	Lockdown_Date	Lockdown_Type	Country_Code
1.00	Afghanistan	37172386.00	NaN	23-Mar-2020 00...	25-Feb-2020 00:00:00	33.94	67.71	NaN	24-Mar-2020 0...	Full	AFG
2.00	Albania	2866376.00	NaN	12-Mar-2020 00...	10-Mar-2020 00:00:00	41.15	NaN	NaN	08-Mar-2020 0...	Full	ALB
3.00	Algeria	42228429.00	2657000.00	13-Mar-2020 00...	26-Feb-2020 00:00:00	28.03	1.66	27.50	24-Mar-2020 0...	Full	DZA
4.00	Andorra	77006.00	3042000.00	23-Mar-2020 00...	03-Mar-2020 00:00:00	42.55	1.60	37.00	16-Mar-2020 0...	Full	AND
5.00	Argentina	44494502.00	6942000.00	09-Mar-2020 00...	04-Mar-2020 00:00:00	-38.42	NaN	30.80	20-Mar-2020 0...	Fu ll	ARG
6.00	Armenia	2951776.00	1652000.00	27-Mar-2020 00...	02-Mar-2020 00:00:00	NaN	NaN	33.90	24-Mar-2020 0...	Full	ARM
7.00	Australia	24982688.00	9246000.00	02-Mar-2020 00...	26-Jan-2020 00:00:00	-25.27	133.78	37.40	25-Mar-2020 0...	Partial	AUS
8.00	Austria	8840521.00	NaN	13-Mar-2020 00...	26-Feb-2020 00:00:00	47.52	14.55	NaN	16-Mar-2020 0...	F ull	AUT
9.00	Azerbaijan	9939800.00	2633000.00	14-Mar-2020 00...	02-Mar-2020 00:00:00	40.14	NaN	30.30	02-Mar-2020 0...	Full	AZE
10.00	Bahamas	385640.00	14000.00	02-Apr-2020 00:...	17-Mar-2020 00:00:00	25.03	-77.40	32.50	17-Apr-2020 00...		BHS
11.00	Bahrain	1569439.00	12045000.00	17-Mar-2020 00...	25-Feb-2020 00:00:00	25.93	NaN	31.20	25-Feb-2020 0...	Full	BHR
12.00	Bangladesh	NaN	14000.00	19-Mar-2020 00...	09-Mar-2020 00:00:00	23.68	90.36	25.60	19-Mar-2020 0...		BGD
13.00	Barbados	286641.00	680000.00	06-Apr-2020 00:...	18-Mar-2020 00:00:00	13.19	NaN	38.50	28-Mar-2020 0...		BRB
14.00	Belarus	NaN	11501600.00	01-Apr-2020 00:...	29-Feb-2020 00:00:00	NaN	27.95	NaN	07-Apr-2020 00...		BLR
15.00	Belgium	NaN	9119000.00	12-Mar-2020 00...	05-Feb-2020 00:00:00	50.50	NaN	NaN	17-Mar-2020 0...	Full	BEL
16.00	Belize	383071.00	489000.00	07-Apr-2020 00:...	24-Mar-2020 00:00:00	NaN	-88.50	23.50	16-Apr-2020 00...	Full	BLZ
17.00	Bolivia	NaN	1142000.00	30-Mar-2020 00...	12-Mar-2020 00:00:00	NaN	-63.59	NaN	12-Mar-2020 0...	Full	BOL
18.00	Bosnia and...	3323929.00	NaN	22-Mar-2020 00...	06-Mar-2020 00:00:00	43.92	NaN	41.00	11-Mar-2020 00...		BIH
19.00	Botswana	NaN	14000.00	01-Apr-2020 00:...	31-Mar-2020 00:00:00	-22.33	NaN	24.40	02-Apr-2020 00...	Partial	BWA
20.00	Brazil	209469333.00	6621000.00	18-Mar-2020 00...	27-Feb-2020 00:00:00	-14.24	-51.93	31.30	17-Mar-2020 0...	Partial	BRA
21.00	Bulgaria	7025037.00	NaN	12-Mar-2020 00...	09-Mar-2020 00:00:00	42.73	25.49	43.50	13-Mar-2020 0...		BGR
22.00	Burkina Faso	19751535.00	144000.00	19-Mar-2020 00...	11-Mar-2020 00:00:00	12.24	NaN	17.00	21-Mar-2020 0...		BFA
23.00	Canada	37057765.00	21134000.00	10-Mar-2020 00...	27-Jan-2020 00:00:00	56.13	-106.35	40.50	16-Mar-2020 0...	Partial	CAN
24.00	Chile	18729160.00	5723000.00	23-Mar-2020 00...	04-Mar-2020 00:00:00	-35.68	NaN	33.70	26-Mar-2020 0...	Full	CHL
25.00	China	1392730000.00	NaN	23-Jan-2020 00...	22-Jan-2020 00:00:00	35.86	NaN	NaN	23-Jan-2020 00...	Full	CHN
26.00	Colombia	NaN	3904000.00	23-Mar-2020 00...	07-Mar-2020 00:00:00	4.57	NaN	30.10	25-Mar-2020 0...	Full	COL
27.00	Congo (Brazza...	NaN	156000.00	03-Apr-2020 00:...	16-Mar-2020 00:00:00	-4.52	21.96	37.00	28-Mar-2020 0...	Partial	COG
28.00	Congo (Kinshasa)	84068091.00	14000.00	22-Mar-2020 00...	12-Mar-2020 00:00:00	NaN	NaN	37.00	31-Mar-2020 0...	Full	COD
29.00	Costa Rica	4999441.00	NaN	20-Mar-2020 00...	07-Mar-2020 00:00:00	9.75	NaN	NaN	15-Mar-2020 0...	Full	CRI
30.00	Croatia	NaN	16645000.00	20-Mar-2020 00...	26-Feb-2020 00:00:00	NaN	NaN	42.60	22-Mar-2020 0...	Partial	HRV
31.00	Cuba	11338138.00	4684000.00	19-Mar-2020 00...	13-Mar-2020 00:00:00	21.52	-77.78	41.10	23-Mar-2020 0...	Full	CUB
32.00	Cyprus	1189265.00	NaN	23-Mar-2020 00...	10-Mar-2020 00:00:00	35.13	NaN	34.90	25-Mar-2020 0...	Full	CYP
33.00	Czechia	10065000.00	NaN	23-Mar-2020 00...	02-Mar-2020 00:00:00	NaN	15.47	NaN	16-Mar-2020 0...	Full	CZE
34.00	Denmark	5793636.00	12749000.00	15-Mar-2020 00...	28-Feb-2020 00:00:00	56.26	NaN	41.60	11-Mar-2020 00...	Full	DNK
35.00	Djibouti	958920.00	14000.00	11-Apr-2020 00:...	19-Mar-2020 00:00:00	11.83	42.59	23.70	23-Mar-2020 0...	Full	DJI
36.00	Dominican Rep...	10627165.00	6569000.00	18-Mar-2020 00...	02-Mar-2020 00:00:00	NaN	-70.16	26.10	17-Mar-2020 0...	Full	DOM
37.00	Ecuador	17084357.00	2535000.00	15-Mar-2020 00...	02-Mar-2020 00:00:00	-1.83	NaN	26.60	24-Mar-2020 0...	Partial	ECU
38.00	Egypt	98423595.00	11196000.00	09-Mar-2020 00...	15-Feb-2020 00:00:00	26.82	NaN	NaN	24-Mar-2020 0...		EGY

Figure 4.19: Step 1. Remove Extra Space Button

Figure 4.20: Step 2. Remove Extra Space Button

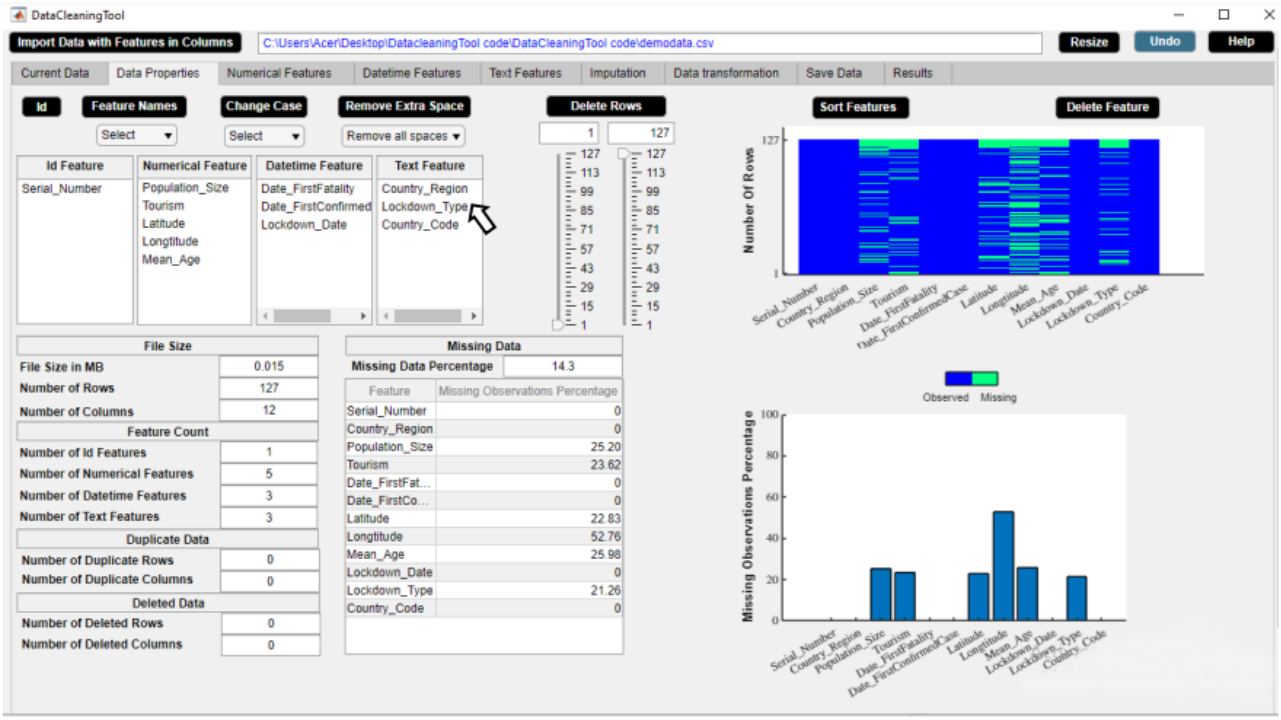


Figure 4.21: Step 3. Remove Extra Space Button

Figure 4.22: Step 4. Remove Extra Space Button

Figure 4.23: Step 5. Remove Extra Space Button

Figure 4.24: Step 6. Remove Extra Space Button

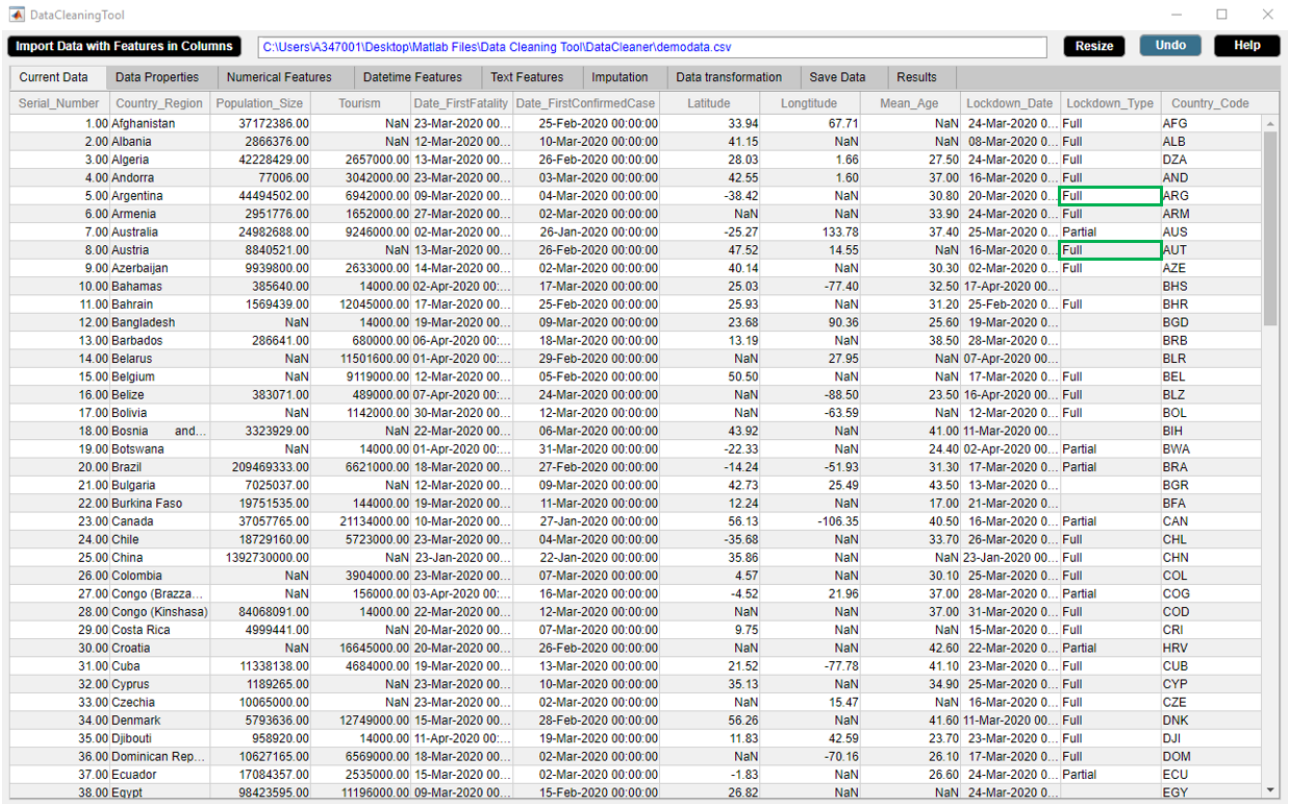
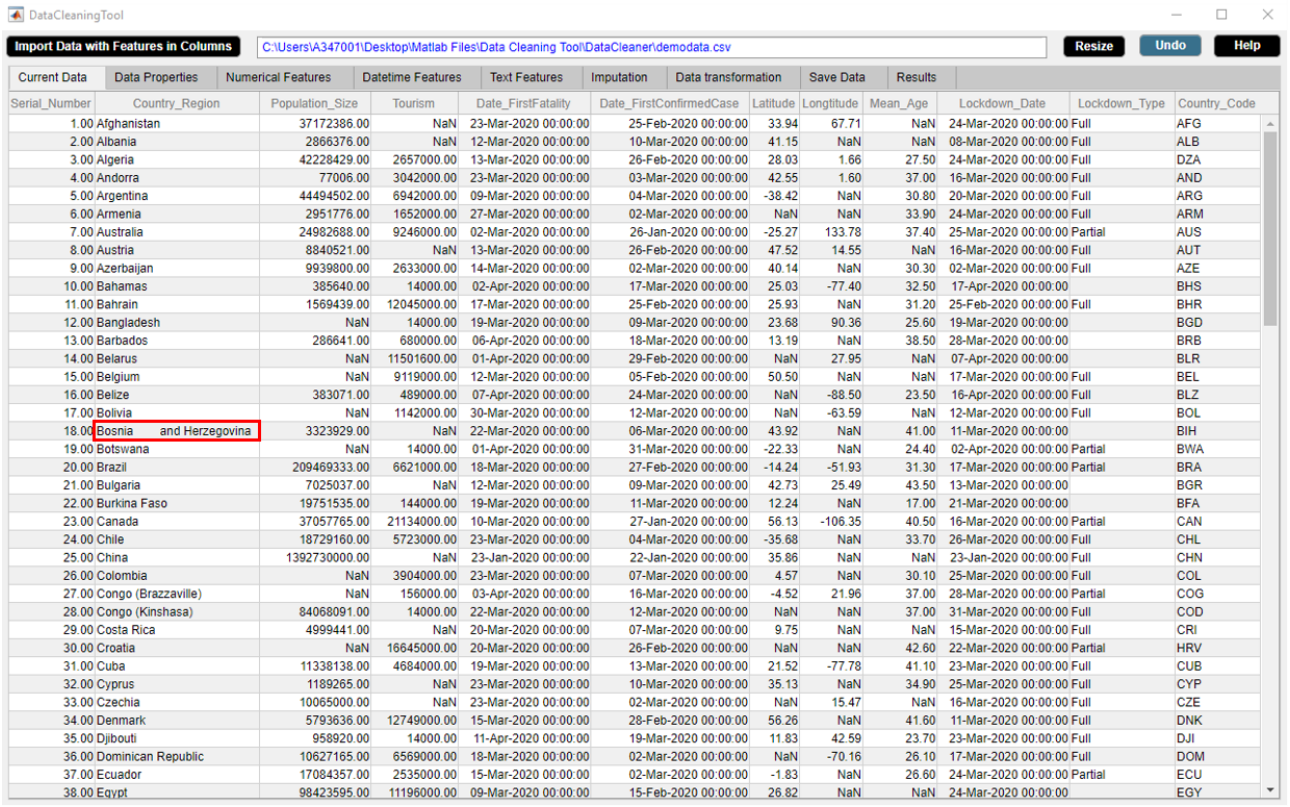


Figure 4.25: Step 7. Remove Extra Space Button

Again, the eighteenth observation of the feature 'Country_region' is 'Bosnia and Herzegovina'. We use **Remove Extra Space** button to remove to single white space in the whole column. Figures 4.26-4.32 illustrate how to use **Remove Extra Space** button to remove to single white space.



Serial_Number	Country_Region	Population_Size	Tourism	Date_FirstFatality	Date_FirstConfirmedCase	Latitude	Longtitude	Mean_Age	Lockdown_Date	Lockdown_Type	Country_Code
1.00	Afghanistan	37172386.00	NaN	23-Mar-2020 00:00:00	25-Feb-2020 00:00:00	33.94	67.71	NaN	24-Mar-2020 00:00:00	Full	AFG
2.00	Albania	2866376.00	NaN	12-Mar-2020 00:00:00	10-Mar-2020 00:00:00	41.15	NaN	NaN	08-Mar-2020 00:00:00	Full	ALB
3.00	Algeria	42228429.00	2657000.00	13-Mar-2020 00:00:00	26-Feb-2020 00:00:00	28.03	1.66	27.50	24-Mar-2020 00:00:00	Full	DZA
4.00	Andorra	77006.00	3042000.00	23-Mar-2020 00:00:00	03-Mar-2020 00:00:00	42.55	1.60	37.00	16-Mar-2020 00:00:00	Full	AND
5.00	Argentina	44494502.00	6942000.00	09-Mar-2020 00:00:00	04-Mar-2020 00:00:00	-38.42	NaN	30.80	20-Mar-2020 00:00:00	Full	ARG
6.00	Armenia	2951776.00	1652000.00	27-Mar-2020 00:00:00	02-Mar-2020 00:00:00	NaN	NaN	33.90	24-Mar-2020 00:00:00	Full	ARM
7.00	Australia	24982688.00	9246000.00	02-Mar-2020 00:00:00	26-Jan-2020 00:00:00	-25.27	133.78	37.40	25-Mar-2020 00:00:00	Partial	AUS
8.00	Austria	8840521.00	NaN	13-Mar-2020 00:00:00	26-Feb-2020 00:00:00	47.52	14.55	NaN	16-Mar-2020 00:00:00	Full	AUT
9.00	Azerbaijan	9939800.00	2633000.00	14-Mar-2020 00:00:00	02-Mar-2020 00:00:00	40.14	NaN	30.30	02-Mar-2020 00:00:00	Full	AZE
10.00	Bahamas	385640.00	14000.00	02-Apr-2020 00:00:00	17-Mar-2020 00:00:00	25.03	-77.40	32.50	17-Apr-2020 00:00:00		BHS
11.00	Bahrain	1569439.00	12045000.00	17-Mar-2020 00:00:00	25-Feb-2020 00:00:00	25.93	NaN	31.20	25-Feb-2020 00:00:00	Full	BHR
12.00	Bangladesh	NaN	14000.00	19-Mar-2020 00:00:00	09-Mar-2020 00:00:00	23.68	90.36	25.60	19-Mar-2020 00:00:00		BGD
13.00	Barbados	286641.00	680000.00	06-Apr-2020 00:00:00	18-Mar-2020 00:00:00	13.19	NaN	38.50	28-Mar-2020 00:00:00		BRB
14.00	Belarus	NaN	11501600.00	01-Apr-2020 00:00:00	29-Feb-2020 00:00:00	NaN	27.95	NaN	07-Apr-2020 00:00:00		BLR
15.00	Belgium	NaN	9119000.00	12-Mar-2020 00:00:00	05-Feb-2020 00:00:00	50.50	NaN	NaN	17-Mar-2020 00:00:00	Full	BEL
16.00	Belize	383071.00	489000.00	07-Apr-2020 00:00:00	24-Mar-2020 00:00:00	NaN	-88.50	23.50	16-Apr-2020 00:00:00	Full	BLZ
17.00	Bolivia	NaN	1142000.00	30-Mar-2020 00:00:00	12-Mar-2020 00:00:00	NaN	-63.59	NaN	12-Mar-2020 00:00:00	Full	BOL
18.00	Bosnia and Herzegovina	3323929.00	NaN	22-Mar-2020 00:00:00	06-Mar-2020 00:00:00	43.92	NaN	41.00	11-Mar-2020 00:00:00		BIH
19.00	Botswana	NaN	14000.00	01-Apr-2020 00:00:00	31-Mar-2020 00:00:00	-22.33	NaN	24.40	02-Apr-2020 00:00:00	Partial	BWA
20.00	Brazil	209469333.00	6621000.00	18-Mar-2020 00:00:00	27-Feb-2020 00:00:00	-14.24	-51.93	31.30	17-Mar-2020 00:00:00	Partial	BRA
21.00	Bulgaria	7025037.00	NaN	12-Mar-2020 00:00:00	09-Mar-2020 00:00:00	42.73	25.49	43.50	13-Mar-2020 00:00:00		BGR
22.00	Burkina Faso	19751535.00	144000.00	19-Mar-2020 00:00:00	11-Mar-2020 00:00:00	12.24	NaN	17.00	21-Mar-2020 00:00:00		BFA
23.00	Canada	37057765.00	21134000.00	10-Mar-2020 00:00:00	27-Jan-2020 00:00:00	56.13	-106.35	40.50	16-Mar-2020 00:00:00	Partial	CAN
24.00	Chile	18729160.00	5723000.00	23-Mar-2020 00:00:00	04-Mar-2020 00:00:00	-35.68	NaN	33.70	26-Mar-2020 00:00:00	Full	CHL
25.00	China	1392730000.00	NaN	23-Jan-2020 00:00:00	22-Jan-2020 00:00:00	35.86	NaN	NaN	23-Jan-2020 00:00:00	Full	CHN
26.00	Colombia	NaN	3904000.00	23-Mar-2020 00:00:00	07-Mar-2020 00:00:00	4.57	NaN	30.10	25-Mar-2020 00:00:00	Full	COL
27.00	Congo (Brazzaville)	NaN	156000.00	03-Apr-2020 00:00:00	16-Mar-2020 00:00:00	-4.52	21.96	37.00	28-Mar-2020 00:00:00	Partial	COG
28.00	Congo (Kinshasa)	84068091.00	14000.00	22-Mar-2020 00:00:00	12-Mar-2020 00:00:00	NaN	NaN	37.00	31-Mar-2020 00:00:00	Full	COD
29.00	Costa Rica	4999441.00	NaN	20-Mar-2020 00:00:00	07-Mar-2020 00:00:00	9.75	NaN	NaN	15-Mar-2020 00:00:00	Full	CRI
30.00	Croatia	NaN	16645000.00	20-Mar-2020 00:00:00	26-Feb-2020 00:00:00	NaN	NaN	42.60	22-Mar-2020 00:00:00	Partial	HRV
31.00	Cuba	11338138.00	4684000.00	19-Mar-2020 00:00:00	13-Mar-2020 00:00:00	21.52	-77.78	41.10	23-Mar-2020 00:00:00	Full	CUB
32.00	Cyprus	1189265.00	NaN	23-Mar-2020 00:00:00	10-Mar-2020 00:00:00	35.13	NaN	34.90	25-Mar-2020 00:00:00	Full	CYP
33.00	Czechia	10065000.00	NaN	23-Mar-2020 00:00:00	02-Mar-2020 00:00:00	NaN	15.47	NaN	16-Mar-2020 00:00:00	Full	CZE
34.00	Denmark	5793636.00	12749000.00	15-Mar-2020 00:00:00	28-Feb-2020 00:00:00	56.26	NaN	41.60	11-Mar-2020 00:00:00	Full	DNK
35.00	Djibouti	958920.00	14000.00	11-Apr-2020 00:00:00	19-Mar-2020 00:00:00	11.83	42.59	23.70	23-Mar-2020 00:00:00	Full	DJI
36.00	Dominican Republic	10627165.00	6569000.00	18-Mar-2020 00:00:00	02-Mar-2020 00:00:00	NaN	-70.16	26.10	17-Mar-2020 00:00:00	Full	DOM
37.00	Ecuador	17084357.00	2535000.00	15-Mar-2020 00:00:00	02-Mar-2020 00:00:00	-1.83	NaN	26.60	24-Mar-2020 00:00:00	Partial	ECU
38.00	Egypt	98423595.00	11196000.00	09-Mar-2020 00:00:00	15-Feb-2020 00:00:00	26.82	NaN	NaN	24-Mar-2020 00:00:00		EGY

Figure 4.26: Step 1. Remove Extra Space Button

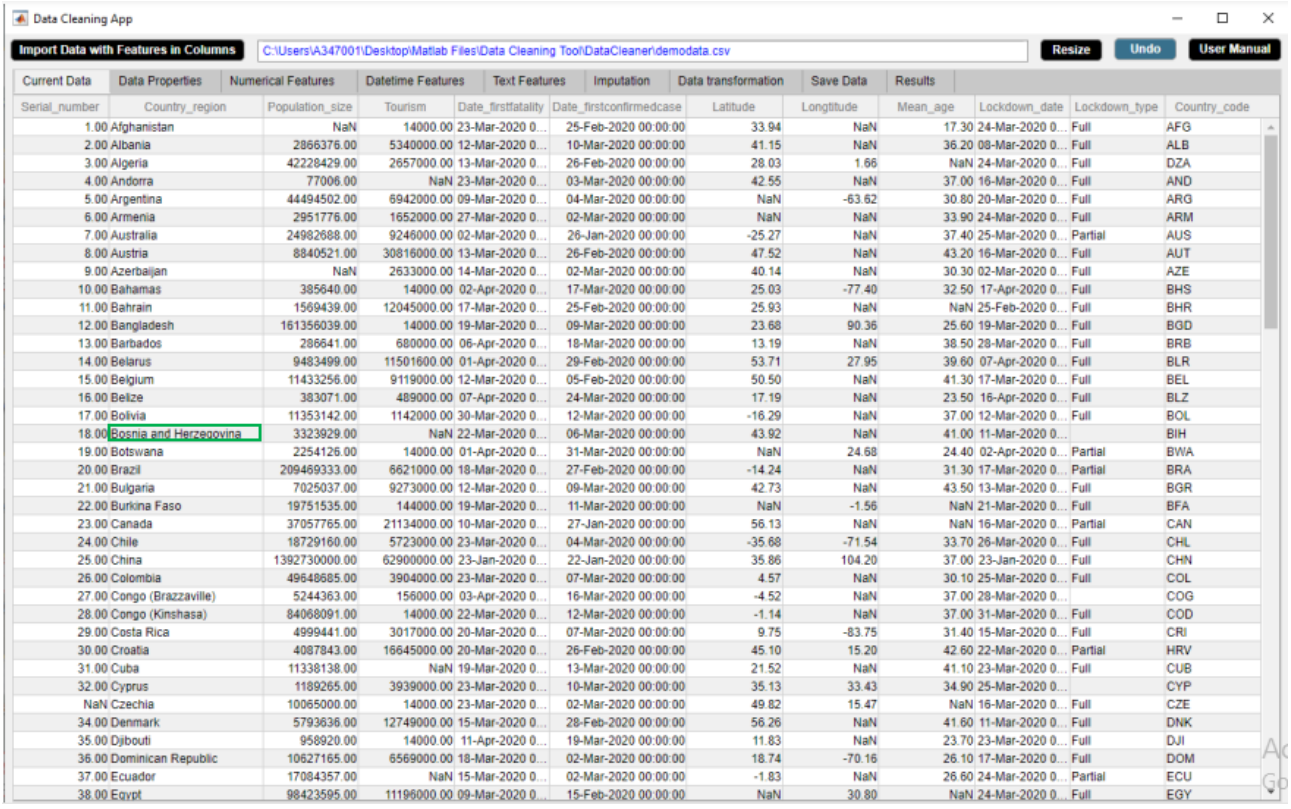
Figure 4.27: Step 2. Remove Extra Space Button

Figure 4.28: Step 3. Remove Extra Space Button

Figure 4.29: Step 4. Remove Extra Space Button

Figure 4.30: Step 5. Remove Extra Space Button

Figure 4.31: Step 6. Remove Extra Space Button



Serial_number	Country_region	Population_size	Tourism	Date_firstfatality	Date_firstconfirmedcase	Latitude	Longtitude	Mean_age	Lockdown_date	Lockdown_type	Country_code
1.00	Afghanistan	NaN	14000.00	23-Mar-2020 0...	25-Feb-2020 00:00:00	33.94	NaN	17.30	24-Mar-2020 0...	Full	AFG
2.00	Albania	2866376.00	5340000.00	12-Mar-2020 0...	10-Mar-2020 00:00:00	41.15	NaN	36.20	08-Mar-2020 0...	Full	ALB
3.00	Algeria	42228429.00	2657000.00	13-Mar-2020 0...	26-Feb-2020 00:00:00	28.03	1.66	NaN	24-Mar-2020 0...	Full	DZA
4.00	Andorra	77006.00	NaN	23-Mar-2020 0...	03-Mar-2020 00:00:00	42.55	NaN	37.00	16-Mar-2020 0...	Full	AND
5.00	Argentina	44494502.00	6942000.00	09-Mar-2020 0...	04-Mar-2020 00:00:00	NaN	-63.62	30.80	20-Mar-2020 0...	Full	ARG
6.00	Armenia	2951776.00	1652000.00	27-Mar-2020 0...	02-Mar-2020 00:00:00	NaN	NaN	33.90	24-Mar-2020 0...	Full	ARM
7.00	Australia	24982688.00	9246000.00	02-Mar-2020 0...	26-Jan-2020 00:00:00	-25.27	NaN	37.40	25-Mar-2020 0...	Partial	AUS
8.00	Austria	8840521.00	30816000.00	13-Mar-2020 0...	26-Feb-2020 00:00:00	47.52	NaN	43.20	16-Mar-2020 0...	Full	AUT
9.00	Azerbaijan	NaN	2633000.00	14-Mar-2020 0...	02-Mar-2020 00:00:00	40.14	NaN	30.30	02-Mar-2020 0...	Full	AZE
10.00	Bahamas	385640.00	14000.00	02-Apr-2020 0...	17-Mar-2020 00:00:00	25.03	-77.40	32.50	17-Apr-2020 0...	Full	BHS
11.00	Bahrain	1569439.00	12045000.00	17-Mar-2020 0...	25-Feb-2020 00:00:00	25.93	NaN	NaN	25-Feb-2020 0...	Full	BHR
12.00	Bangladesh	161356039.00	14000.00	19-Mar-2020 0...	09-Mar-2020 00:00:00	23.68	90.36	25.60	19-Mar-2020 0...	Full	BGD
13.00	Barbados	286641.00	680000.00	06-Apr-2020 0...	18-Mar-2020 00:00:00	13.19	NaN	38.50	28-Mar-2020 0...	Full	BRB
14.00	Belarus	9483499.00	11501600.00	01-Apr-2020 0...	29-Feb-2020 00:00:00	53.71	27.95	39.60	07-Apr-2020 0...	Full	BLR
15.00	Belgium	11433256.00	9119000.00	12-Mar-2020 0...	05-Feb-2020 00:00:00	50.50	NaN	41.30	17-Mar-2020 0...	Full	BEL
16.00	Belize	383071.00	489000.00	07-Apr-2020 0...	24-Mar-2020 00:00:00	17.19	NaN	23.50	16-Apr-2020 0...	Full	BLZ
17.00	Bolivia	11353142.00	1142000.00	30-Mar-2020 0...	12-Mar-2020 00:00:00	-16.29	NaN	37.00	12-Mar-2020 0...	Full	BOL
18.00	Bosnia and Herzegovina	3323929.00	NaN	22-Mar-2020 0...	06-Mar-2020 00:00:00	43.92	NaN	41.00	11-Mar-2020 0...		BIH
19.00	Botswana	2254126.00	14000.00	01-Apr-2020 0...	31-Mar-2020 00:00:00	NaN	24.68	24.40	02-Apr-2020 0...	Partial	BWA
20.00	Brazil	209469333.00	6621000.00	18-Mar-2020 0...	27-Feb-2020 00:00:00	-14.24	NaN	31.30	17-Mar-2020 0...	Partial	BRA
21.00	Bulgaria	7025037.00	9273000.00	12-Mar-2020 0...	09-Mar-2020 00:00:00	42.73	NaN	43.50	13-Mar-2020 0...	Full	BGR
22.00	Burkina Faso	19751535.00	144000.00	19-Mar-2020 0...	11-Mar-2020 00:00:00	NaN	-1.56	NaN	21-Mar-2020 0...	Full	BFA
23.00	Canada	37057765.00	21134000.00	10-Mar-2020 0...	27-Jan-2020 00:00:00	56.13	NaN	NaN	16-Mar-2020 0...	Partial	CAN
24.00	Chile	18729160.00	5723000.00	23-Mar-2020 0...	04-Mar-2020 00:00:00	-35.68	-71.54	33.70	26-Mar-2020 0...	Full	CHL
25.00	China	1392730000.00	62900000.00	23-Jan-2020 0...	22-Jan-2020 00:00:00	35.86	104.20	37.00	23-Jan-2020 0...	Full	CHN
26.00	Colombia	49648685.00	3904000.00	23-Mar-2020 0...	07-Mar-2020 00:00:00	4.57	NaN	30.10	25-Mar-2020 0...	Full	COL
27.00	Congo (Brazzaville)	5244363.00	156000.00	03-Apr-2020 0...	16-Mar-2020 00:00:00	-4.52	NaN	37.00	28-Mar-2020 0...		COG
28.00	Congo (Kinshasa)	84068091.00	14000.00	22-Mar-2020 0...	12-Mar-2020 00:00:00	-1.14	NaN	37.00	31-Mar-2020 0...	Full	COD
29.00	Costa Rica	4999441.00	3017000.00	20-Mar-2020 0...	07-Mar-2020 00:00:00	9.75	-83.75	31.40	15-Mar-2020 0...	Full	CRI
30.00	Croatia	4087843.00	16645000.00	20-Mar-2020 0...	26-Feb-2020 00:00:00	45.10	15.20	42.60	22-Mar-2020 0...	Partial	HRV
31.00	Cuba	11338138.00	NaN	19-Mar-2020 0...	13-Mar-2020 00:00:00	21.52	NaN	41.10	23-Mar-2020 0...	Full	CUB
32.00	Cyprus	1189265.00	3939000.00	23-Mar-2020 0...	10-Mar-2020 00:00:00	35.13	33.43	34.90	25-Mar-2020 0...		CYP
NaN	Czechia	10065000.00	14000.00	23-Mar-2020 0...	02-Mar-2020 00:00:00	49.82	15.47	NaN	16-Mar-2020 0...	Full	CZE
34.00	Denmark	5793636.00	12749000.00	15-Mar-2020 0...	28-Feb-2020 00:00:00	56.26	NaN	41.60	11-Mar-2020 0...	Full	DNK
35.00	Djibouti	958920.00	14000.00	11-Apr-2020 0...	19-Mar-2020 00:00:00	11.83	NaN	23.70	23-Mar-2020 0...	Full	DJI
36.00	Dominican Republic	10627165.00	6569000.00	18-Mar-2020 0...	02-Mar-2020 00:00:00	18.74	-70.16	26.10	17-Mar-2020 0...	Full	DOM
37.00	Ecuador	17084357.00	NaN	15-Mar-2020 0...	02-Mar-2020 00:00:00	-1.83	NaN	26.60	24-Mar-2020 0...	Partial	ECU
38.00	Egypt	98423595.00	11196000.00	09-Mar-2020 0...	15-Feb-2020 00:00:00	NaN	30.80	NaN	24-Mar-2020 0...	Full	EGY

Figure 4.32: Step 7. Remove Extra Space Button

4.2.5 Delete Rows Button

Deletes rows from data.

Application

- Delete rows containing a large number of missing observations.

Example

Step 1: Select minimum row number from minimum slider and maximum row number from maximum slider.

Step 2: Click **Delete Rows** button.

Step 3: **Delete Rows** button in use turns grey in color.

Step 4: **Delete Rows** button returns back to its original color once it completes its task.

The example data contains a large number of missing values in the last 7 rows. We use **Delete Rows** button to delete the last 7 rows of the data. Figures 4.33-4.36 illustrate how to use **Delete Rows** button.

Figure 4.33: Step 1. Delete Rows Button

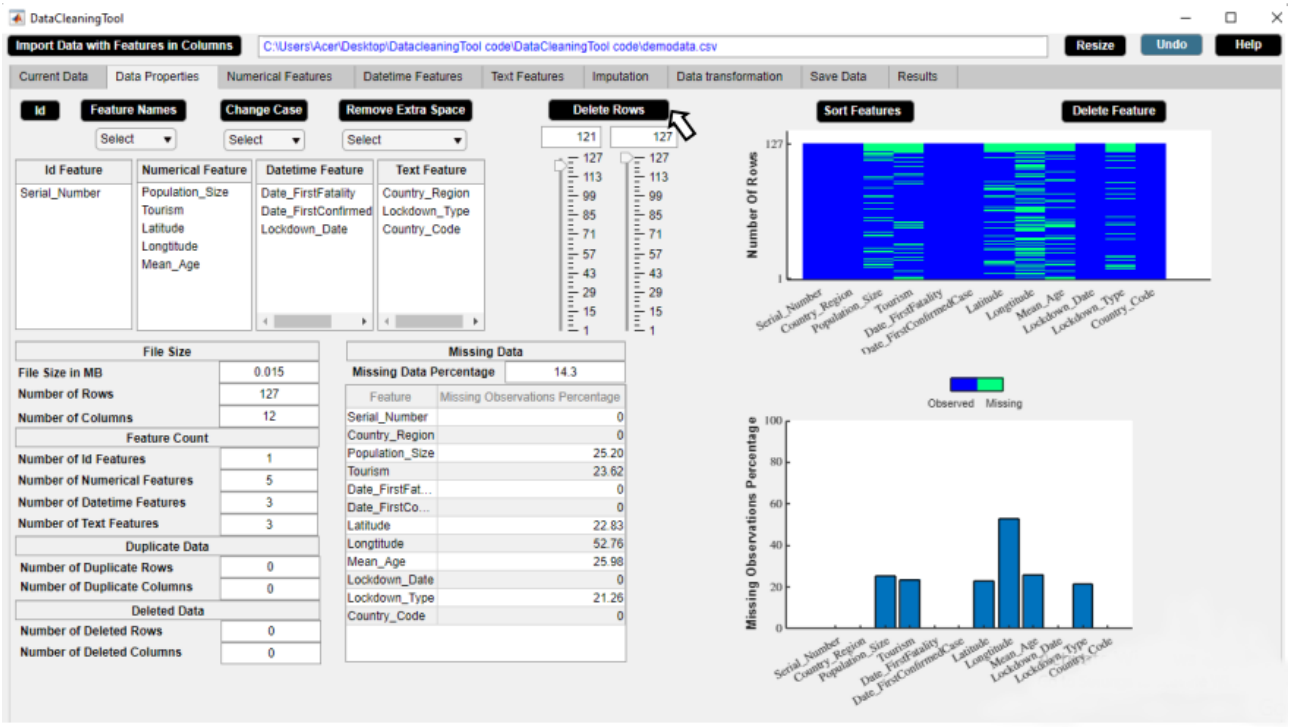


Figure 4.34: Step 2. Delete Rows Button

Figure 4.35: Step 3. Delete Rows Button

Figure 4.36: Step 4. Delete Rows Button

4.2.6 Sort Features Button

Sorts features in ascending order by missing observations percentage.

Example

Step 1: Click **Sort Features** button.

Step 2: **Sort Features** button in use turns grey in color.

Step 3: **Sort Features** button returns back to its original color once it completes its task.

We use **Sort Features** button to sort the features of the example data by increasing missing observations percentage. Figures 4.37-4.39 illustrate how to use **Sort Features** button.

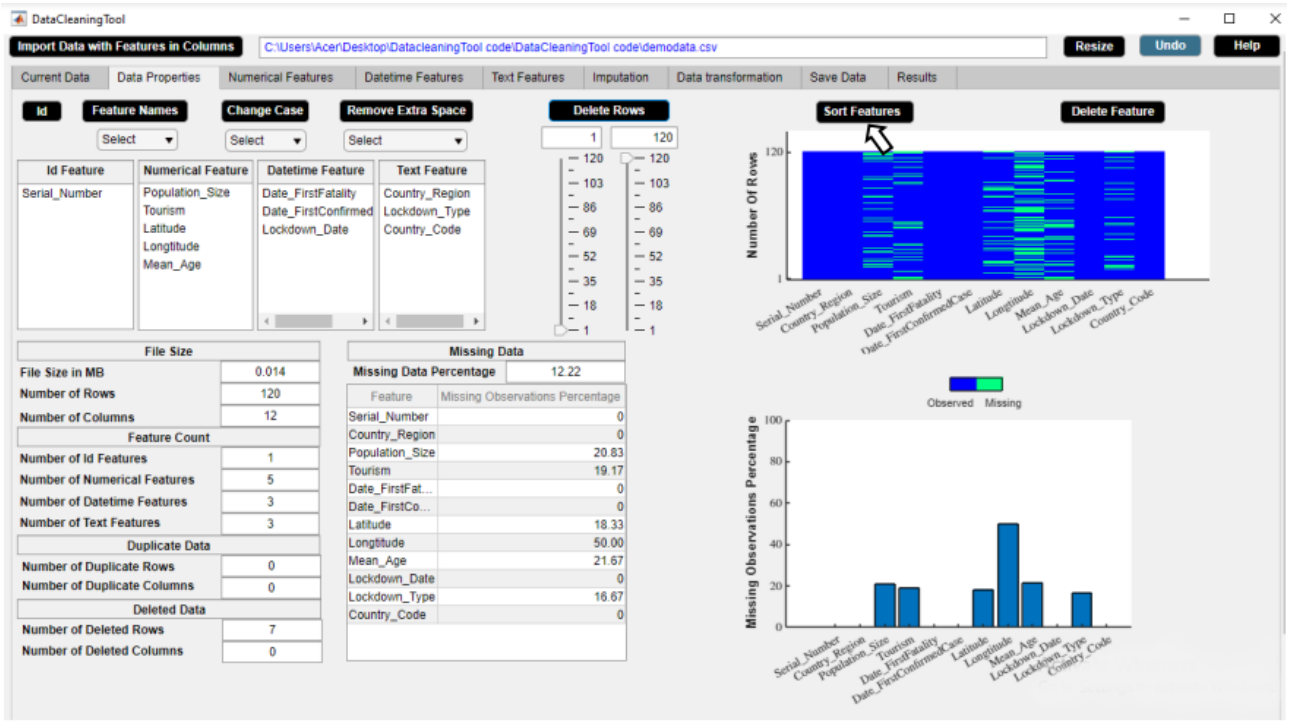


Figure 4.37: Step 1. Sort Features Button

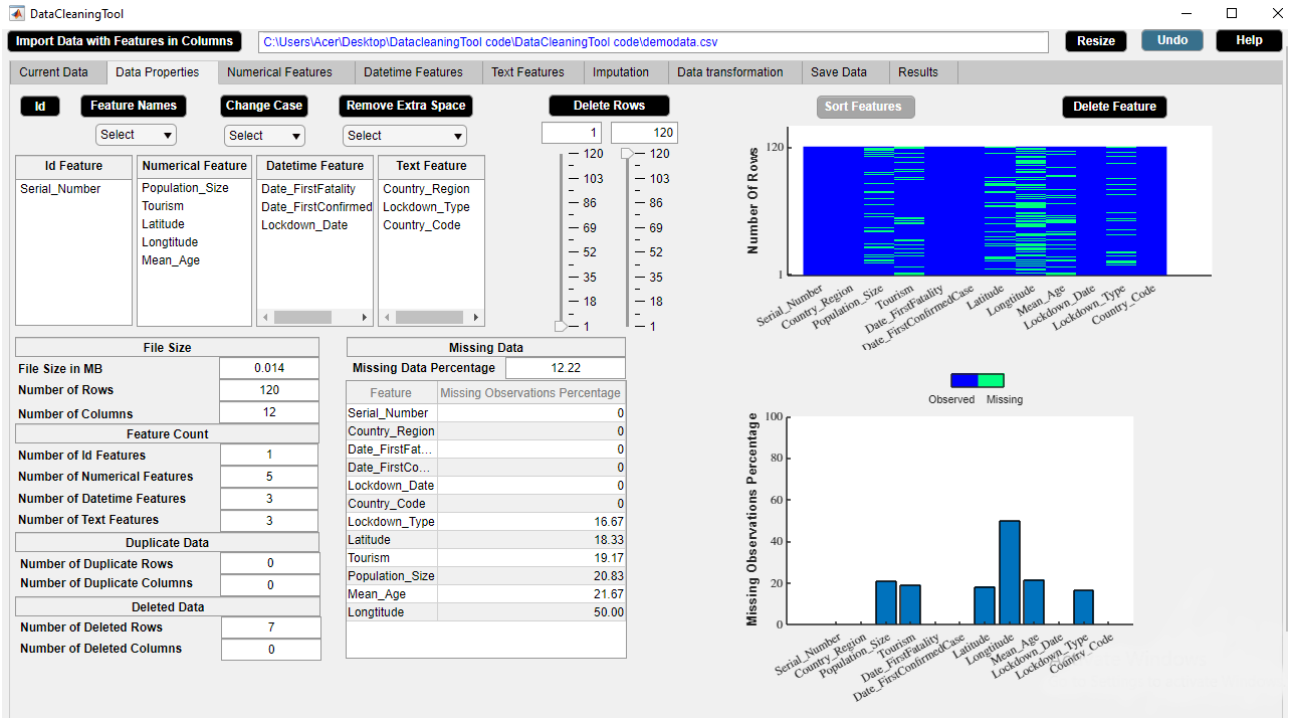


Figure 4.38: Step 2. Sort Features Button

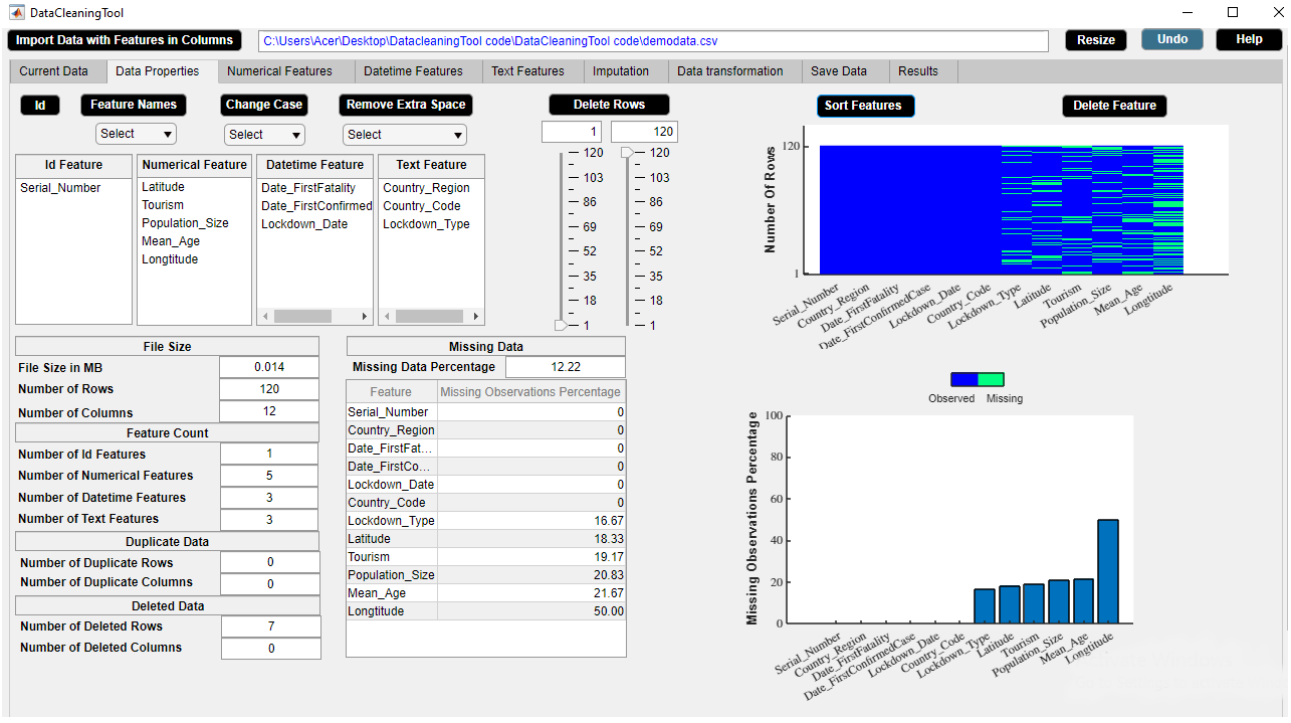


Figure 4.39: Step 3. Sort Features Button

4.2.7 Delete Feature Button

Delete a feature from data.

Application

- Delete an unwanted or irrelevant feature.
- Delete a feature containing a large number of missing observations.

Example

Step 1: Select a feature from **Feature** column of missing observations percentage table.

Step 2: Click **Delete Feature** button.

Step 3: **Delete Feature** button in use turns grey in color.

Step 4: **Delete Feature** button returns back to its original color once it completes its task.

From a data analyst's point of view, 'Country_Code' is an irrelevant feature in the example data. We use **Delete Feature** button to delete 'Country_Code' feature. Figures 4.40-4.43 illustrate how to use **Delete Feature** button.

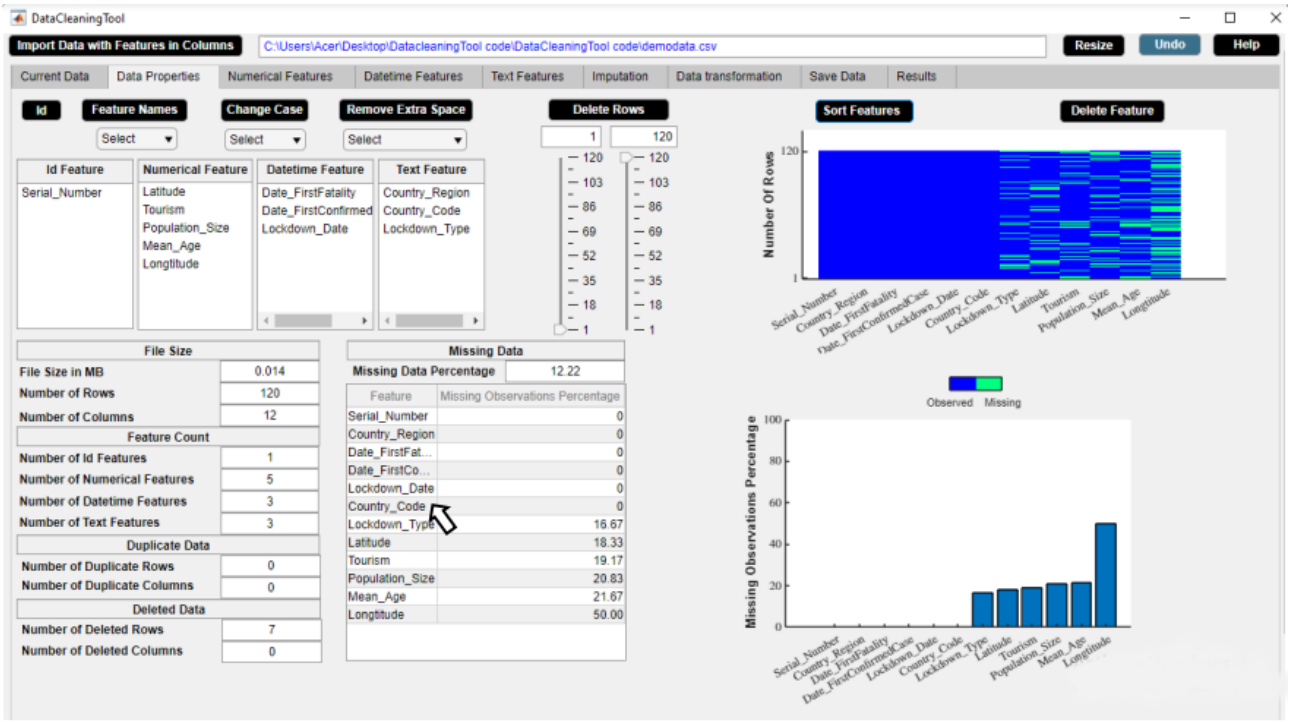


Figure 4.40: Step 1. Delete Feature Button

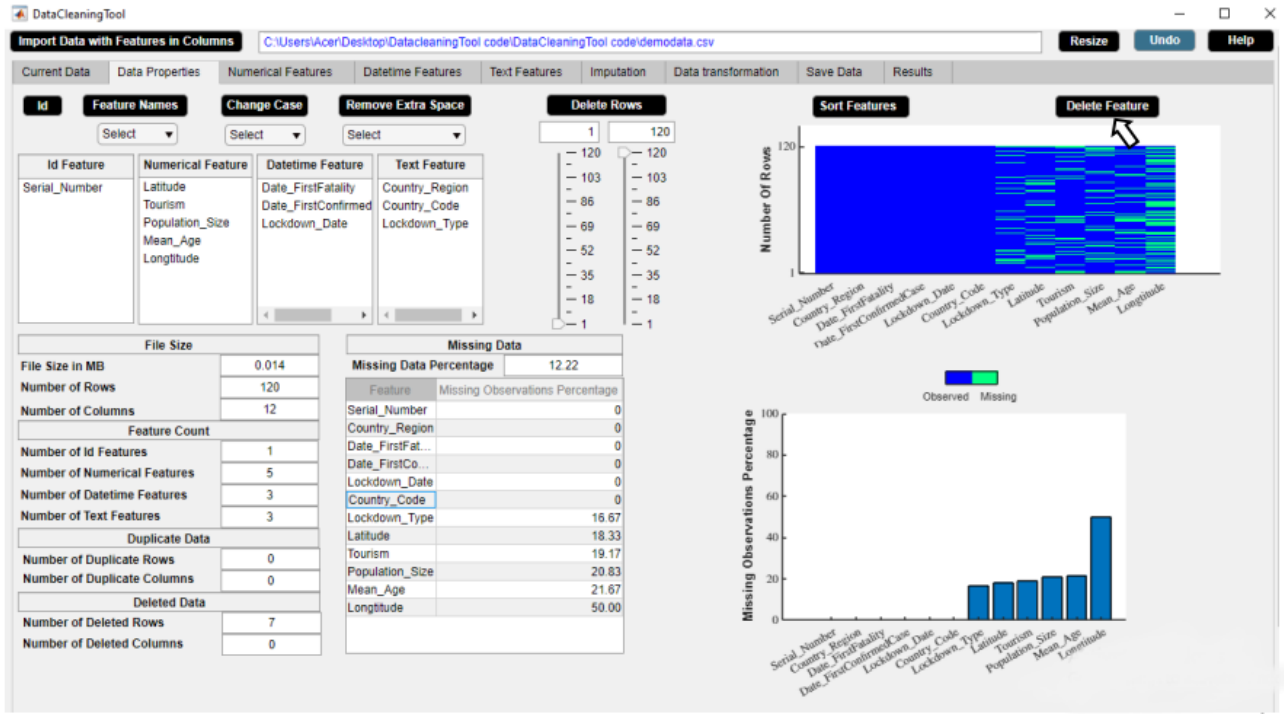


Figure 4.41: Step 2. Delete Feature Button

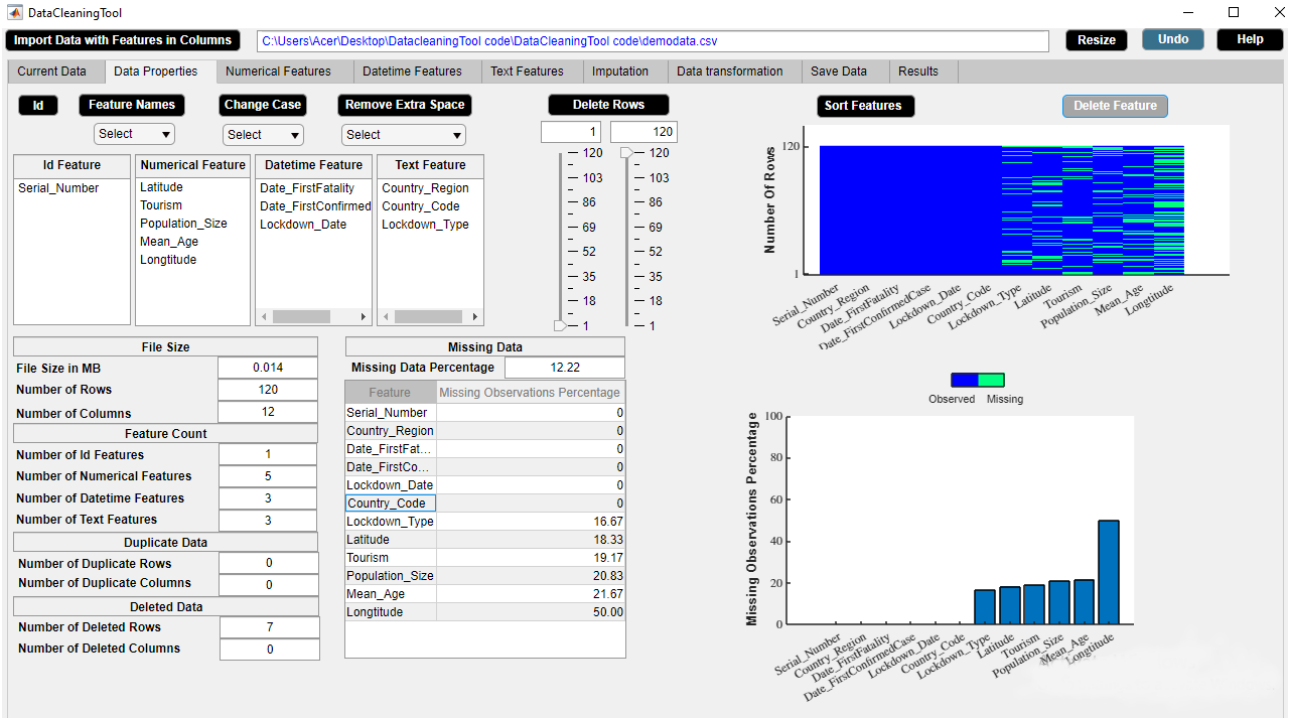


Figure 4.42: Step 3. Delete Feature Button

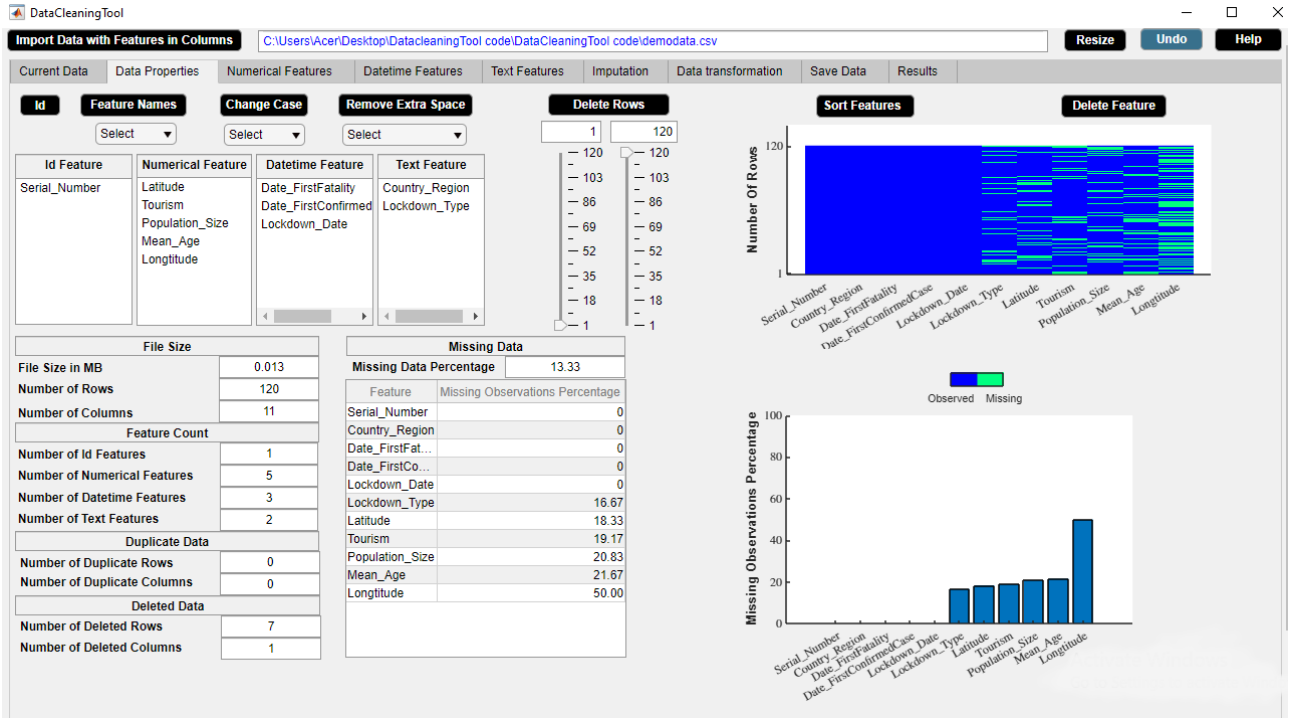


Figure 4.43: Step 4. Delete Feature Button

4.3 Numerical Features Widget

The Numerical Features widget displays statistical description of the numerical data. The Numerical Features widget is shown in figure 4.44. The properties of the Numerical Features widget are as follows.

- The widget shows the descriptive statistics of each numerical feature of the data such as minimum observation and maximum observation of the feature. Descriptive statistics of a feature gives a quantitative description of a feature.
- The widget shows the duplicate observations present in each numerical feature and the missing observations percentage of each numerical feature. Duplicate observation can be an error in the data and could possibly influence later analyses of the data.
- Cross-field validation constraint and range constraint can be set in the widget. This will result in some unwanted numerical observations.
- The statistical information of the numerical data in the widget gets updated after each activity.

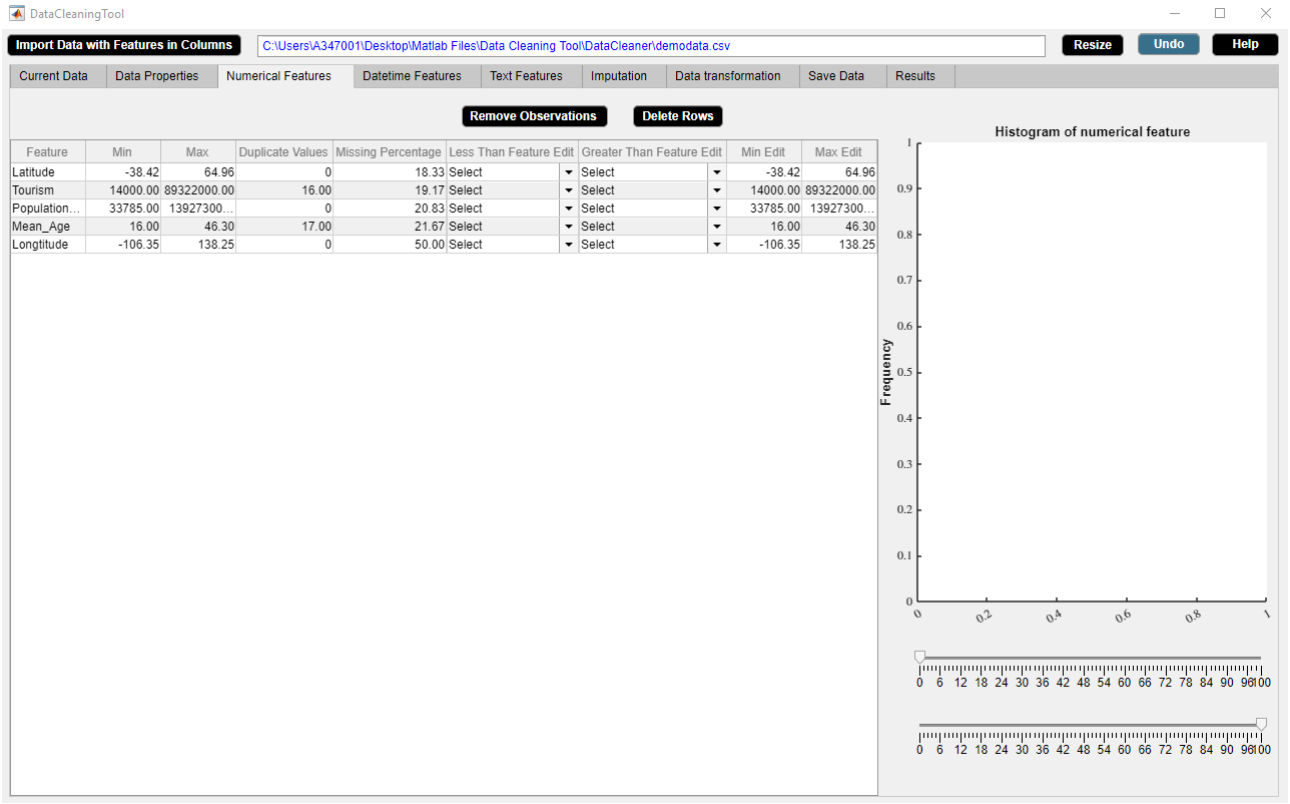


Figure 4.44: Numerical Features Widget.

4.3.1 Numerical Feature Cell Selection Button

Displays histogram of a numerical feature.

Application

- Outlier visualization technique.

Example

Step 1: Select a numerical feature from **Feature** column of the numerical features descriptive statistics table.

Step 2: A histogram of the selected numerical feature appears in the right side of the **Numerical Features** widget and the sliders get updated accordingly.

We use **Numerical Feature Cell Selection** button to visualize the histogram of 'Population_Size' feature. Figures 4.45-4.46 illustrate how to use **Numerical Feature Cell Selection** button.

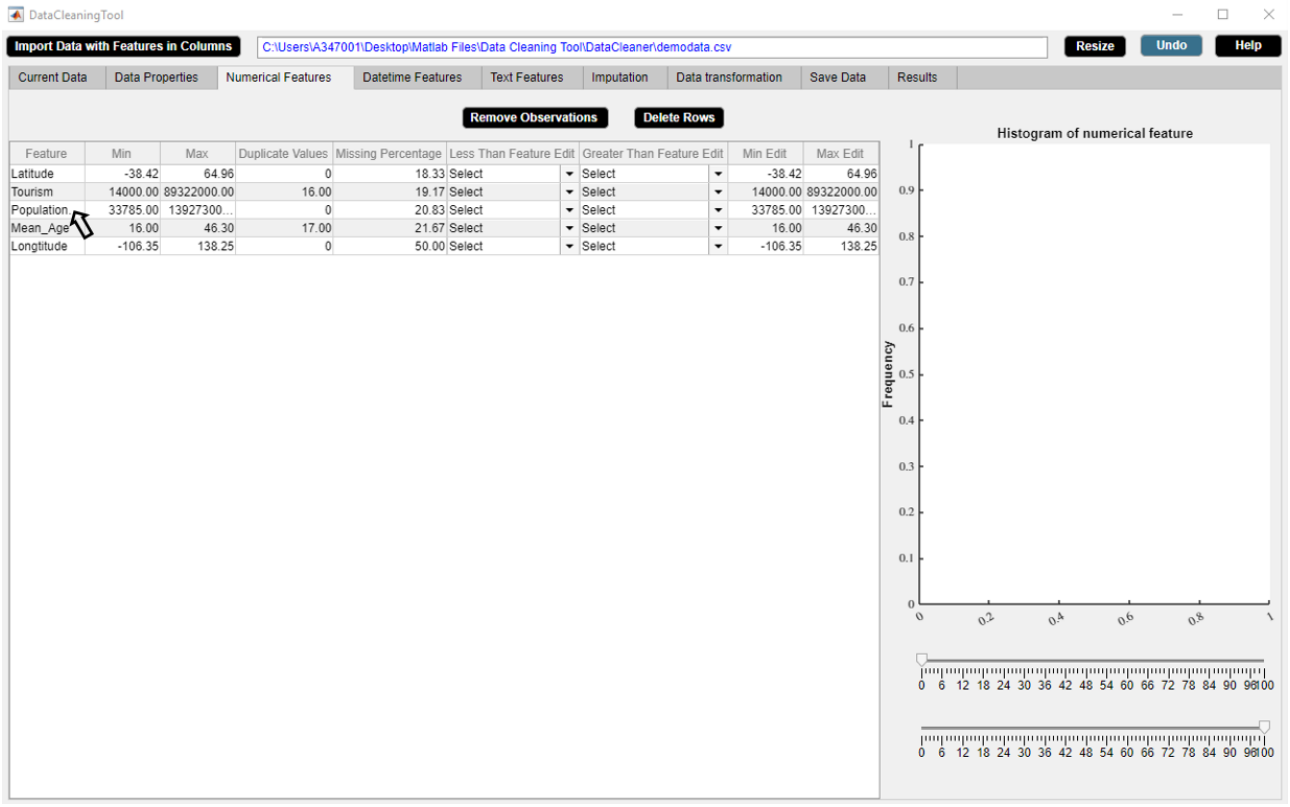


Figure 4.45: Step 1. Numerical Feature Cell Selection Button

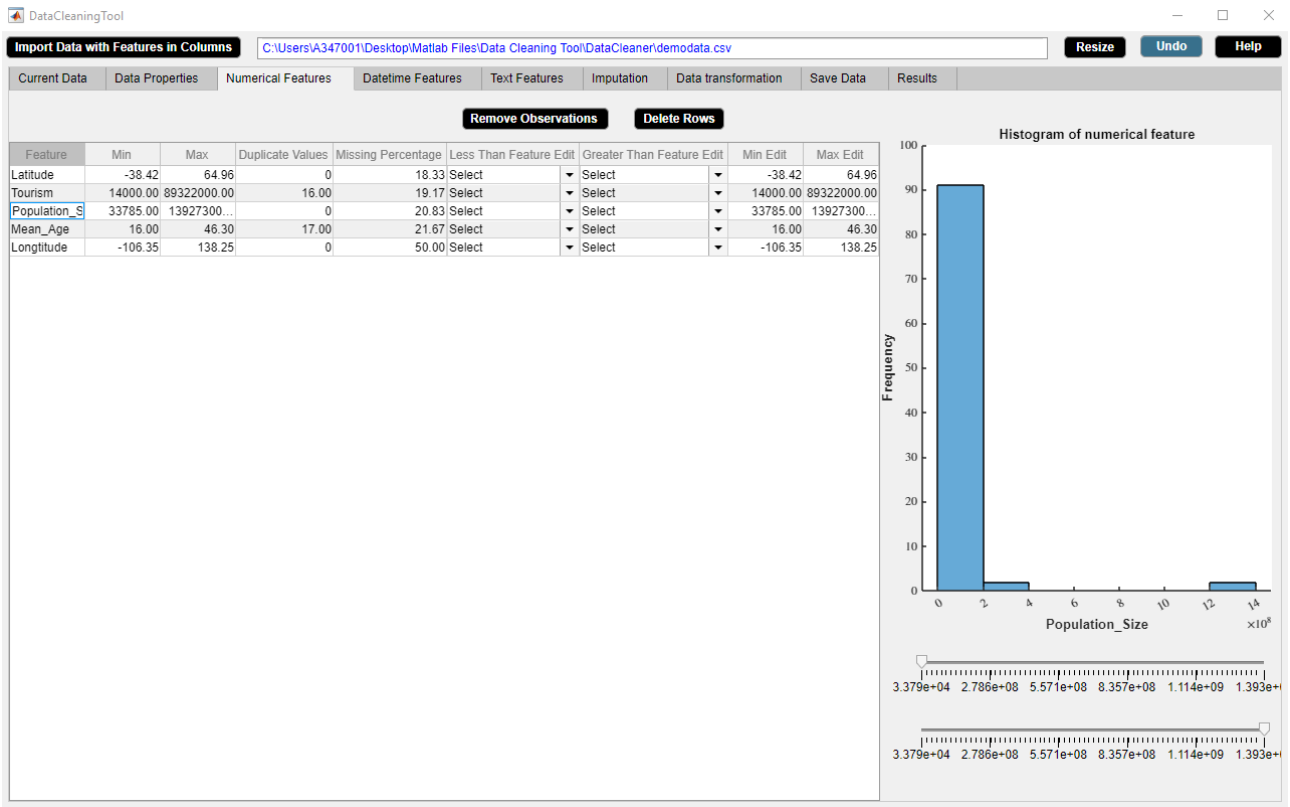


Figure 4.46: Step 2. Numerical Feature Cell Selection Button

4.3.2 Remove Observations Button

Replaces unwanted numerical observations by missing values.

Application

- Removes unwanted or irrelevant observations.

Example

Step 1: Set cross-field validation constraint from **Less Than Feature Edit** dropdown menu or **Greater Than Feature Edit** dropdown menu or set range constraint from **Min Edit** box or **Max Edit** box in the **Numerical Features** widget.

Step 2: Click **Remove Observations** button.

Step 3: **Remove Observations** button in use turns grey in color.

Step 4: **Remove Observations** button returns back to its original color once it completes its task.

We wish to prepare the data for analysis for the countries whose 'Population_Size' is greater than 'tourism'. We use **Remove Observations** button to extract data for the countries whose 'Population_Size' is greater than 'Tourism'. Figures 4.47-4.50 illustrate how to use **Remove Observations** button.

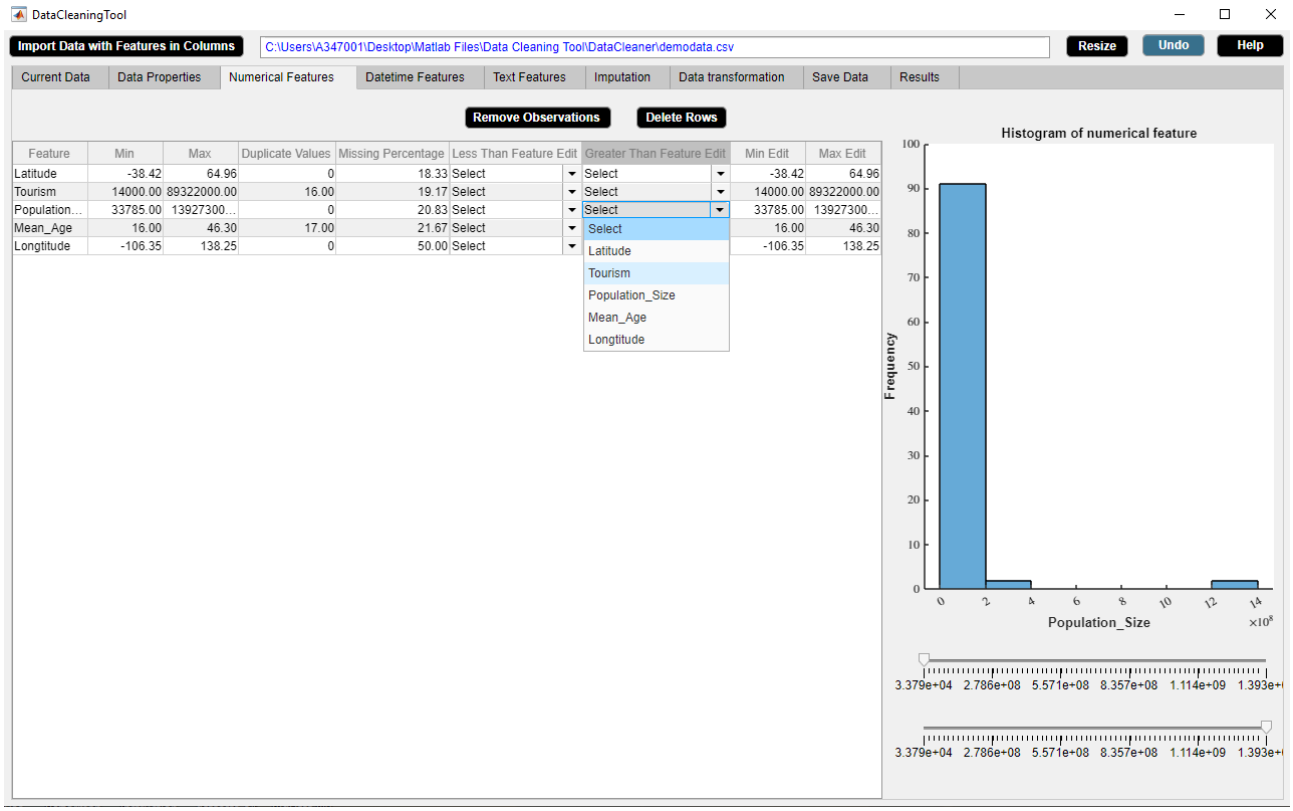


Figure 4.47: Step 1. Remove Observations Button

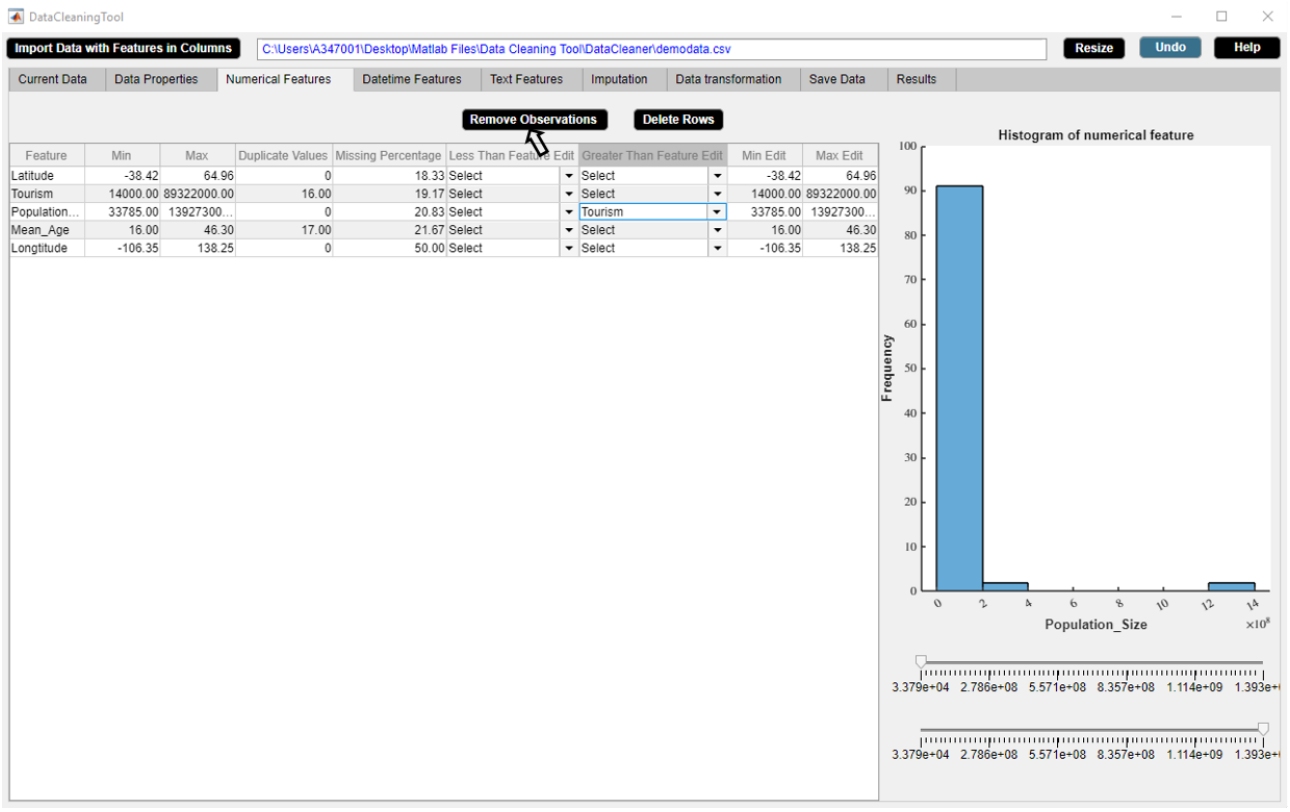


Figure 4.48: Step 2. Remove Observations Button

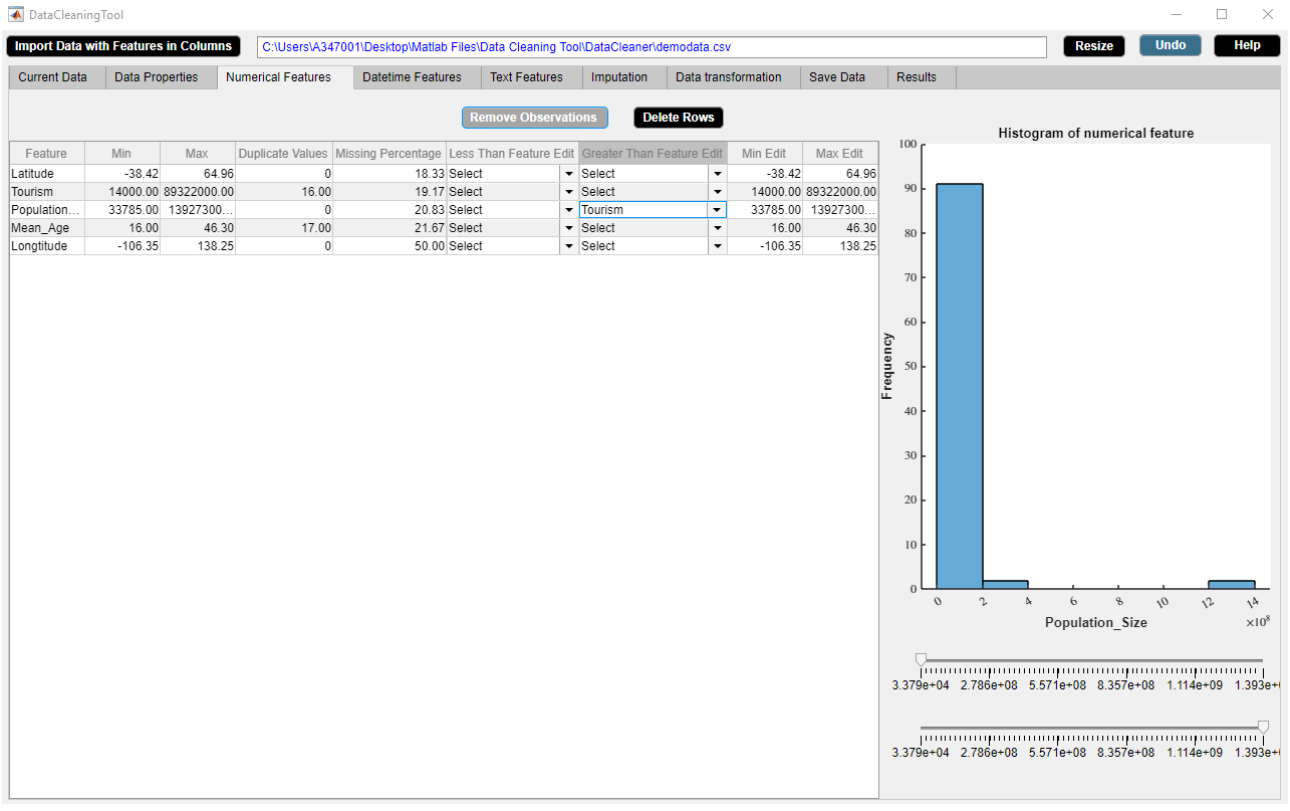


Figure 4.49: Step 3. Remove Observations Button

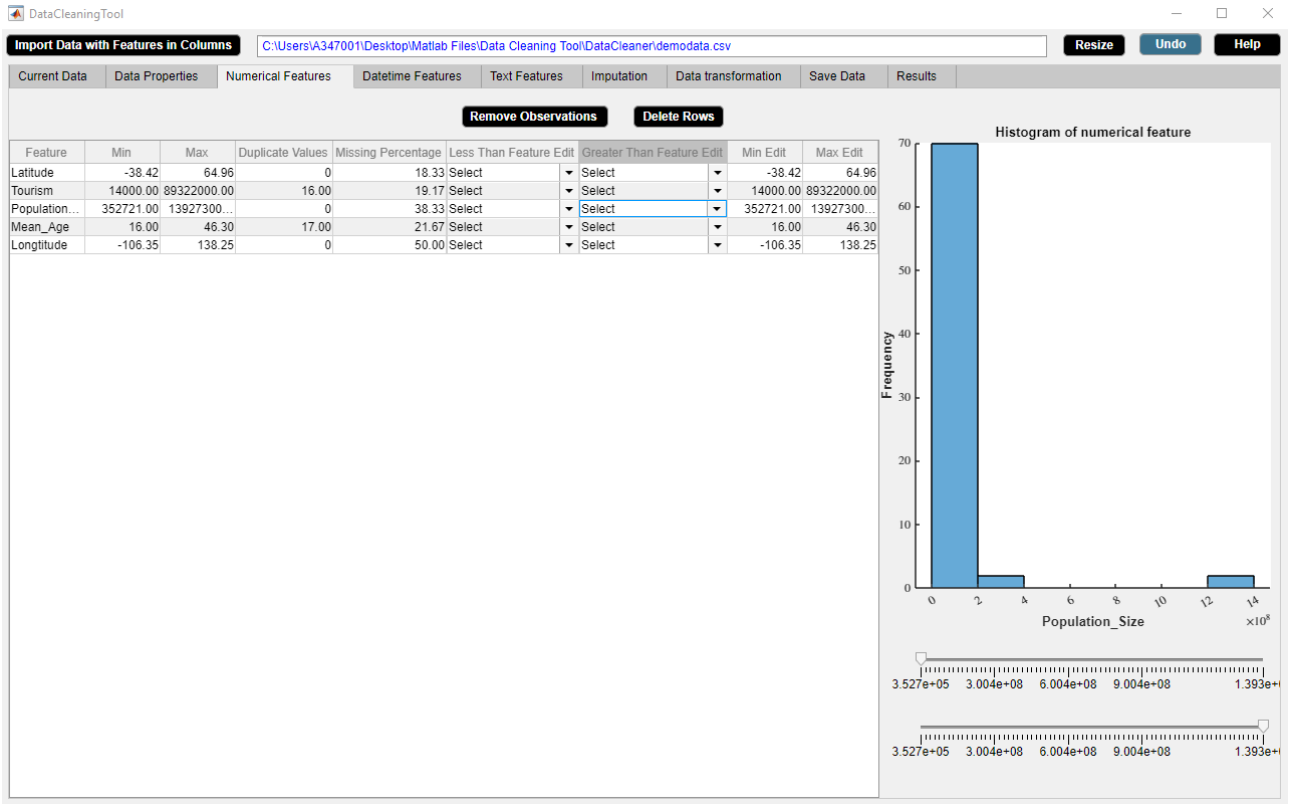


Figure 4.50: Step 4. Remove Observations Button

4.3.3 Delete Rows Button

Deletes rows with unwanted numerical observations.

Application

- Delete unwanted or irrelevant rows.
- Delete rows containing a large number of missing observations.

Example

Step 1: Select a numerical feature from **Feature** column of the numerical features descriptive statistics table.

Step 2: A histogram of the selected numerical feature appears in the right side of the **Numerical Features** widget and the sliders get updated accordingly. Set cross-field validation constraint from **Less Than Feature Edit** dropdown menu or **Greater Than Feature Edit** dropdown menu or set range constraint from **Min Edit** box or **Max Edit** box of the numerical features descriptive statistics table in the **Numerical Features** widget. Also, minimum value and maximum value can be selected from sliders.

Step 3: Click **Delete Rows** button.

Step 4: **Delete Rows** button in use turns grey in color.

Step 5: **Delete Rows** button returns back to its original color once it completes its task.

We wish to prepare the data for analysis for the countries whose maximum 'Mean_age' is 45. We use **Delete Rows** button to extract data for the countries whose maximum 'Mean_age' is 45. Figures 4.51-4.55 illustrate how to use **Delete Rows** button.

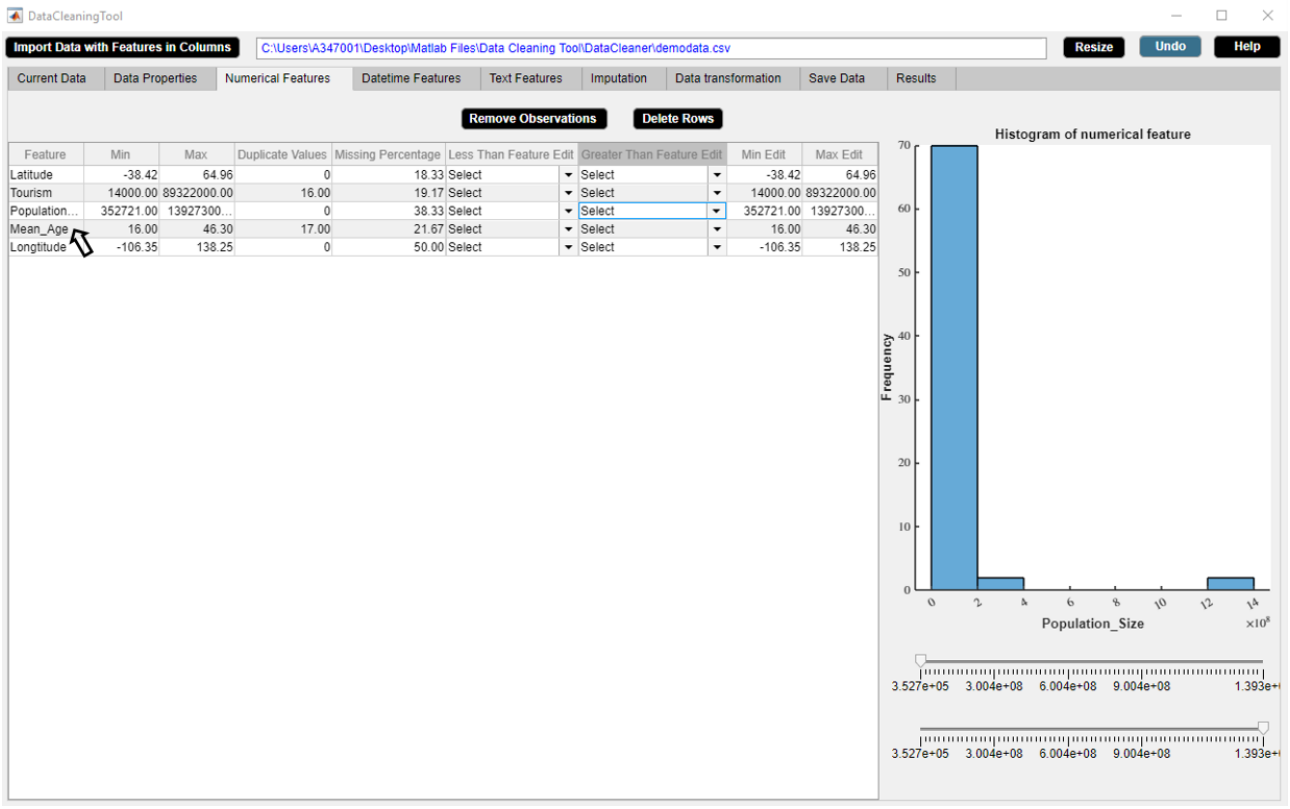


Figure 4.51: Step 1. Delete Rows Button

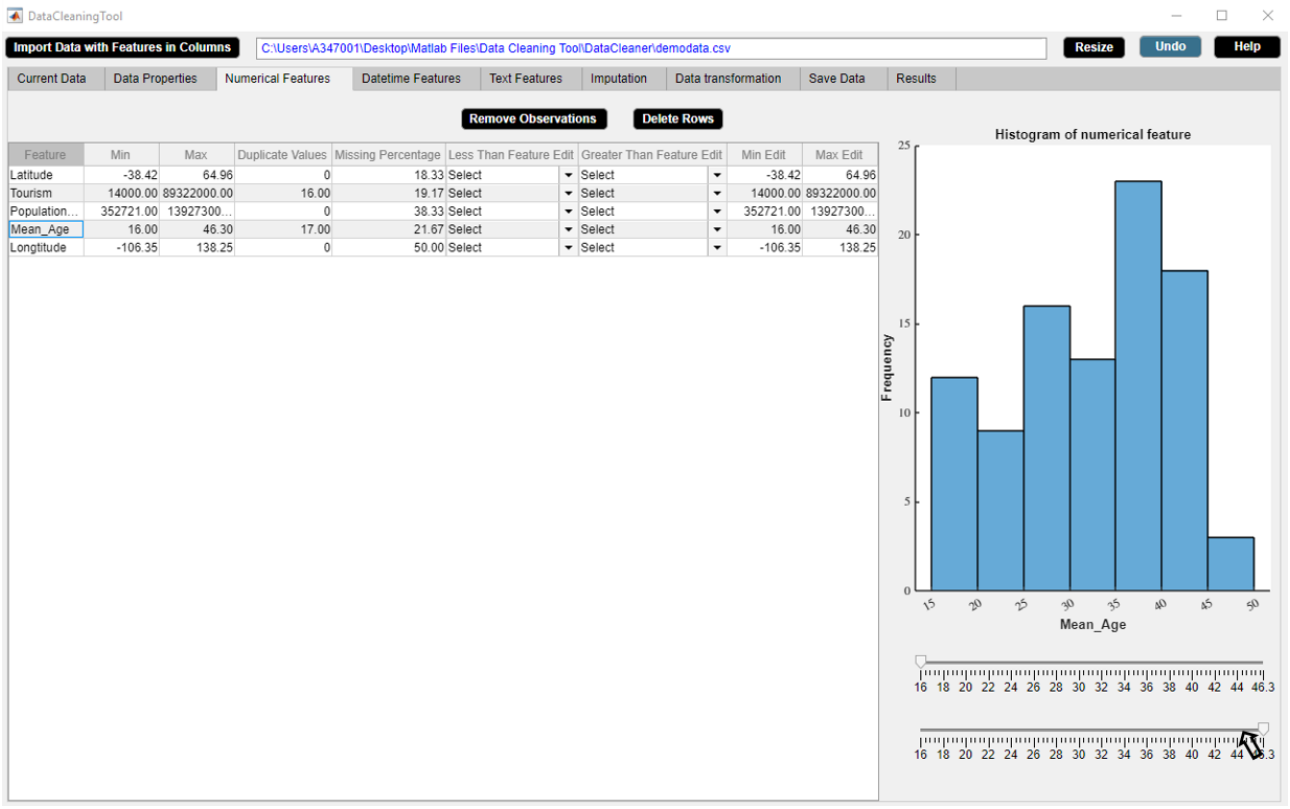


Figure 4.52: Step 2. Delete Rows Button

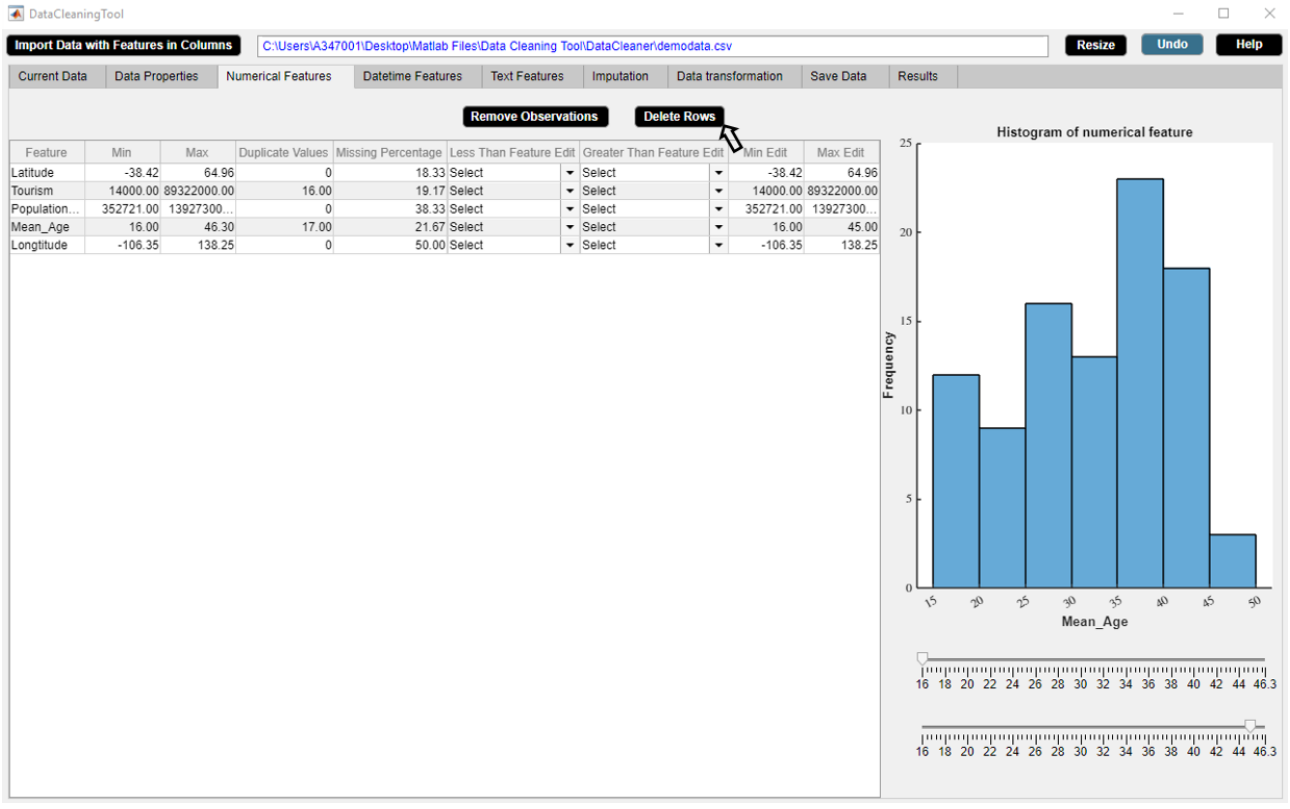


Figure 4.53: Step 3. Delete Rows Button

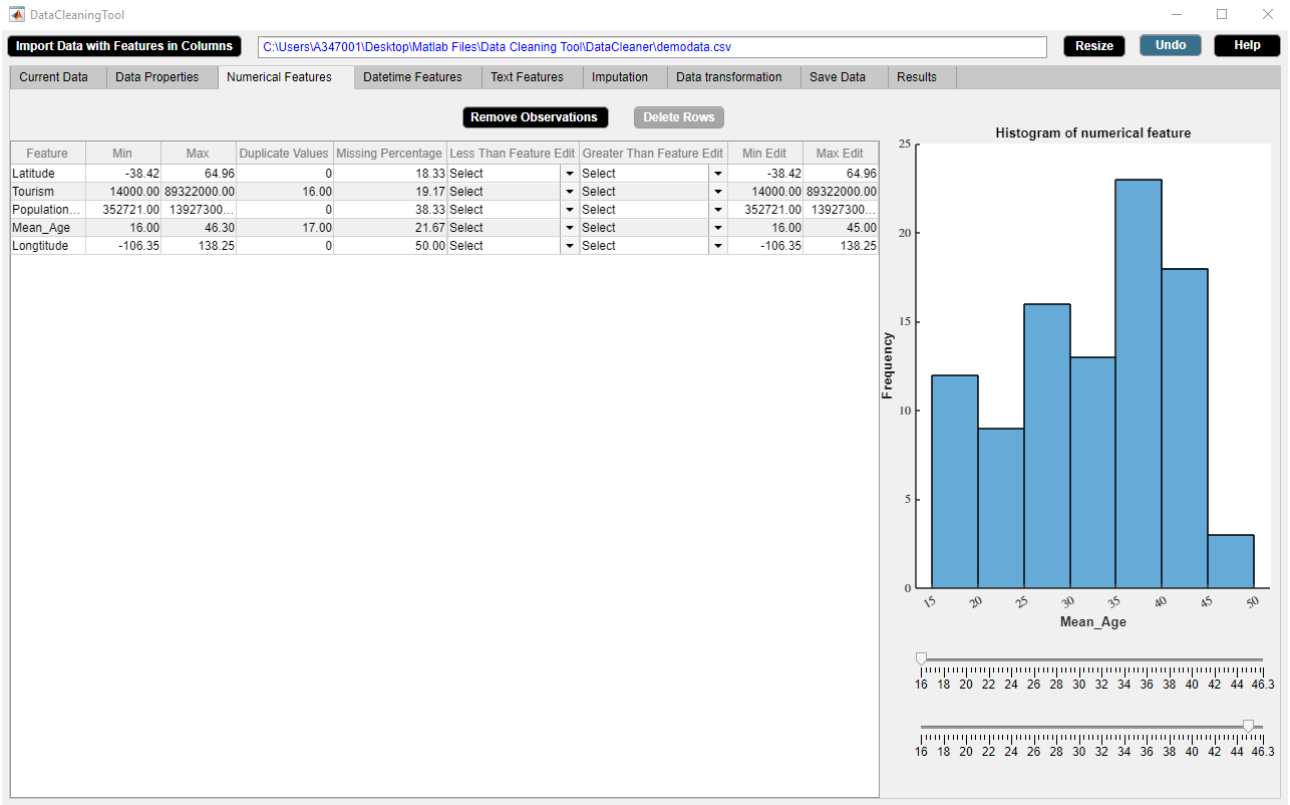


Figure 4.54: Step 4. Delete Rows Button

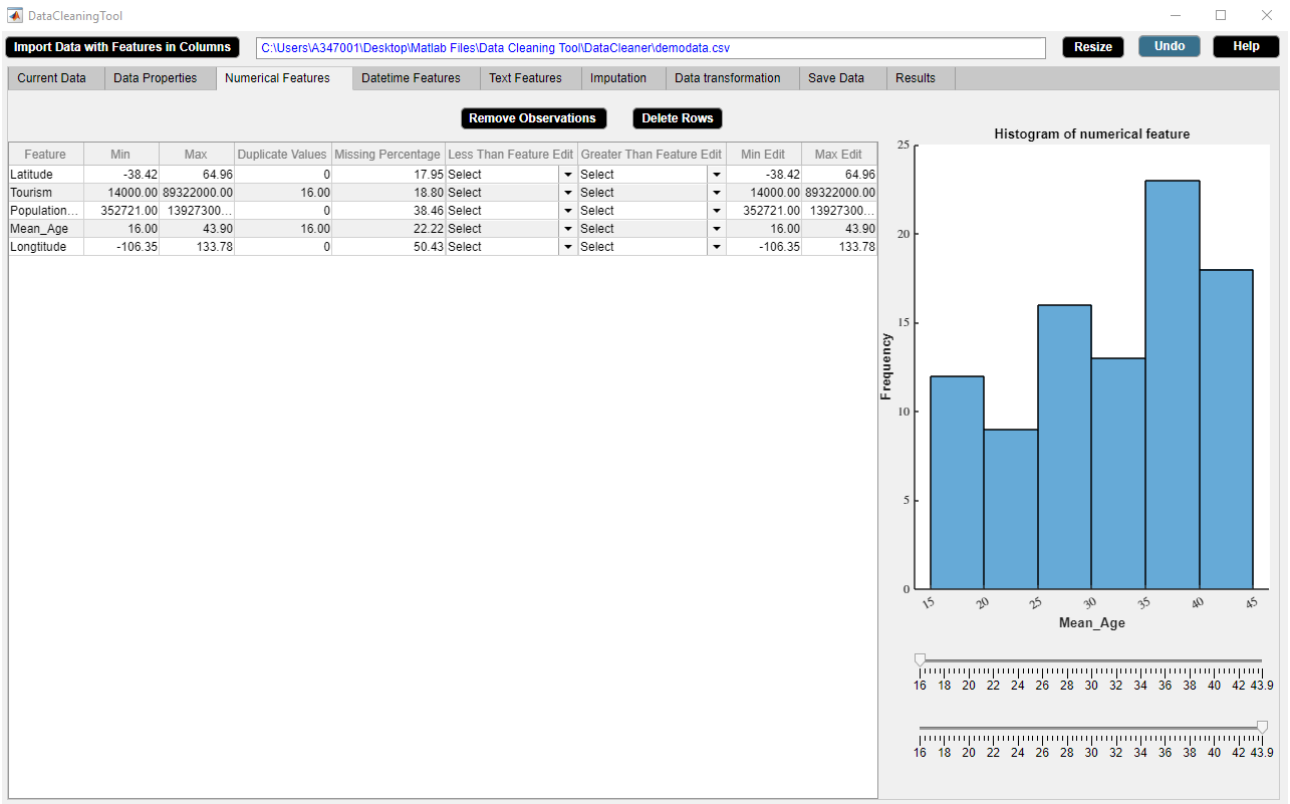


Figure 4.55: Step 5. Delete Rows Button

4.4 Datetime Features Widget

The Datetime Features widget displays statistical description of the datetime data. The Datetime Features widget is shown in figure 4.56. The properties of the Datetime Features widget are as follows.

- The widget shows the descriptive statistics of each datetime feature of the data such as minimum observation and maximum observation of the feature.
- The widget also shows the missing observations percentage of each datetime feature.
- Datetime format can be changed.
- Cross-field validation constraint and range constraint can be set in the widget for each datetime feature. This will result in some unwanted datetime observations.
- The statistical information of the datetime data in the widget gets updated after each activity.

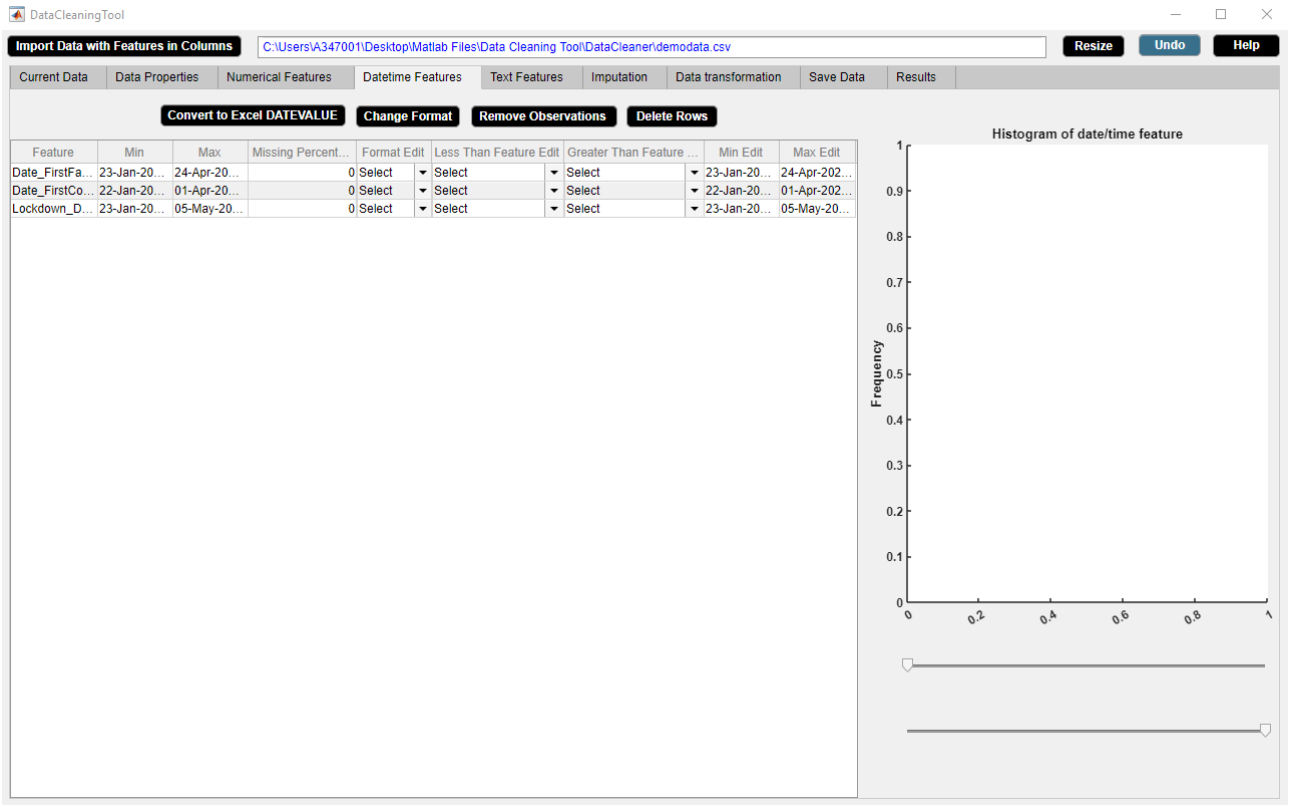


Figure 4.56: Datetime Features Widget.

4.4.1 Datetime Feature Cell Selection Button

Displays histogram of a datetime feature.

Application

- Outlier visualization technique.

Example

Step 1: Select a datetime feature from **Feature** column of the datetime features descriptive statistics table.

Step 2: A histogram of the selected datetime feature appears in the right side of the **Datetime Features** widget and the sliders get updated accordingly.

We use **Datetime Feature Cell Selection** button to visualize the histogram of 'Date_FirstConfirmedCase' feature. Figures 4.57-4.58 illustrate how to use **Datetime Feature Cell Selection** button.

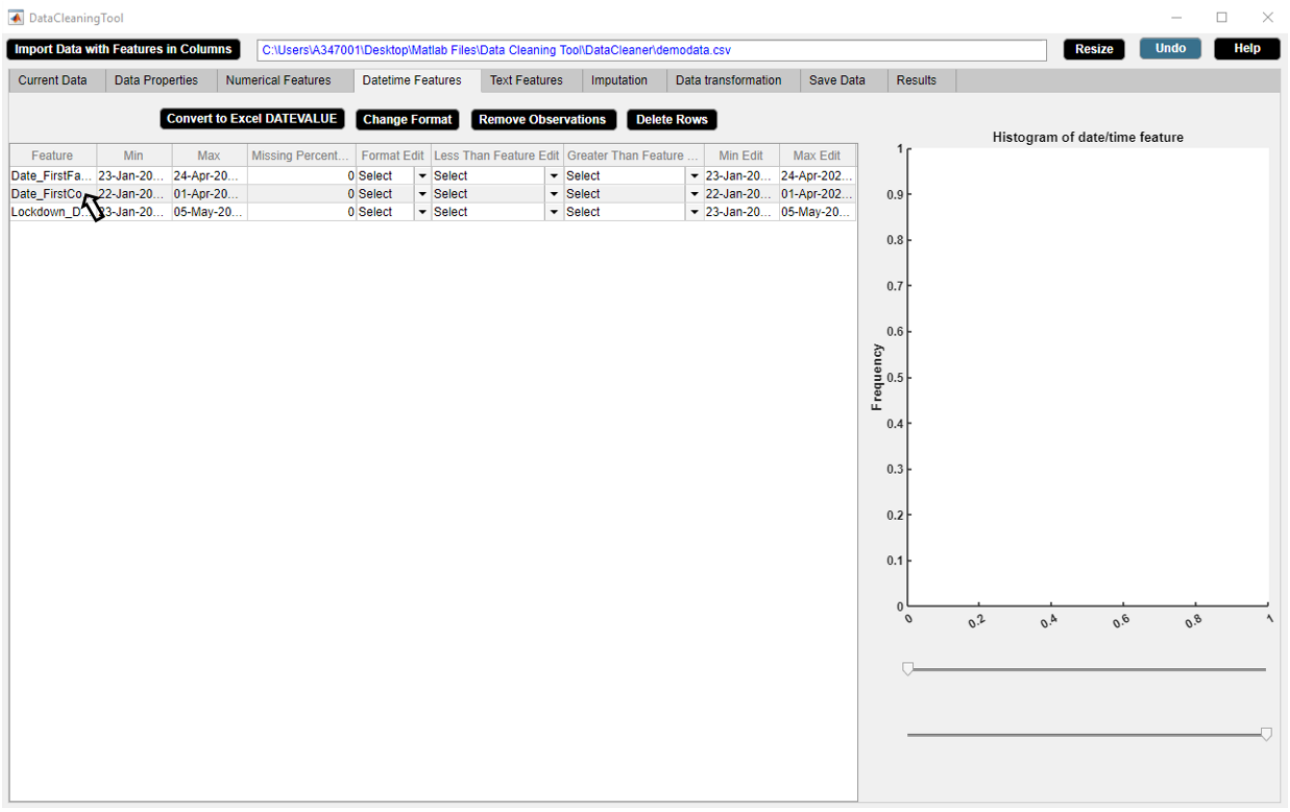


Figure 4.57: Step 1. Datetime Feature Cell Selection Button

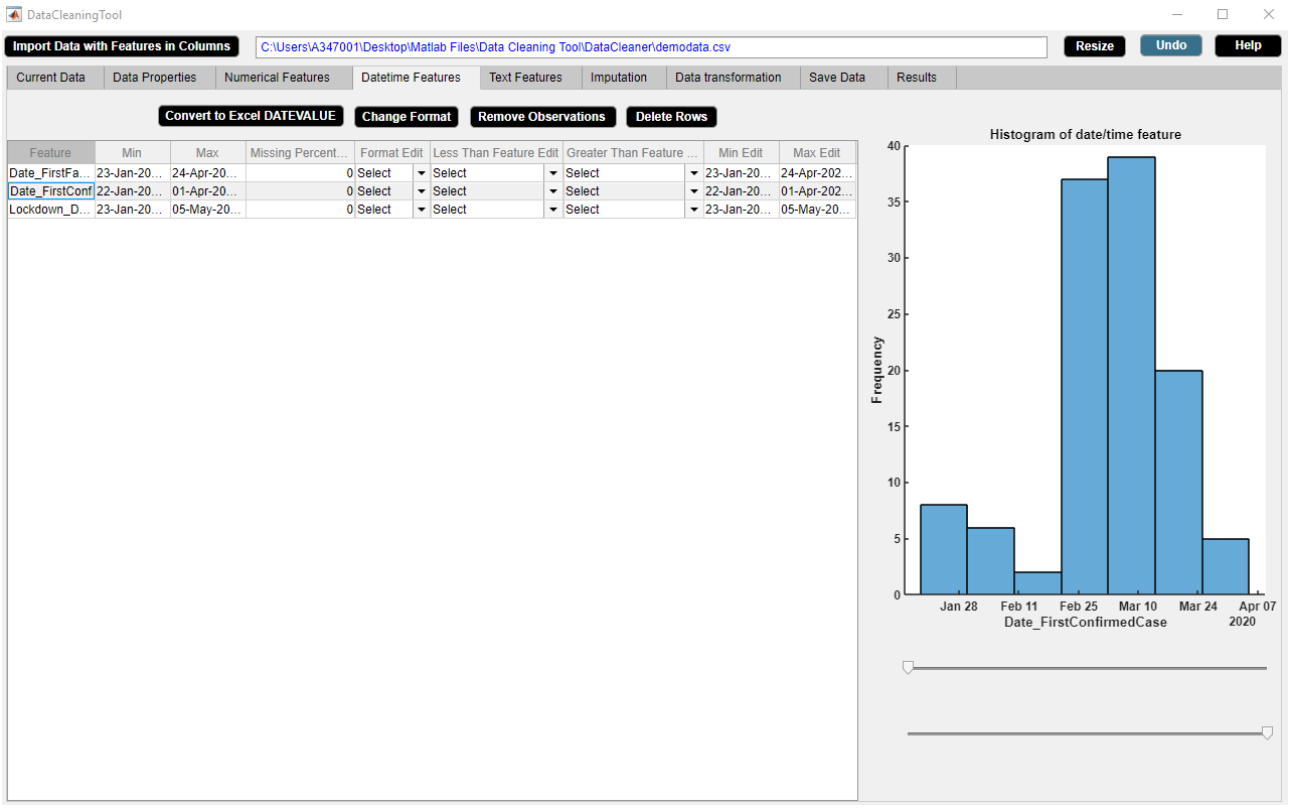


Figure 4.58: Step 2. Datetime Feature Cell Selection Button

4.4.2 Convert To Excel DATEVALUE Button

Converts datetime to Excel DATEVALUE. First it transforms datetime to Matlab serial date number and then to Excel serial date number. MATLAB date numbers start from January 1, 0000 A.D., and hence there is a difference of 693960 relative to the Excel date system which uses January 1, 1900, as starting point.

4.4.3 Change Format Button

Changes datetime format.

Example

Step 1: Select a datetime format from **Format Edit** dropdown menu of the datetime features descriptive statistics table.

Step 2: Click **Change Format** button.

Step 3: **Change Format** button in use turns grey in color.

Step 4: **Change Format** button returns back to its original color once it completes its task.

Step 5: Check the datetime format in the **Current Data** widget.

We use **Change Format** button to change the datetime format of all the datetime features to 'yyyy-MM-dd HH:mm:ss'. Figures 4.59-4.63 illustrate how to use **Change Format** button.

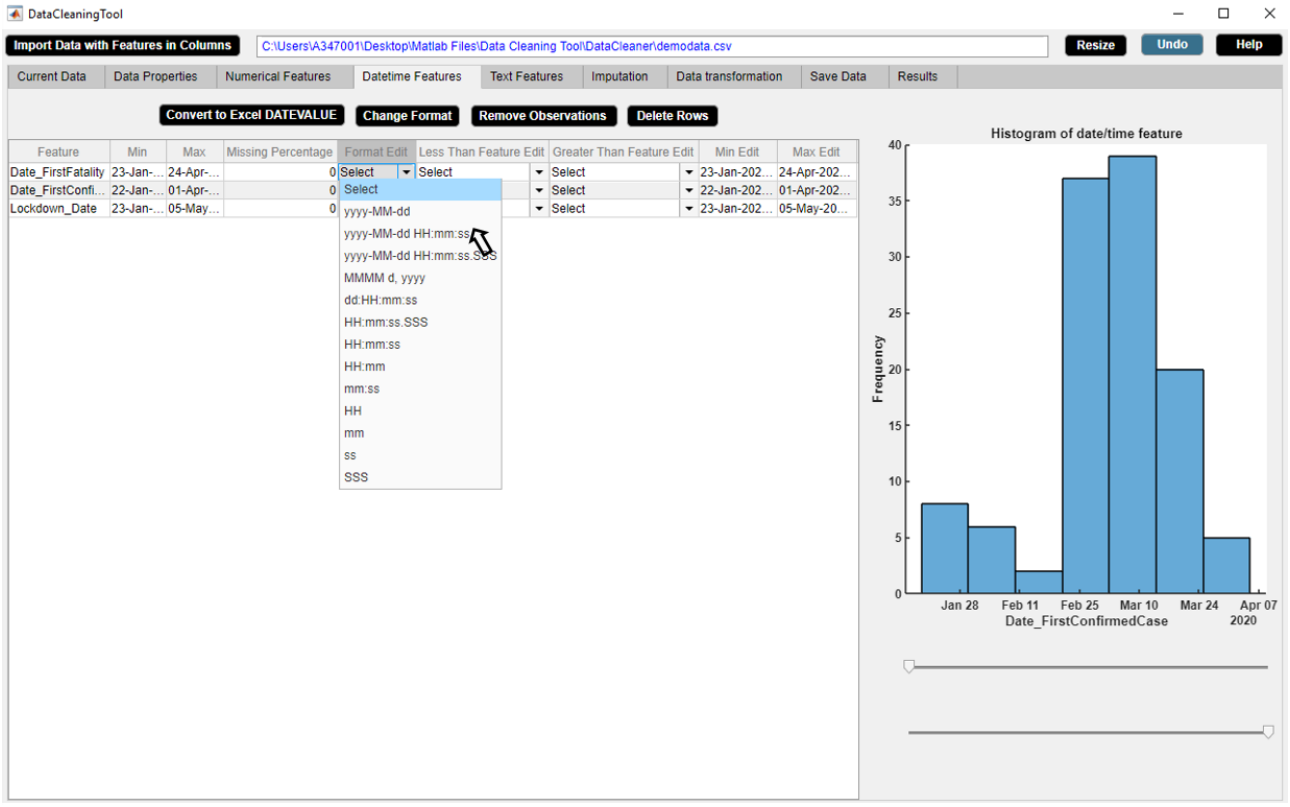


Figure 4.59: Step 1. Change Format Button

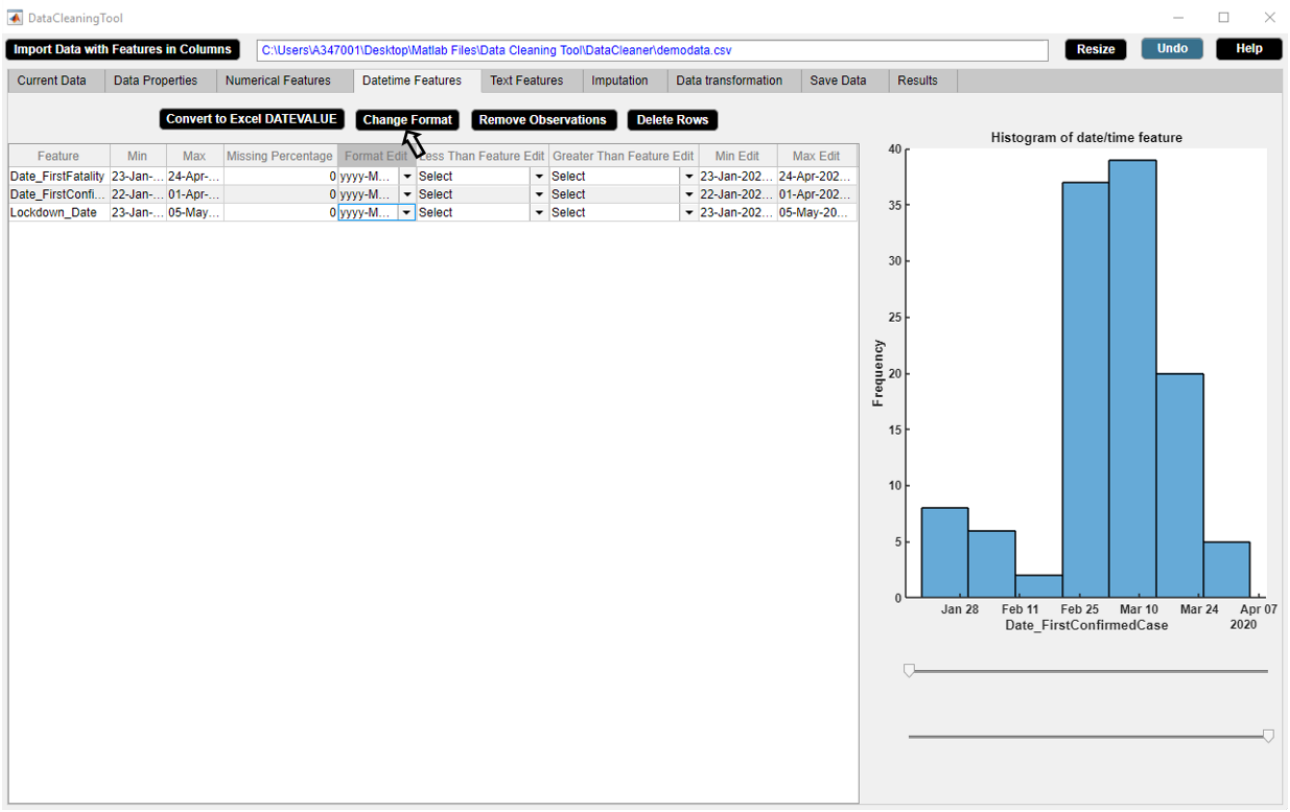


Figure 4.60: Step 2. Change Format Button

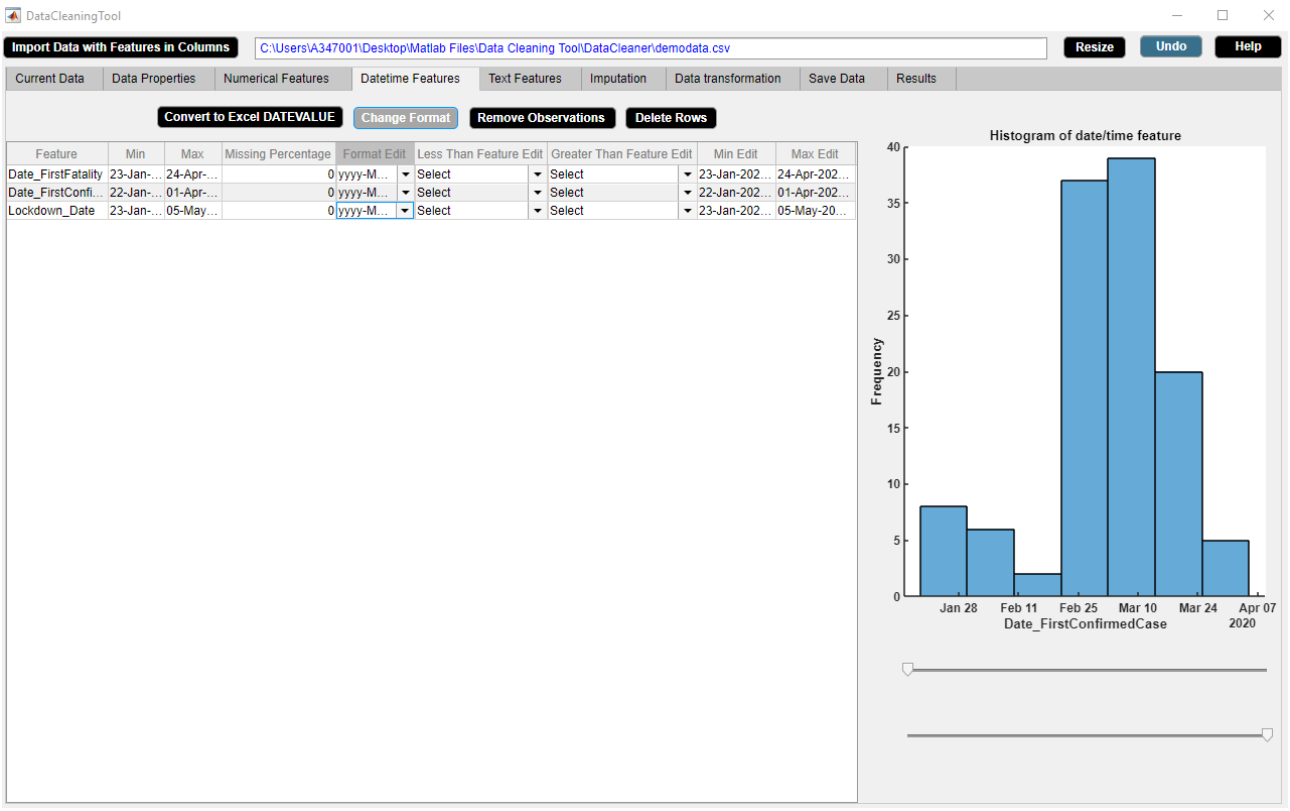


Figure 4.61: Step 3. Change Format Button

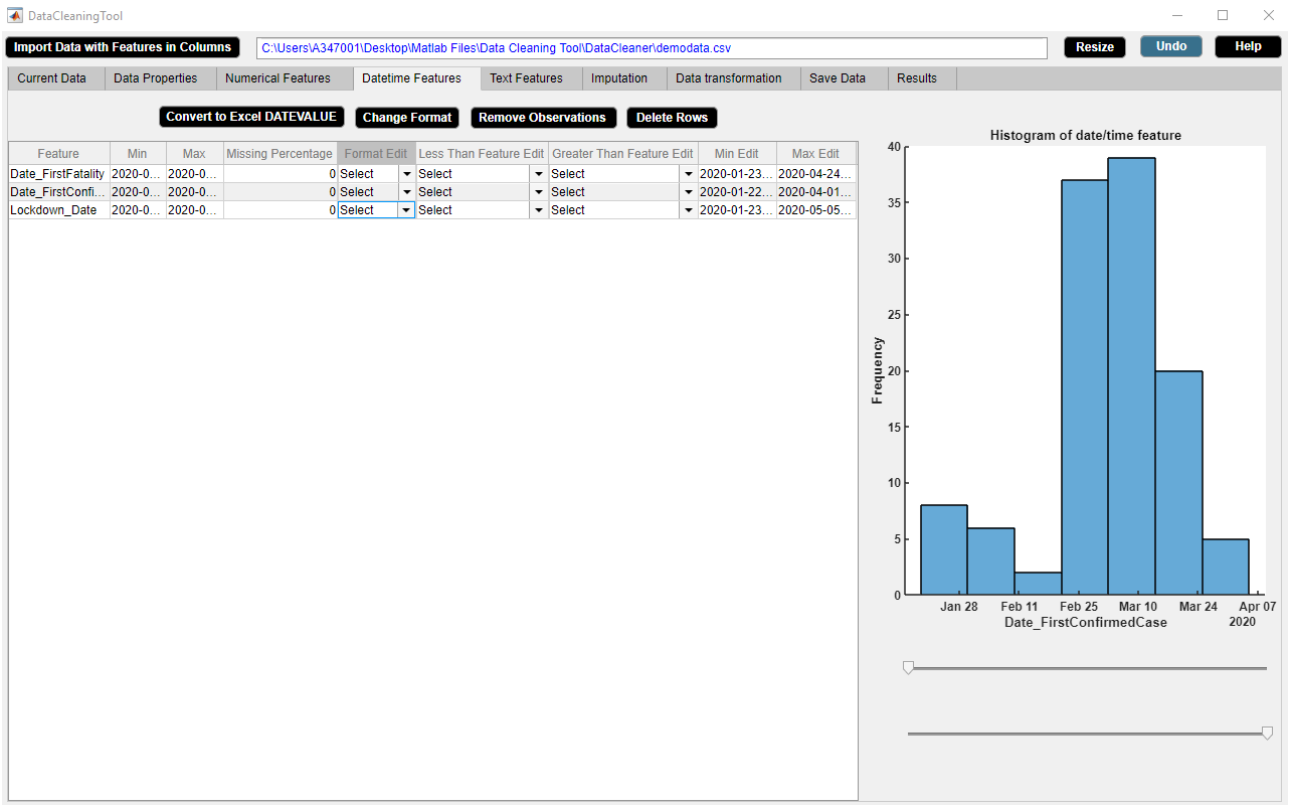


Figure 4.62: Step 4. Change Format Button

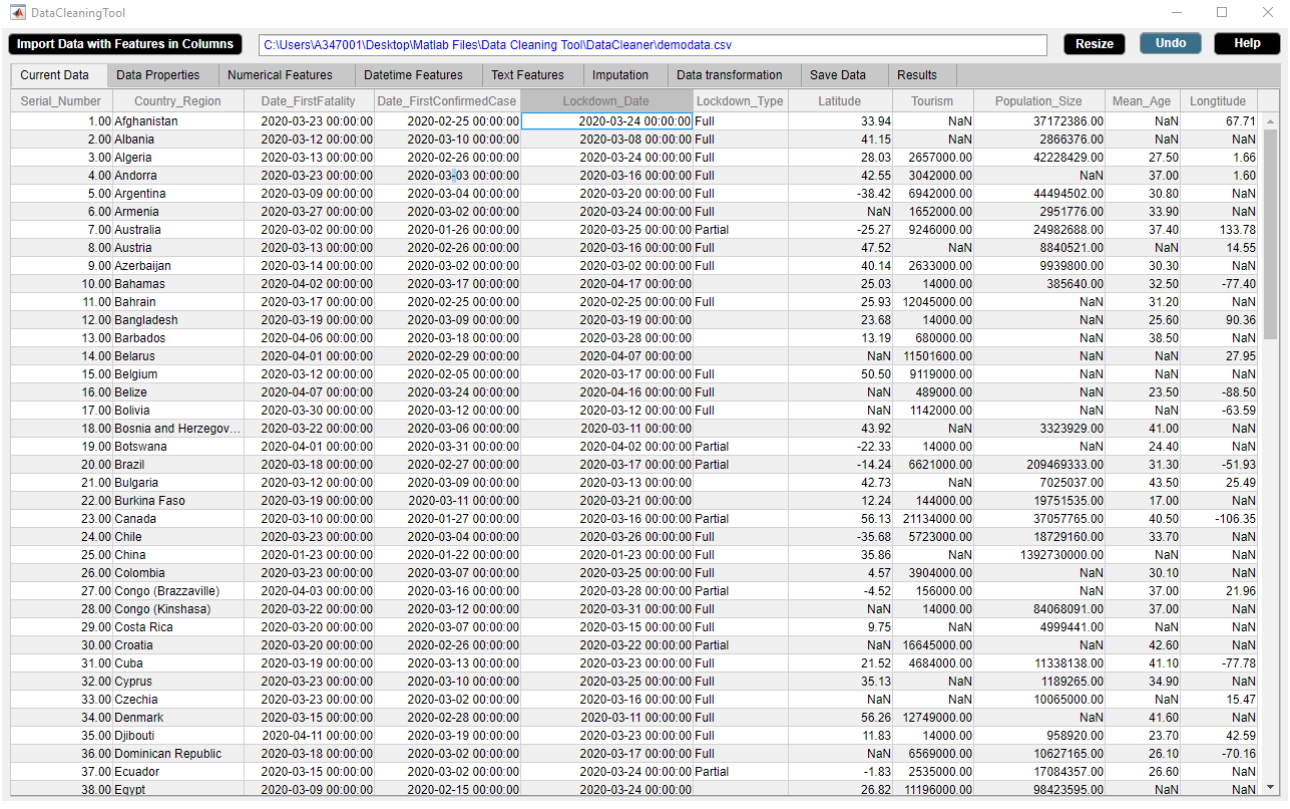


Figure 4.63: Step 5. Change Format Button

4.4.4 Remove Observations Button

Replaces unwanted datetime observations by missing values.

Application

- Remove unwanted or irrelevant observations.

4.4.5 Delete Rows Button

Deletes rows with unwanted datetime observations.

Application

- Delete unwanted or irrelevant rows.
- Delete rows containing a large number of missing observations.

Example

Step 1: Set cross-field validation constraint from **Less Than Feature Edit** dropdown menu or **Greater Than Feature Edit** dropdown menu or set range constraint from **Min Edit** box or **Max Edit** box of the datetime features descriptive statistics table in the **Datetime Features** widget.

Step 2: Click **Delete Rows** button.

Step 3: **Delete Rows** button in use turns grey in color.

Step 4: **Delete Rows** button returns back to its original color once it completes its task.

We wish to prepare the data for analysis for the countries whose 'Date_FirstConfirmedCase' is less than 'Date_FirstFatality'. We use **Delete Rows** button to extract data for the countries whose 'Date_FirstConfirmedCase' is less than 'Date_FirstFatality'. Figures 4.64-4.67 illustrate how to use **Delete Rows** button.

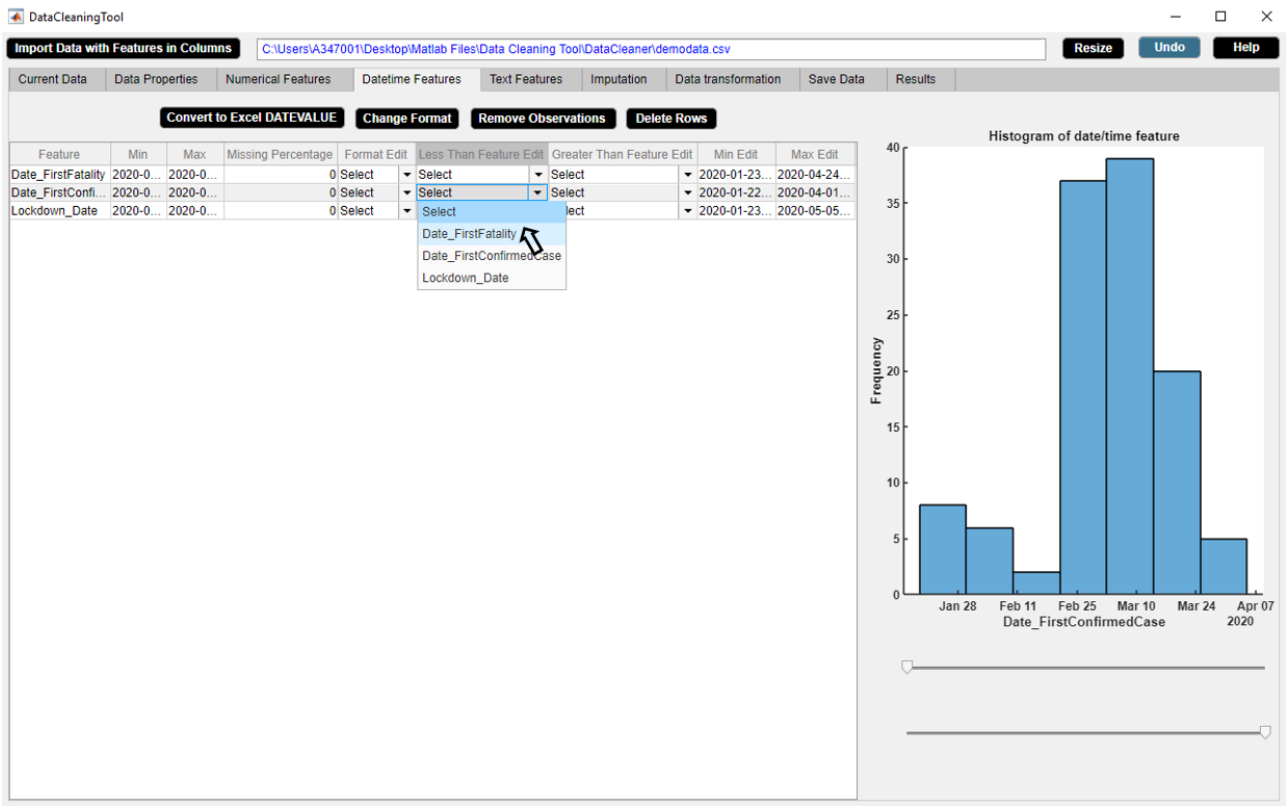


Figure 4.64: Step 1. Delete Rows Button

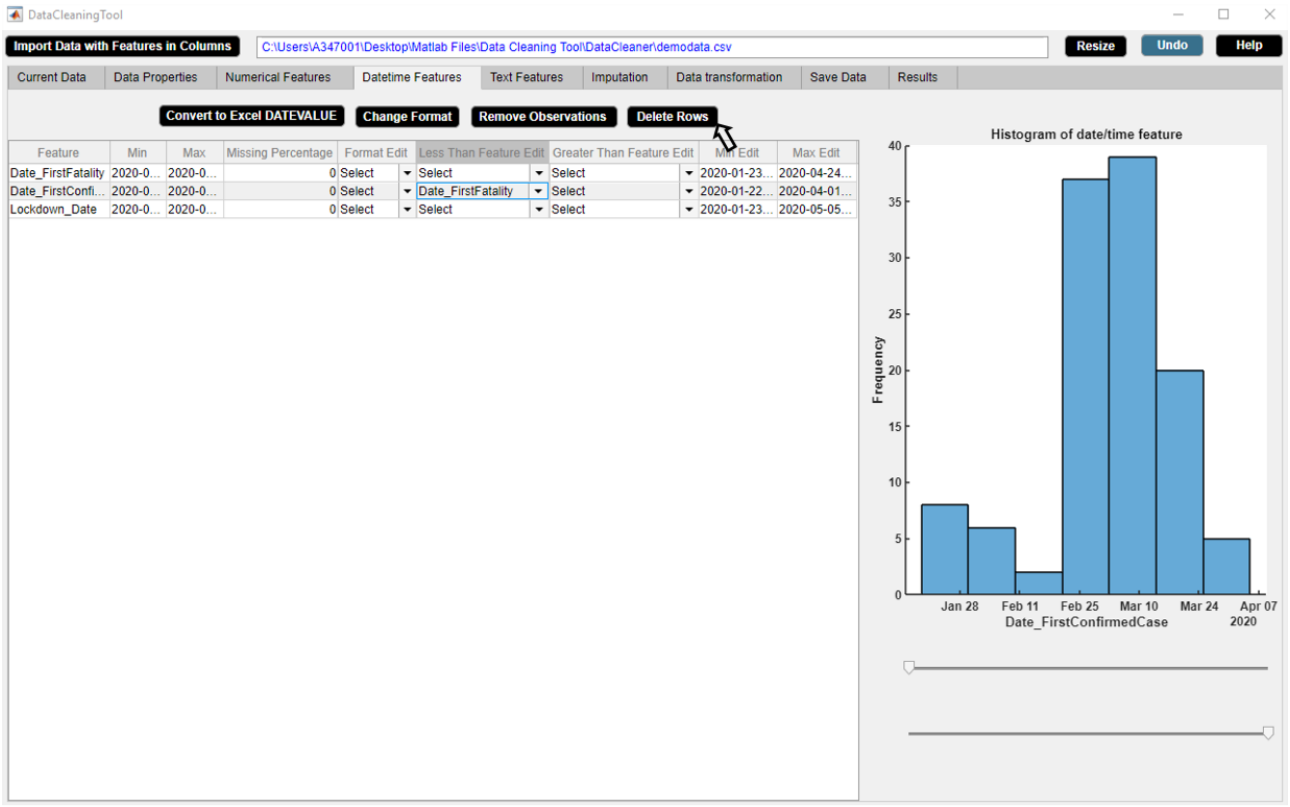


Figure 4.65: Step 2. Delete Rows Button

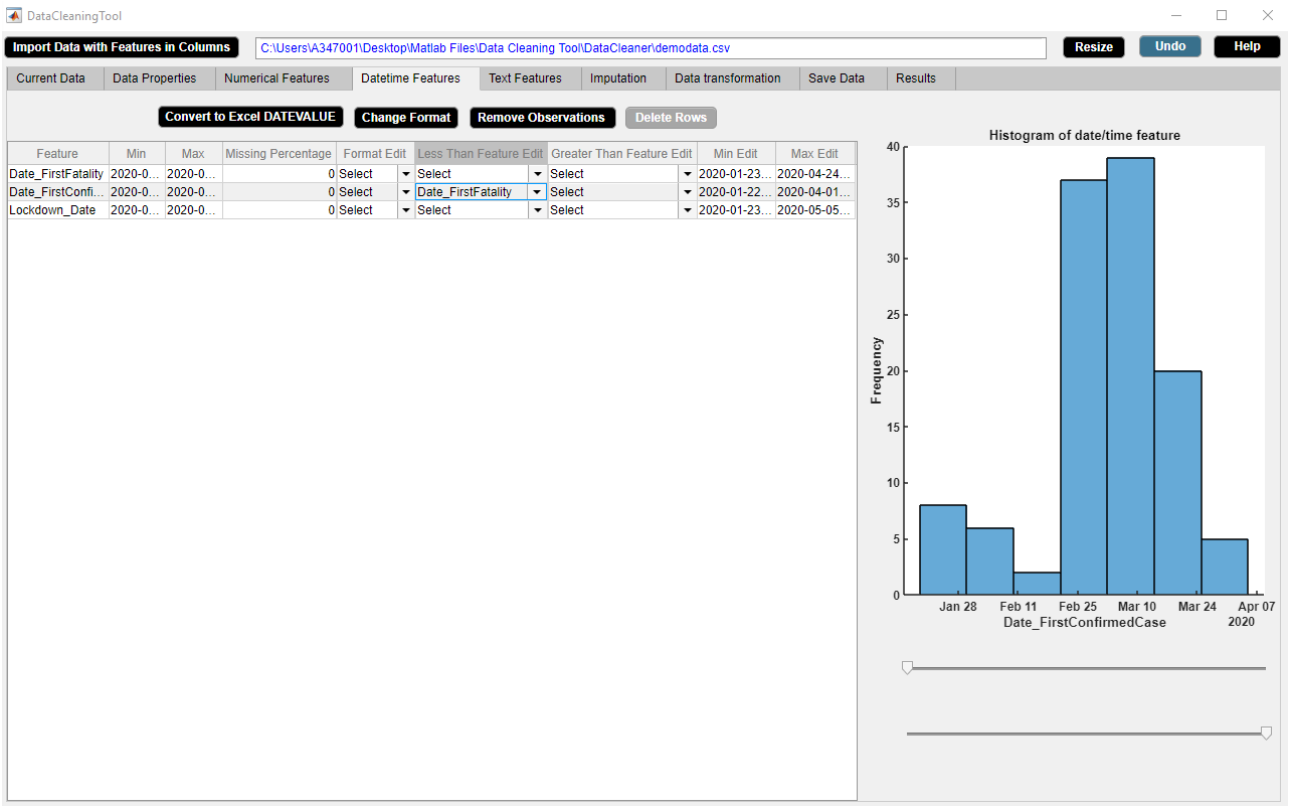


Figure 4.66: Step 3. Delete Rows Button

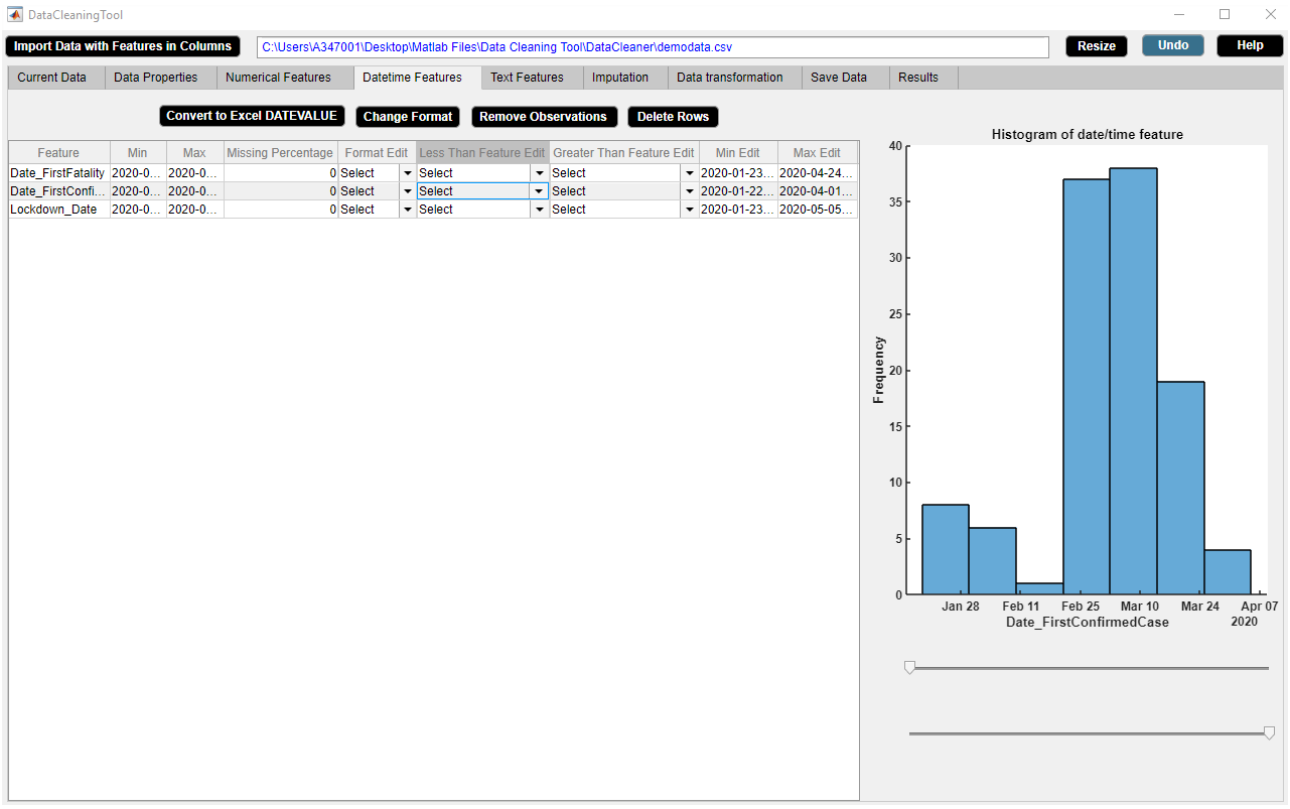


Figure 4.67: Step 4. Delete Rows Button

4.5 Text Features Widget

The Text Features widget displays statistical description of the text data. The Text Features widget is shown in figure 4.68. The properties of the Text Features widget are as follows.

- The widget shows the descriptive statistics of each text feature of the data such as categories and categories count of the feature.
- The widget also shows the missing observations percentage of each text feature.
- The statistical information of the text data in the widget gets updated after each activity.

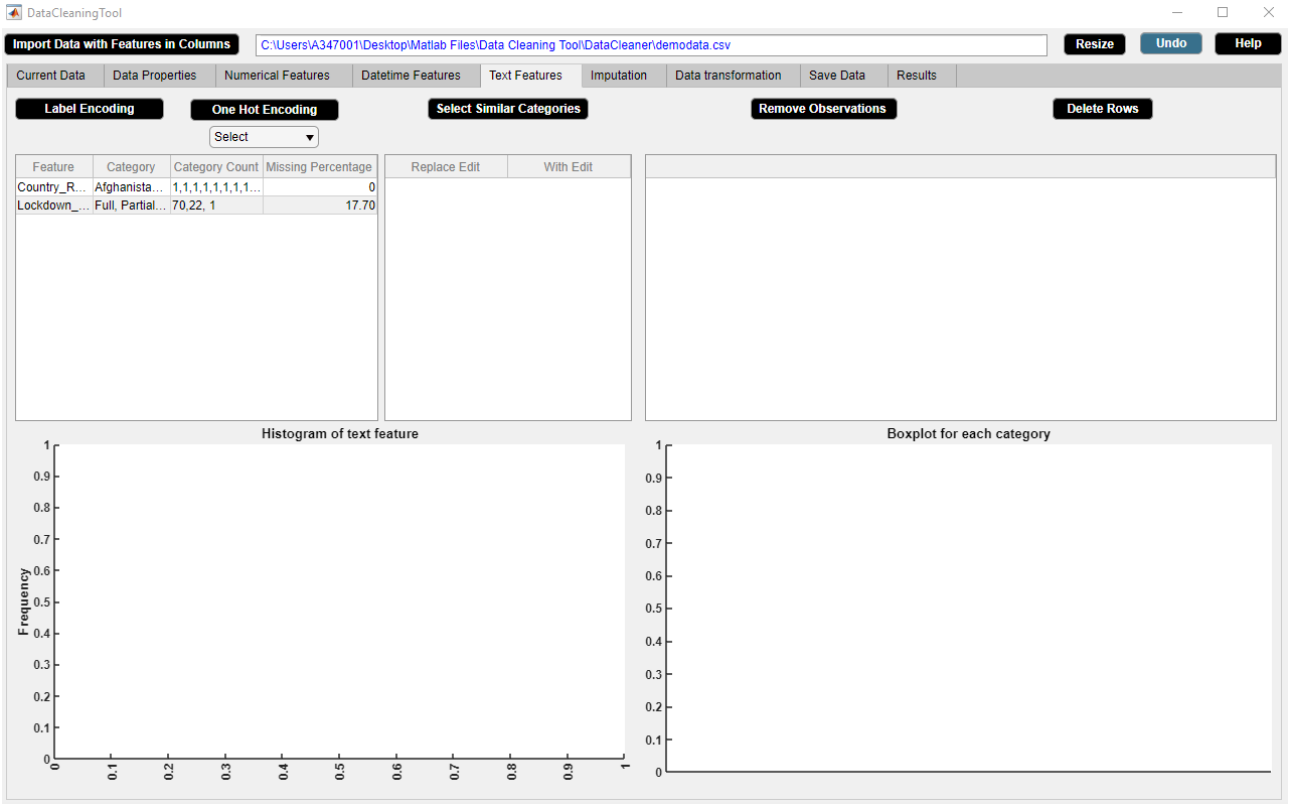


Figure 4.68: Text Features Widget.

4.5.1 Select Similar Categories Button

Replaces categories with similar ones.

Example

Step 1: Select a text feature from feature column of the text features descriptive statistics table.

Step 2: Select similar category from **With Edit** dropdown menu.

Step 3: Click **Select Similar Categories** button.

Step 4: **Select Similar Categories** button in use turns grey in color.

Step 5: **Select Similar Categories** button returns back to its original color once it completes its task.

We use **Select Similar Categories** button to refer 'Total' as 'Full' in the example data. Figures 4.69-4.73 illustrate how to use **Select Similar Categories** button.

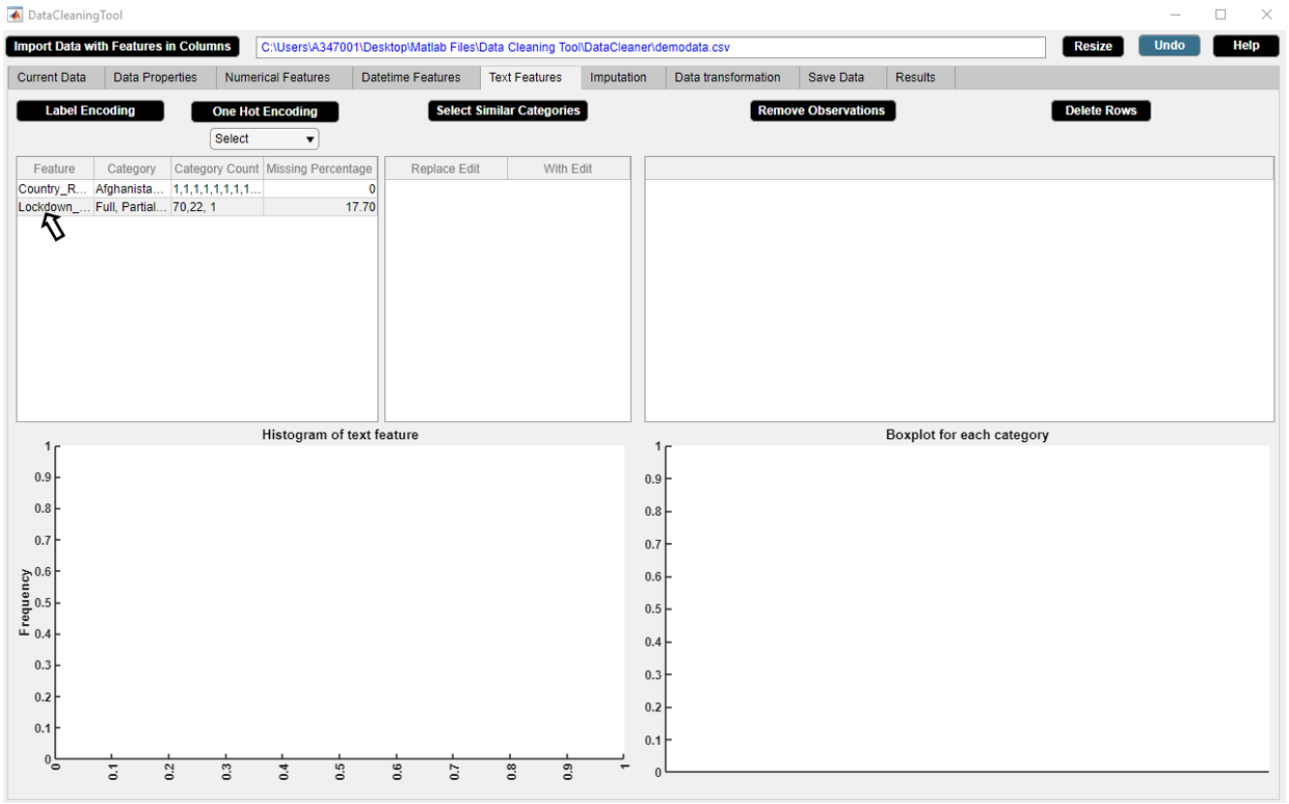


Figure 4.69: Step 1. Select Similar Categories Button

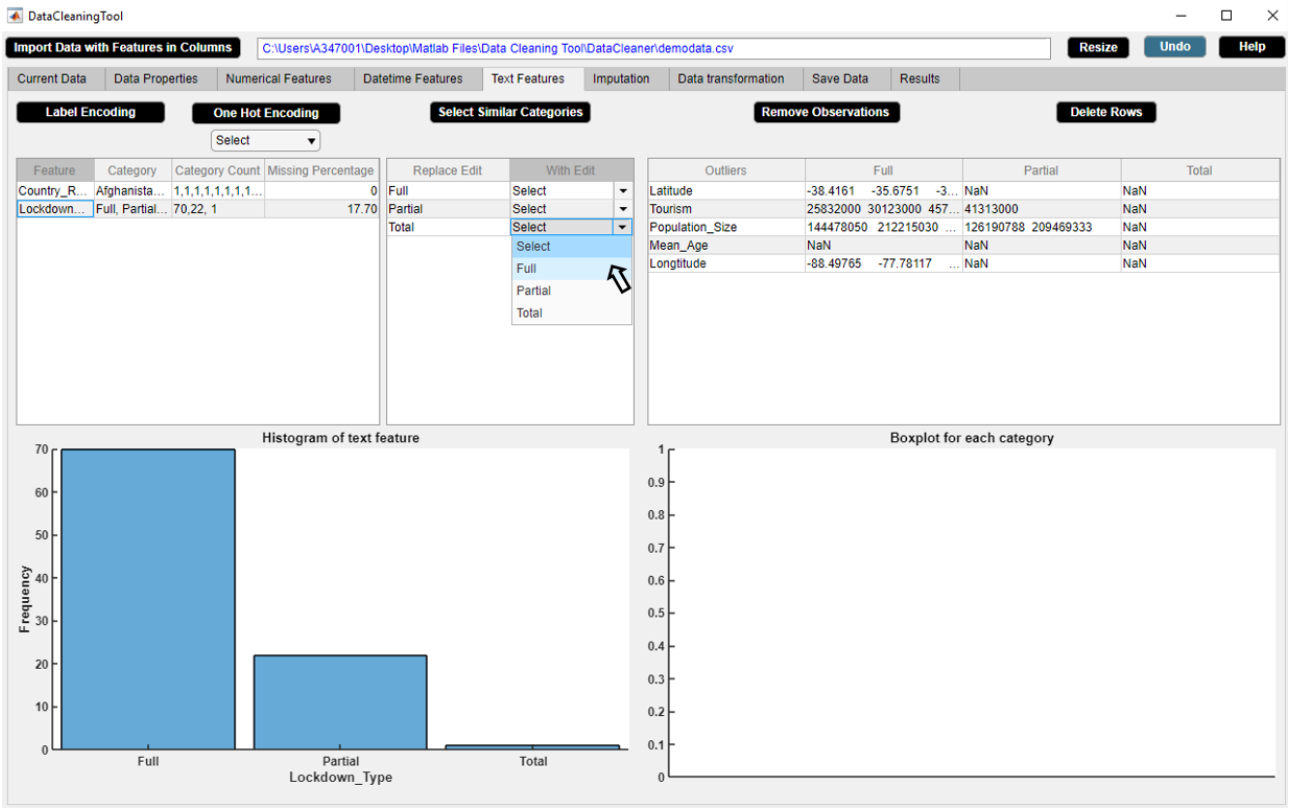


Figure 4.70: Step 2. Select Similar Categories Button

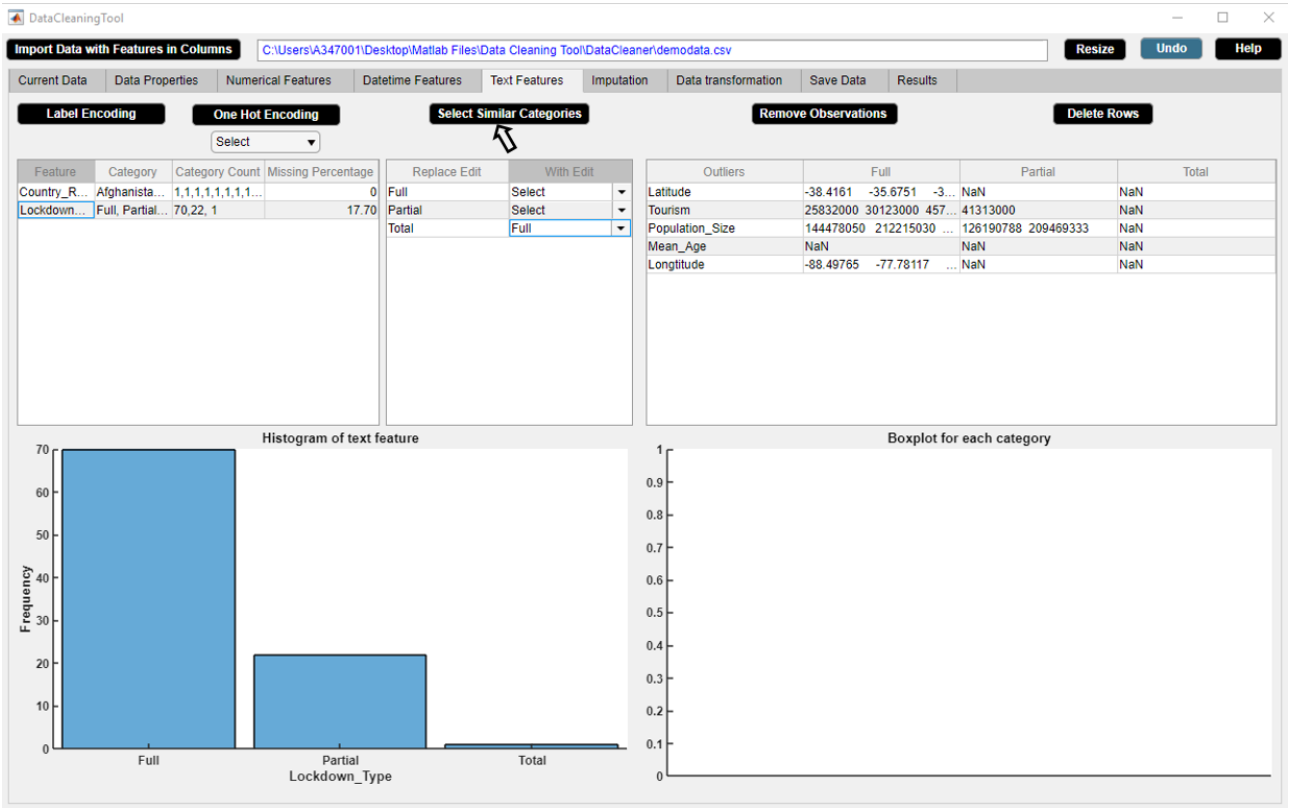


Figure 4.71: Step 3. Select Similar Categories Button

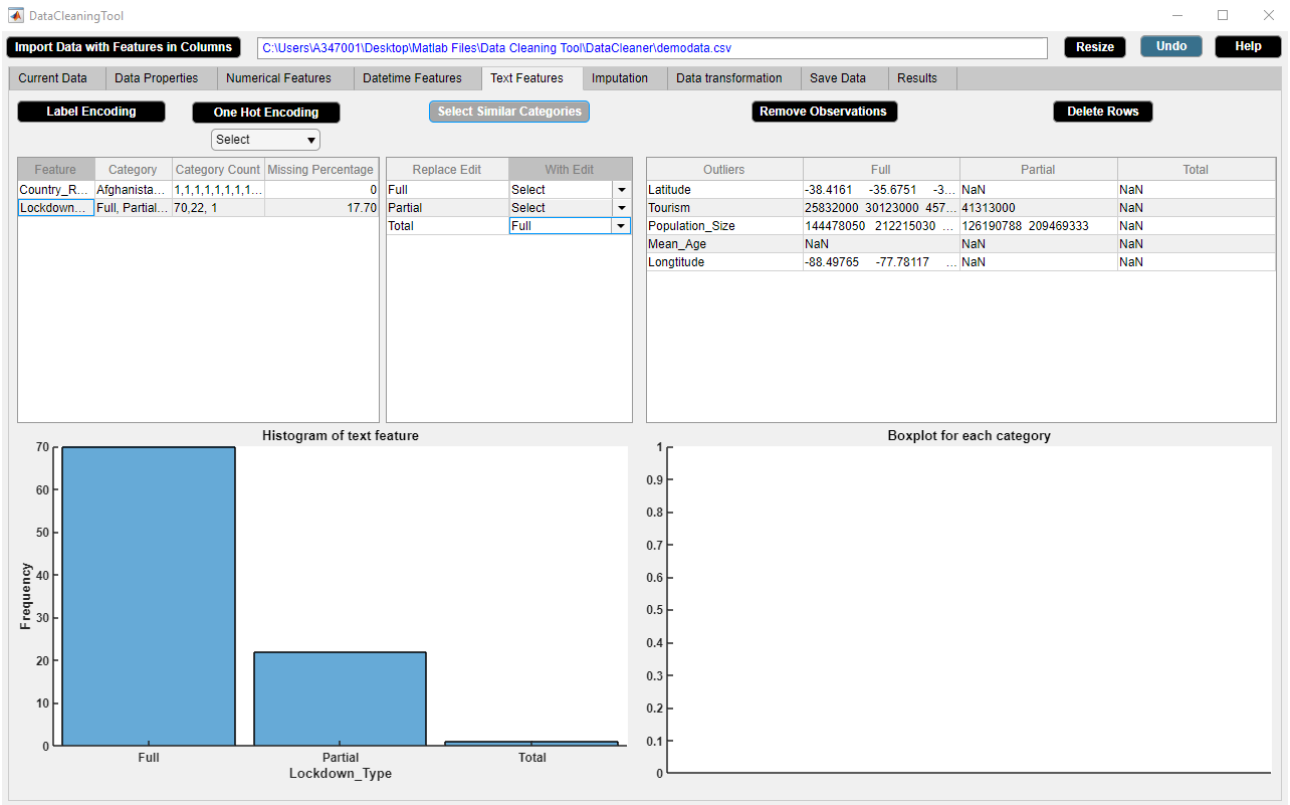


Figure 4.72: Step 4. Select Similar Categories Button

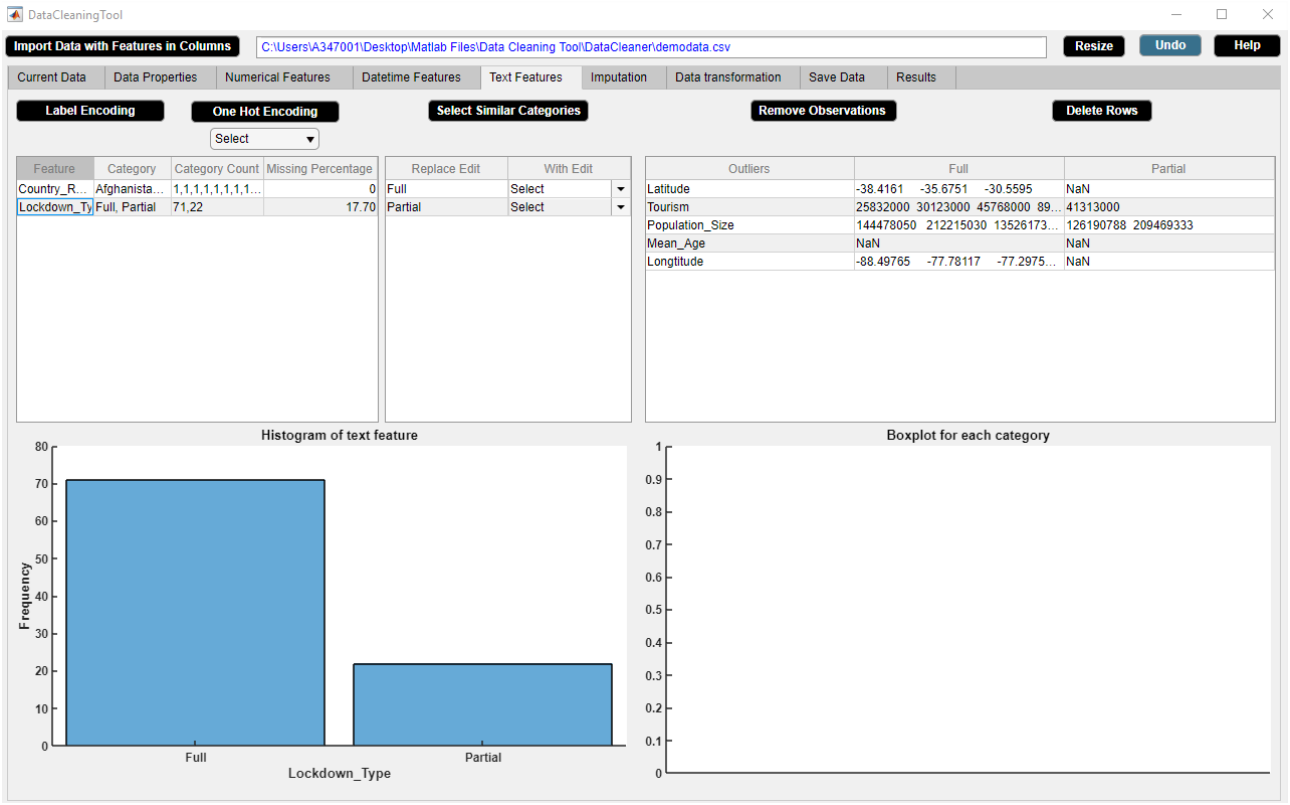


Figure 4.73: Step 5. Select Similar Categories Button

4.5.2 Text Feature Cell Selection Button

Displays histogram of a text feature.

Application

- Outlier visualization technique.

Example

Step 1: Select a text feature from **Feature** column of the text features descriptive statistics table.

Step 2: A histogram of the selected text feature appears in the lower left side of the **Text Features** widget. Select a numerical feature from **Outliers** column of the right hand side table.

Step 3: A box plot of the selected numerical feature versus the selected text feature appears in the lower right side of the **Text Features** widget.

We use **Text Feature Cell Selection** button to visualize the histogram of 'Lockdown_Type' feature and the box plot of 'Mean_Age' versus 'Lockdown_Type'. It can be seen from the histogram of 'Lockdown_Type' that there are more countries with 'Full' lockdown rather than with 'Partial' lockdown. It can be seen from the box plot of 'Mean_Age' versus 'Lockdown_Type' that 'Mean_Age' of the population is larger for the countries with 'Full' lockdown rather than for the countries with 'Partial' lockdown. Figures 4.74-4.76 illustrate how to use **Text Feature Cell Selection** button.

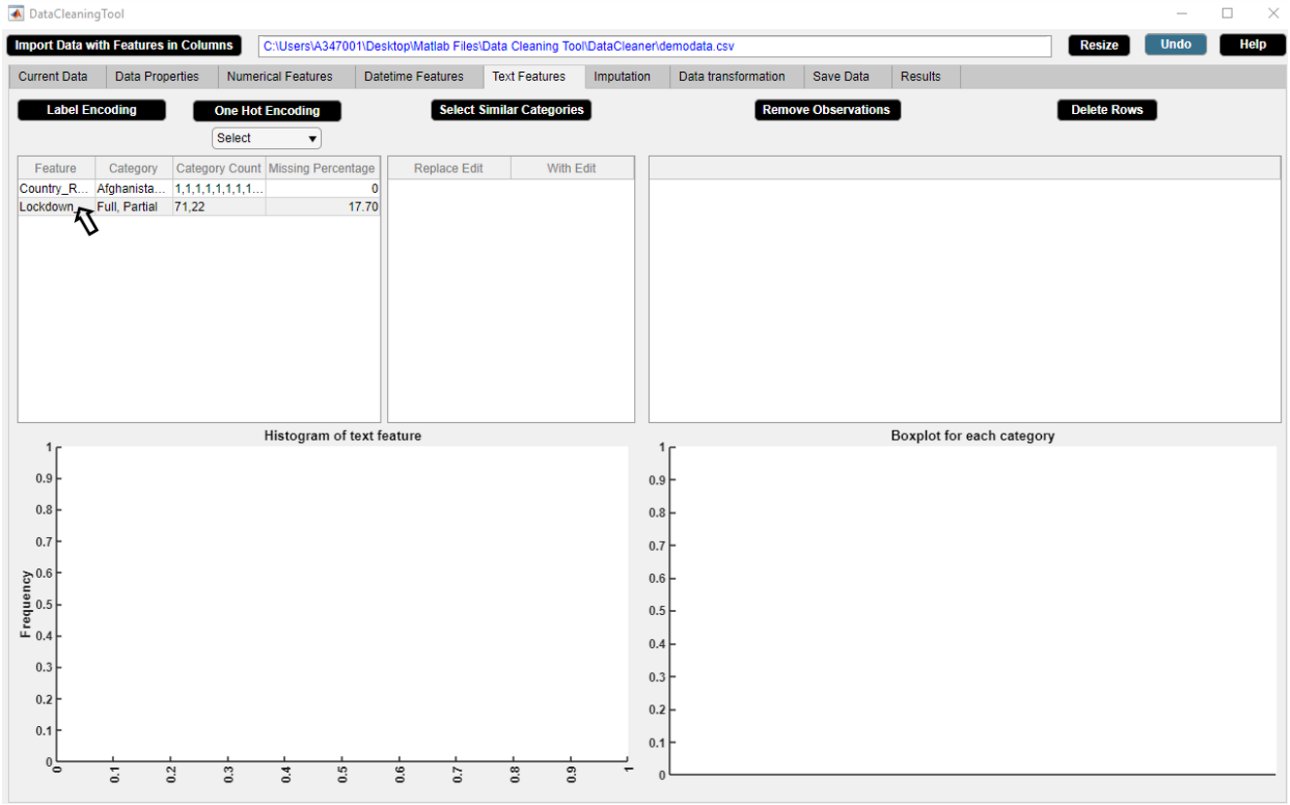


Figure 4.74: Step 1. Text Feature Cell Selection Button

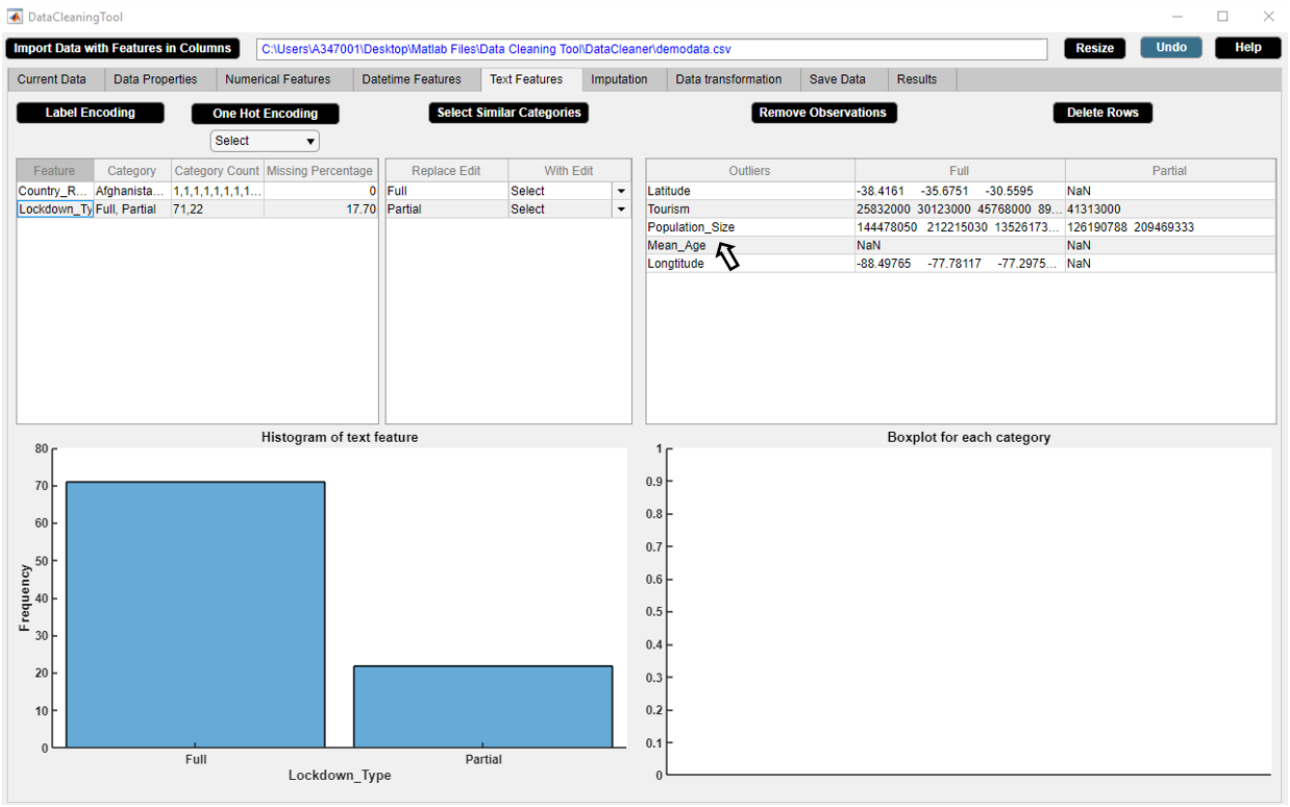


Figure 4.75: Step 2. Text Feature Cell Selection Button

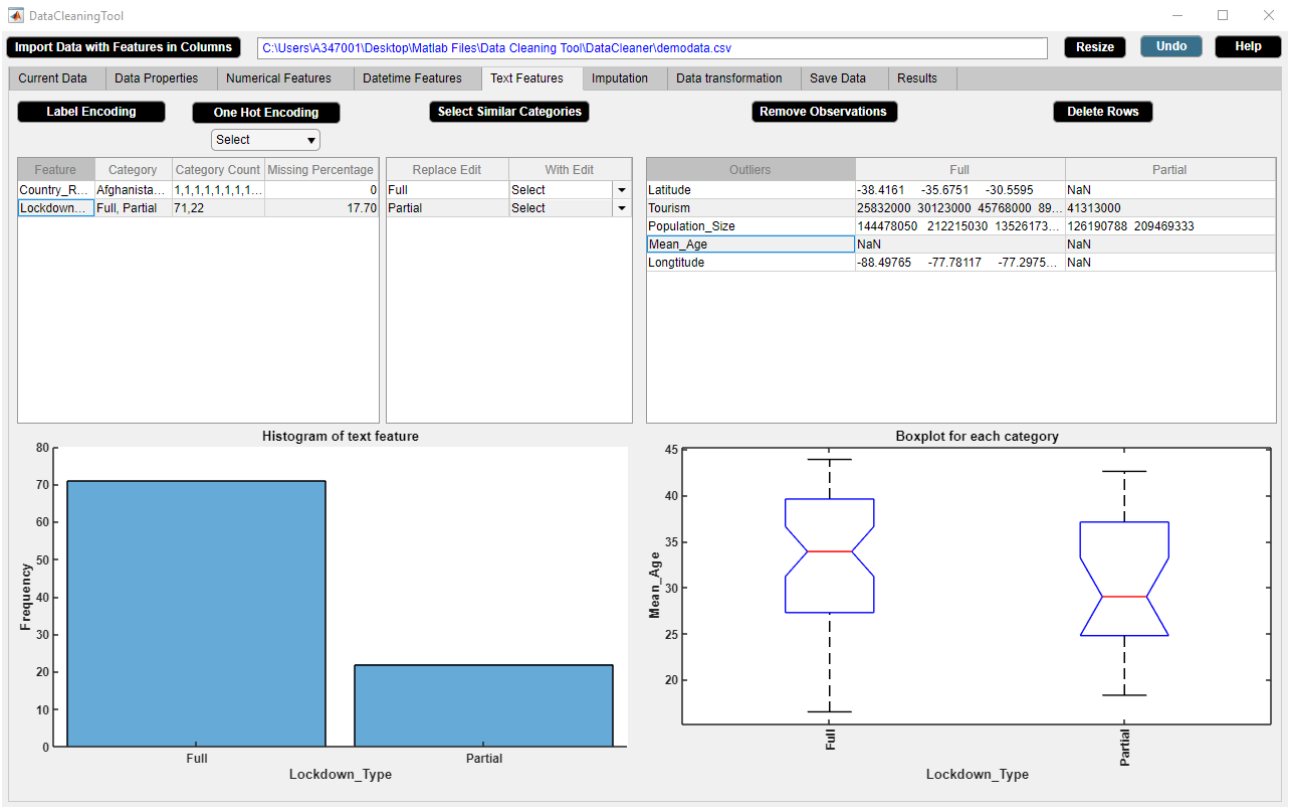


Figure 4.76: Step 3. Text Feature Cell Selection Button

4.5.3 Label Encoding Button

Assigns each category of a categorical feature a value from 0 to n-1 where n is the number of categories. Note that label encoding is an encoding approach usually for handling ordinal categorical features.

Example

Step 1: Select a categorical feature from **Feature** column of the text features descriptive statistics table.

Step 2: Click **Label Encoding** button.

Step 3: **Label Encoding** button in use turns grey in color.

Step 4: **Label Encoding** button returns back to its original color once it completes its task.

Step 5: Check the change in **Current Data** widget.

We use **Label Encoding** button if we wish to label encode the categorical feature 'Lockdown_Type'. Figures 4.77-4.81 illustrate how to use **Label Encoding** button.

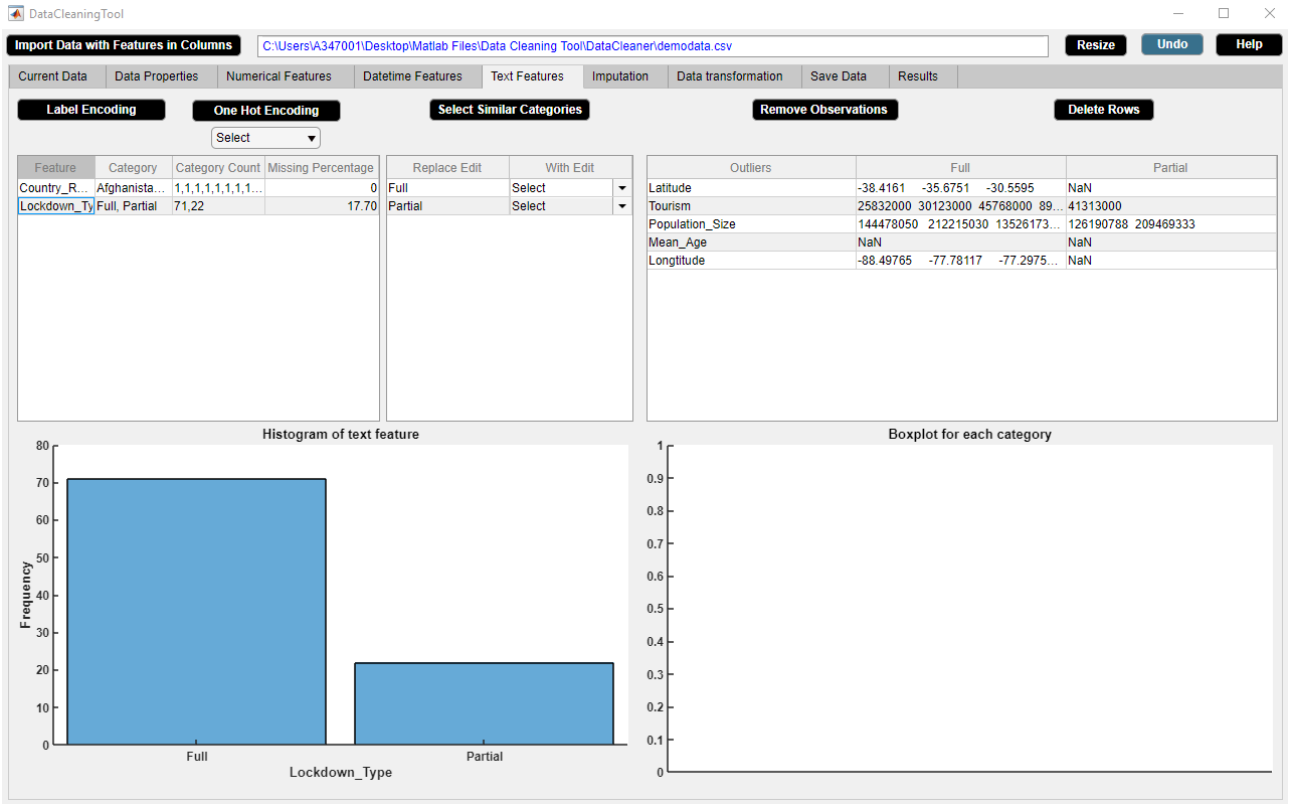


Figure 4.77: Step 1. Label Encoding Button

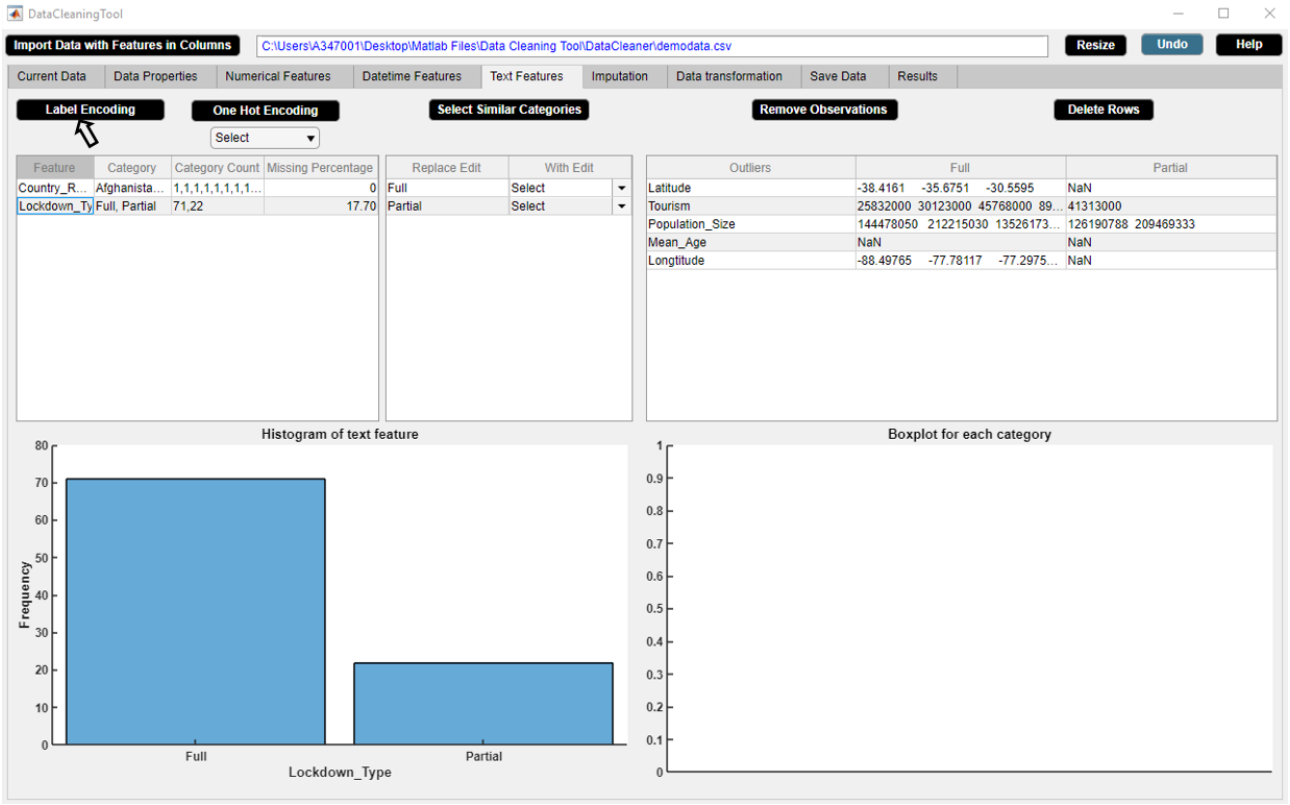


Figure 4.78: Step 2. Label Encoding Button

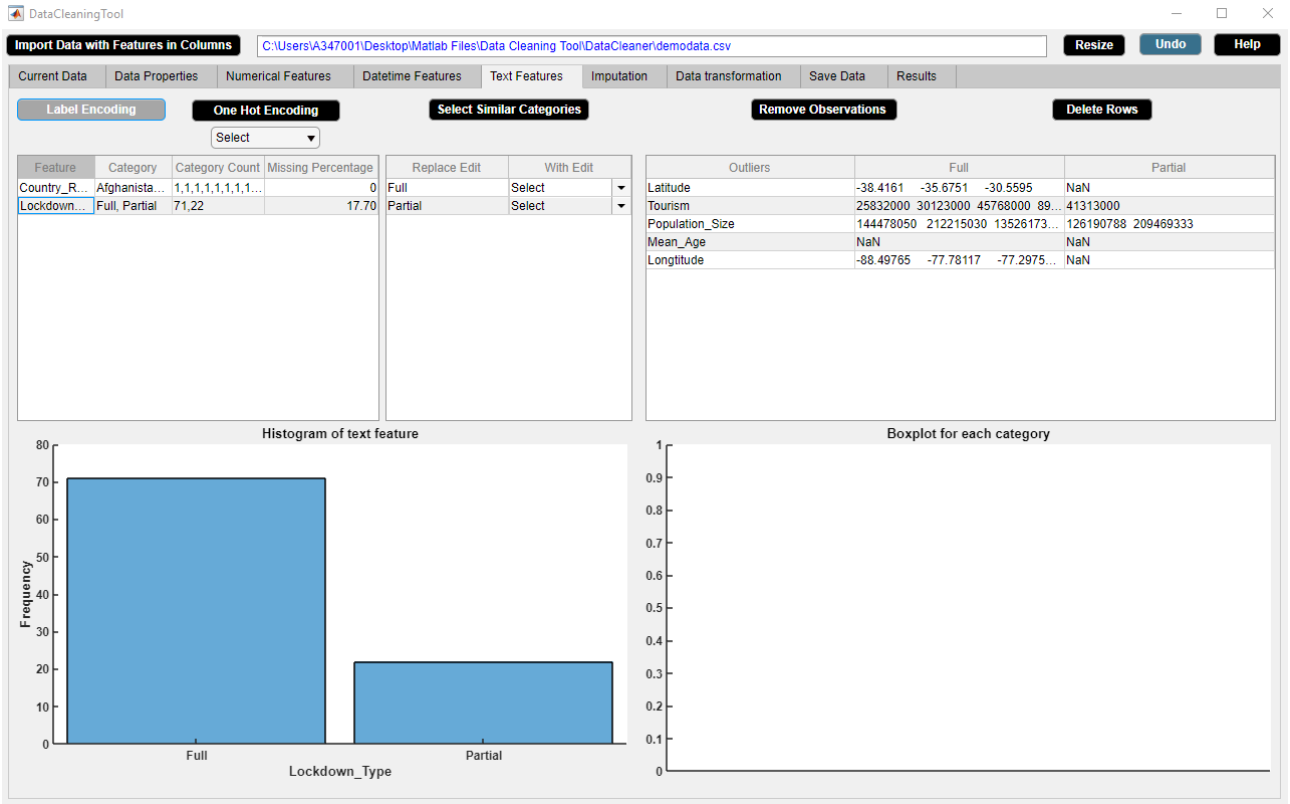


Figure 4.79: Step 3. Label Encoding Button

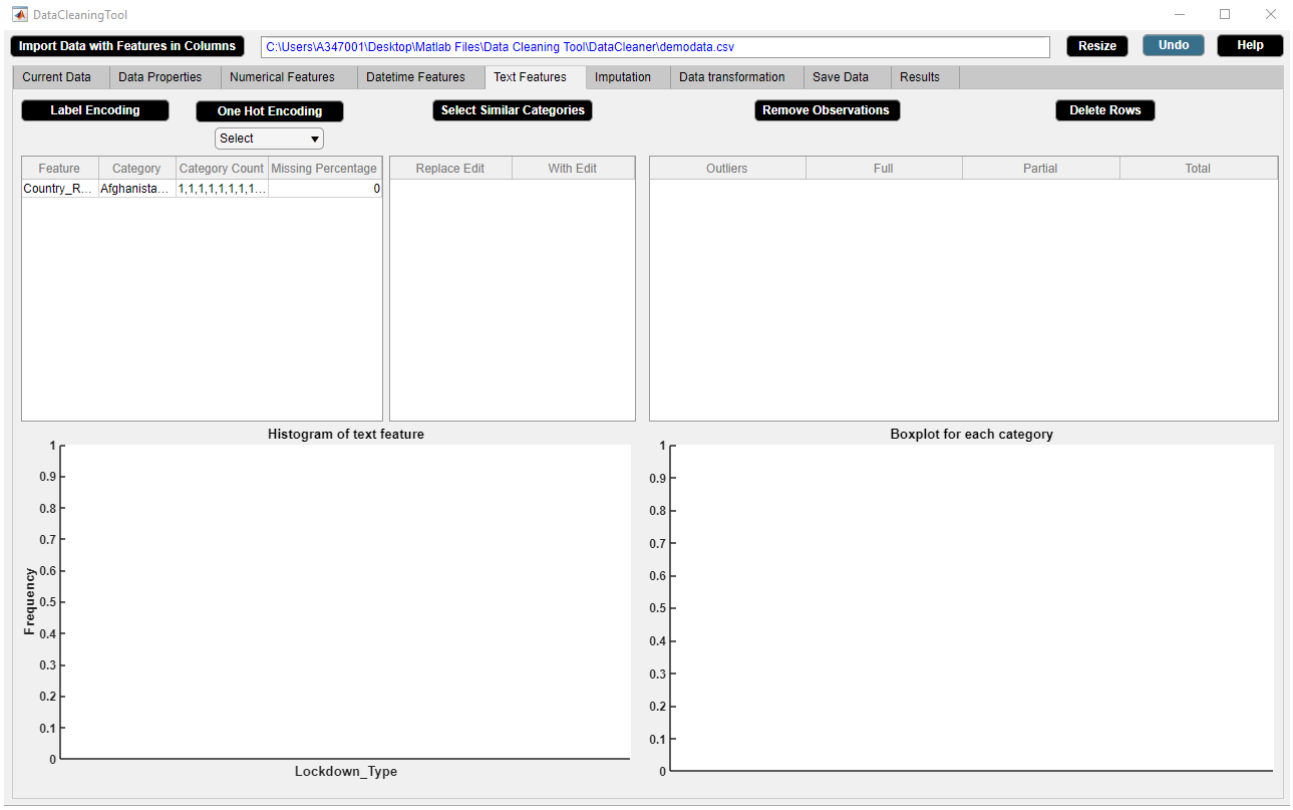


Figure 4.80: Step 4. Label Encoding Button

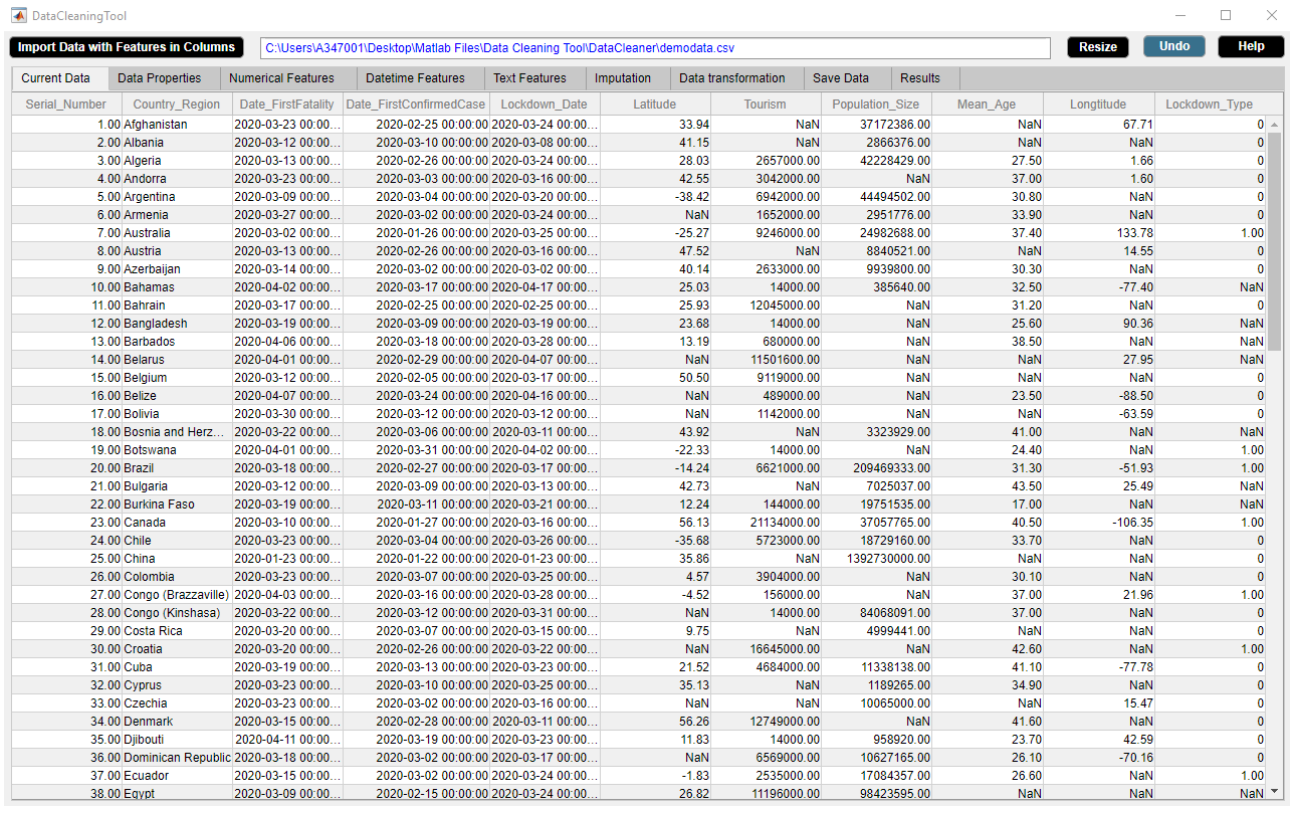


Figure 4.81: Step 5. Label Encoding Button

4.5.4 One Hot Encoding Button

Transforms n categories to either n or n-1 dummy variables for a categorical feature. Note that one hot encoding is an encoding approach usually for handling nominal categorical features.

Example

Step 1: Select a categorical feature from **Feature** column of the text features descriptive statistics table.

Step 2: Select any one option from **One Hot Encoding** dropdown menu. We transform n categories of a categorical feature to n dummy variables for methods such as singular value decomposition whereas n-1 dummy variables for methods such as regression.

Step 3: Click **One Hot Encoding** button.

Step 4: **One Hot Encoding** button in use turns grey in color.

Step 5: **One Hot Encoding** button returns back to its original color once it completes its task.

Step 6: Check the change in **Current Data** widget.

We use **One Hot Encoding** button if we wish to one hot encode the categorical feature 'Country_Region'. Figures 4.82-4.87 illustrate how to use **One Hot Encoding** button.

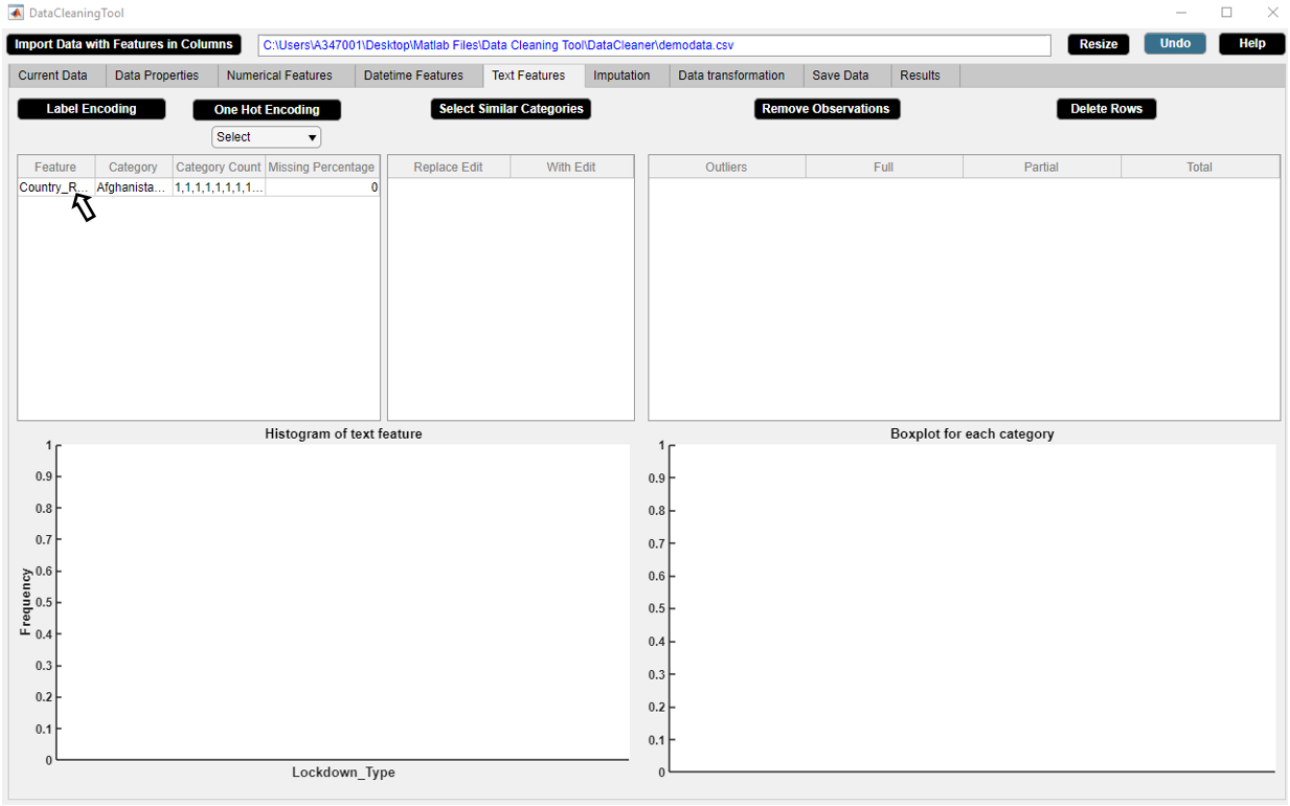


Figure 4.82: Step 1. One Hot Encoding Button

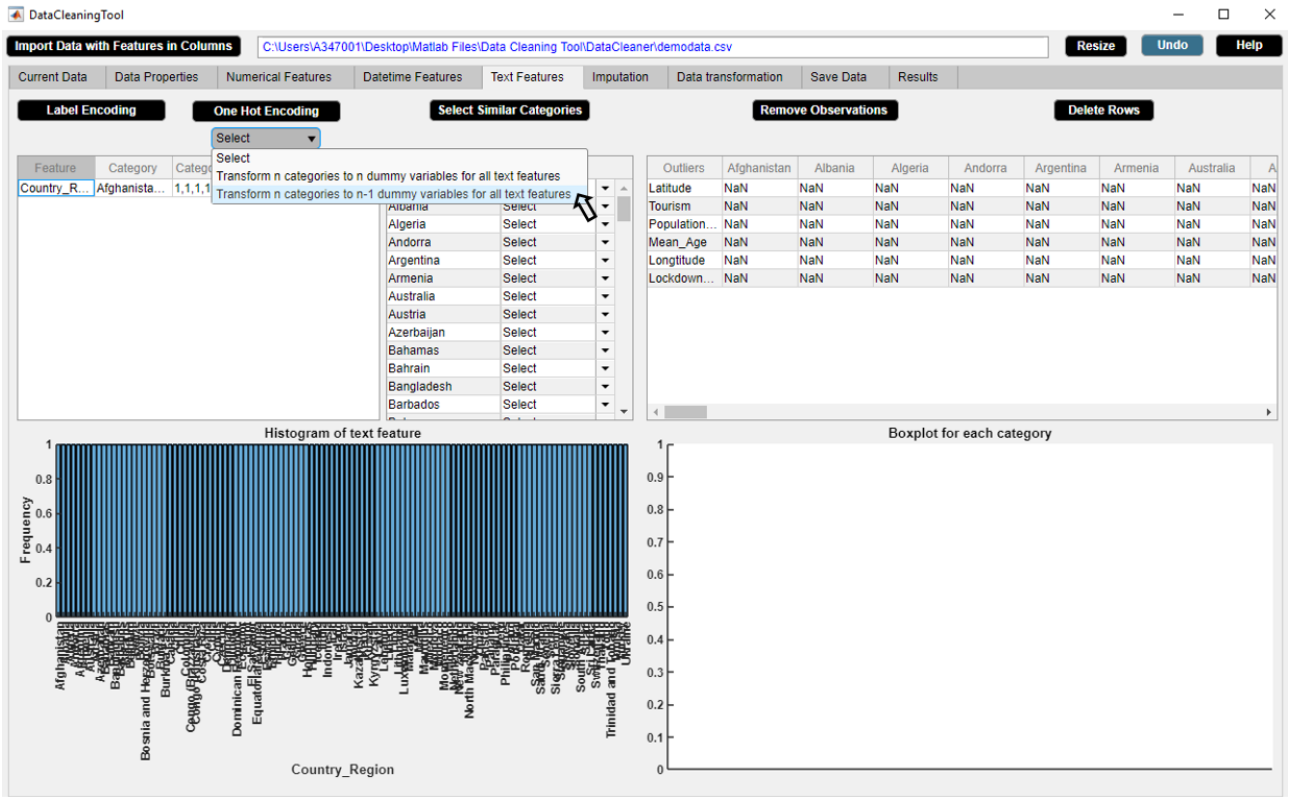


Figure 4.83: Step 2. One Hot Encoding Button

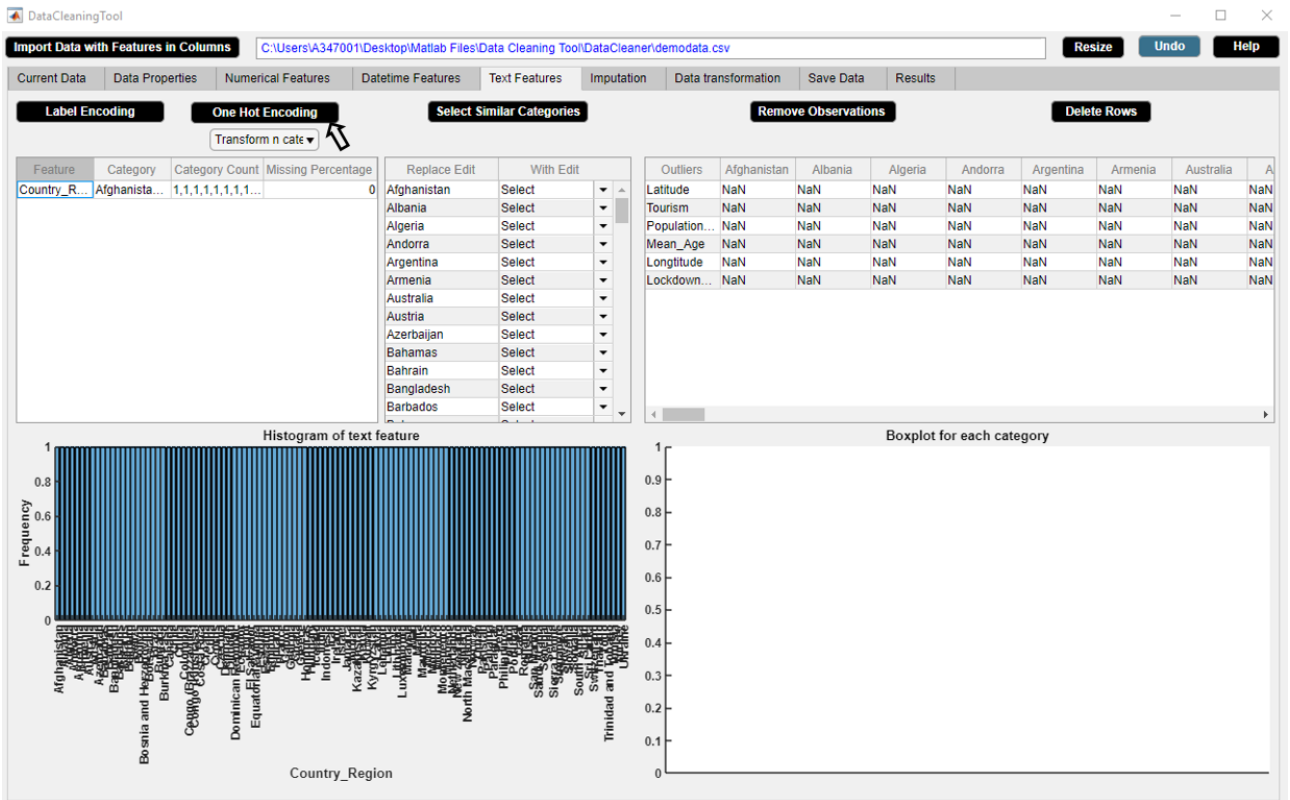


Figure 4.84: Step 3. One Hot Encoding Button

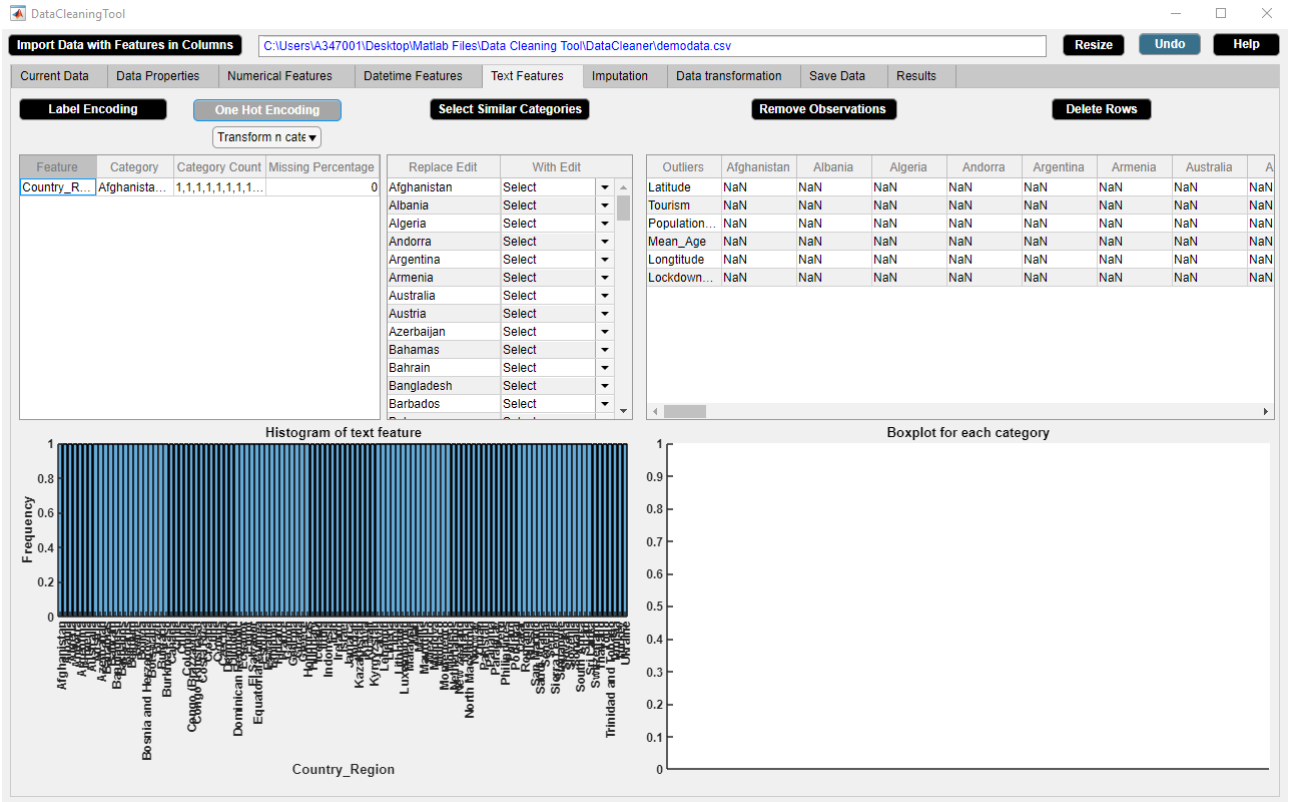


Figure 4.85: Step 4. One Hot Encoding Button

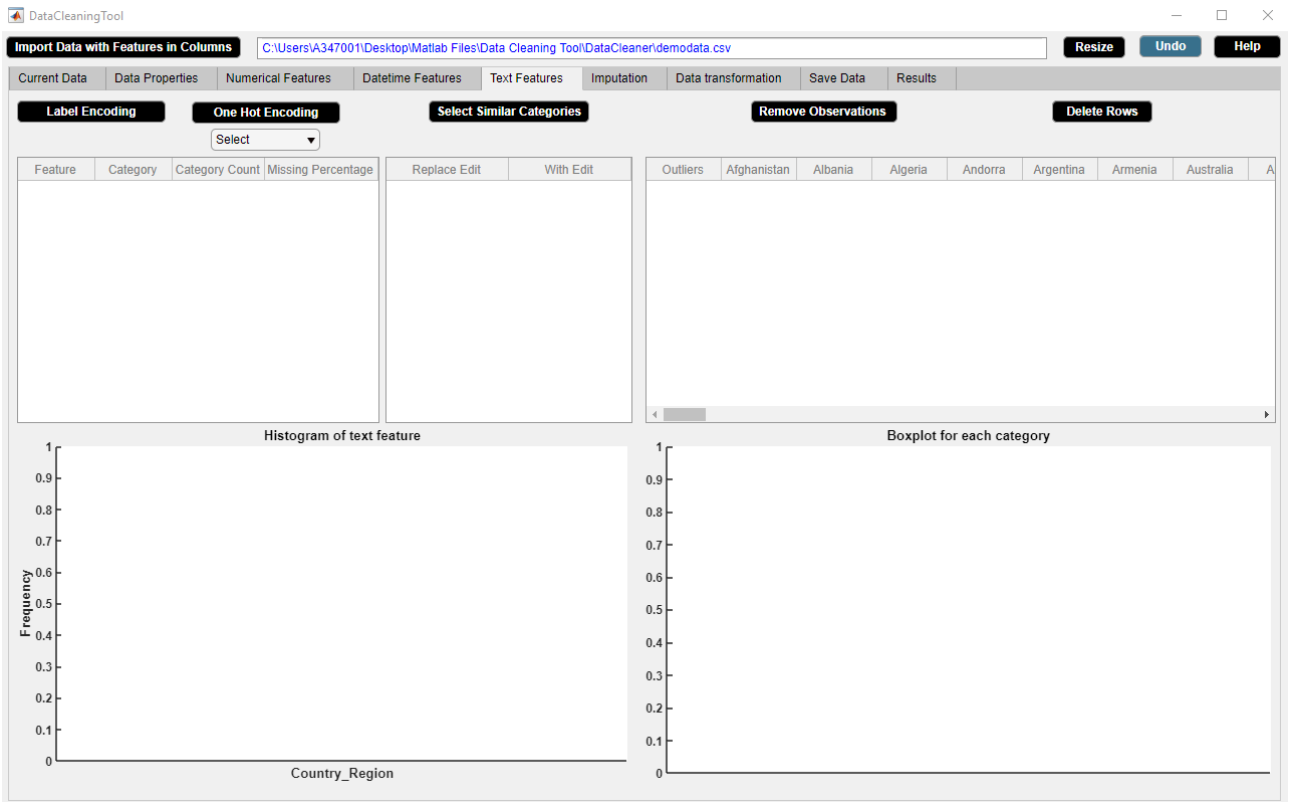


Figure 4.86: Step 5. One Hot Encoding Button

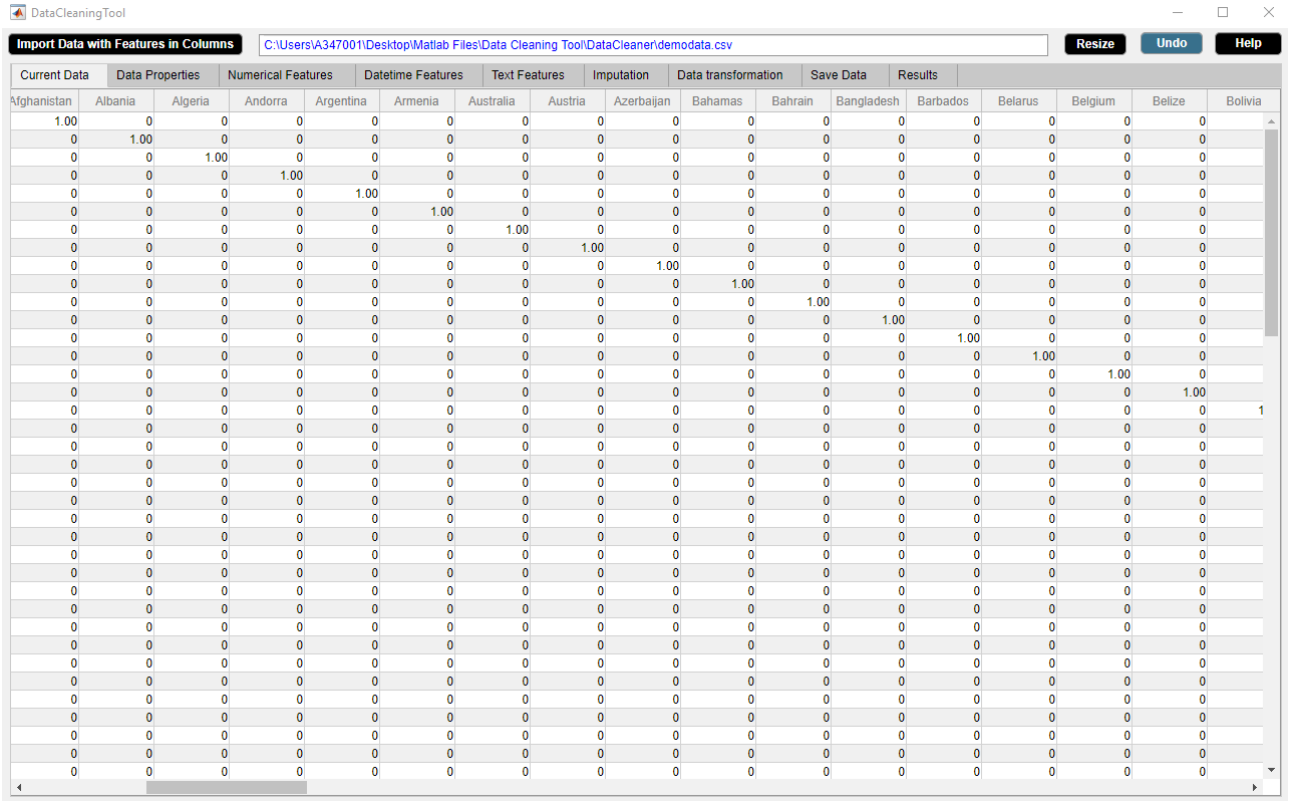


Figure 4.87: Step 6. One Hot Encoding Button

4.5.5 Remove Observations Button

Replaces outliers by missing values.

Application

- Removes outliers.

4.5.6 Delete Rows Button

Deletes rows with outliers.

Application

- Deletes rows containing outliers.

4.6 Imputation Widget

The Imputation widget displays information about the missing data and the expected error of imputation for numerical and categorical features. The Imputation widget is shown in figure 4.88. The properties of the Imputation widget are as follows.

- The widgets shows information about missing data such as percentage of missing data, expected error of imputation for numerical and categorical features. The performance analysis results of the missForest method discussed in chapter 4 is used to predict the expected error of imputation for numerical and categorical features for the specific ratio of data and percentage of missing data.
- The widget also presents the missing observations percentage table and the missingness plot.
- If datetime observations are missing, a message stating that datetime imputation is possible appears in red color in the lower side of the Imputation widget.
- The information of the missing data in the widget gets updated after each activity.

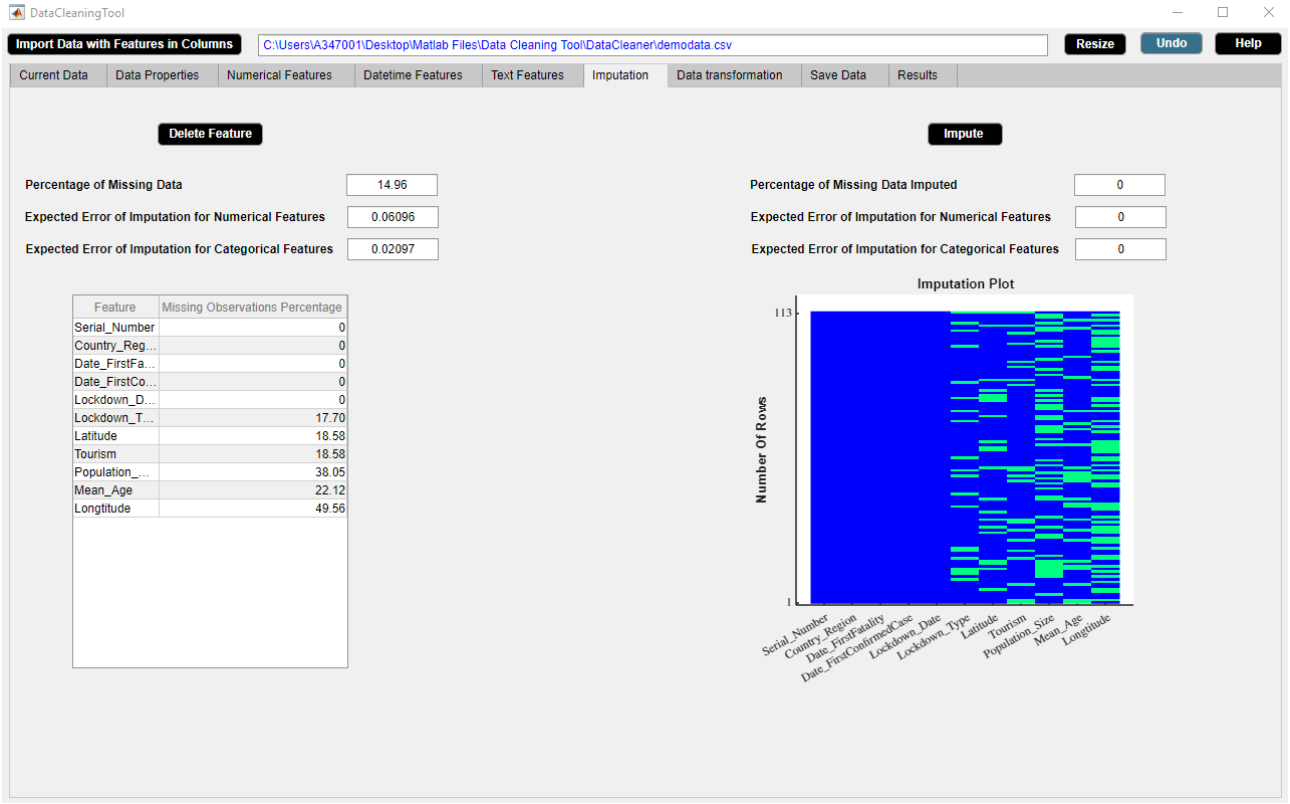


Figure 4.88: Imputation Widget.

4.6.1 Delete Feature Button

Delete a feature from data.

Application

- Delete an unwanted or irrelevant feature.
- Delete a feature containing a large number of missing observations.

Example

Step 1: Select a feature from **Feature** column of missing observations percentage table.

Step 2: Click **Delete Feature** button.

Step 3: **Delete Feature** button in use turns grey in color.

Step 4: **Delete Feature** button returns back to its original color once it completes its task.

In the example data, 'Longitude' has a large number of missing values. We use **Delete Feature** button to delete 'Longitude' feature. Figures 4.89-4.92 illustrate how to use **Delete Feature** button.

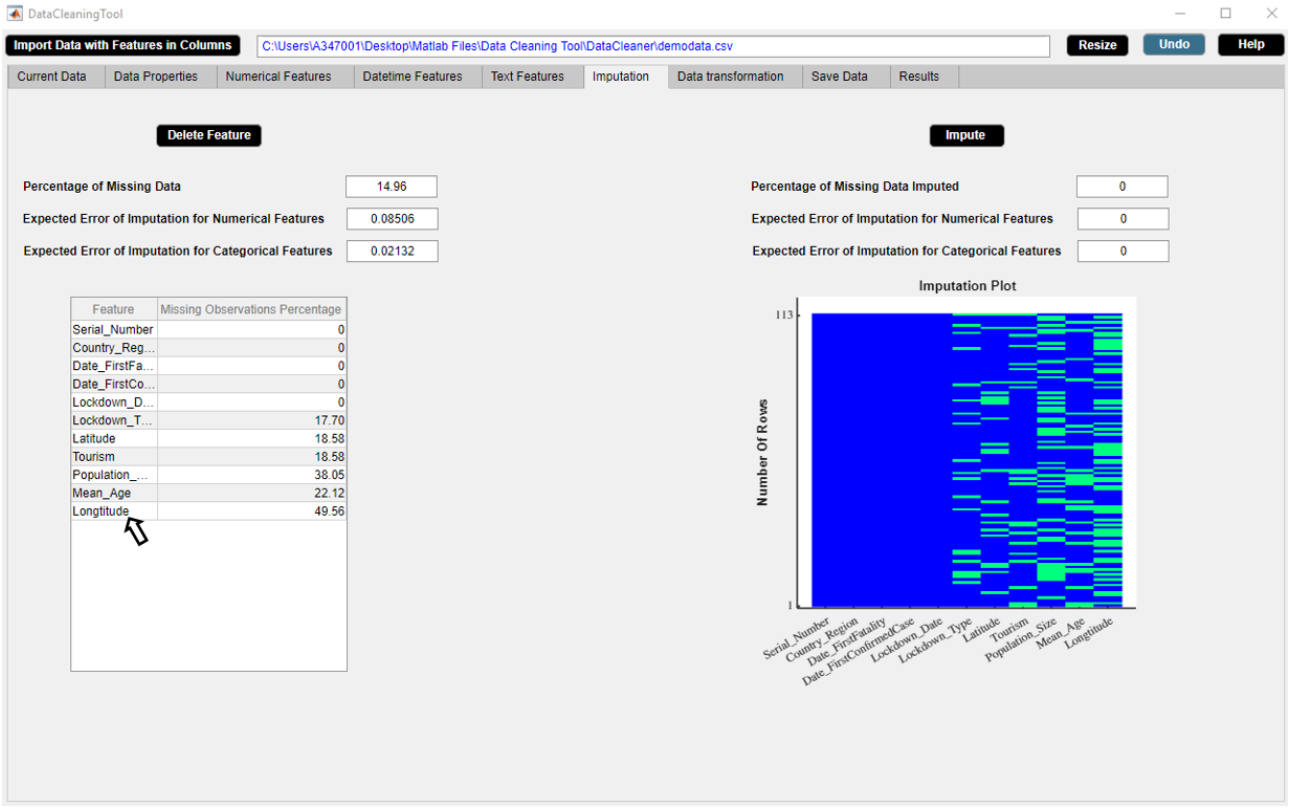


Figure 4.89: Step 1. Delete Feature Button

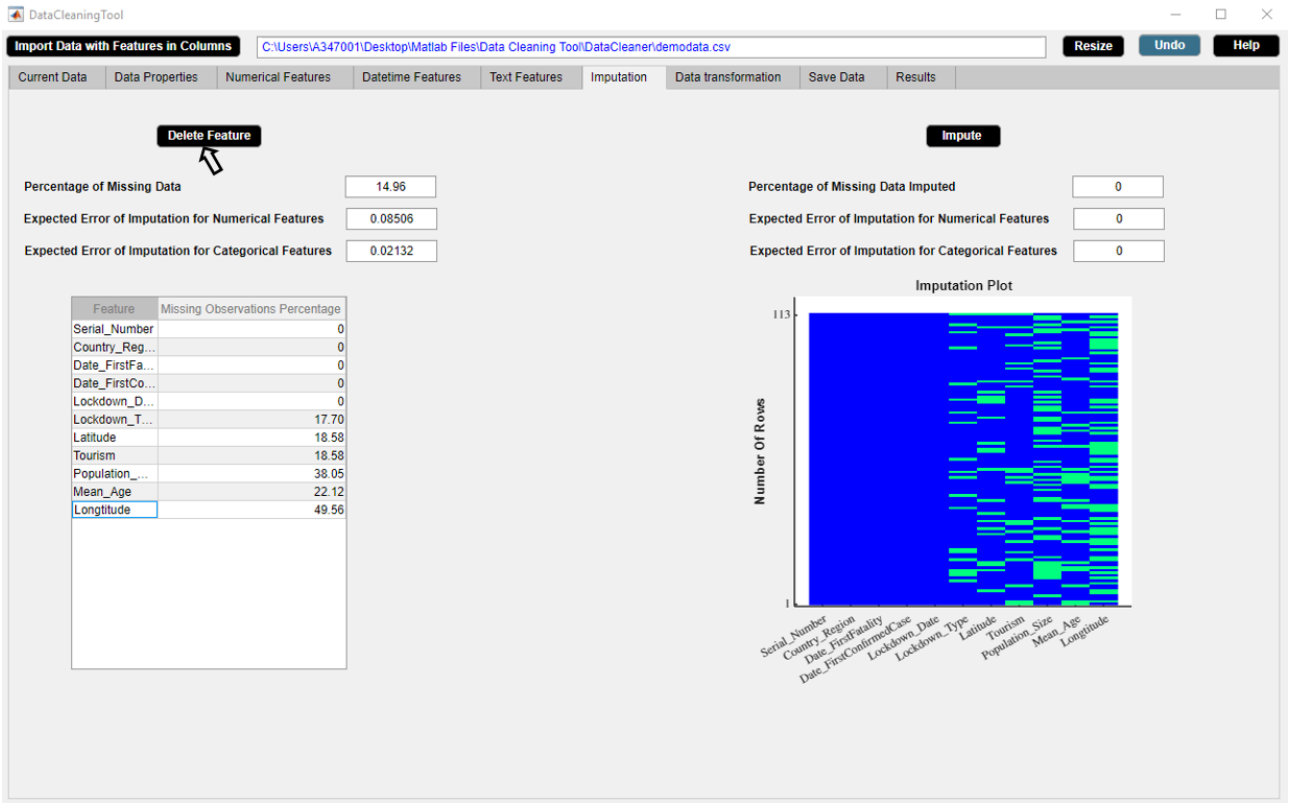


Figure 4.90: Step 2. Delete Feature Button

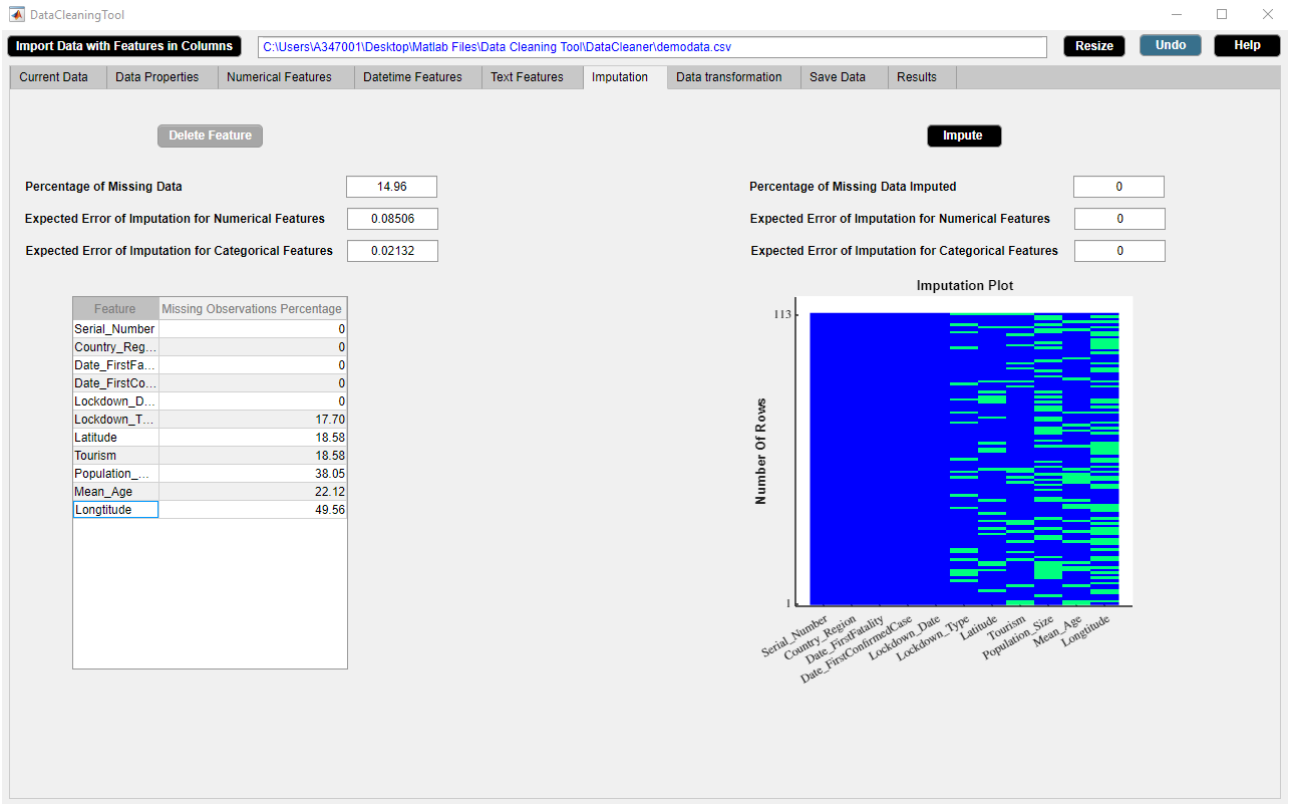


Figure 4.91: Step 3. Delete Feature Button

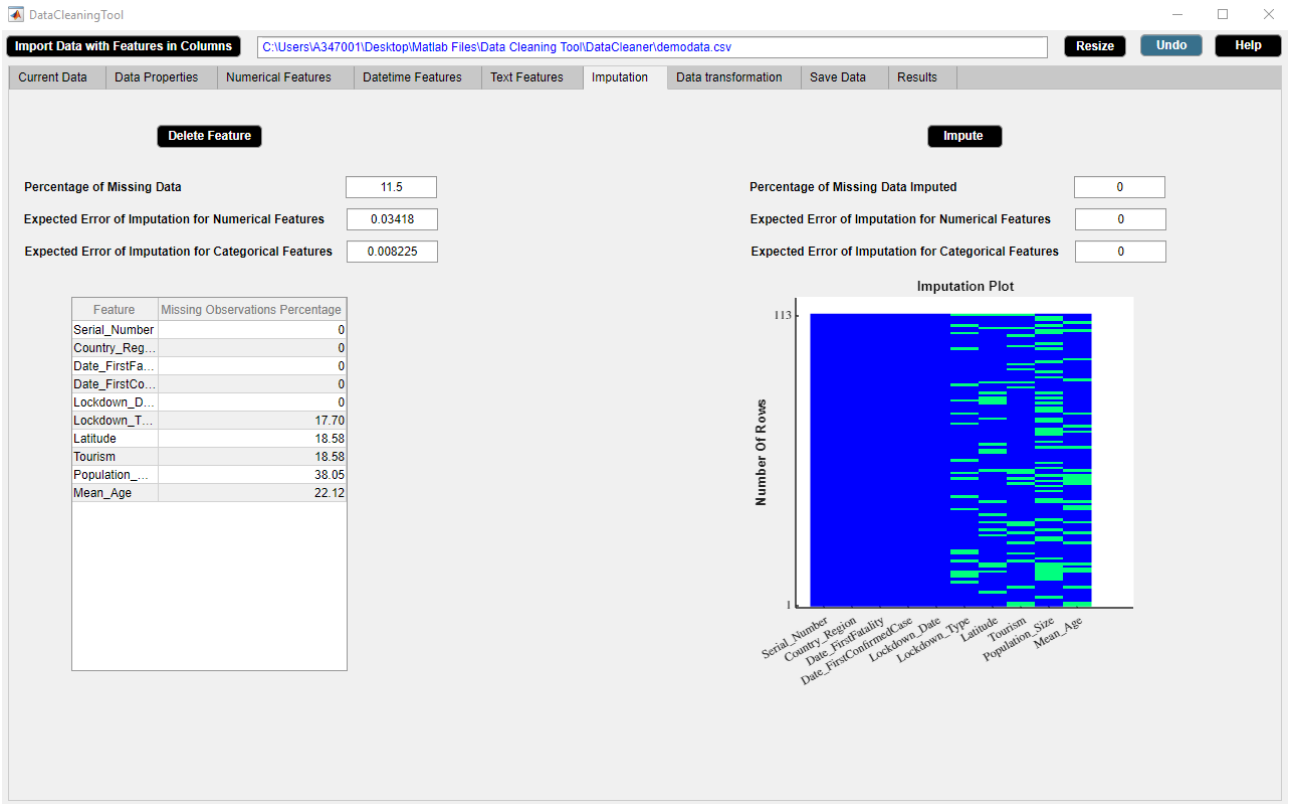


Figure 4.92: Step 4. Delete Feature Button

4.6.2 Impute Button

Replaces missing values by estimated ones using missForest algorithm.

Application

- Impute missing observations.

Example

Step 1: Click **Impute** button.

Step 2: **Impute** button in use turns grey in color. If datetime observations are missing, a message stating that datetime imputation is not possible appears in red color in the lower side of the **Imputation** widget.

Step 3: **Impute** button returns back to its original color once it completes its task.

We use **Impute** button to impute missing values in the example data. Figures 4.93-4.95 illustrate how to use **Impute** button.

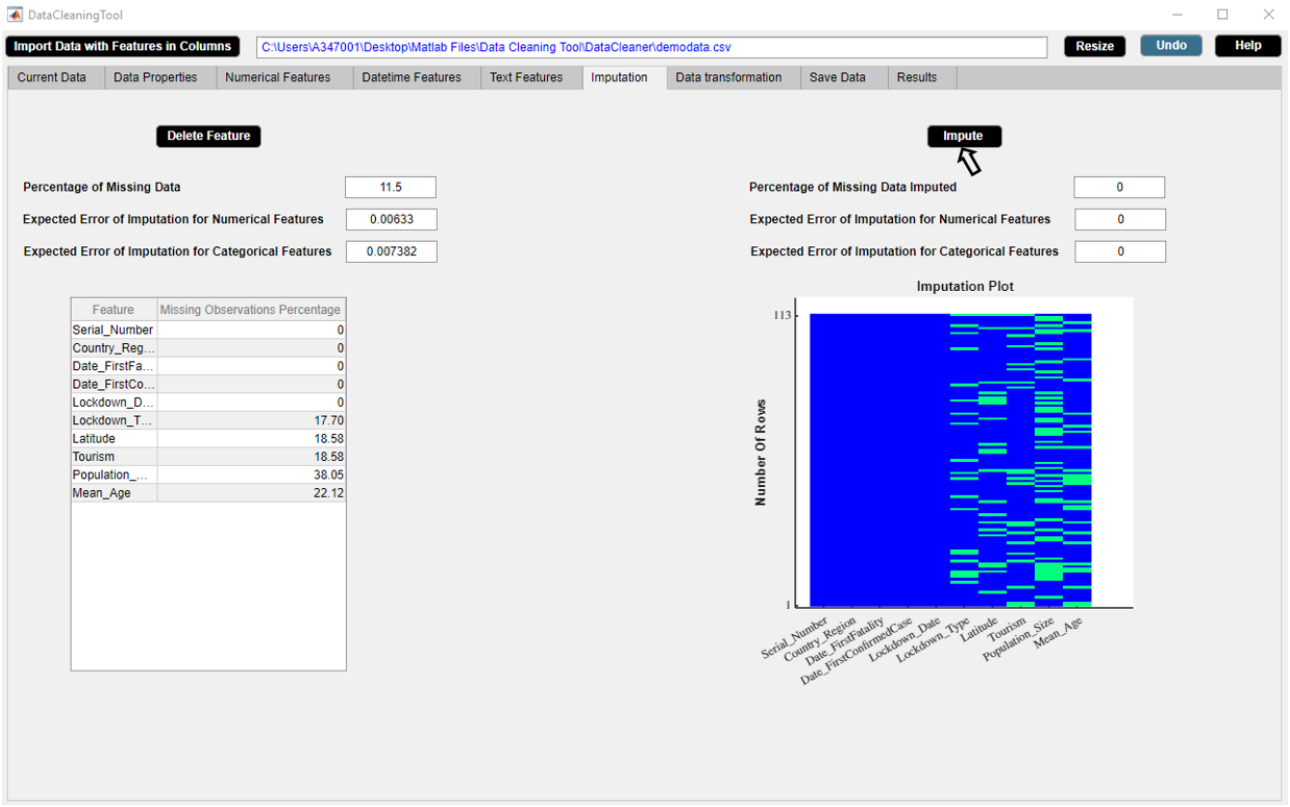


Figure 4.93: Step 1. Impute Button

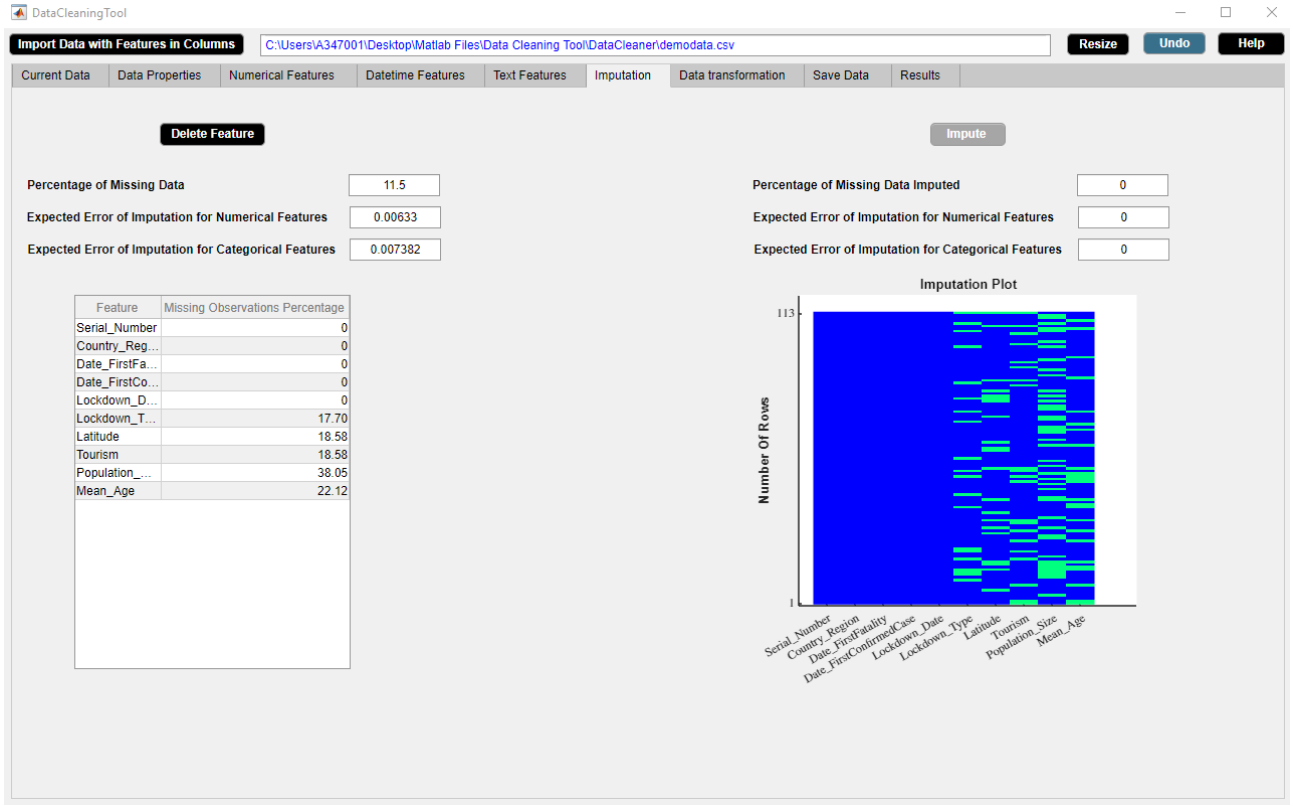


Figure 4.94: Step 2. Impute Button

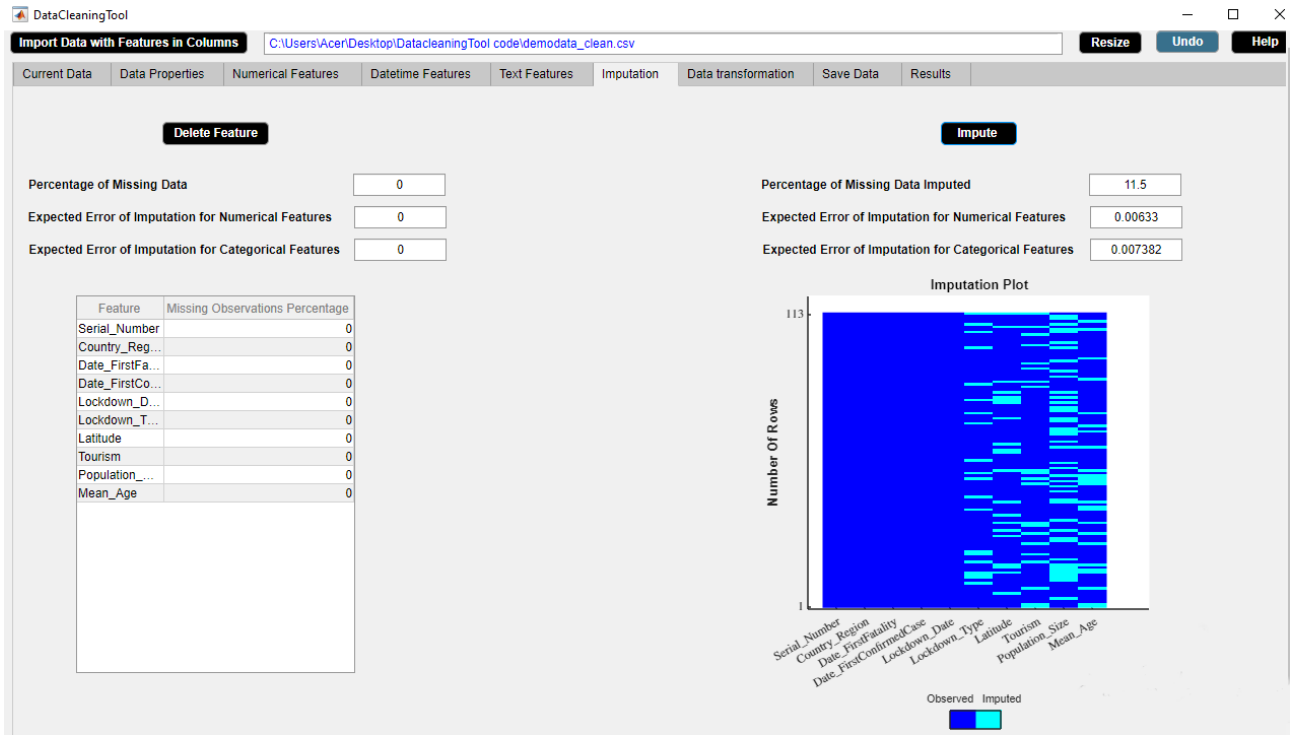


Figure 4.95: Step 3. Impute Button

4.7 Data Transformation Widget

The Data Transformation widget displays the numerical features of the data on which data transformation can only be applied. The Data Transformation widget is shown in figure 4.96. The properties of the Data Transformation widget are as follows.

- The widget presents the numerical features of the data.
- The numerical features of the data in the widget gets updated after each activity.

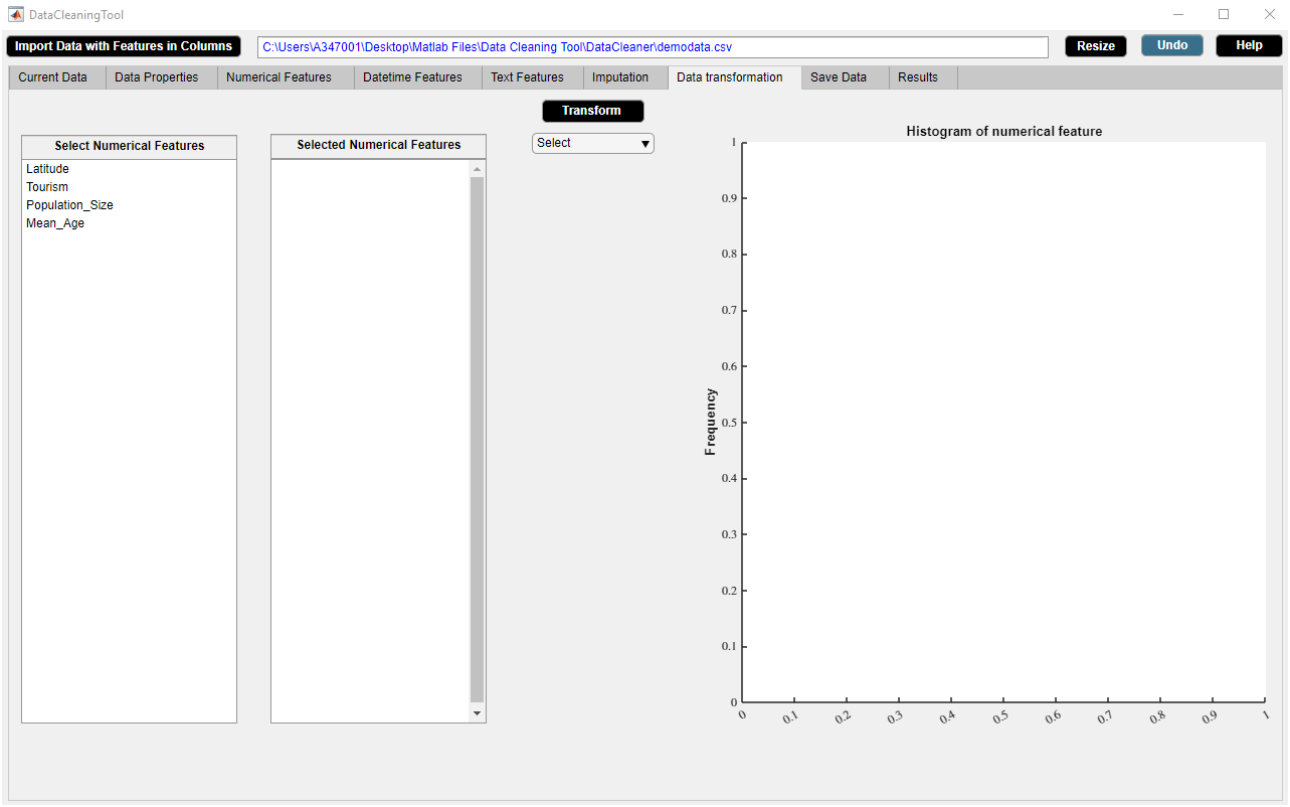


Figure 4.96: Data Transformation Widget.

4.7.1 Transform Button

Standardize or normalize or logarithm or exponential or square root or inverse transform selected numerical features.

Application

- Outliers.

Example

Step 1: Select numerical feature/features from **Select Numerical Features** list box. Select an option from **Transform** dropdown menu. Here 'mean 0 and standard deviation' represents standardize, 'between 0 and 1' represents normalize, 'ln' represents natural logarithm transform, 'log10' represents logarithm base 10 transform, 'log2' represents logarithm base 2 transform, 'exp' represents natural exponential transform, 'sqrt' represents square root transform and 'reciprocal' represents inverse transform.

Step 2: Click **Transform** button.

Step 3: **Transform** button in use turns grey in color.

Step 4: **Transform** button returns back to its original color once it completes its task. A message regarding the percentage increase in missing data due to data transformation appears in red color in the lower side of the **Data Transformation** widget. Select the numerical feature from **Selected Numerical Features** list box.

Step 5: A histogram of the selected numerical feature appears in the right hand side of the **Data Transformation** widget.

We use **Transform** button to logarithmize 'Population_Size' in the example data. When we logarithmize 'Population_Size', the distribution becomes symmetric. Figures 4.97-4.101 illustrate how to use **Transform** button.

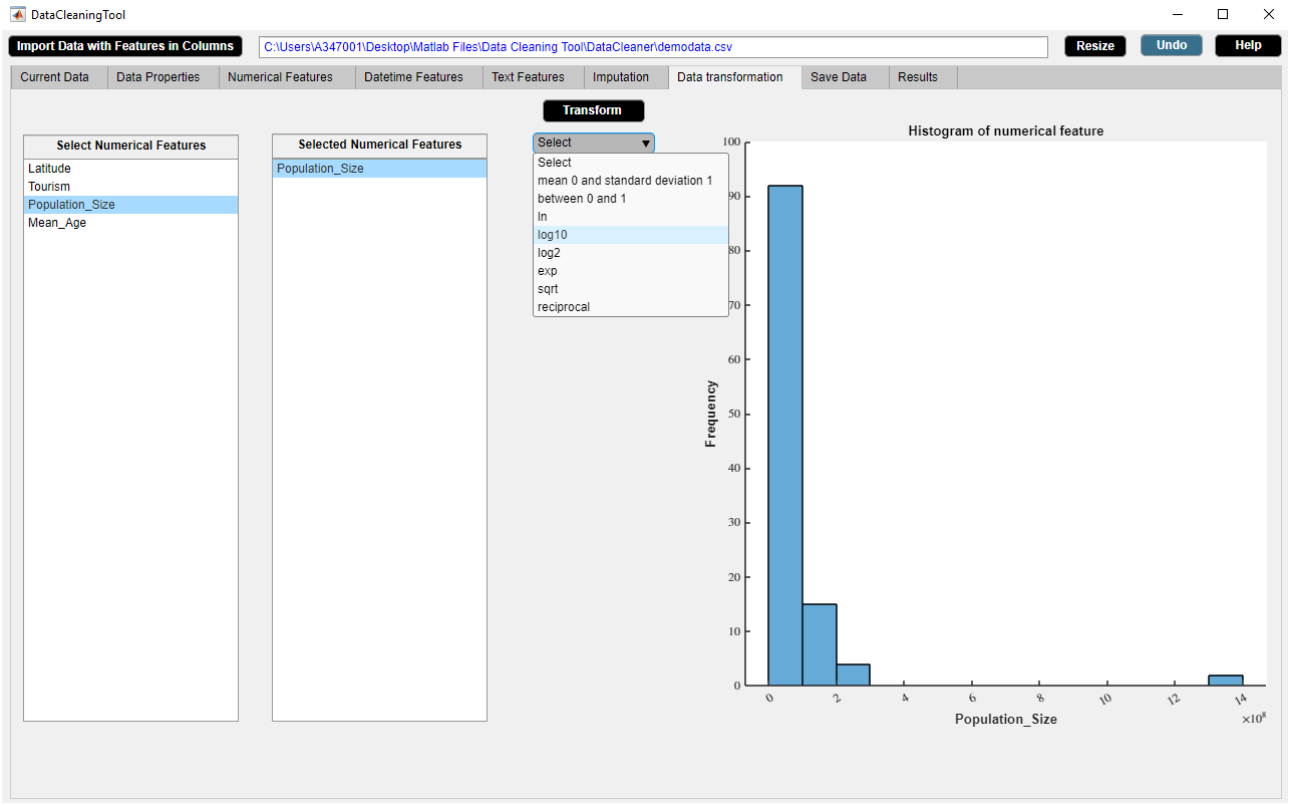


Figure 4.97: Step 1. Transform Button

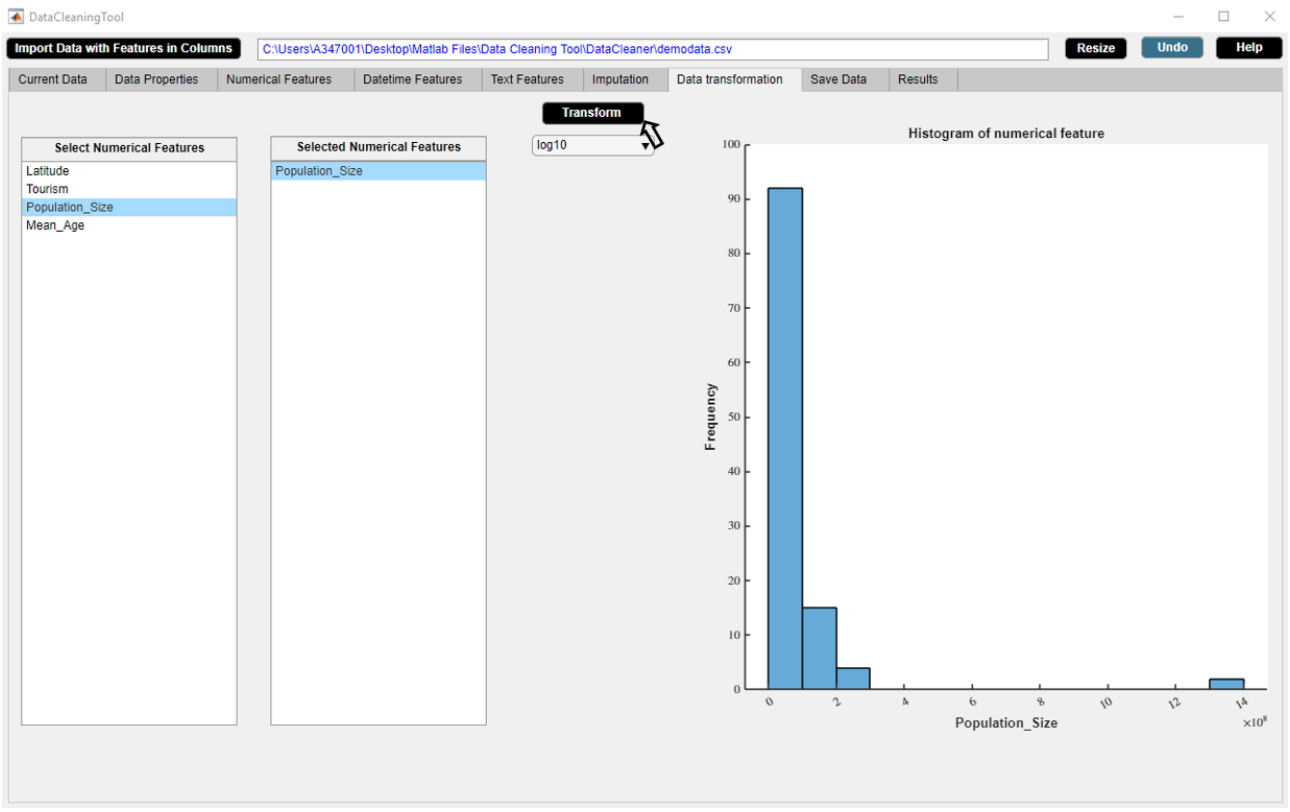


Figure 4.98: Step 2. Transform Button

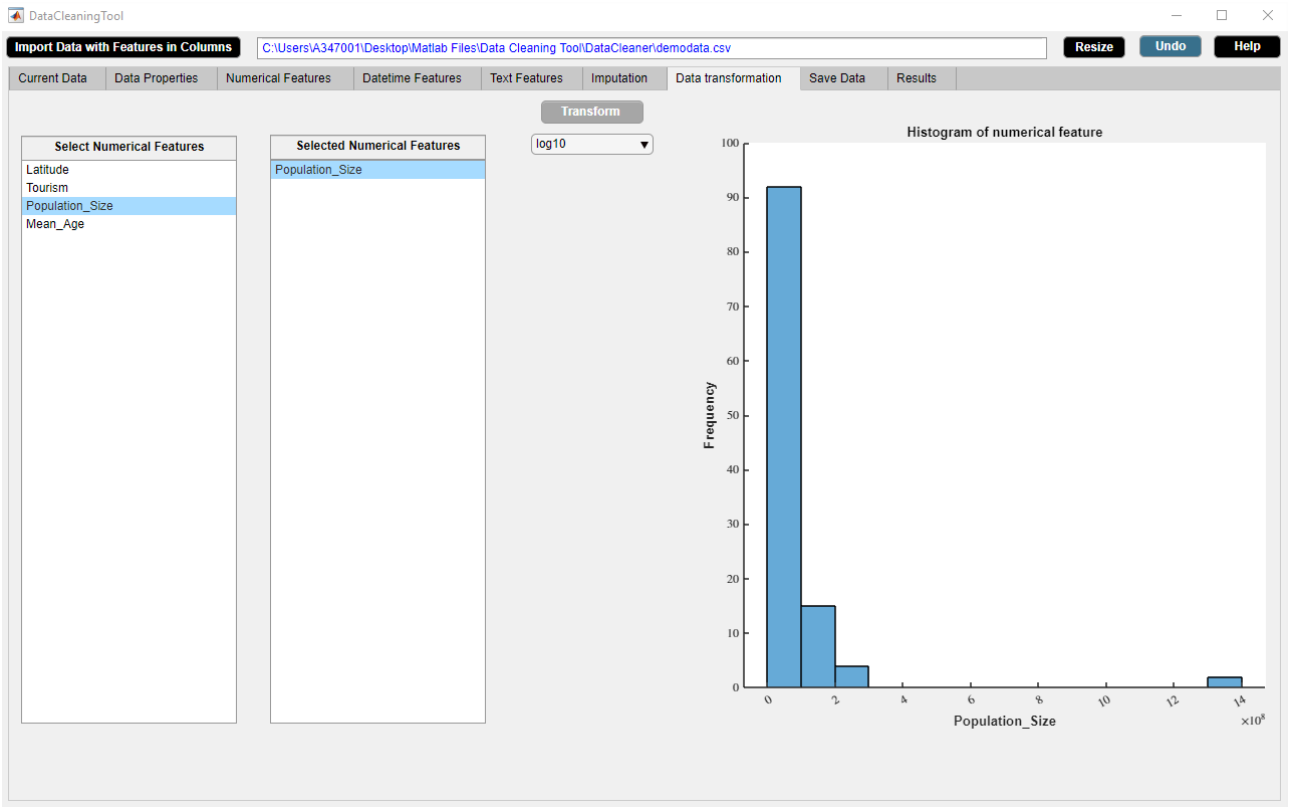


Figure 4.99: Step 3. Transform Button

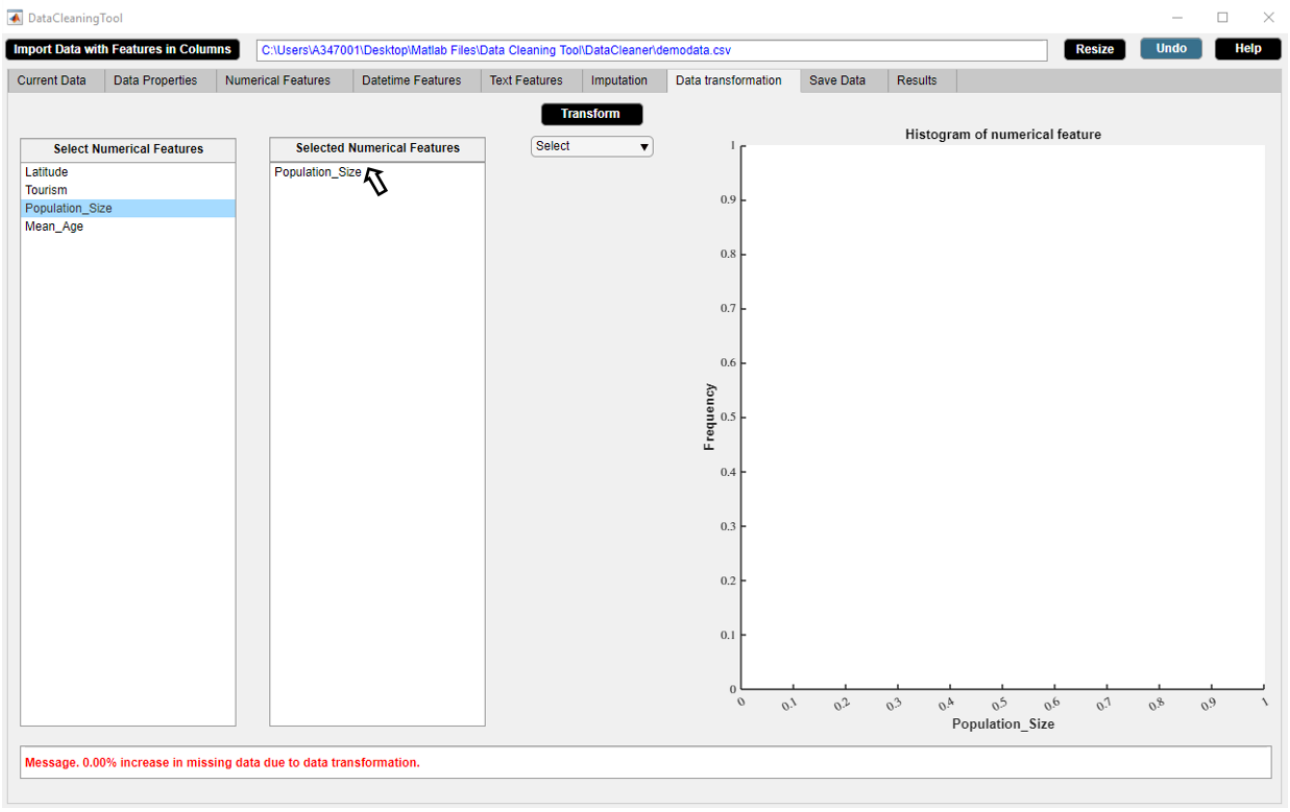


Figure 4.100: Step 4. Transform Button

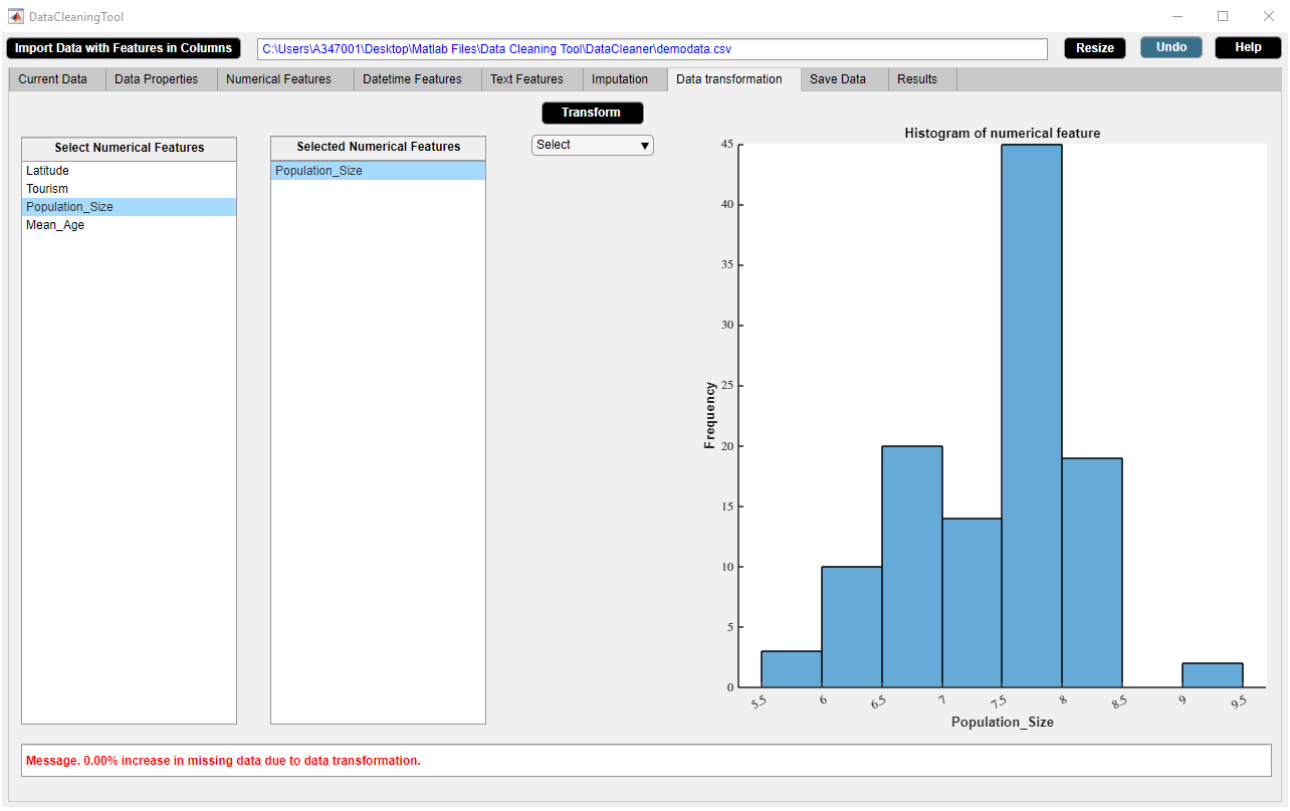


Figure 4.101: Step 5. Transform Button

4.8 Save Data

The Save Data widget displays the full paths of the saved files. The Save Data widget is shown in figure 4.102. The properties of the Save Data widget are as follows.

- The widget saves data in csv or xlsx format after data cleaning.
- Data can be saved for multiple times after each activity.
- The full paths of the saved files are displayed.

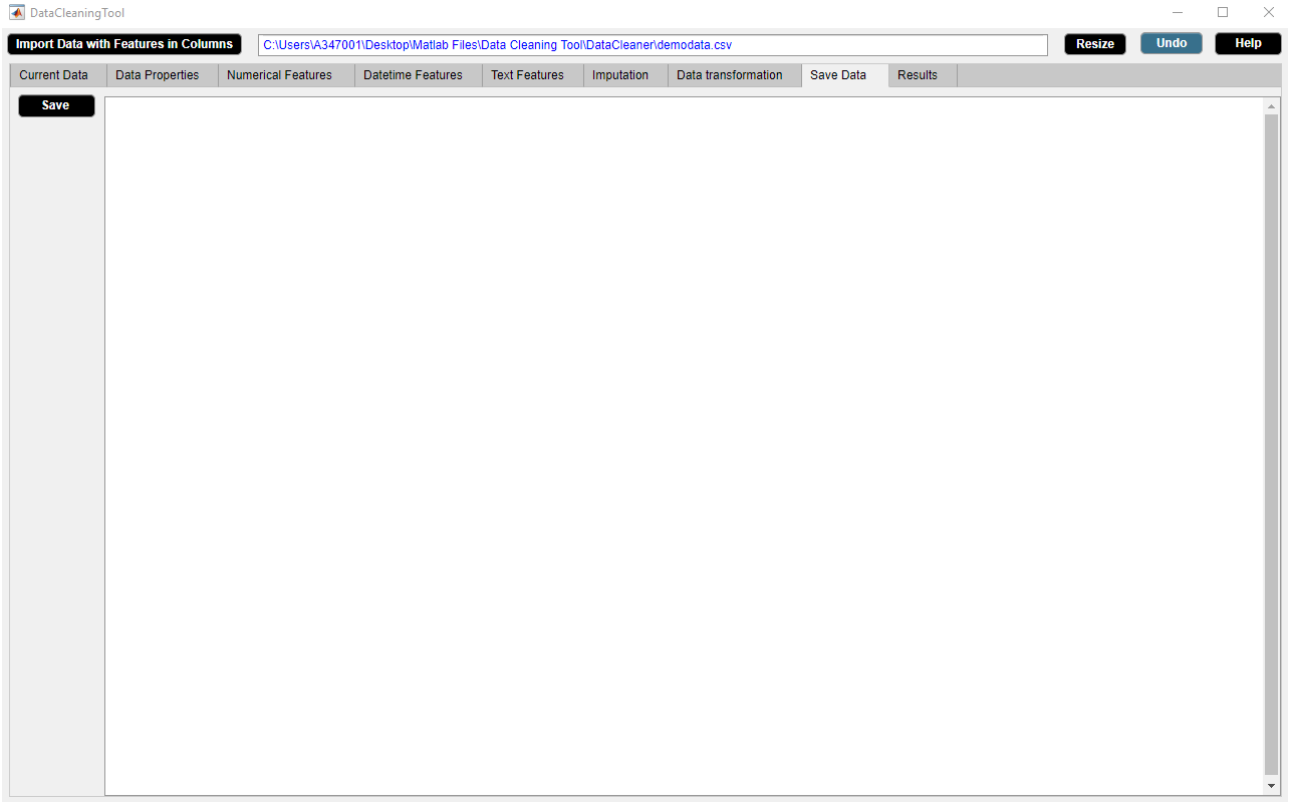


Figure 4.102: Save Data Widget.

4.8.1 Save Button

Saves as comma-separated (.csv) or Excel (.xlsx) file.

Example

Step 1: Click **Save** button.

Step 2: **Save** button in use turns grey in color.

Step 3: **Save** button returns back to its original color once it completes its task.

We use **Save** button to save the example data in csv format. Figures 4.103-4.106 illustrate how to use **Save** button.

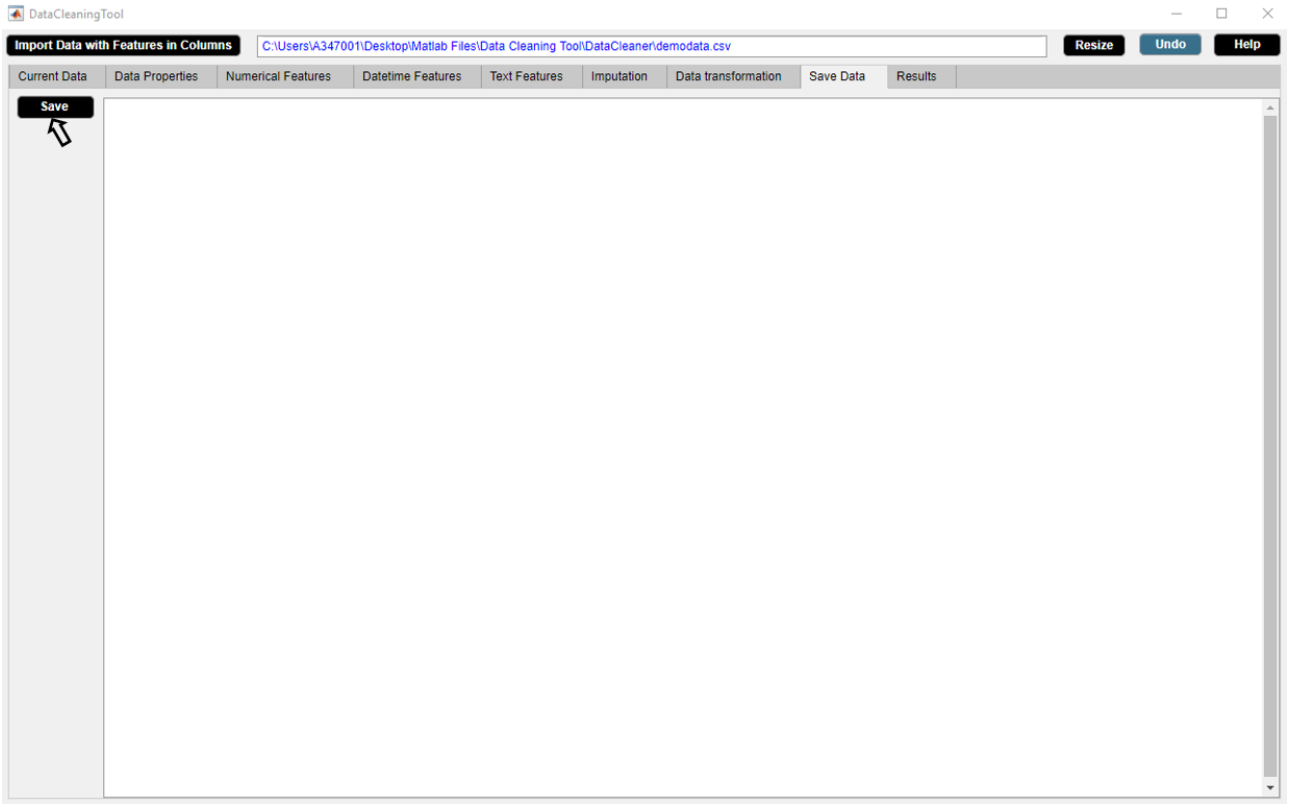


Figure 4.103: Step 1. Save Button

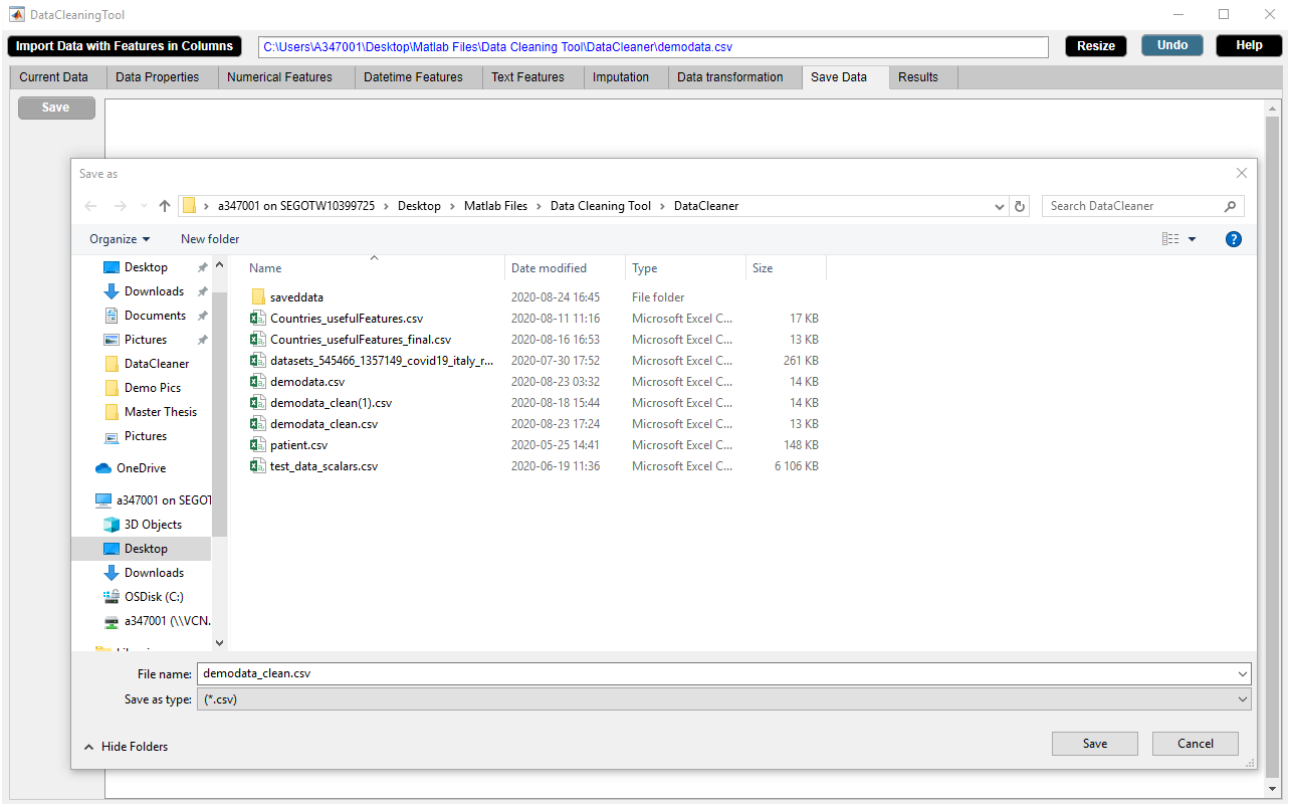


Figure 4.104: Step 2. Save Button

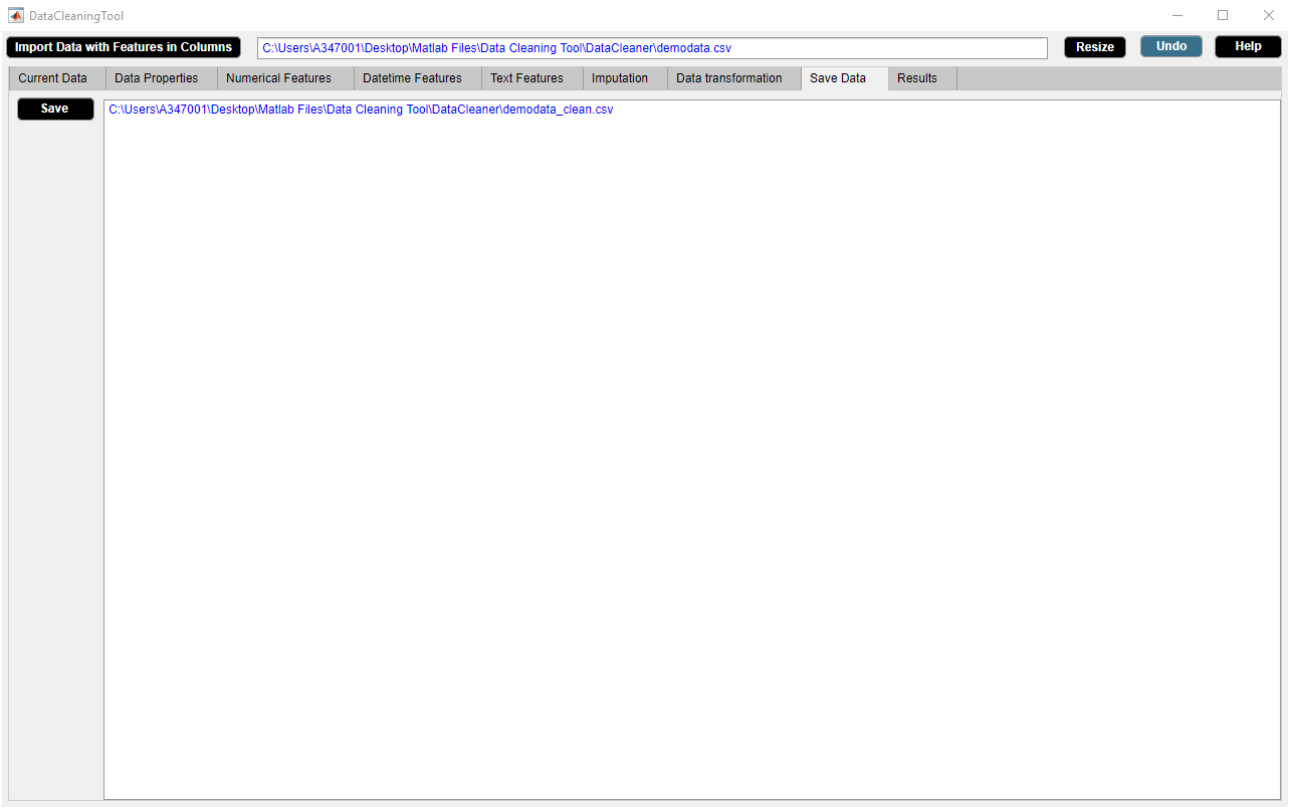


Figure 4.105: Step 3. Save Button

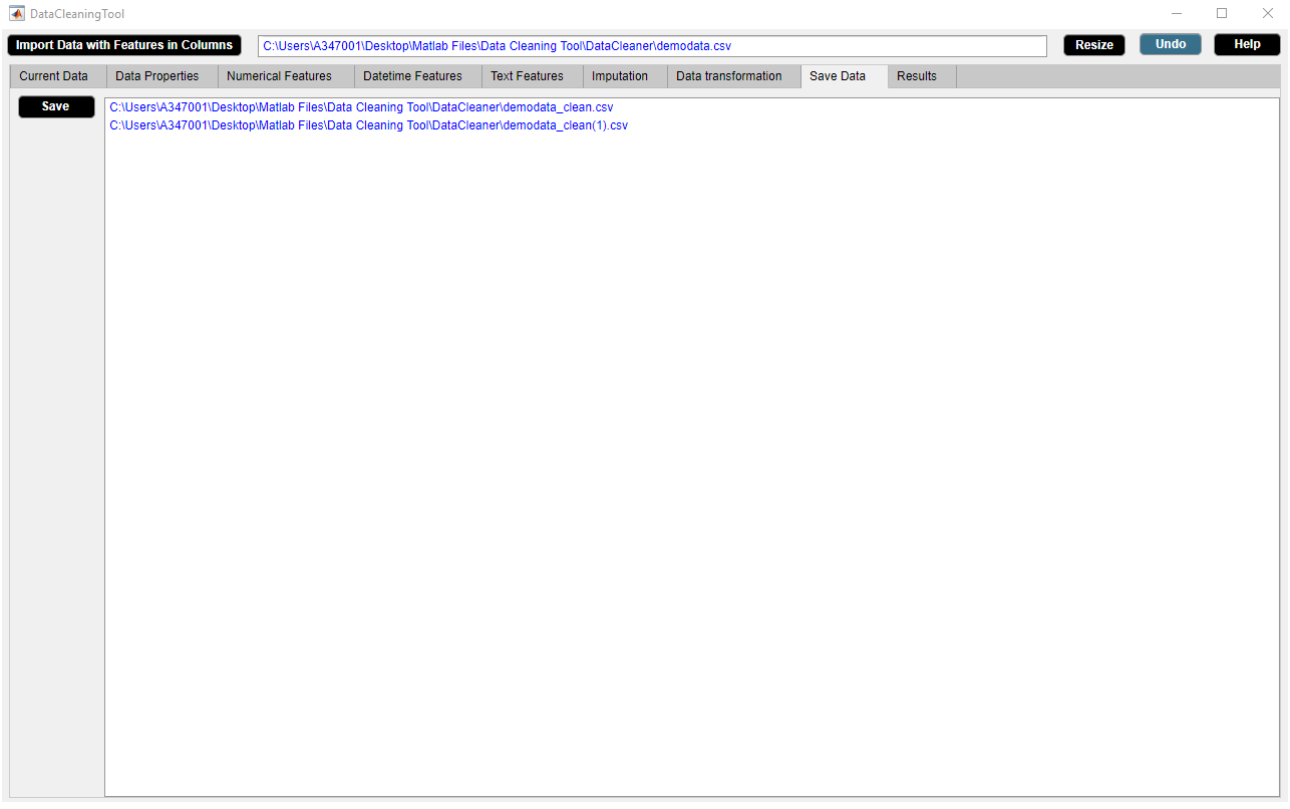


Figure 4.106: Step 4. Save Button

4.9 Results

The Results widget displays information about the final report. The Results widget is shown in figure 4.107. The properties of the Results widget are as follows.

- The widget generates results in pdf format after data cleaning. The results contains a detailed report of all the changes made in DataCleaningTool.
- Results can be generated containing a detailed report of specific changes made in DataCleaningTool.
- Results can be generated for multiple times after each activity.
- The full paths of the results are displayed.

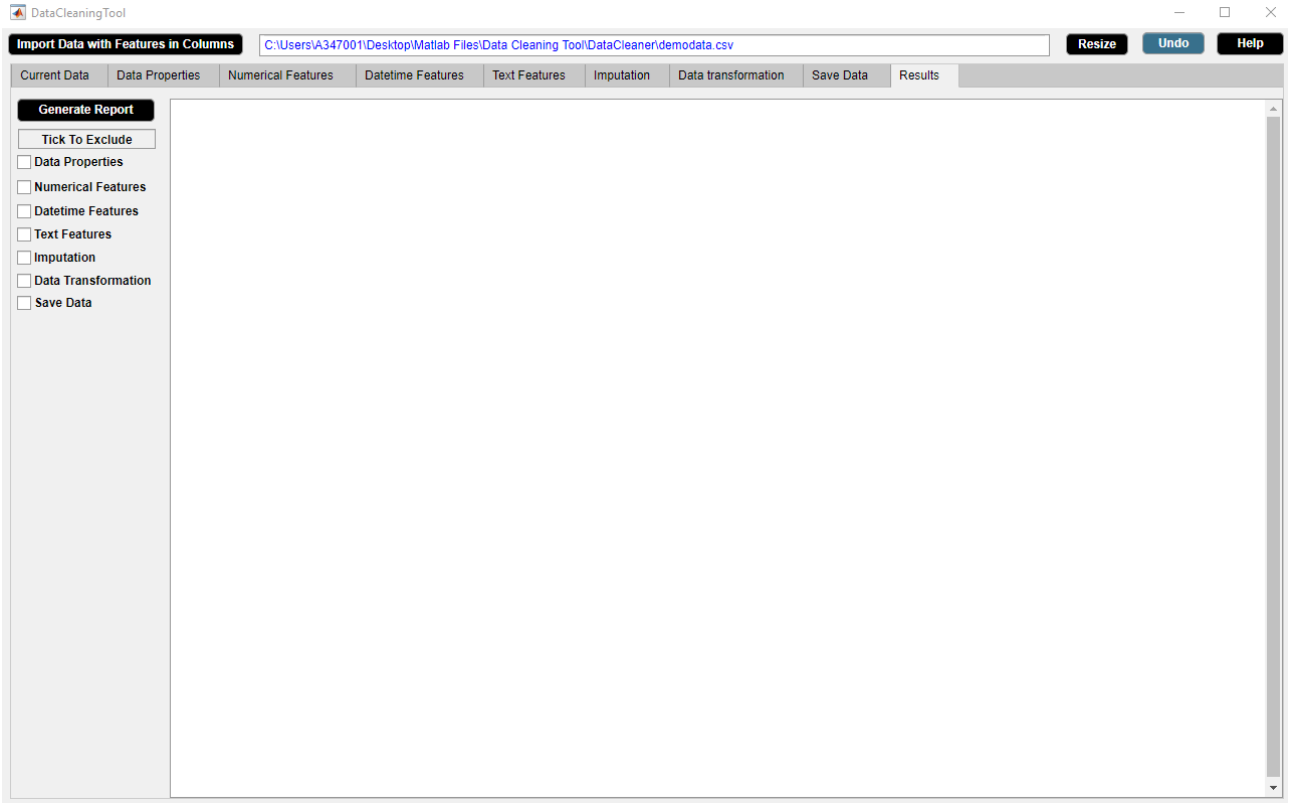


Figure 4.107: Results Widget.

4.9.1 Generate Report Button

Generate pdf file containing results.

Example

Step 1: Click **Generate Report** button.

Step 2: **Generate Report** button in use turns grey in color.

Step 3: **Generate Report** button returns back to its original color once it completes its task.

We use **Generate Report** button to save the example data in csv format. Figures 4.108-4.111 illustrate how to use **Generate Report** button.

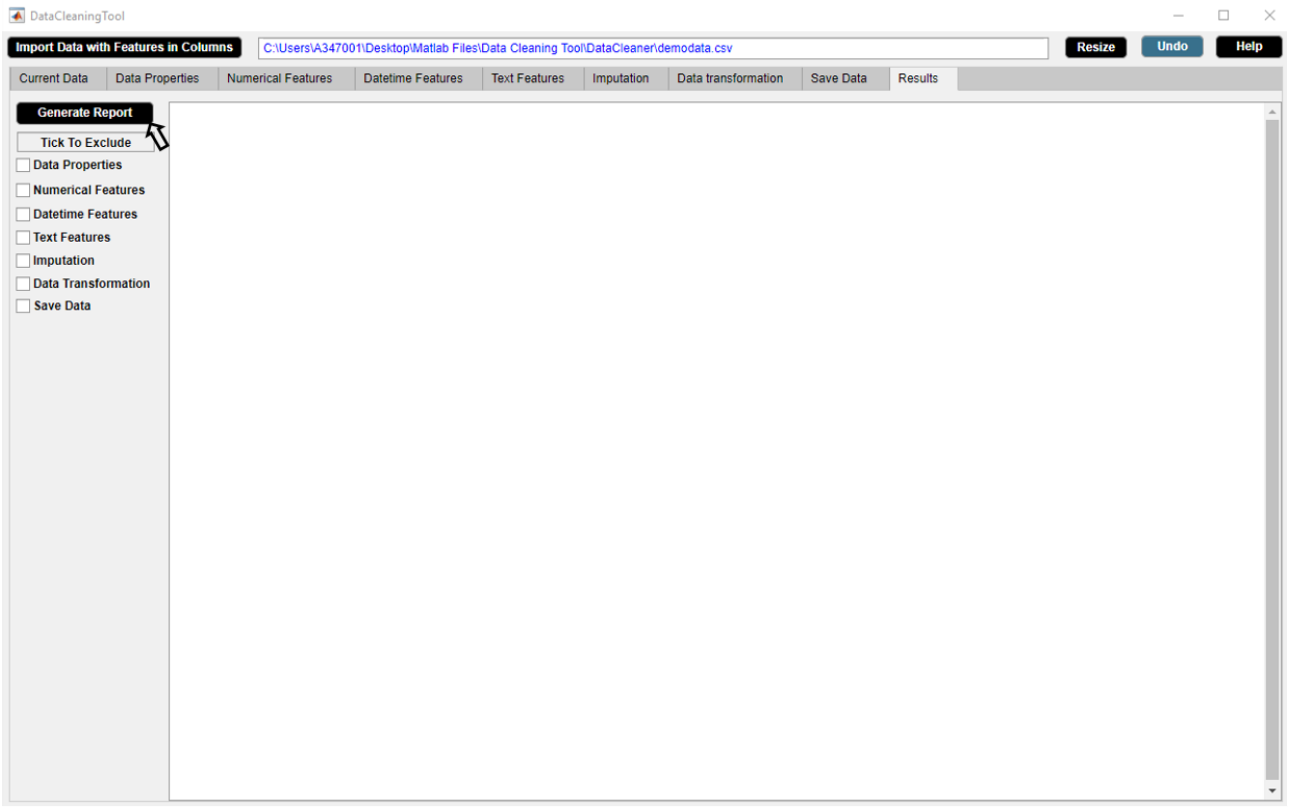


Figure 4.108: Step 1. Generate Report Button

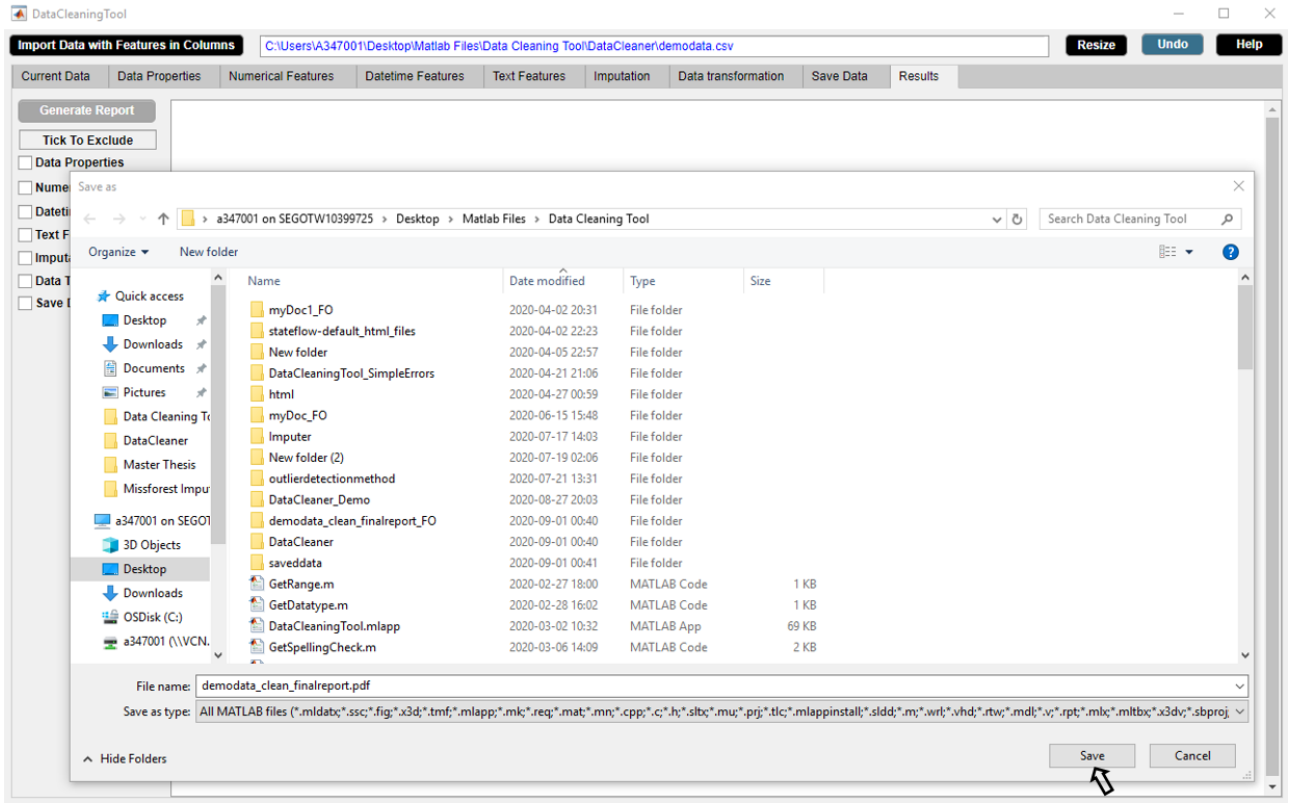


Figure 4.109: Step 2. Generate Report Button

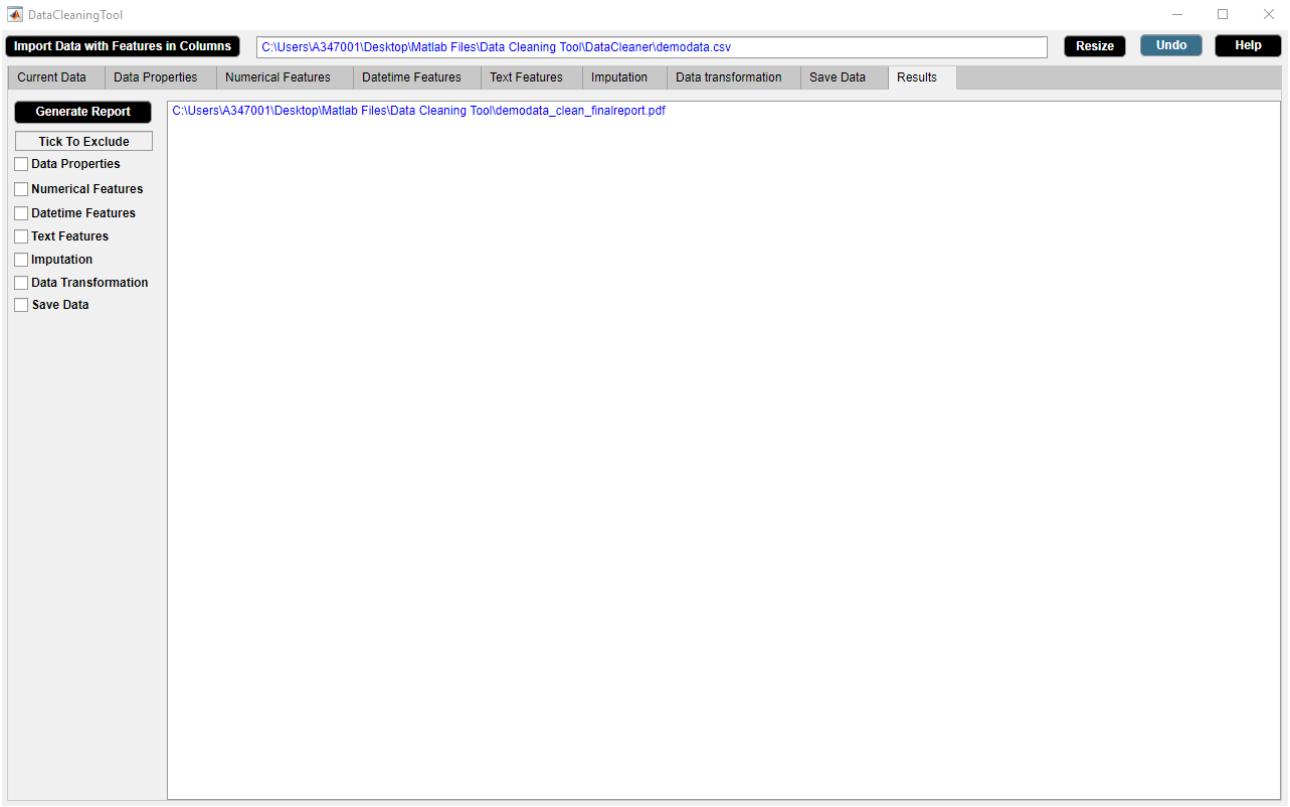


Figure 4.110: Step 3. Generate Report Button

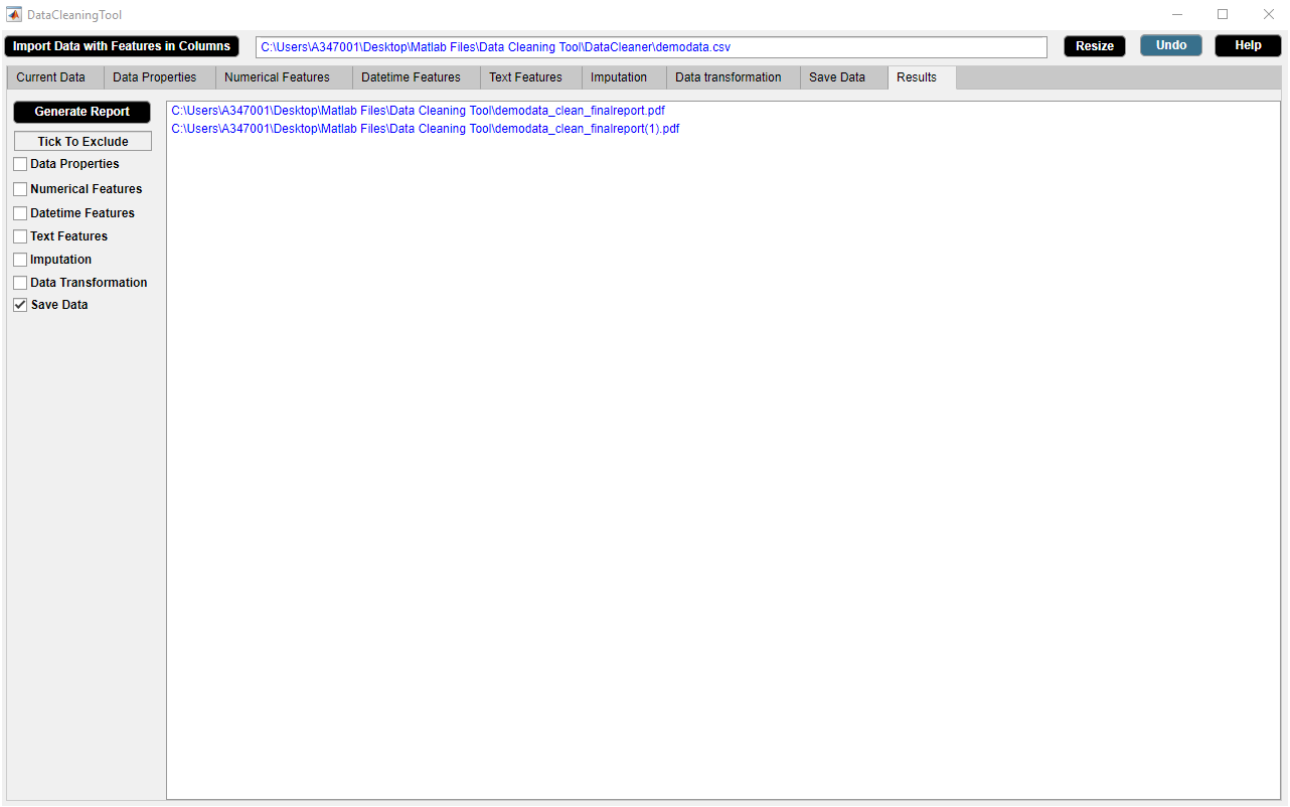


Figure 4.111: Step 4. Generate Report Button

Chapter 5

Other Attributes

Other attributes include the following three buttons which are present in the upper right side of the DataCleaningTool 3.1.

5.1 Resize Button

Resizes the DataCleaningTool to a reduced size.

5.2 Undo Button

Performs the last activity and all the widgets get updated accordingly.

5.3 Help Button

Generates user manual of DataCleaningTool in pdf format.

References

[1] *Set command window output display format - matlab format [internet]. mathworks.com. 2020 [cited 7 september 2020]. available from: Https://www.mathworks.com/help/matlab/ref/format.html.*

[2] D. J. Stekhoven and P. Bühlmann, "Missforest - non-parametric missing value imputation for mixed-type data", *Bioinformatics*, vol. 28 1, pp. 112–8, 2012.

[3] *Standard score [internet]. en.wikipedia.org. 2020 [cited 14 september 2020]. available from: Https://en.wikipedia.org/wiki/standard_score.*

[4] *Covid-19 useful features by country [internet]. kaggle.com. 2020 [cited 9 september 2020]. available from: Https://www.kaggle.com/ishivinal/covid19-useful-features-by-country.*