UCDAVIS HEALTH | Clinical and Translational Science Center

Instruction Manual for the UC Davis CTSC Data Loofah

Thank you for choosing to use the UC Davis CTSC Data Loofah.

The Data Loofah is an investigative tool used to assess the quality of data prior to statistical analysis. Even with the best care and data management practices, data can accumulate errors that can be difficult to catch and may negatively impact the analysis process and results. This is where the Data Loofah comes in. The Data Loofah will summarize variables and outcomes to help you identify potential errors like extreme, nonsensical, inconsistent and missing values, and incorrectly categorized variables.

The purpose of the Data Loofah is to facilitate identifying data errors; it does not support correcting the errors and does not conduct statistical analyses. While the Data Loofah will generate summary statistics, the purpose of these summaries is to identify data errors and should not be used for reporting scientific results.

Access the Data Loofah here: https://dataloofah.ucdavis.edu/

NOTE: The Data Loofah can only be accessed by using a UC Davis/UC Davis Health computer or a UC Davis/UC Davis Health VPN.

Additional statistics resources:

https://health.ucdavis.edu/ctsc/area/biostatistics/other-resources.html

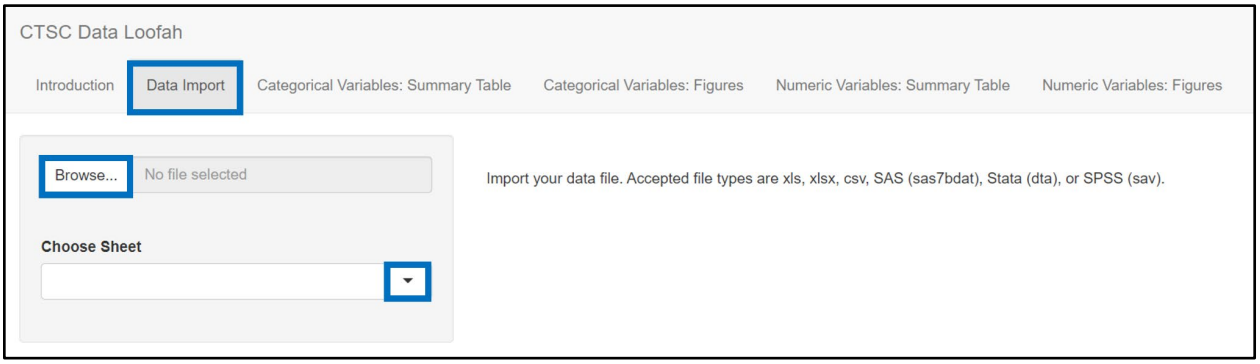
For inquiries, please email dataloofah@ucdavis.edu.

January 2023

Instructions:

1. **Uploading your Data**
   Upload your data by clicking on the “Data Import” tab. Select “Browse” to choose your file, then select the sheet number you would like to upload in the drop-down list.

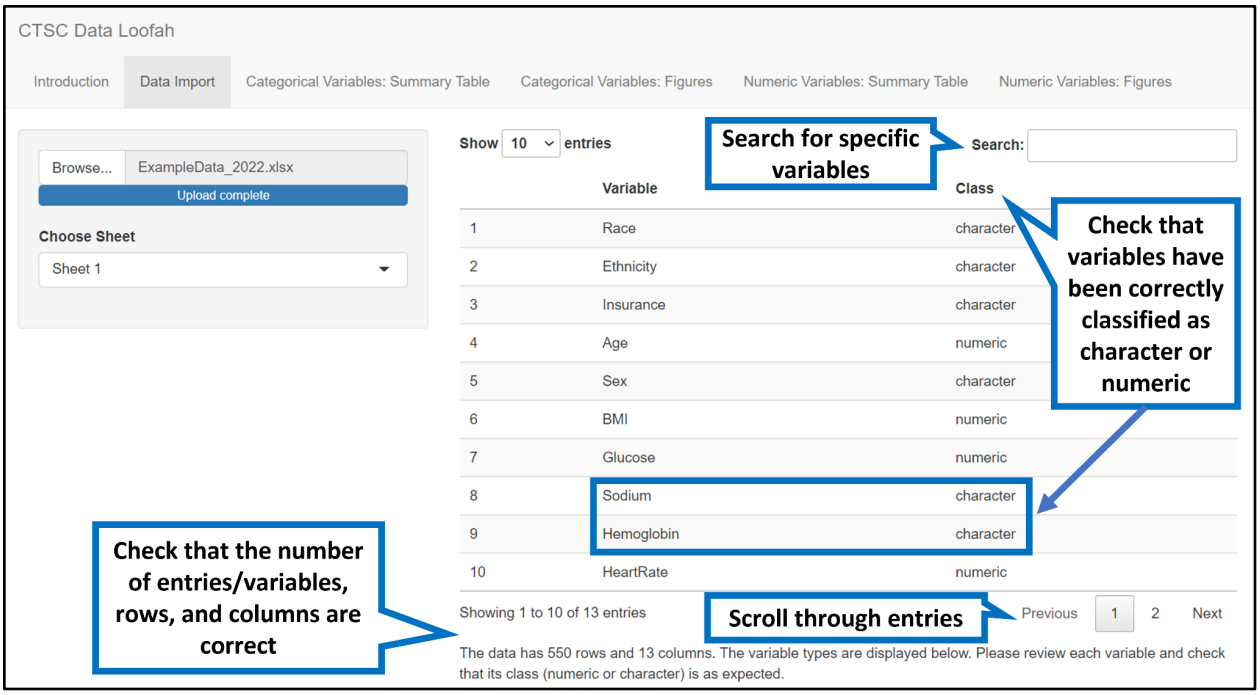
NOTE: Accepted file types include xls, xlsx, csv, SAS (sas7bdat), Stata (dta), and SPSS (sav). Please make sure that your data sheet does not contain any summary statistics or figures.



2. **Viewing your Upload Summary**
   Once your upload is complete, you will be presented with a summary of the variables identified and what class they were categorized as (character or numeric).

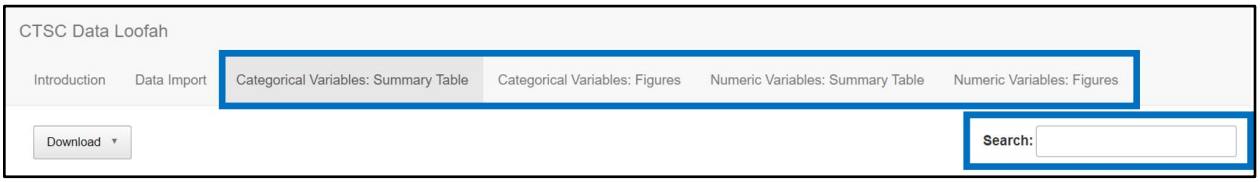
   You can view the number of entries, rows, and columns. You can also look up variable names using the search bar or scrolling through the pages of entries.

   NOTE: Date variables are not easily recognized by Data Loofah and should be ignored while reviewing data.

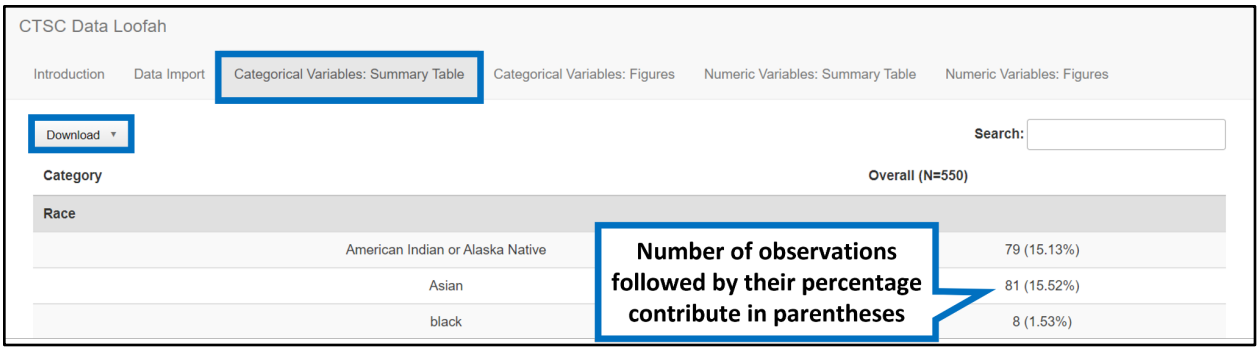


3. Viewing your Data

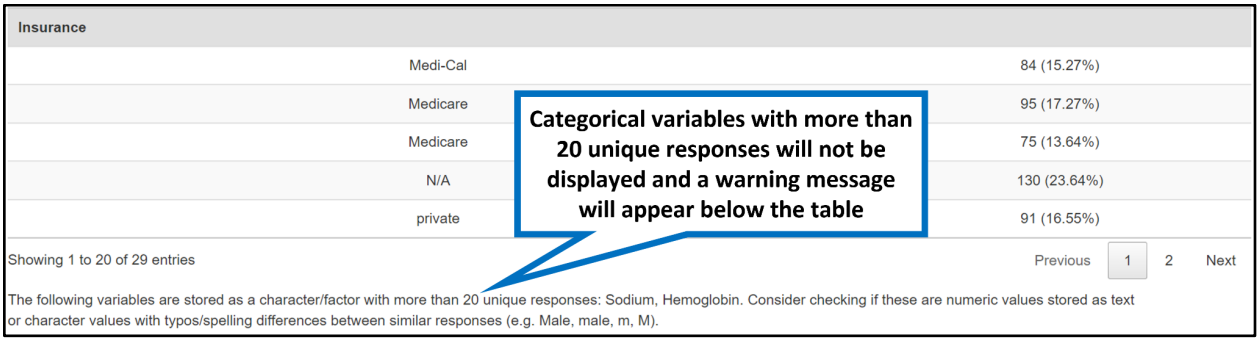
Your data will be presented in tables and figures by variable class - categorical and numeric. You can look up a specific variable by name or values using the search bar or scrolling through the pages of entries.



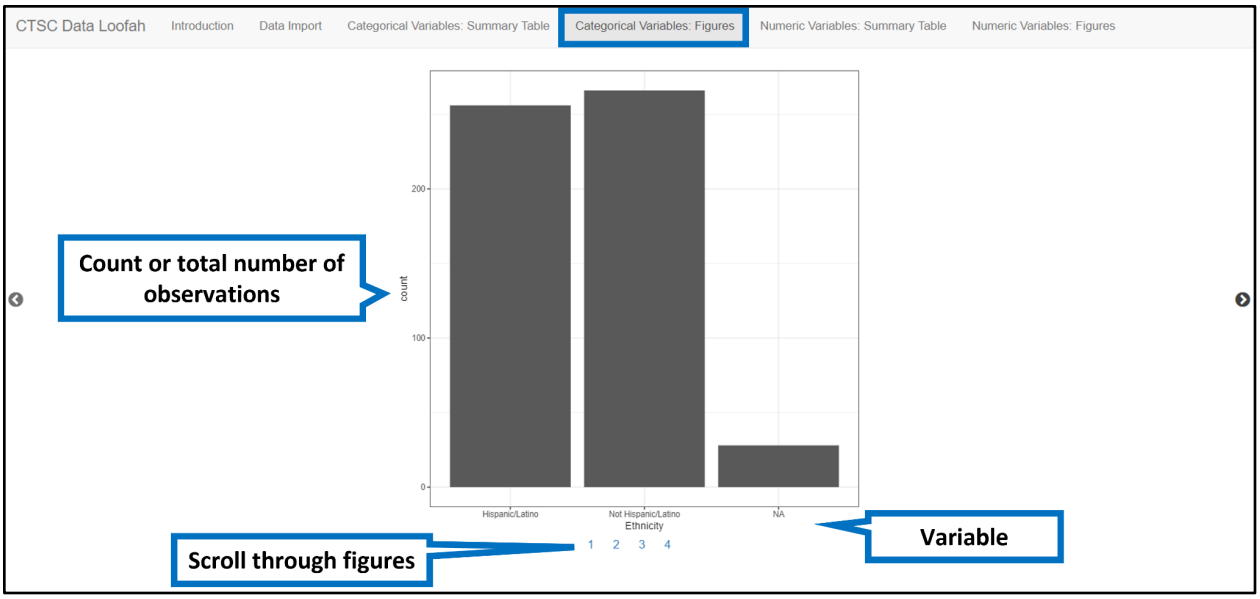
- **Categorical Variables: Summary Tables**
  Categorical variables will be presented by variable name and the total number of observations “n”, followed by their percentage contribution in parentheses. You can download the summary tables by selecting the ‘download’ button.



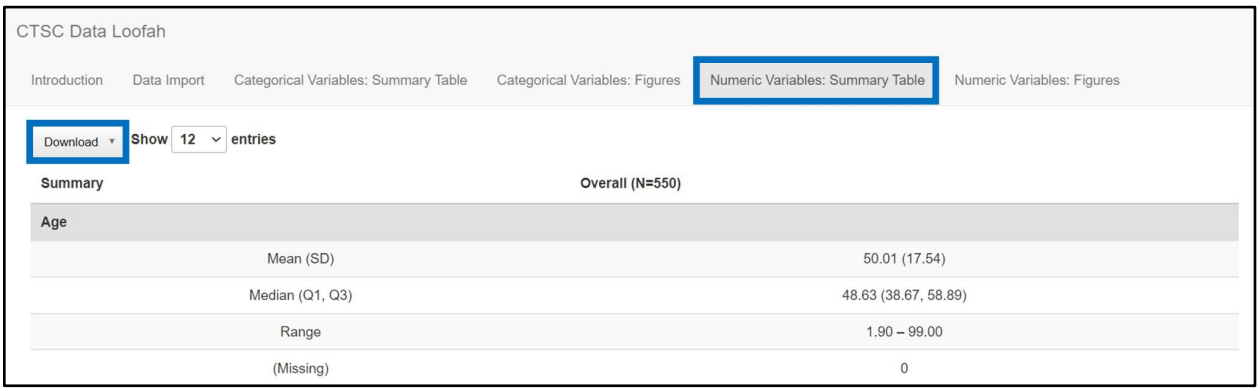
NOTE: Categorical variables with more than 20 unique responses will not be displayed, and a warning message will appear at the bottom of the table. This can indicate that a numeric variable has been classified as a categorical variable (see image below).



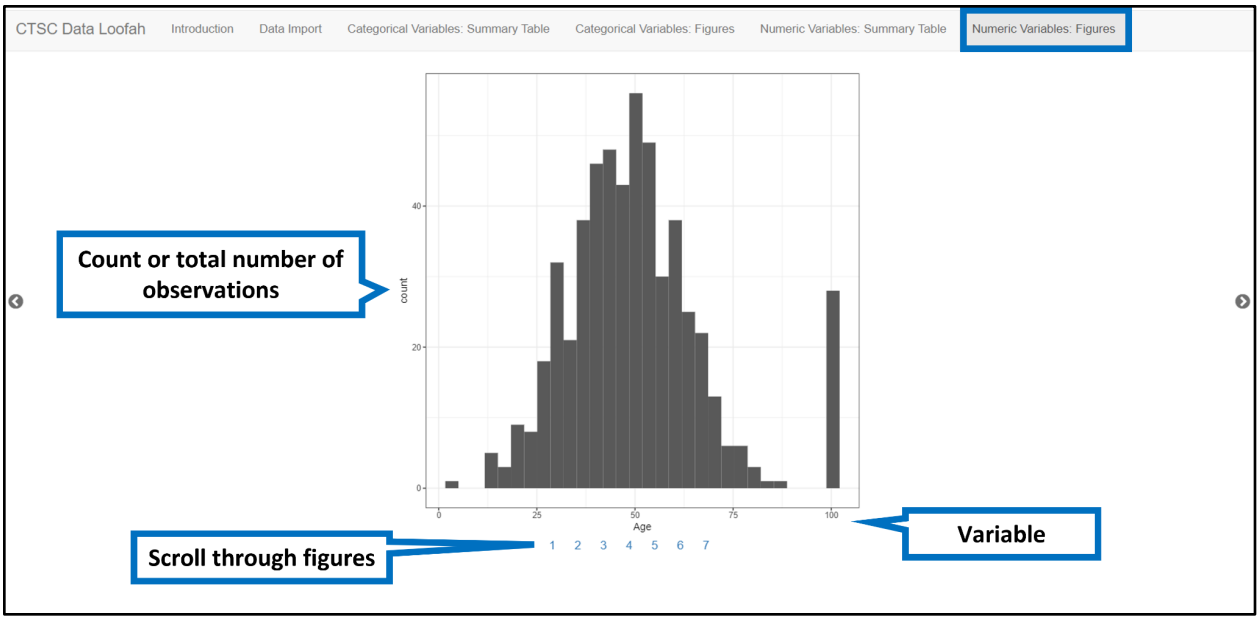
- **Categorical Variables: Figures**
  This is the same data shown in the “Categorical Variables: Summary Tables”; however, here categorical variables will be presented graphically with the variable’s unique categories on the x-axis and the total number of observations on the y-axis.



- **Numeric Variables: Summary Tables**
  Summary statistics (mean, median, range, and number of missing) will be presented by variable name for each numeric variable. You can download the summary tables by selecting the 'download' button.



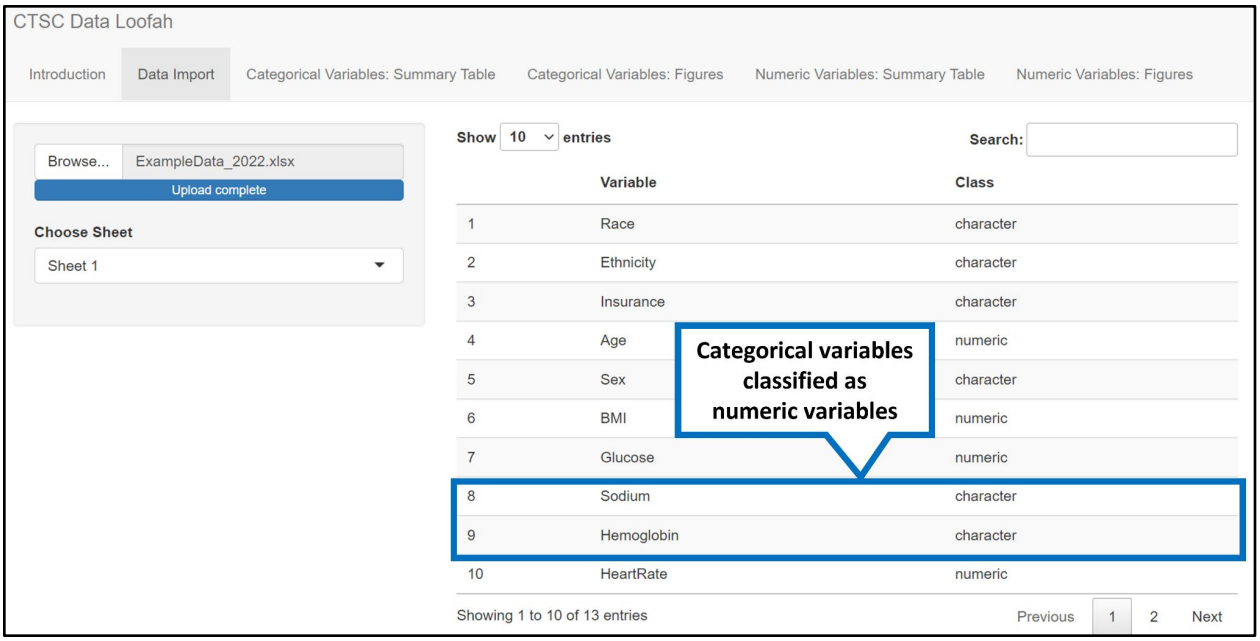
- **Numeric Variable: Figures**
  This is the same data as the "Numeric Variables: Summary Tables"; however, numeric variables will be presented graphically with the variable's values on the x-axis and the total number of observations on the y-axis.

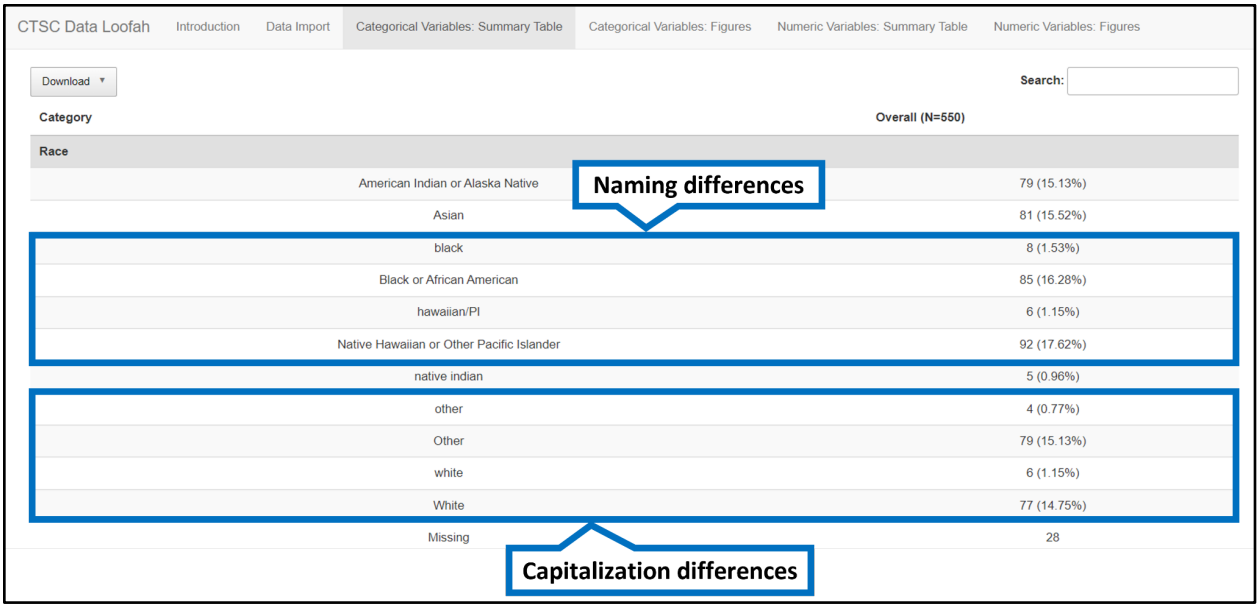


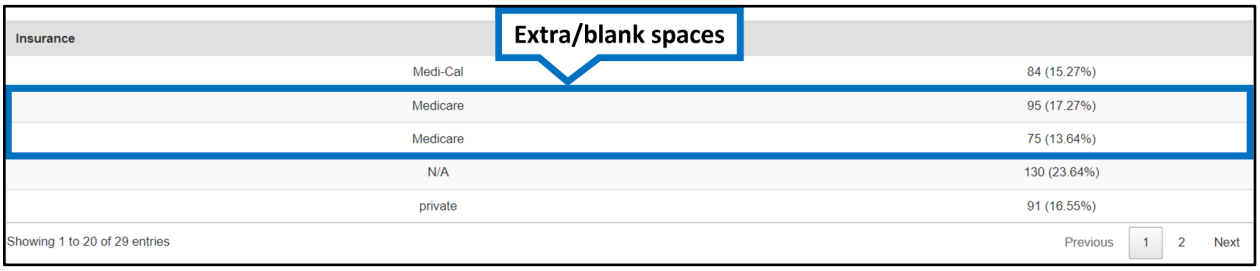
4. Checking your Data for Errors

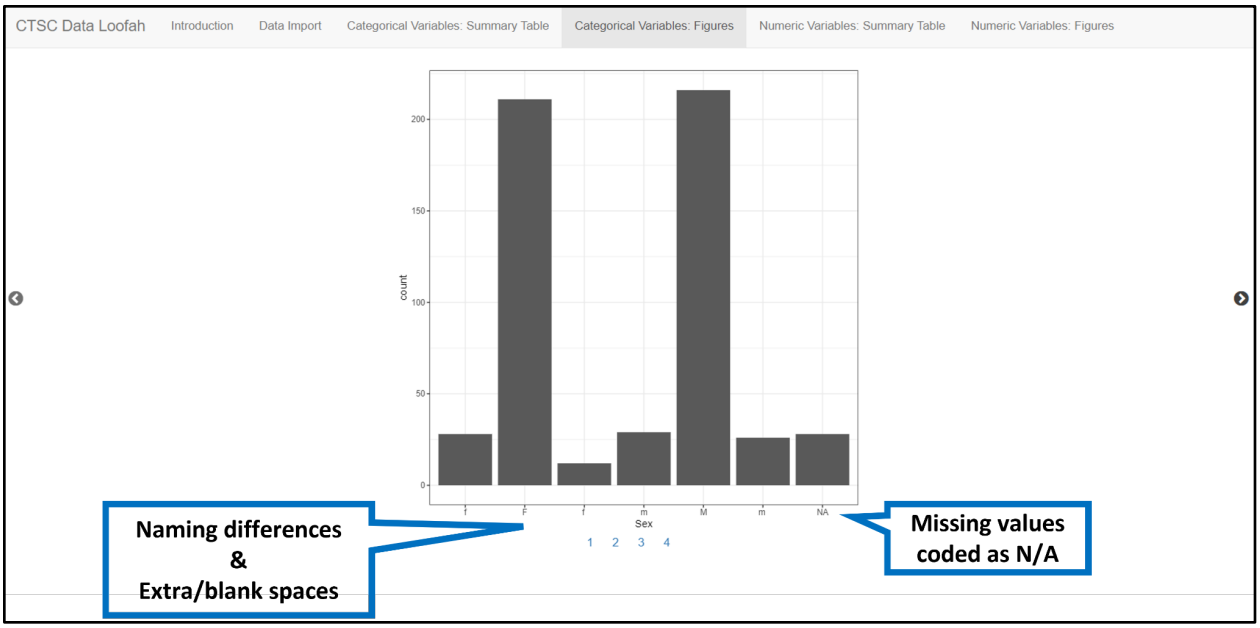
- **Common errors for Categorical Variables**:
  - numeric variables classified as categorical variables
  - different naming designations for the same response (e.g., female, f)
  - typos/spelling differences for the same responses
  - capitalization differences for the same responses (e.g., male, Male)
  - extra/blank spaces, particularly at the end of the entry
  - missing values are coded using a numeric value or N/A.

Examples:


CTSC Data Loofah
Introduction
Data Import
Categorical Variables: Summary Table
Categorical Variables: Figures
Numeric Variables: Summary Table
Numeric Variables: Figures
Browse...
ExampleData_2022.xlsx
Upload complete
Choose Sheet
Sheet 1
Show 10 entries
Search:
Variable
Class
1 Race character
2 Ethnicity character
3 Insurance character
4 Age numeric
5 Sex character
6 BMI numeric
7 Glucose numeric
8 Sodium character
9 Hemoglobin character
10 HeartRate numeric
Categorical variables classified as numeric variables
Showing 1 to 10 of 13 entries
Previous
1
2
Next

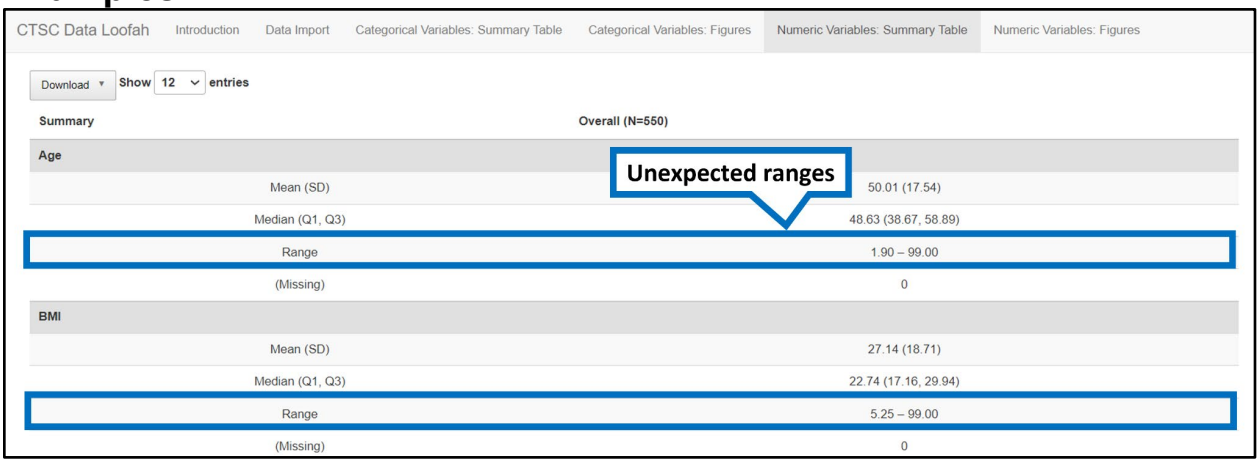

CTSC Data Loofah
Introduction
Data Import
Categorical Variables: Summary Table
Categorical Variables: Figures
Numeric Variables: Summary Table
Numeric Variables: Figures
Download
Search:
Category
Overall (N=550)
Race
American Indian or Alaska Native 79 (15.13%)
Asian 81 (15.52%)
black 8 (1.53%)
Black or African American 85 (16.28%)
hawaiian/PI 6 (1.15%)
Native Hawaiian or Other Pacific Islander 92 (17.62%)
native indian 5 (0.96%)
other 4 (0.77%)
Other 79 (15.13%)
white 6 (1.15%)
White 77 (14.75%)
Missing 28
Naming differences
Capitalization differences


Insurance
Medi-Cal 84 (15.27%)
Medicare 95 (17.27%)
Medicare 75 (13.64%)
N/A 130 (23.64%)
private 91 (16.55%)
Extra/blank spaces
Showing 1 to 20 of 29 entries
Previous
1
2
Next

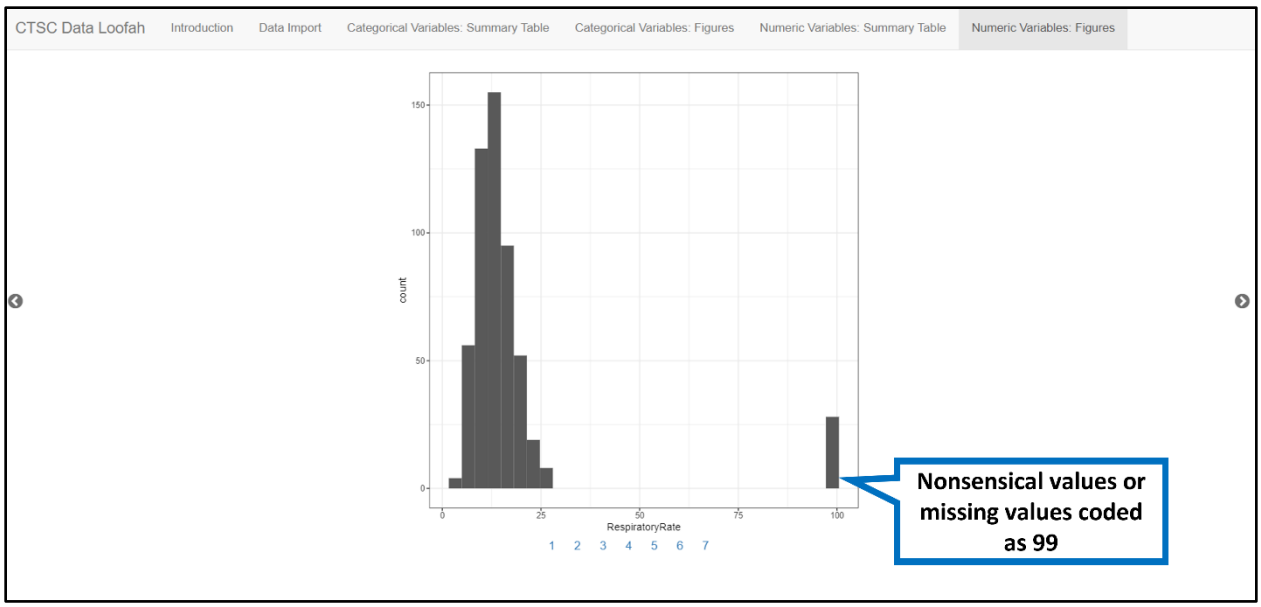


- **Common errors for Numeric Variables**:
  - unexpected mean, median, or range
  - nonsensical values (e.g., outliers or different units)
  - missing values are coded using a numeric value that is within the range of the variable (e.g., 99)

Examples:



Summary	Overall (N=550)
Age	
Mean (SD)	50.01 (17.54)
Median (Q1, Q3)	48.63 (38.67, 58.89)
Range	1.90 – 99.00
(Missing)	0
BMI	
Mean (SD)	27.14 (18.71)
Median (Q1, Q3)	22.74 (17.16, 29.94)
Range	5.25 – 99.00
(Missing)	0



5. **Once you have identified any errors, you need to make the appropriate corrections to your data file. You can then re-upload the corrected data file to ensure all corrections were made.**