preliminary REPORT ON Data analysis for brc-stream  
Drinking water quality of kuttanad

17/02/2025

|  |
| --- |
| Notes to client |
| * As mentioned in the conclusion, there are no inferences made from the data analysis plots and tables. * Please go through the plots and tables, and let me know of any errors or confusions. |

# Overview

## Data Profiling

|  |  |
| --- | --- |
|  | * A MS Excel template was provided to the client for the data entry and successfully entered all the survey responses in the required format. * The raw data (“.xlsx” format) containing coded responses of questionnaire survey conducted at the three Panchayats such as Thakazhy, Edathua, Thalavady of Kuttanad Taluk of Alappuzha district, Kerala, India * The data analysis is requested to done using R-programming language by the client, however for the sake of complexity of raw data, analyst had to choose Python programming language. * The raw-uncleaned data contained 31221 rows and 54 columns, however most of the rows were empty containing no data and with numerous blank rows in-between valid data rows. Also there were wrong data entries, mis-matched responses and missing entries. * For the sake of logical data analysis, out of 54 columns, only 38 columns were decided to taken for the analysis. |

## Data Cleaning

|  |  |
| --- | --- |
|  | * As noted in the data profiling, the raw data were bit-messy and a through manual check was needed. * Manually entered missing entries (Eg: Panchayat code) * Removed all blank rows and corrected the wrong formatting of responses (Eg: & -separated multiple responses, missing comma between multiple responses). * The final cleaned dataset had shape of 393 rows and 54 columns. |

## Data Reduction

|  |  |
| --- | --- |
|  | Cleaned data had rows that are least important in analysis, so analyst had to choose logical columns that can deduce logical relationships. Thus chose only 38 columns. |

## Data Transformation

|  |  |
| --- | --- |
|  | * Column expansion: There were columns which had multiple responses. So to do the data analysis, analysts had to expand those columns. * Grouping: Since the analysis is Panchayat-wise, analyst grouped the data Panchayat-wise using the Panchayat code. * Data were split into categorical (coded responses) and numerical (numerical value) due to its nature. * Now that the data is ready for the analysis |