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 |

## Data Analysis

|  |  |
| --- | --- |
|  | Since the data mostly consists of categorical columns, analyst had to choose for methods applicable for this kind of data. |

* 1. **Descriptive Statistics**

**5.1.1 Measure of central tendency**: Only “Mode” from the measure of central tendency can be used here.

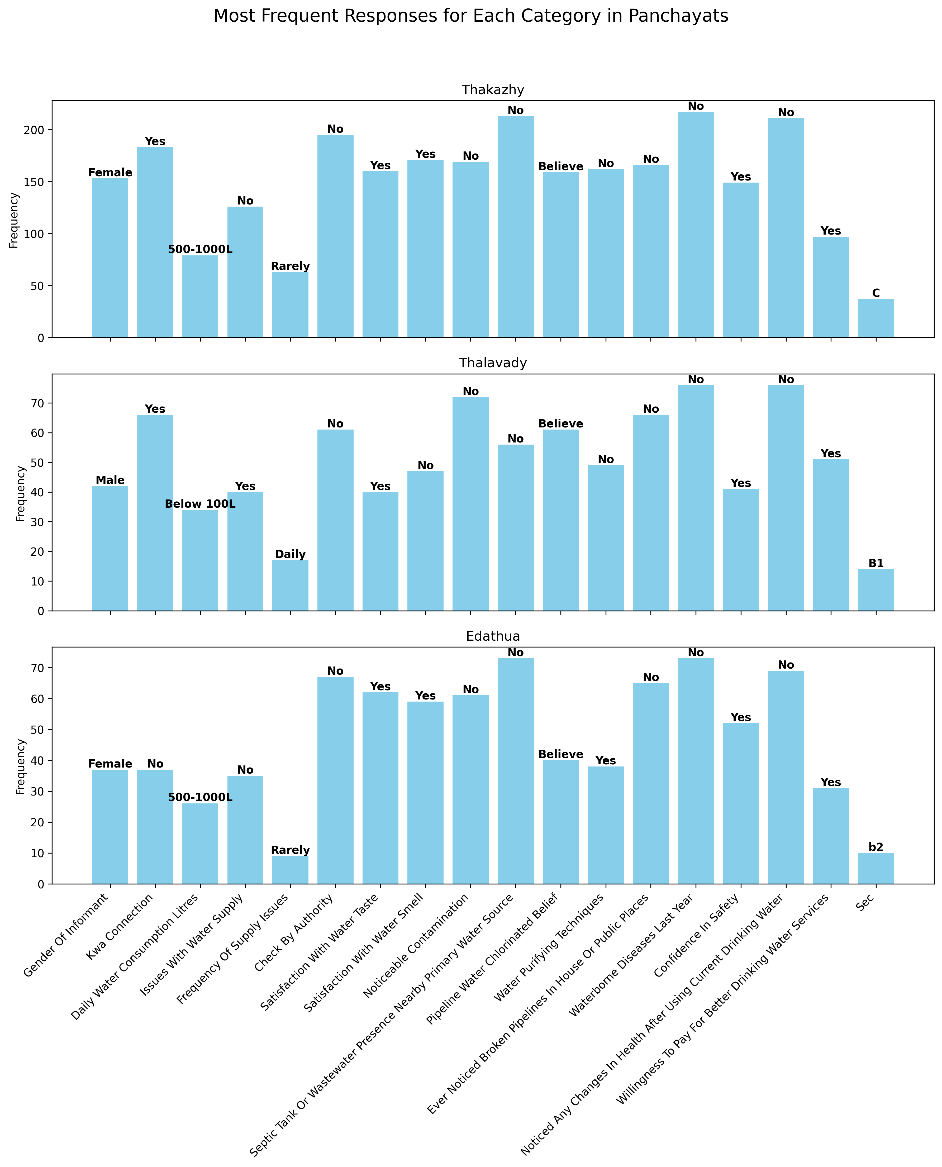
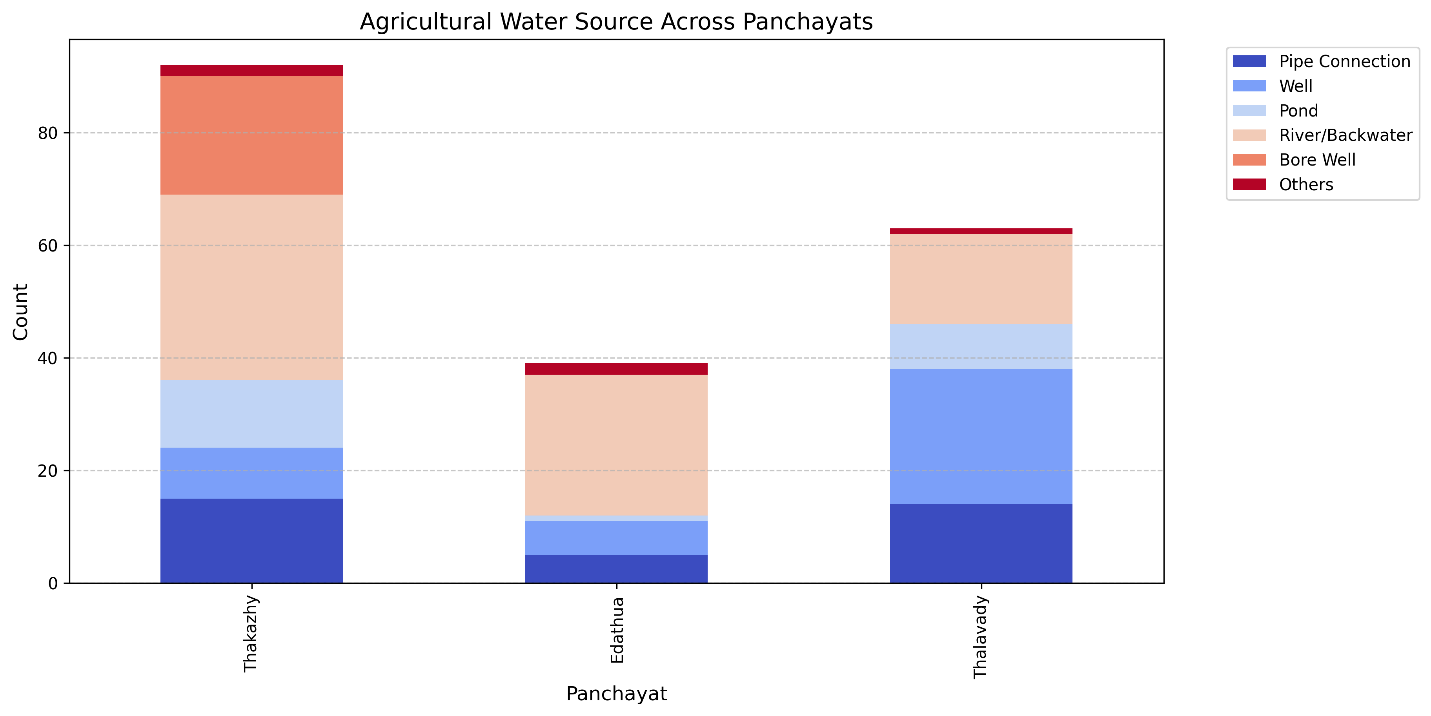
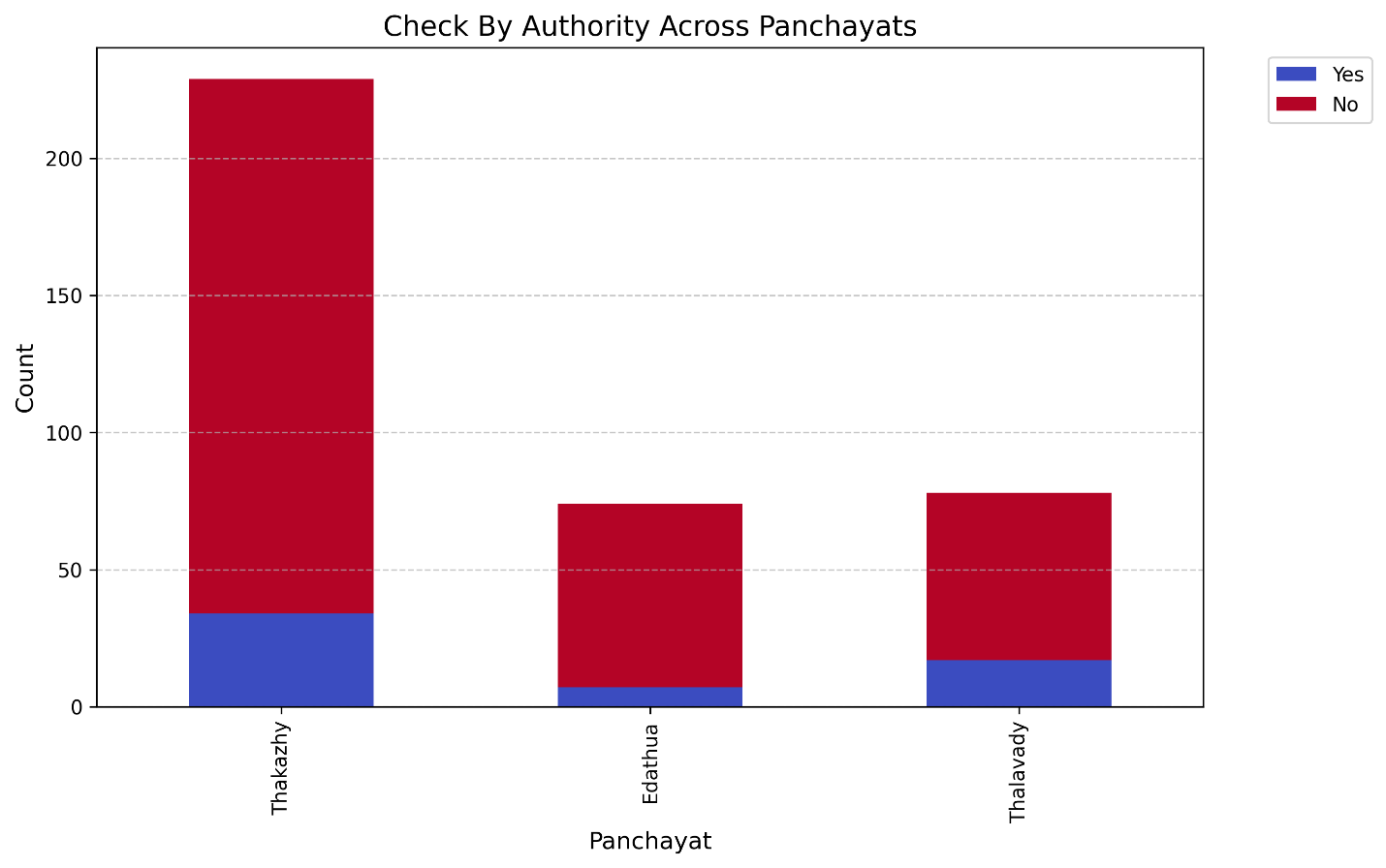


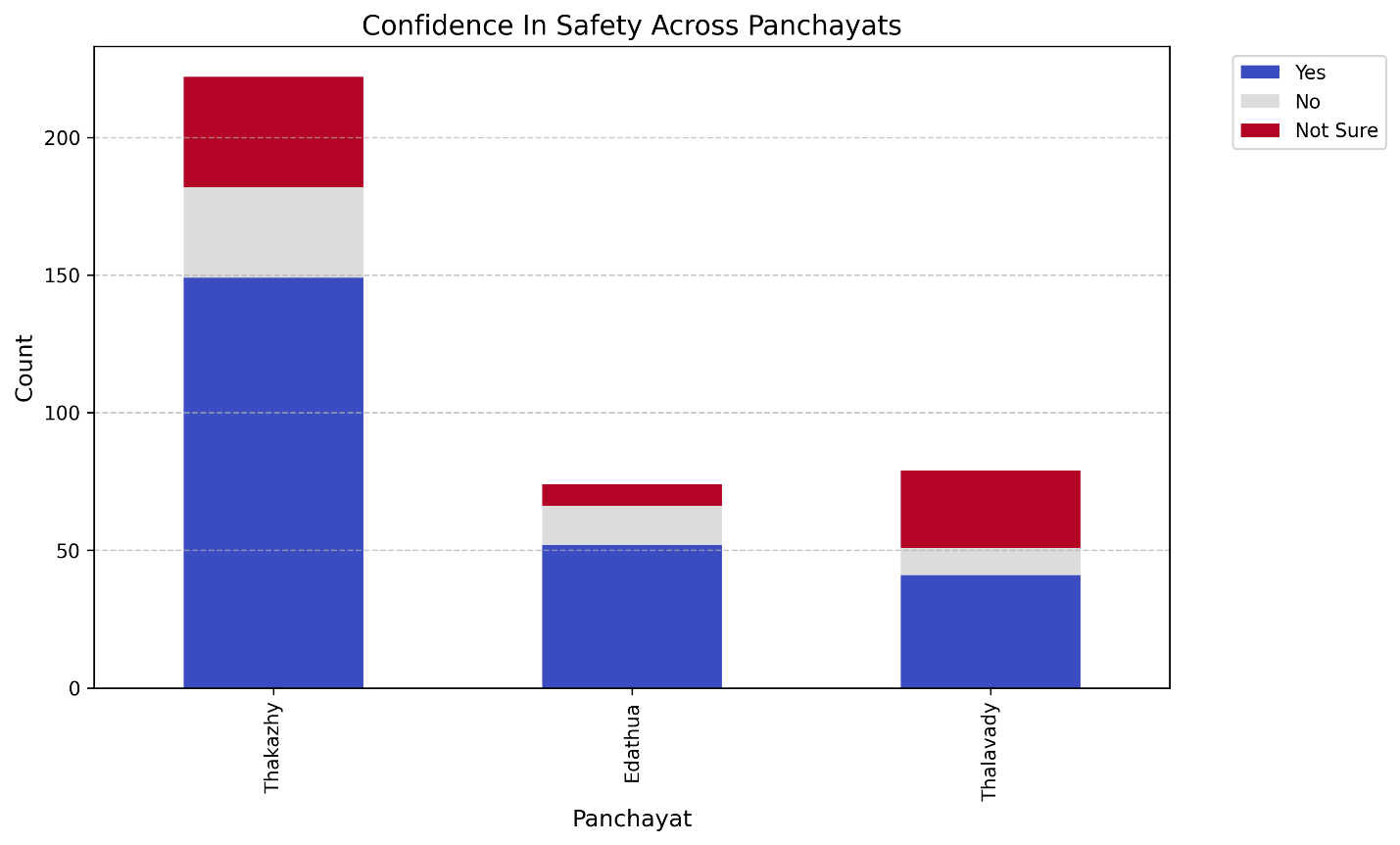
Fig1: Mode analysis

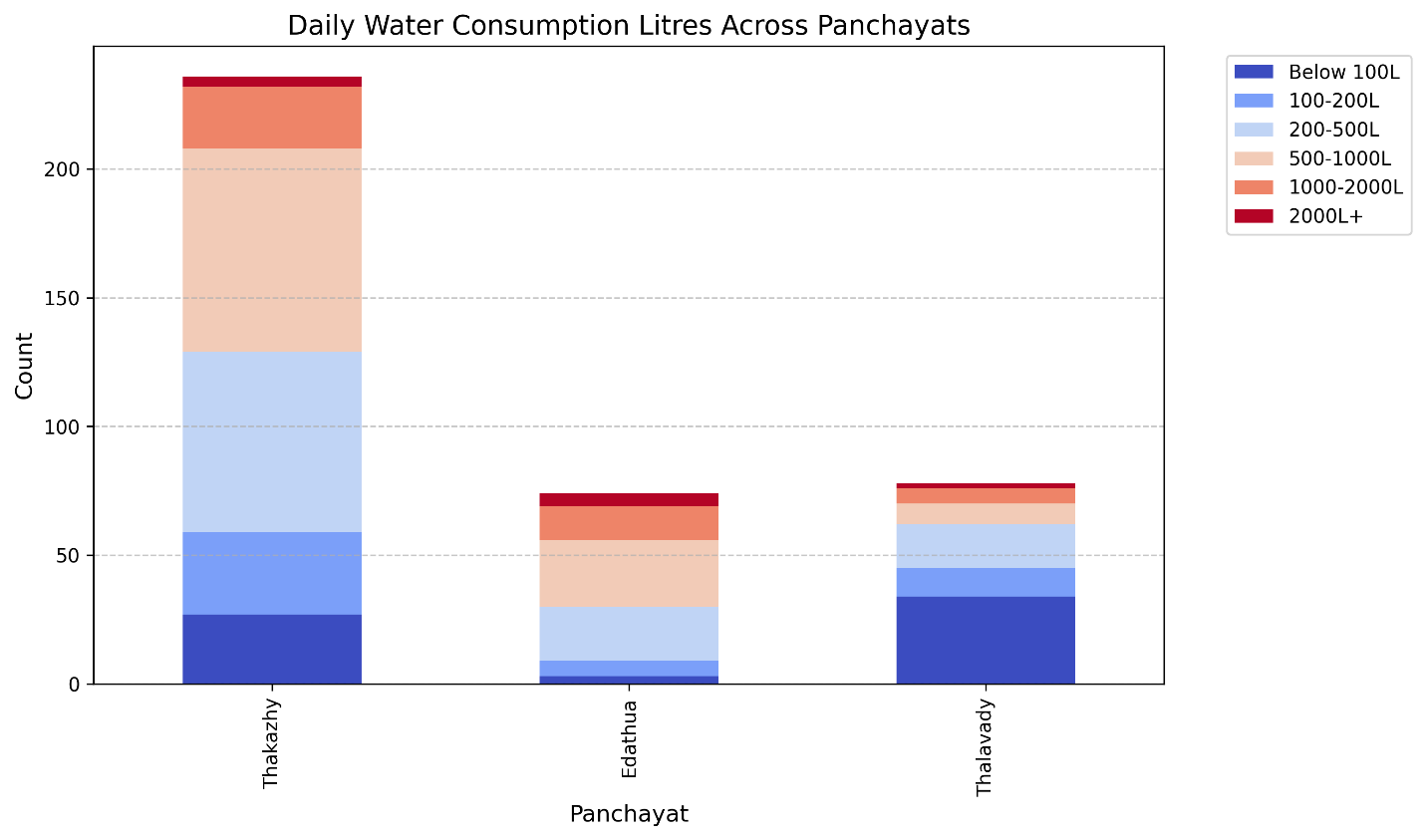
The Fig1 represents the mode analysis carried out on the data. That means they are the responses that are recorded in more times than others.

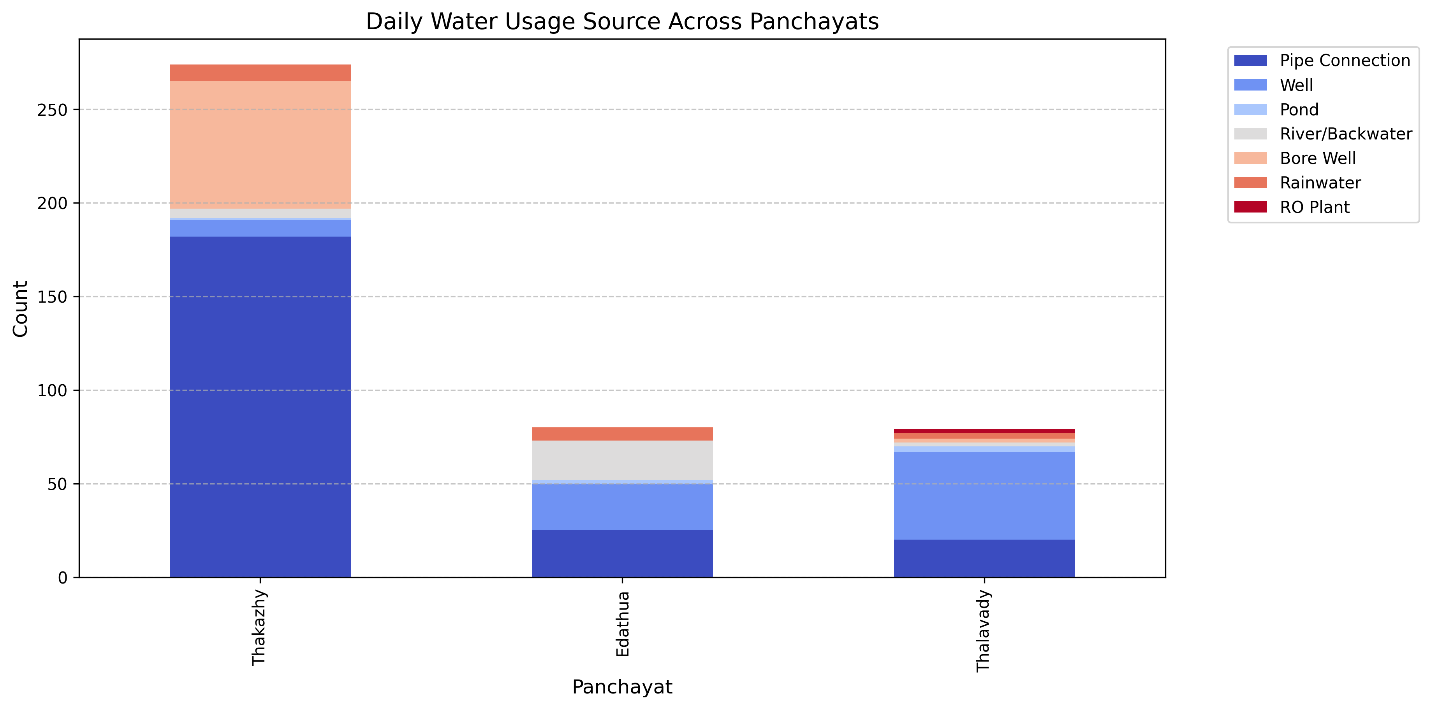
* + 1. **Frequency (count) distribution**

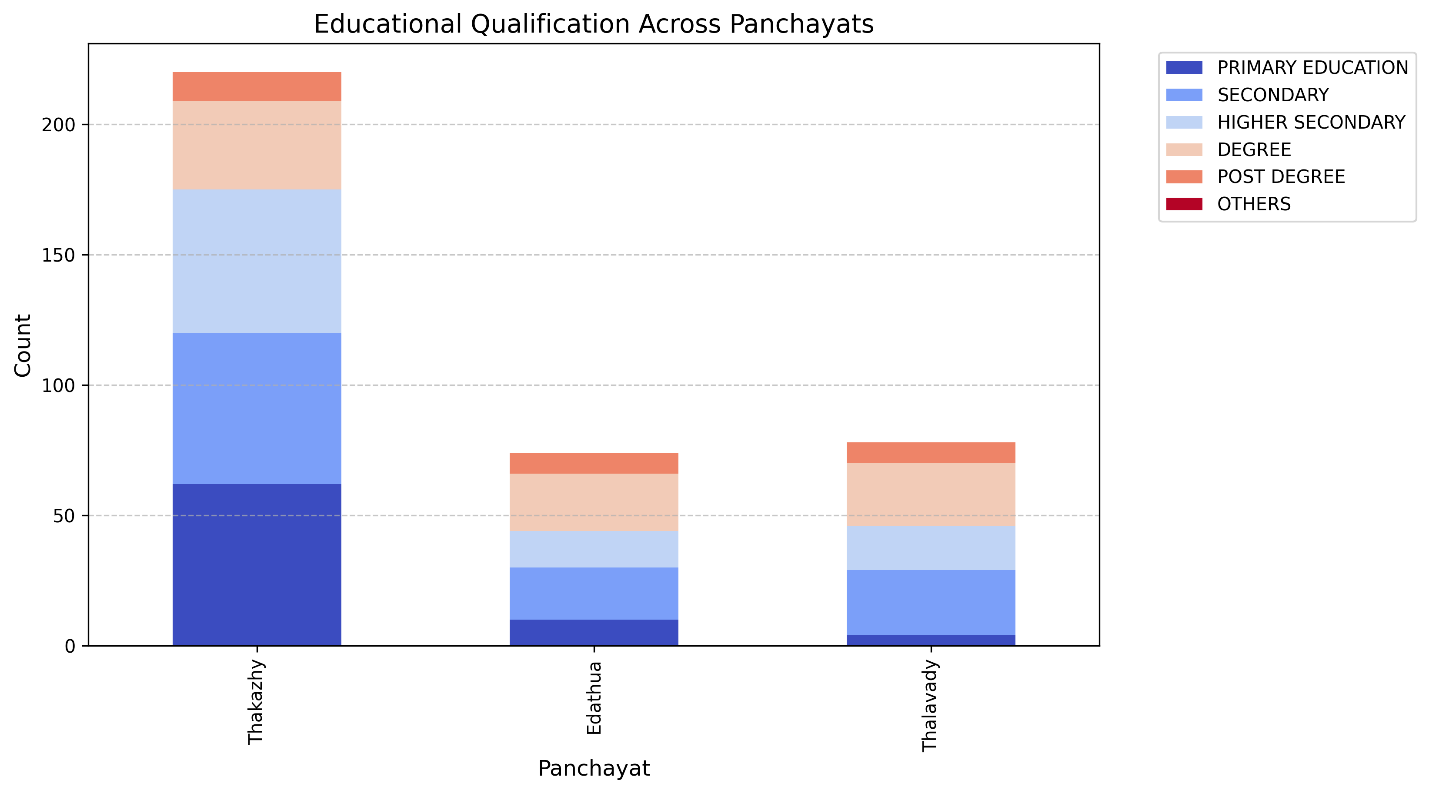
****

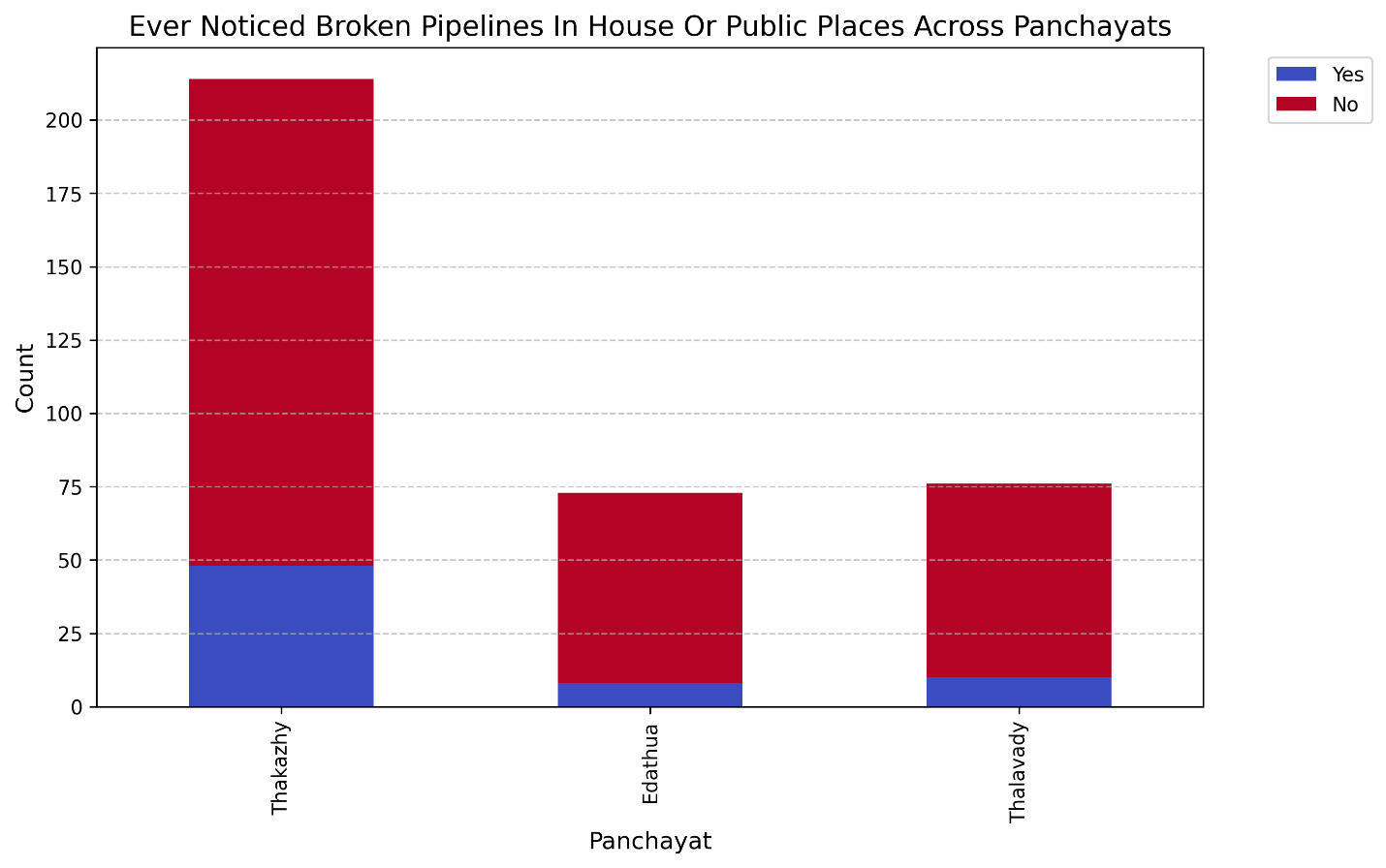
****

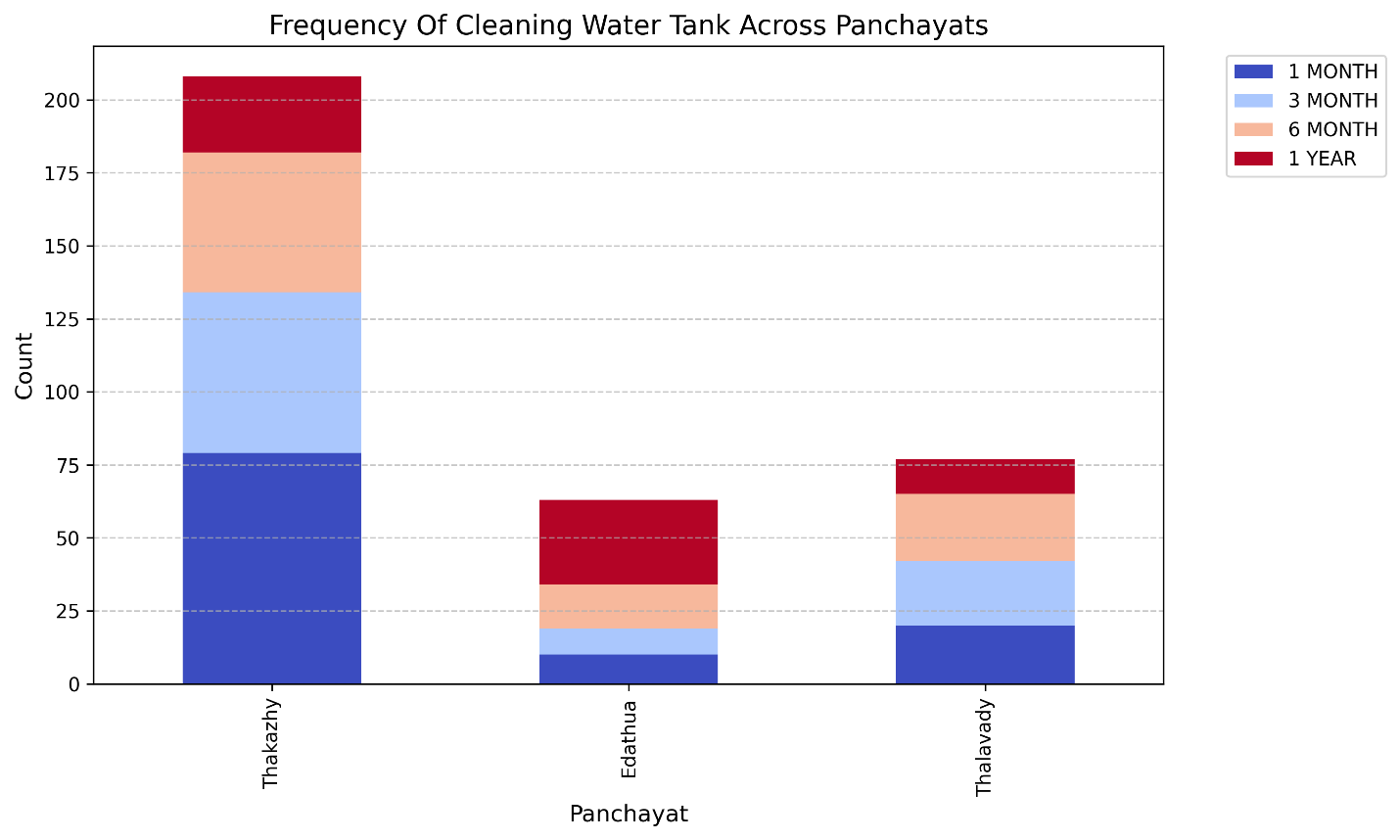
****

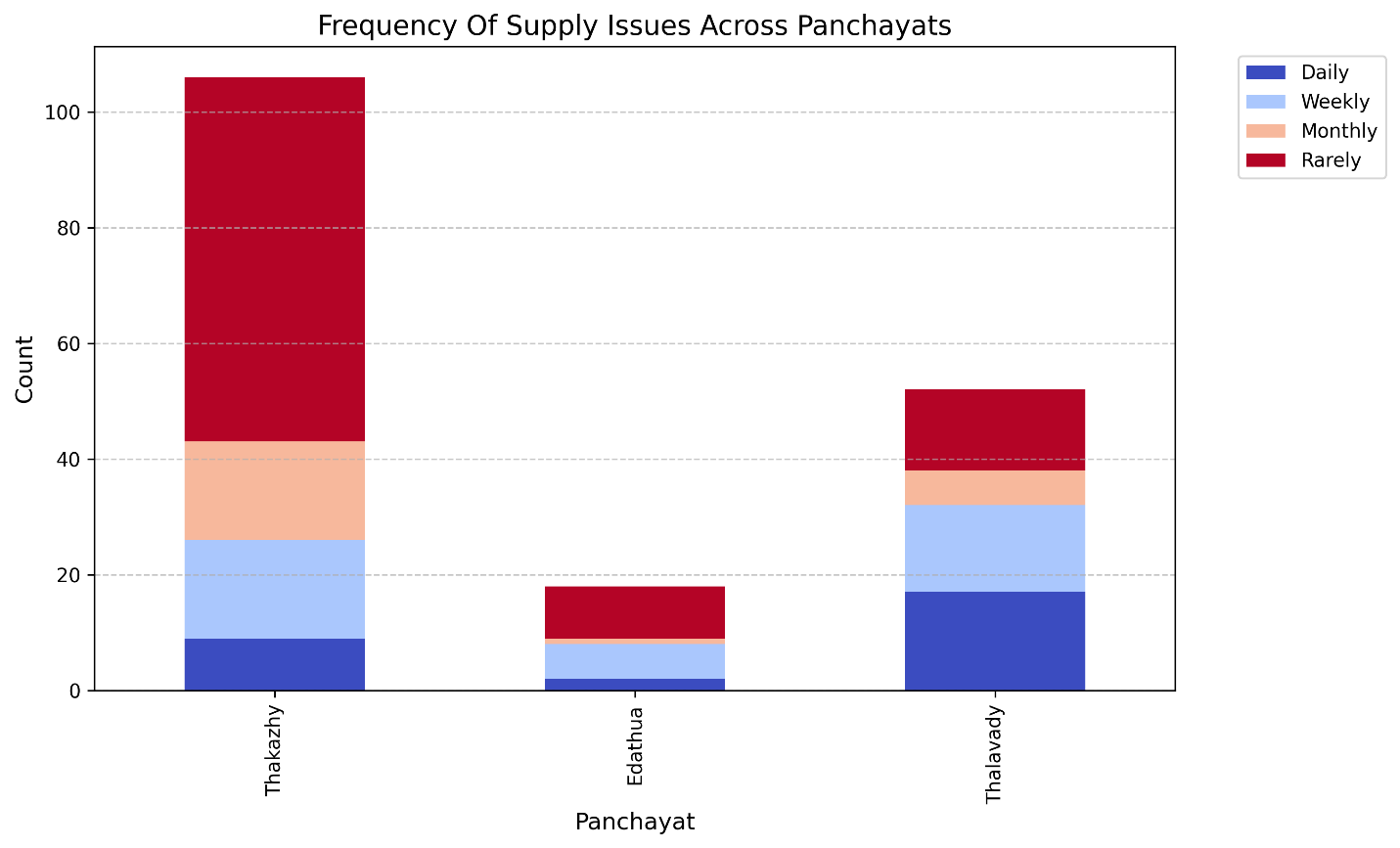
****

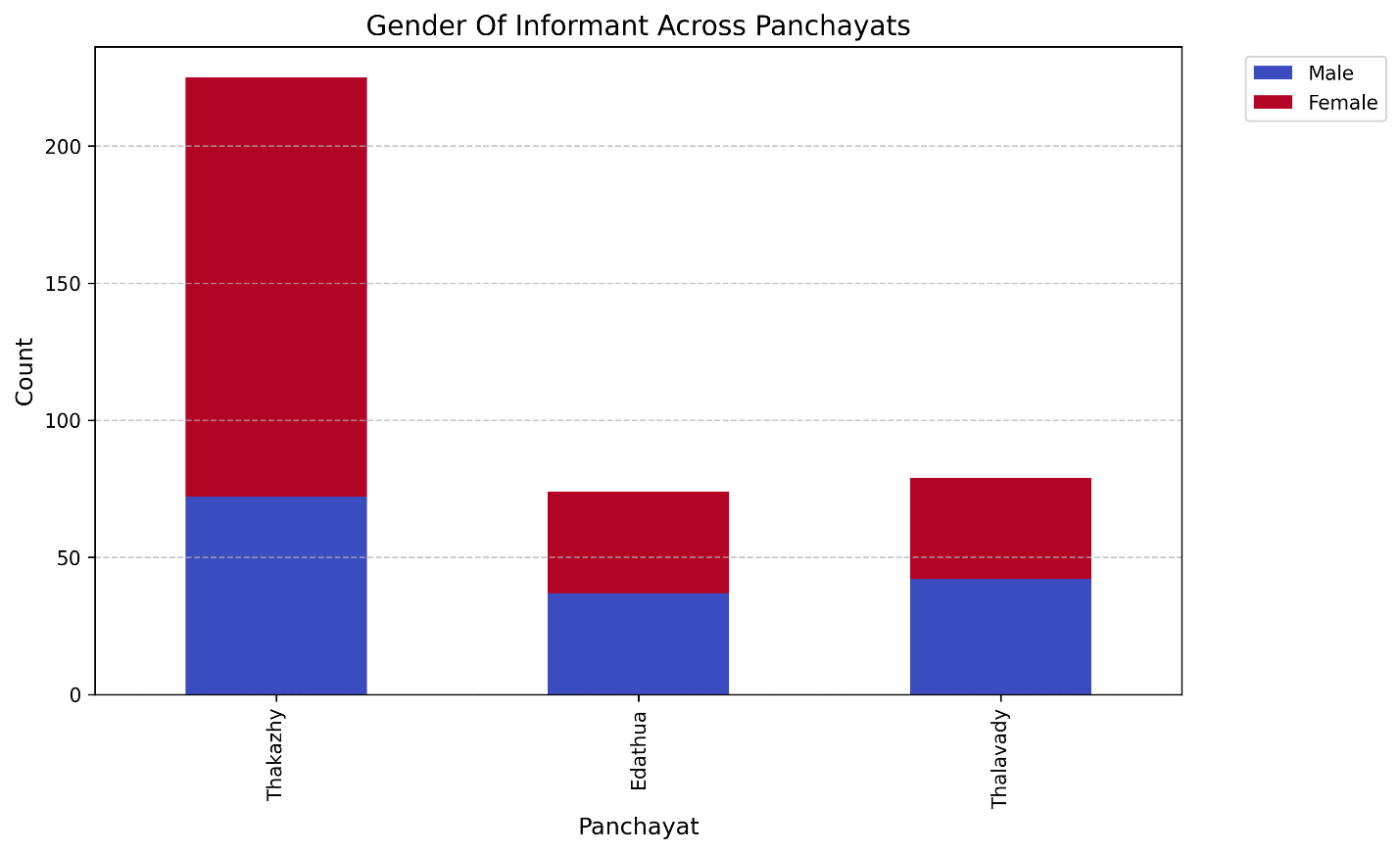
****

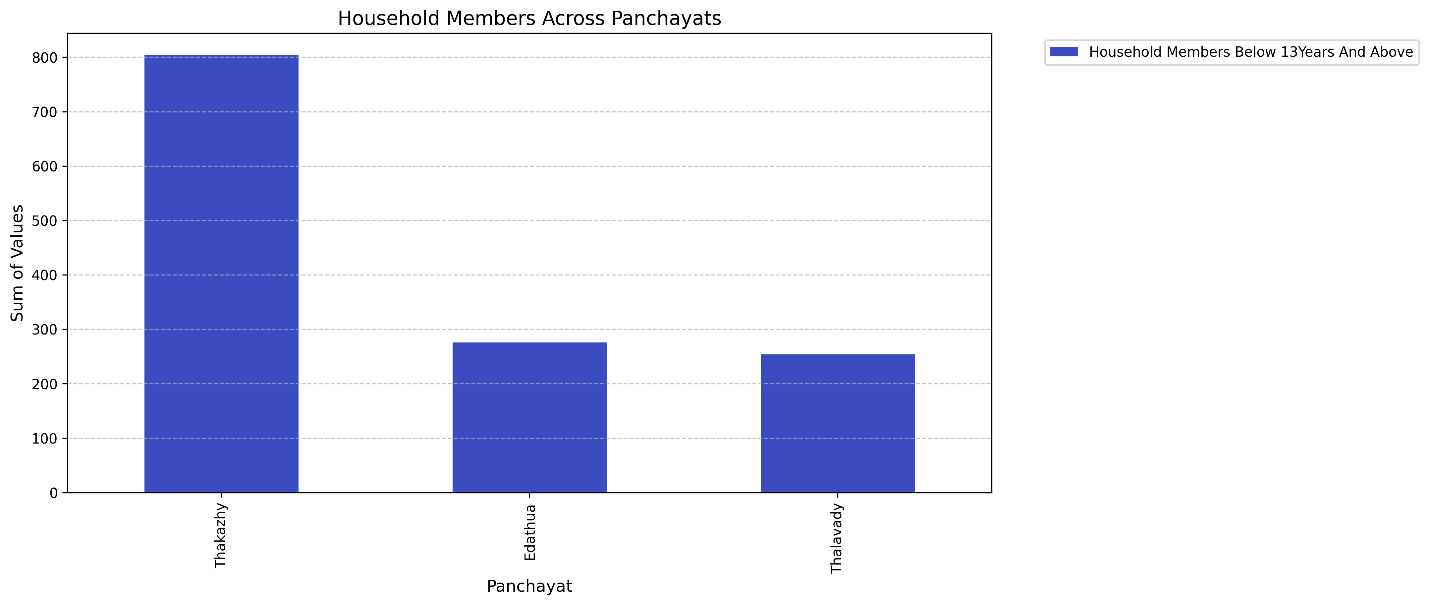
****

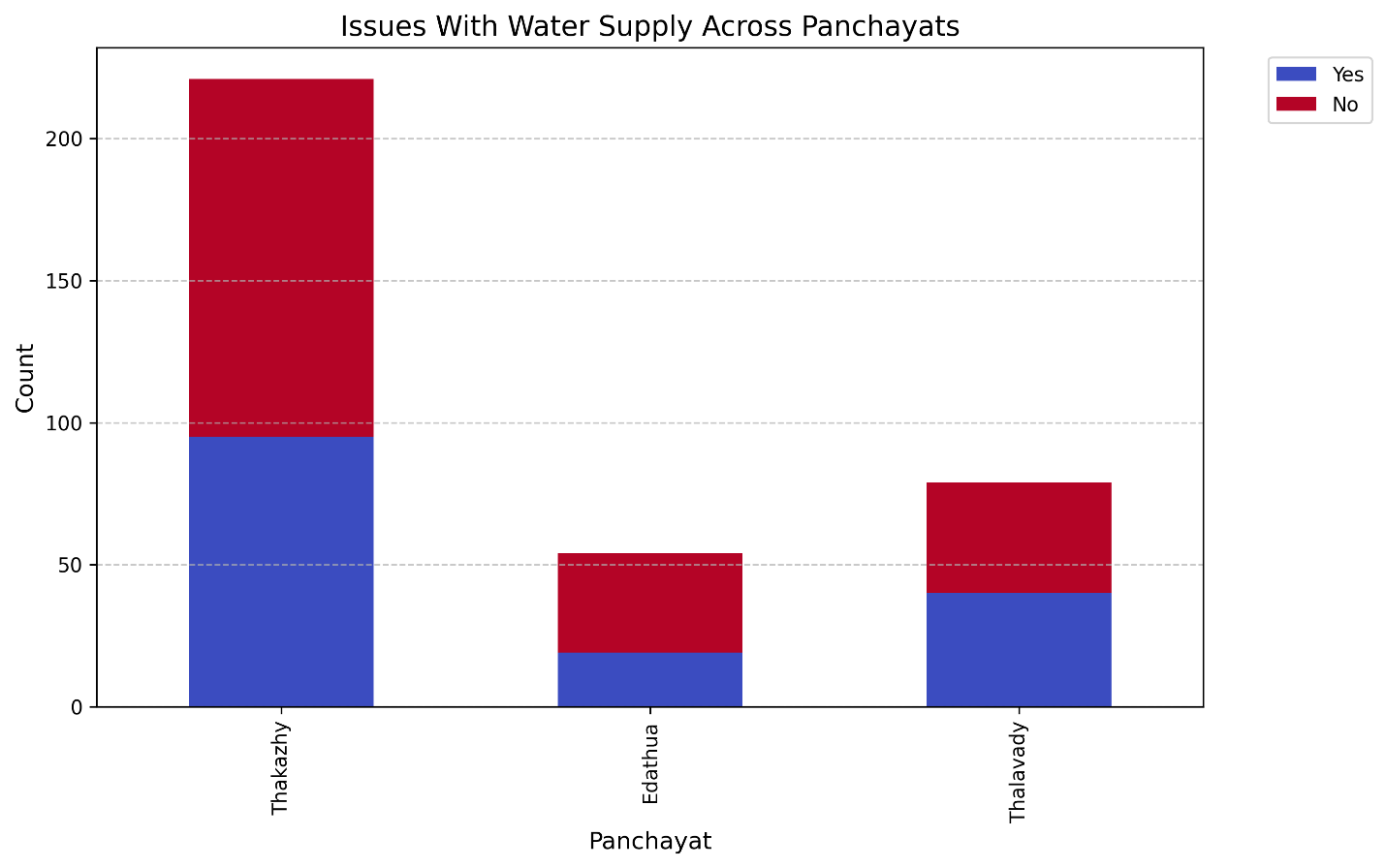
****

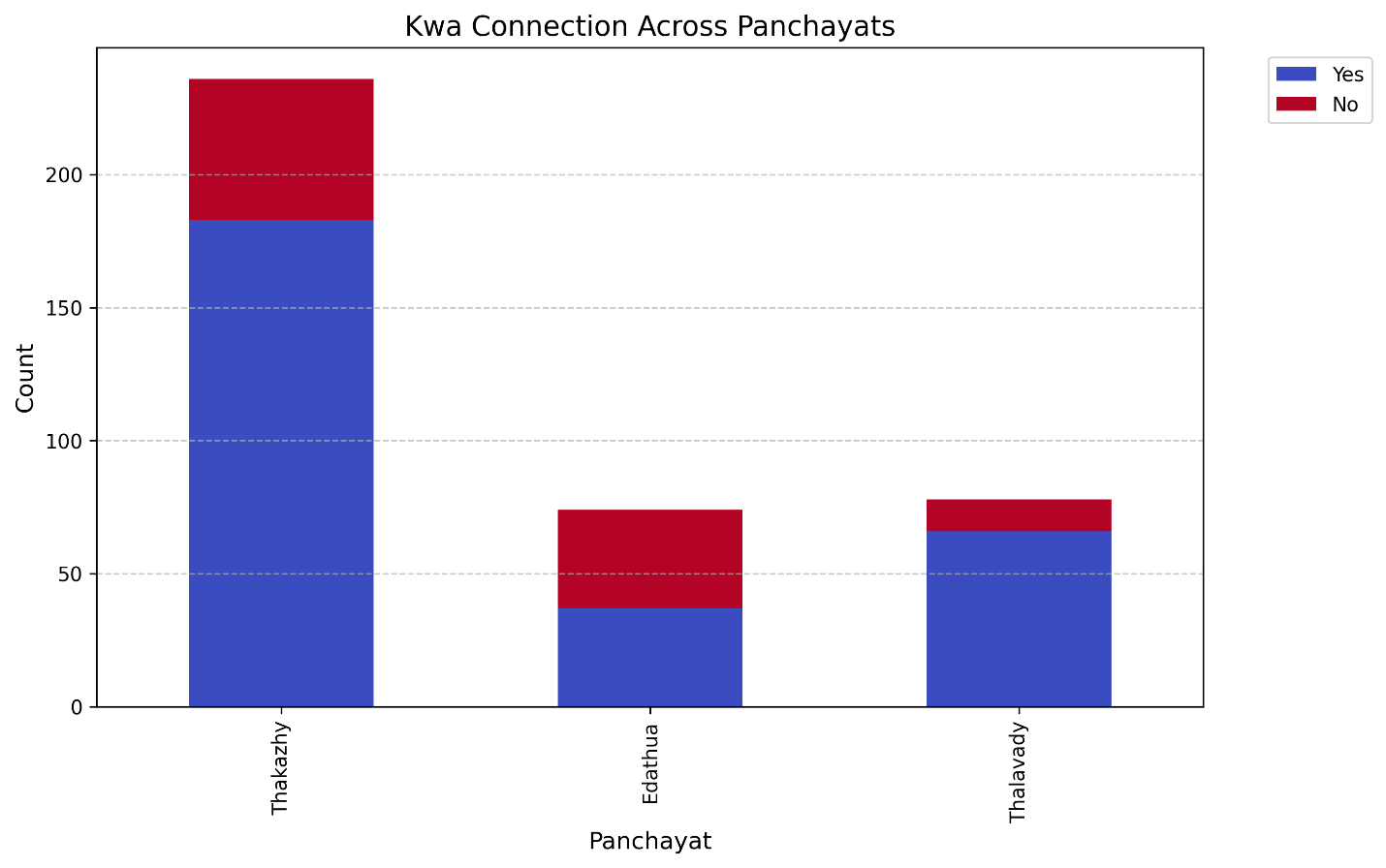
****

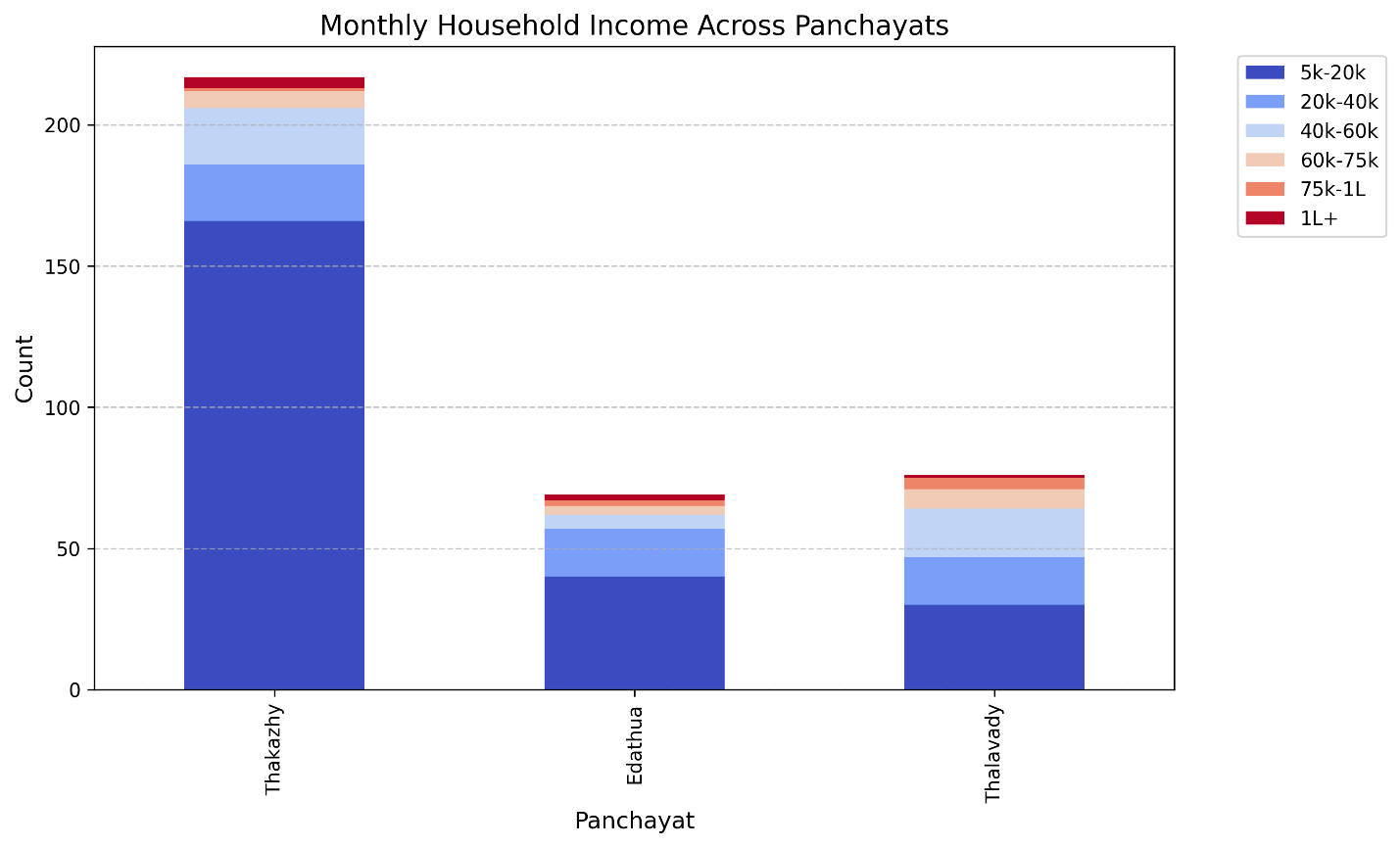
****

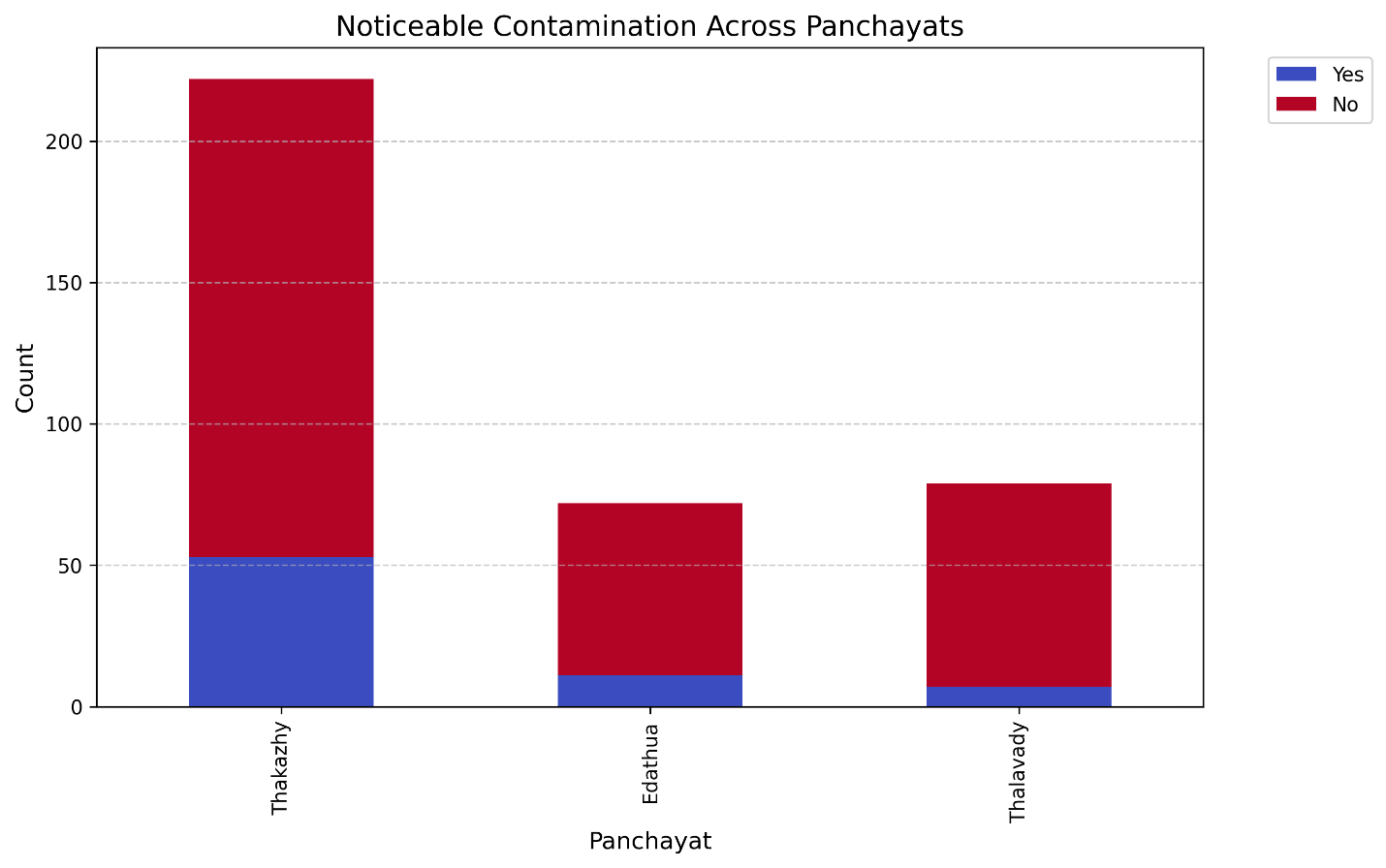
****

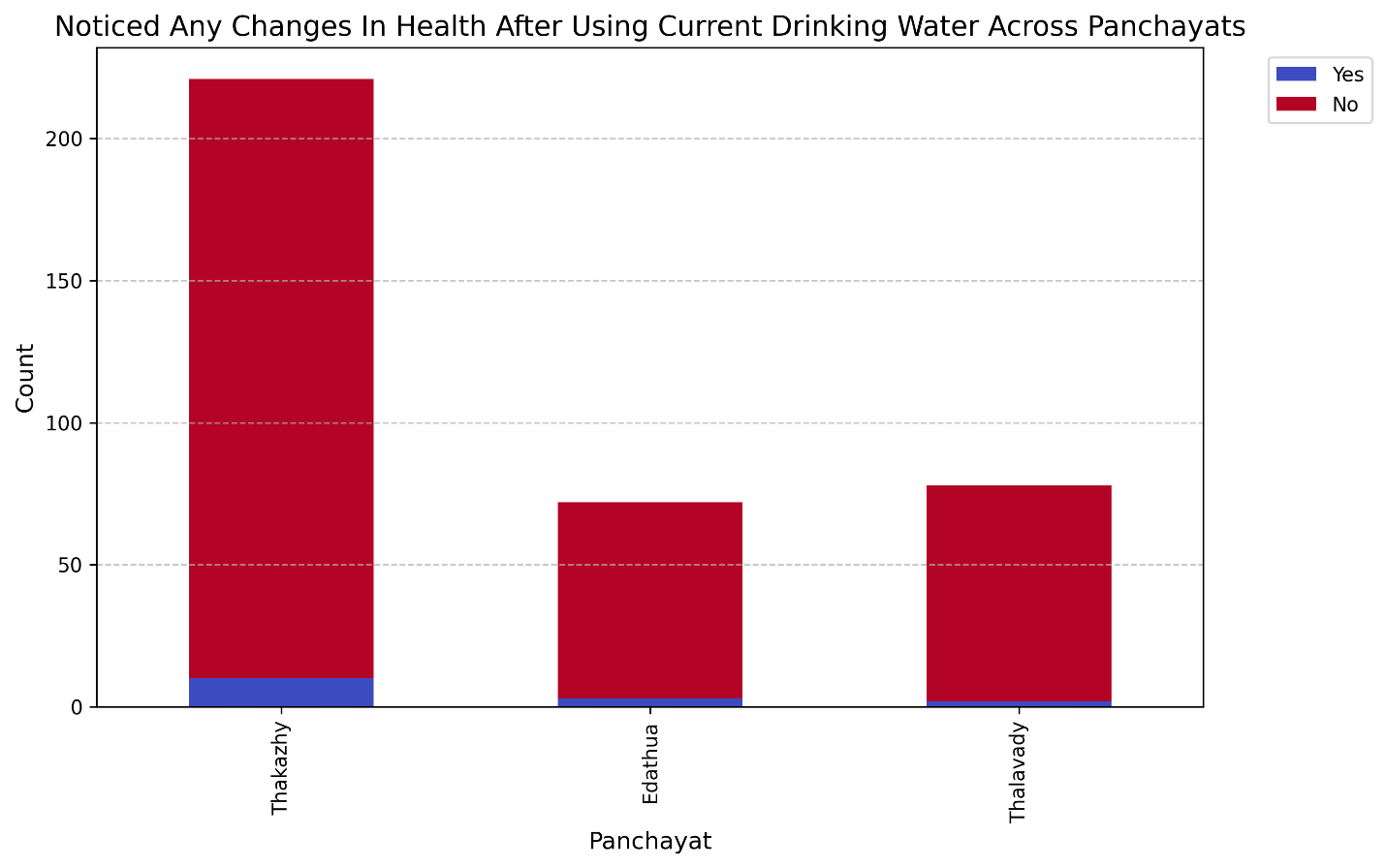
****

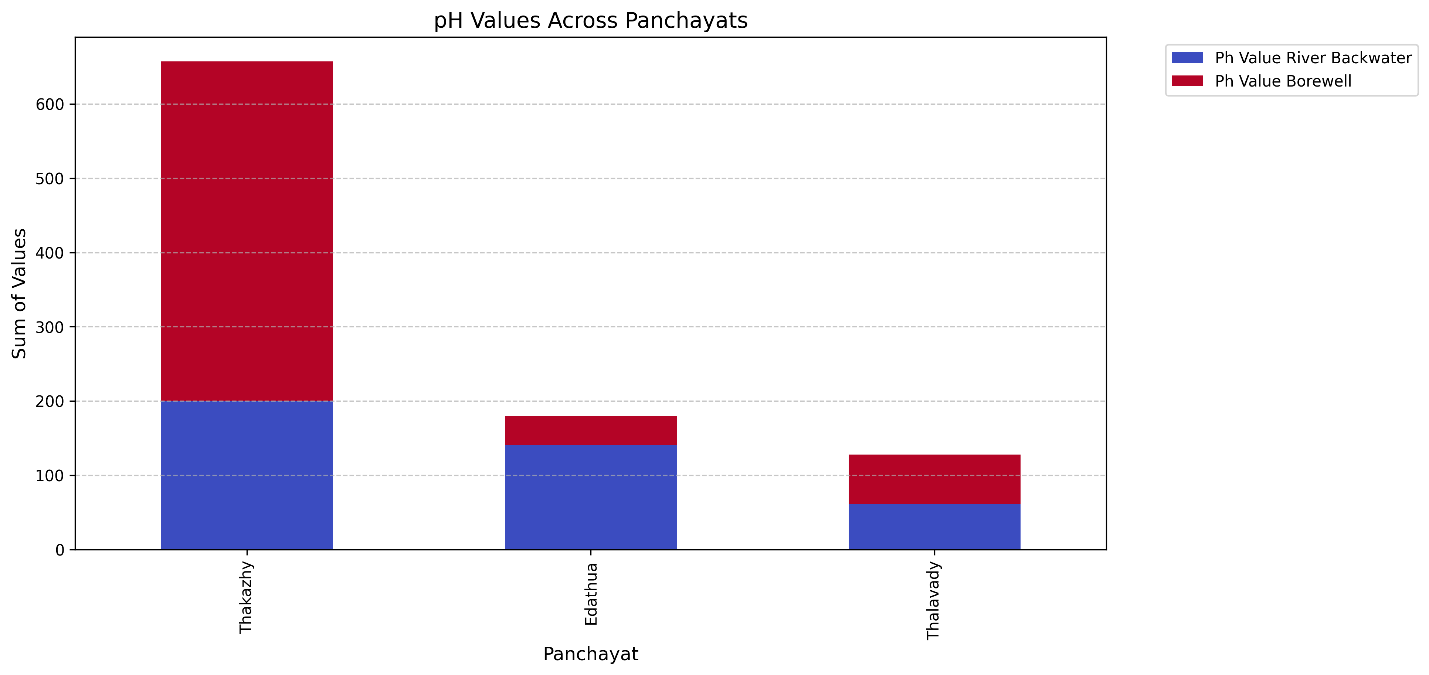
****

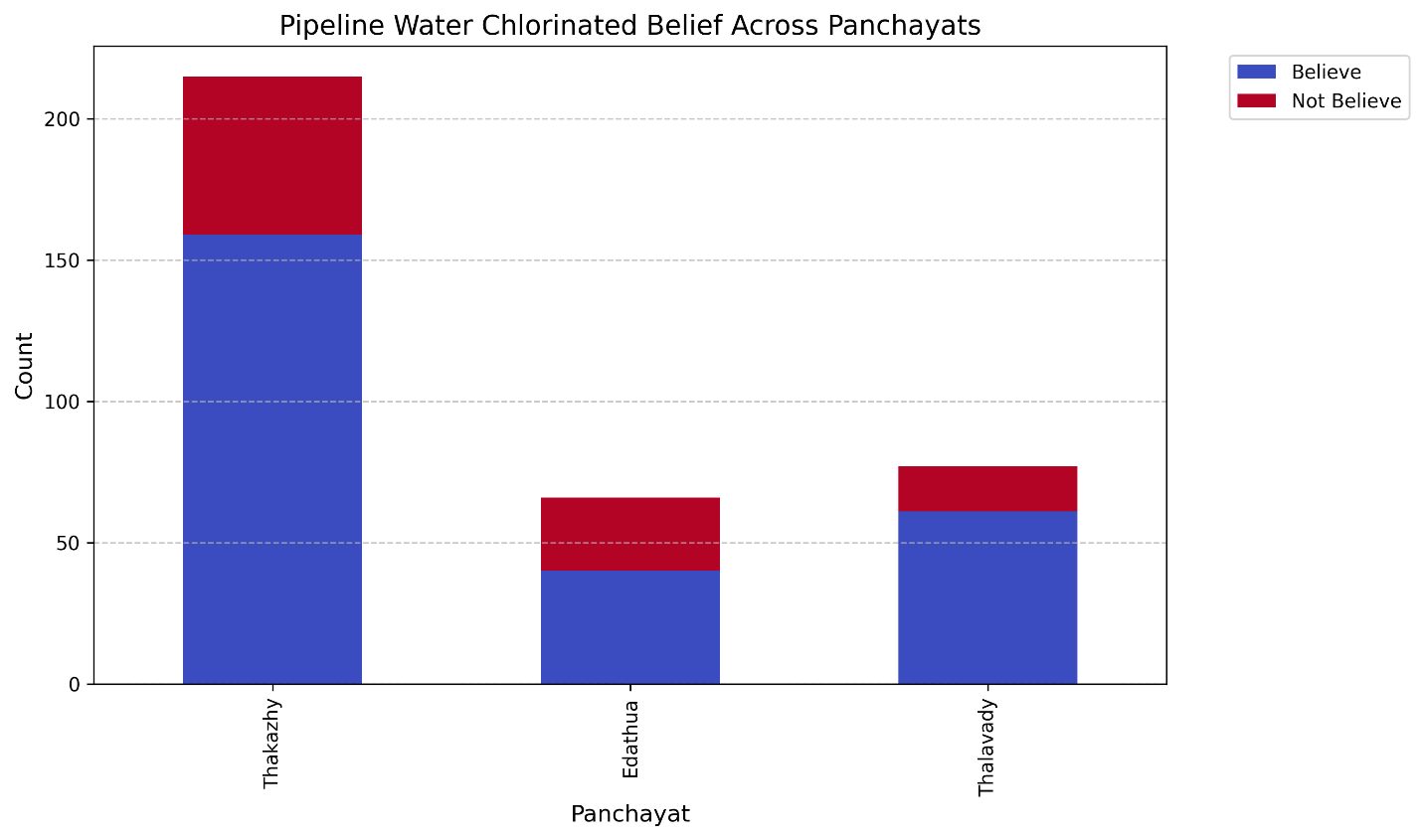
****

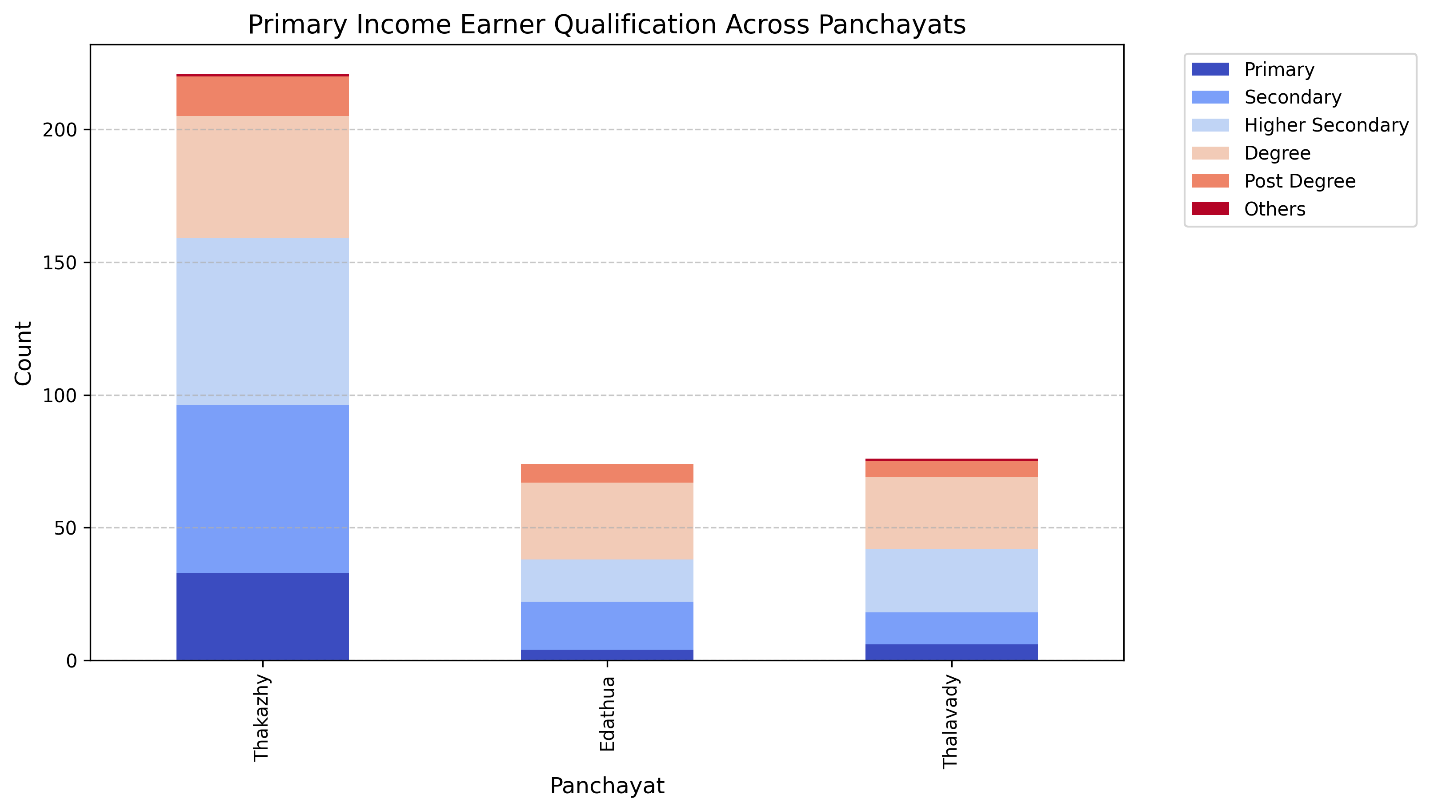
****

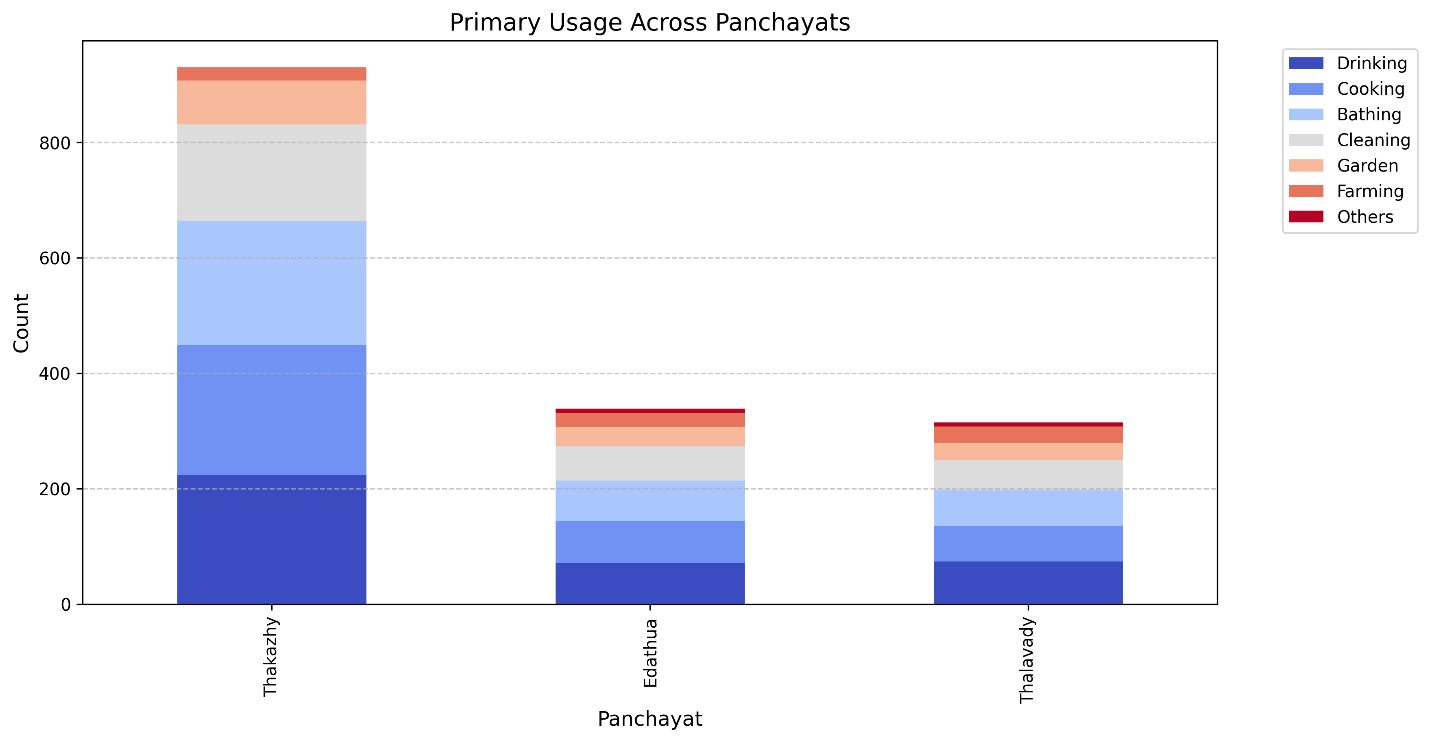
****

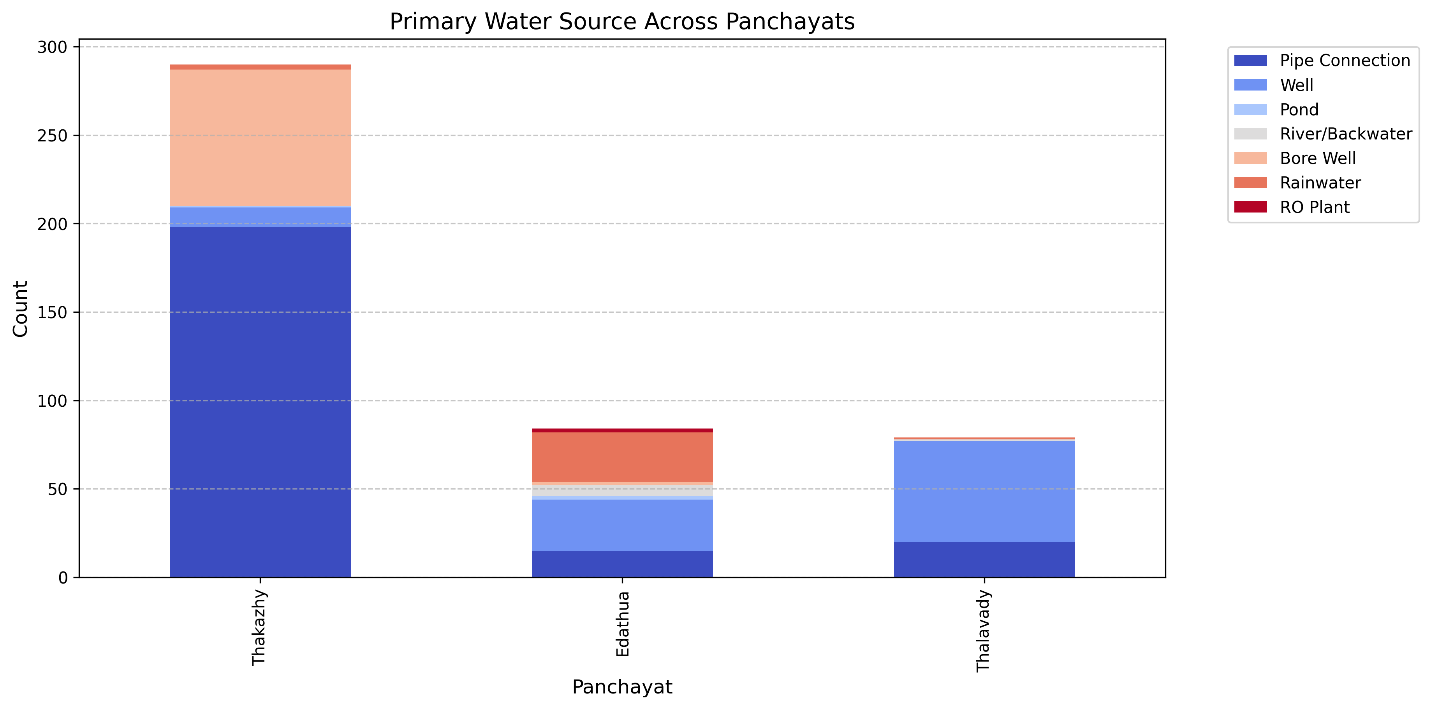
****

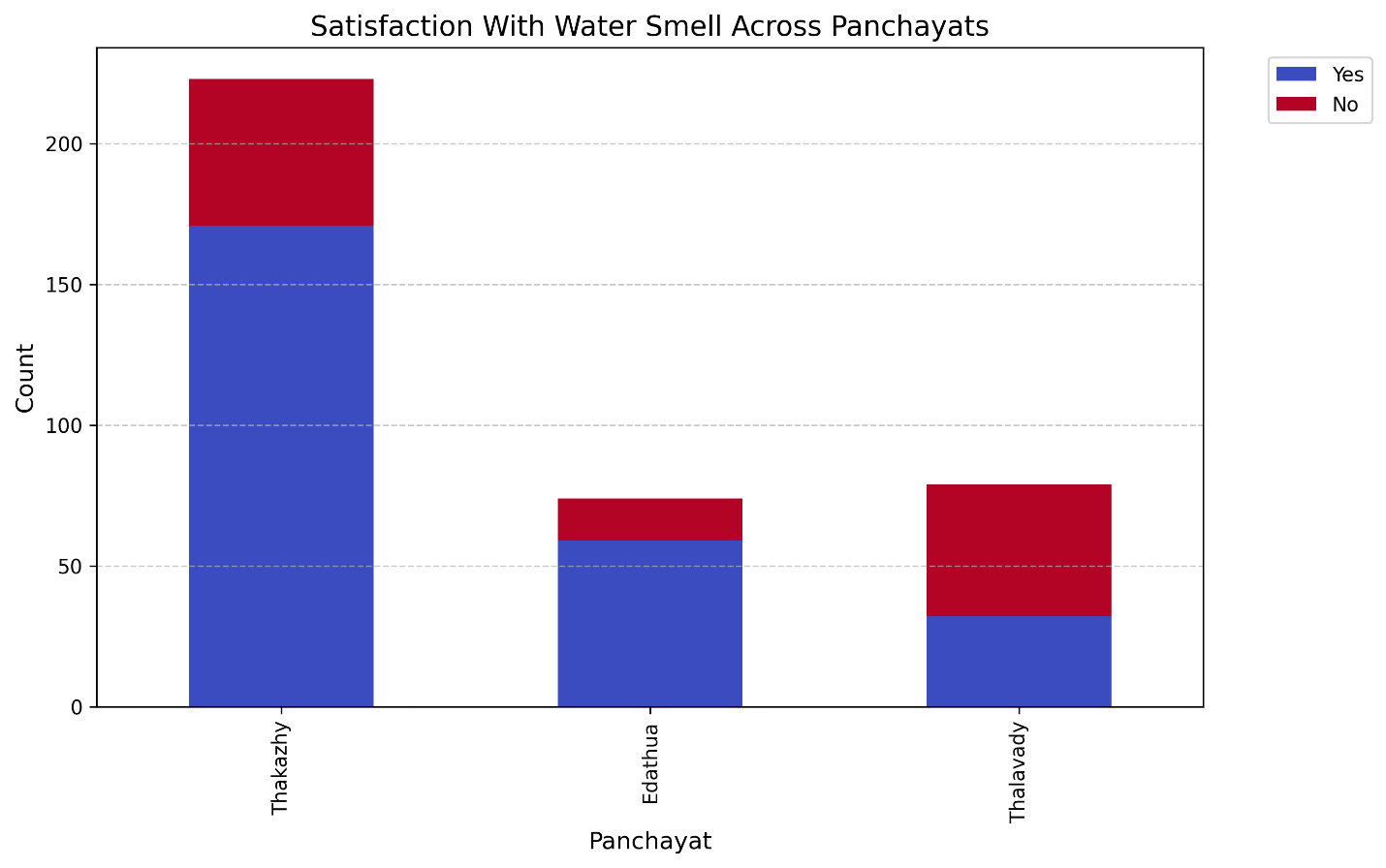
****

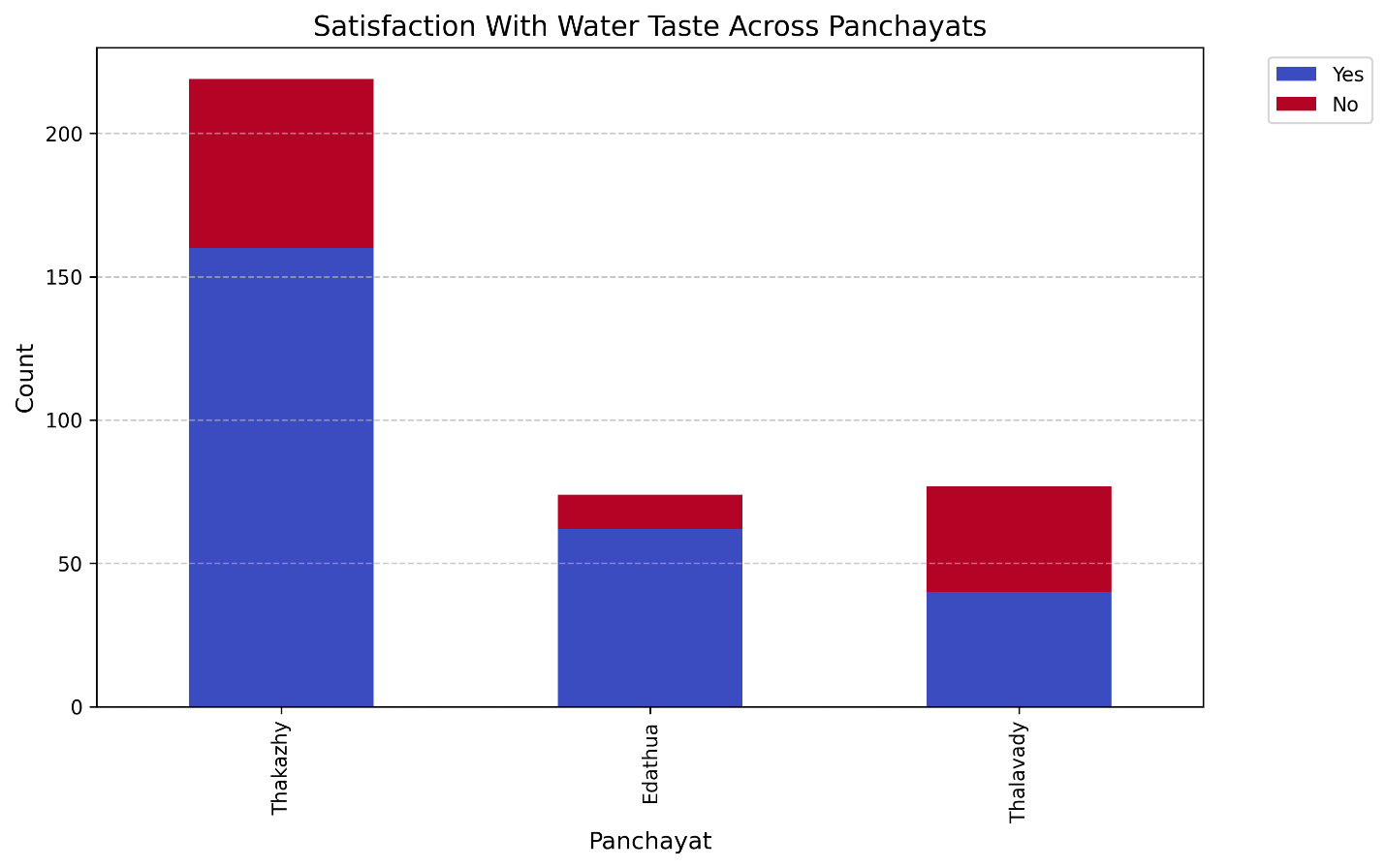
****

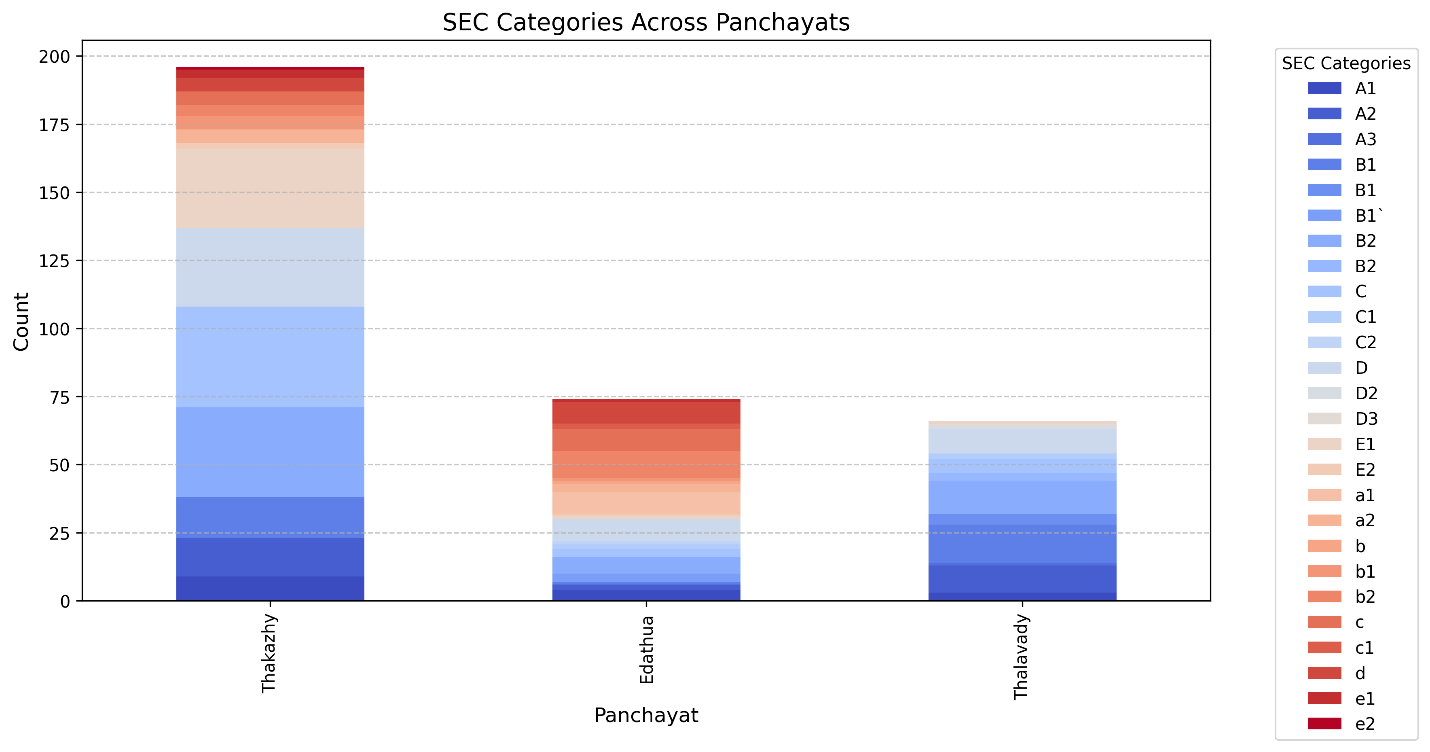
****

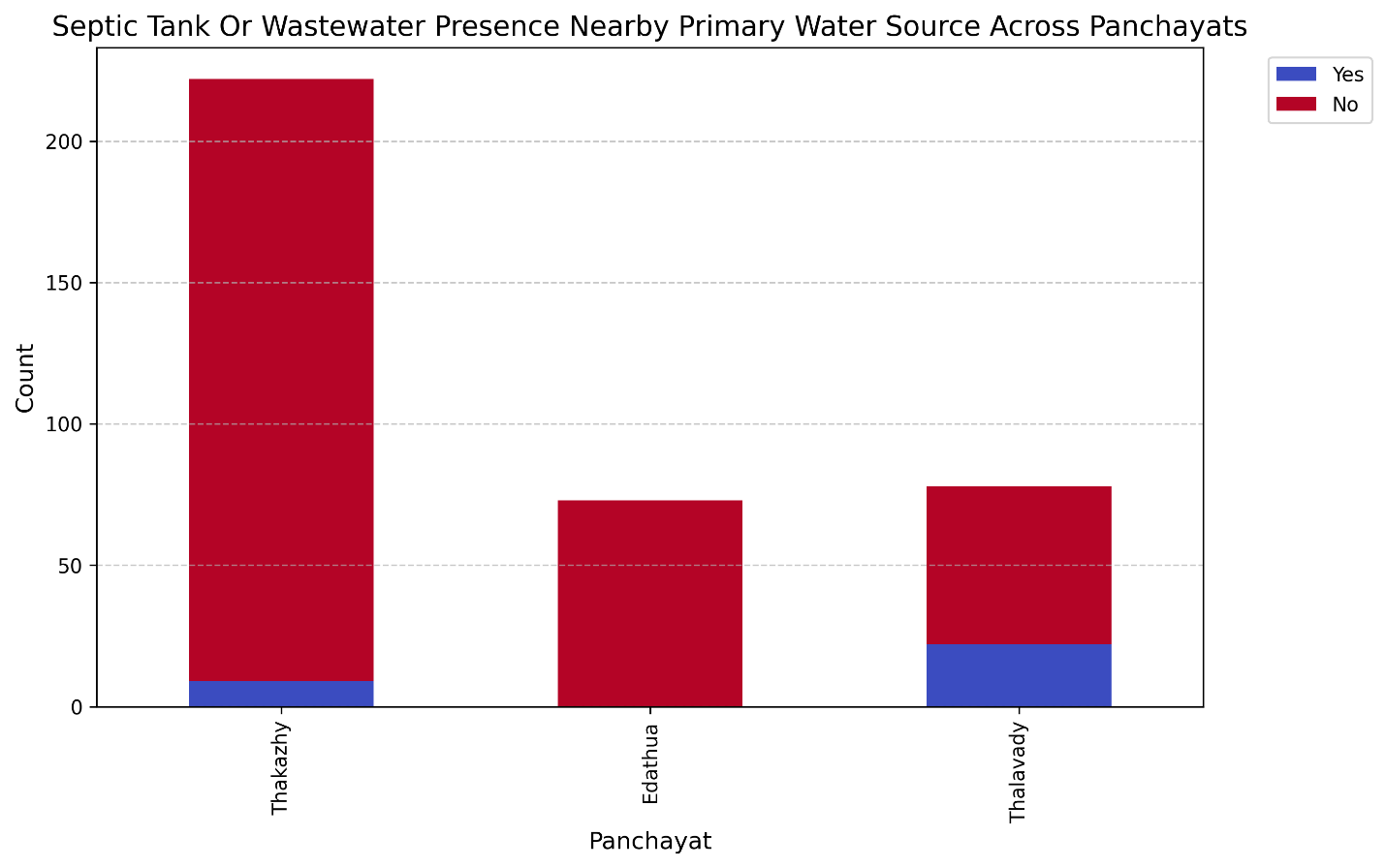
****

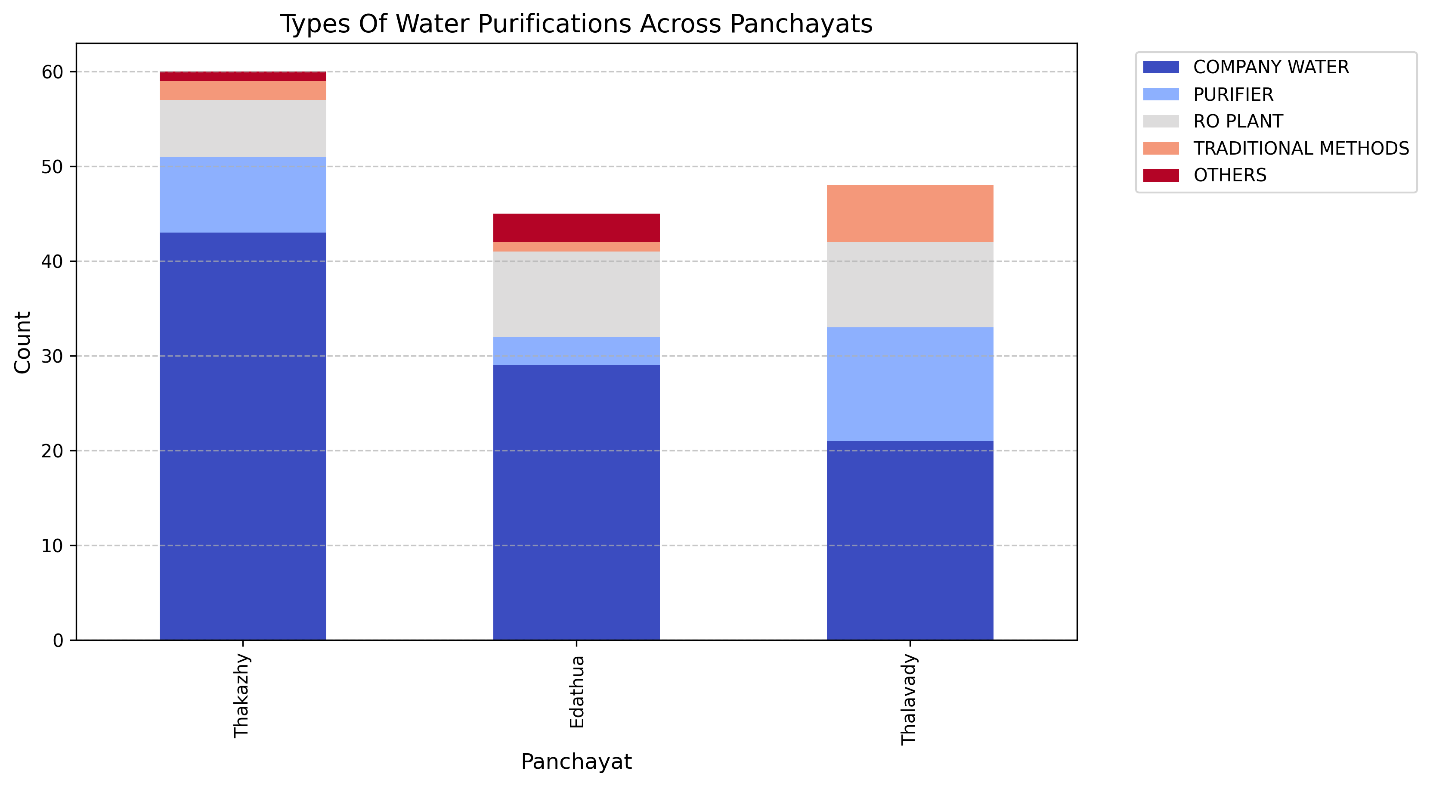
****

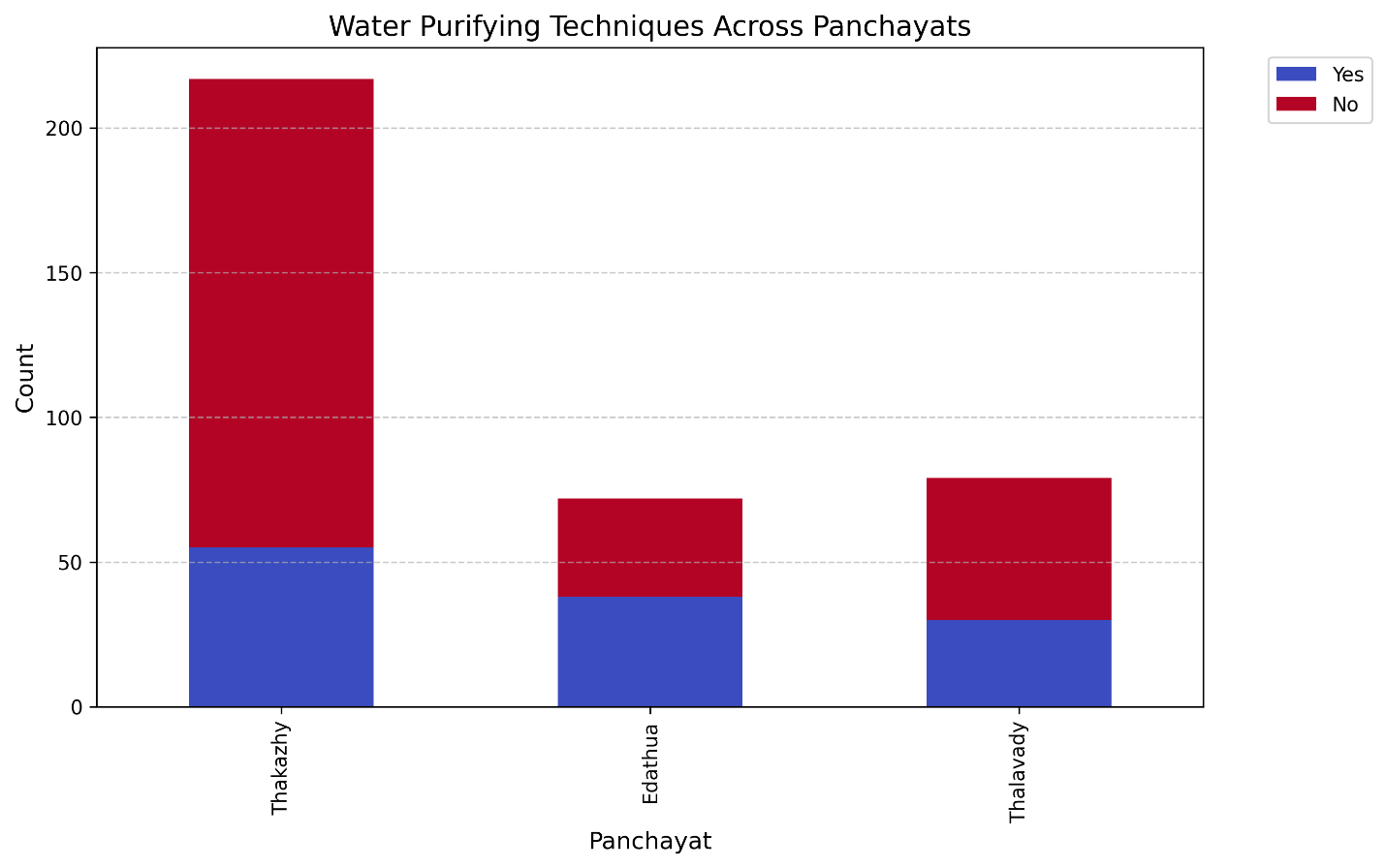
****

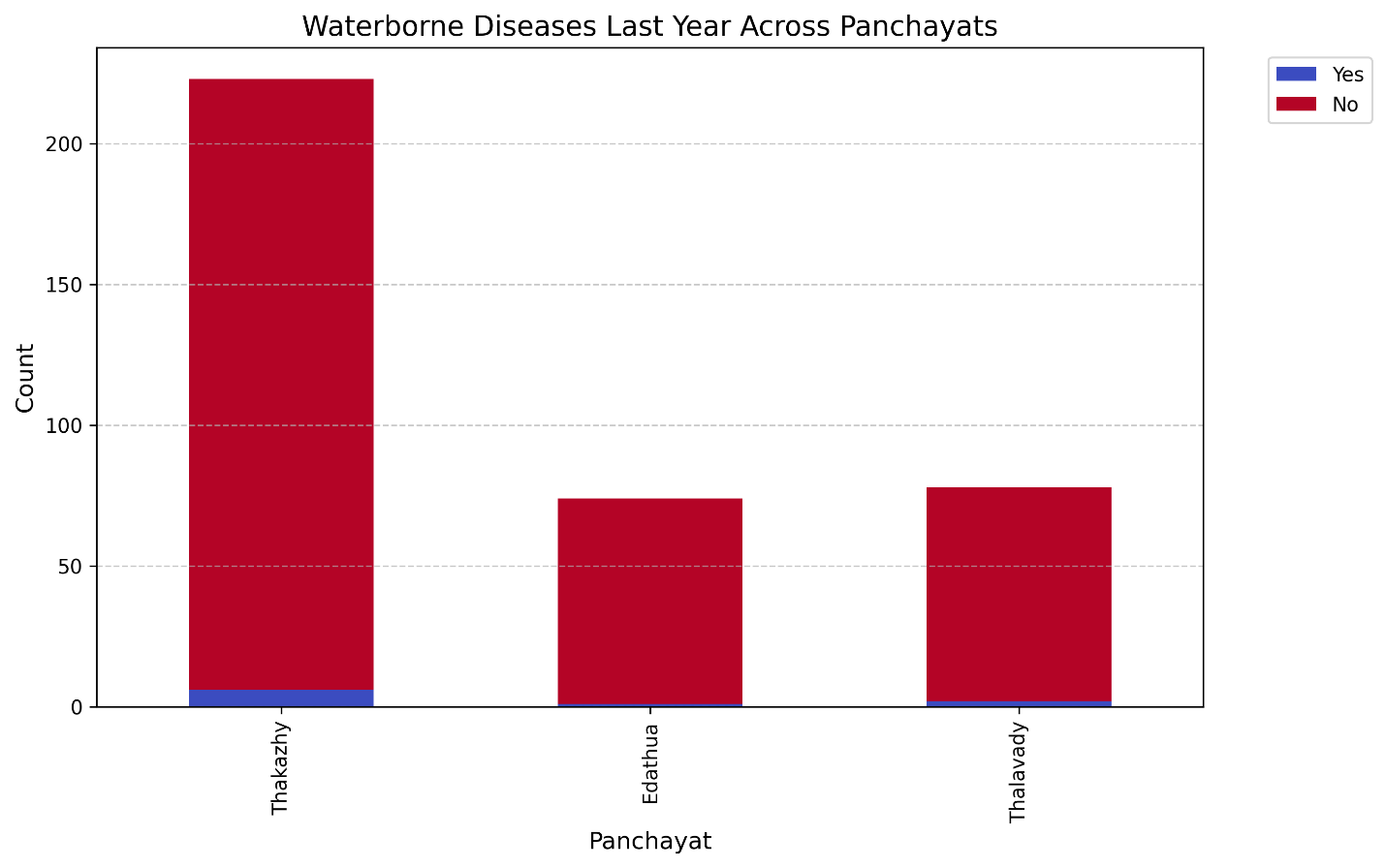
****

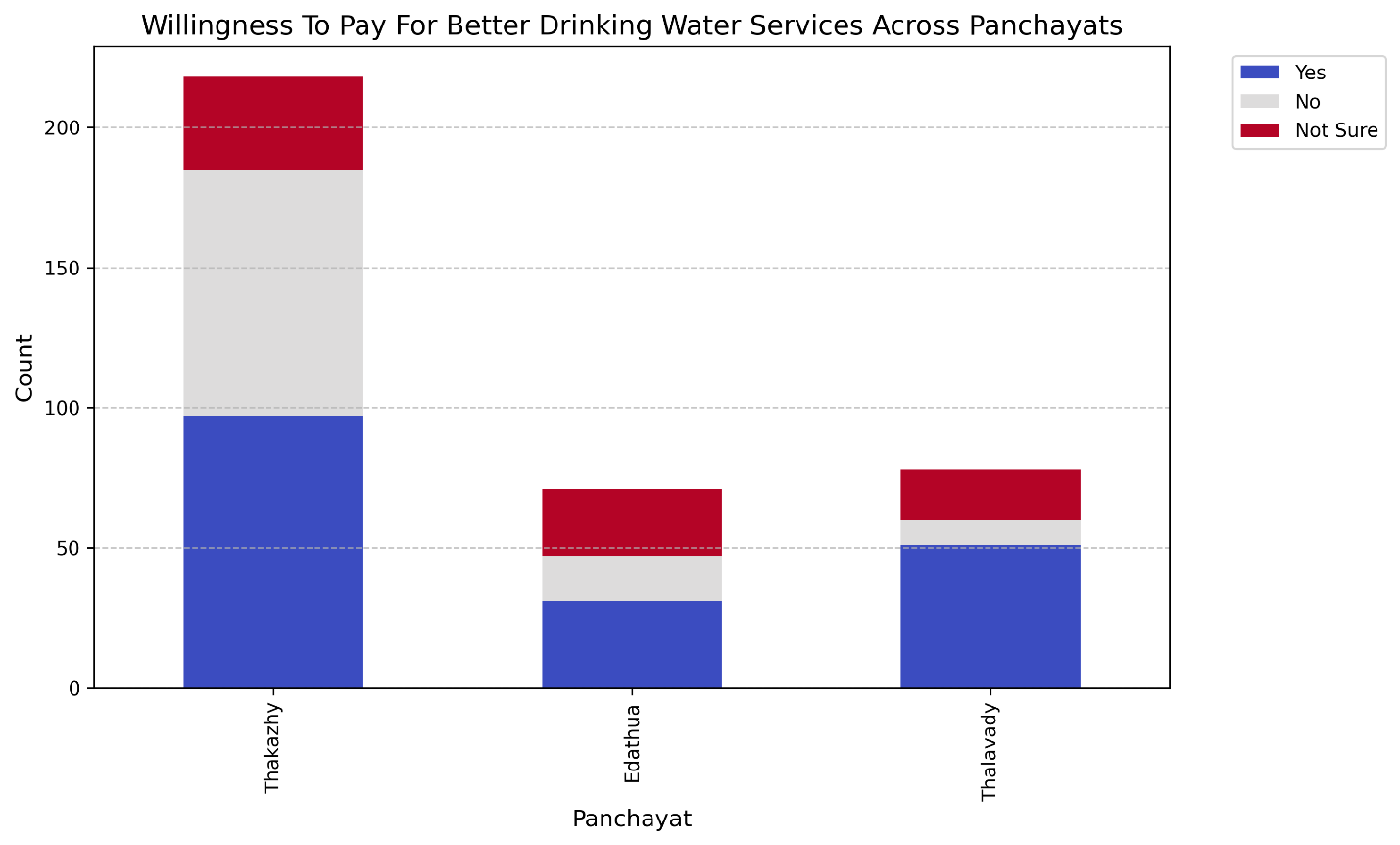
****

****

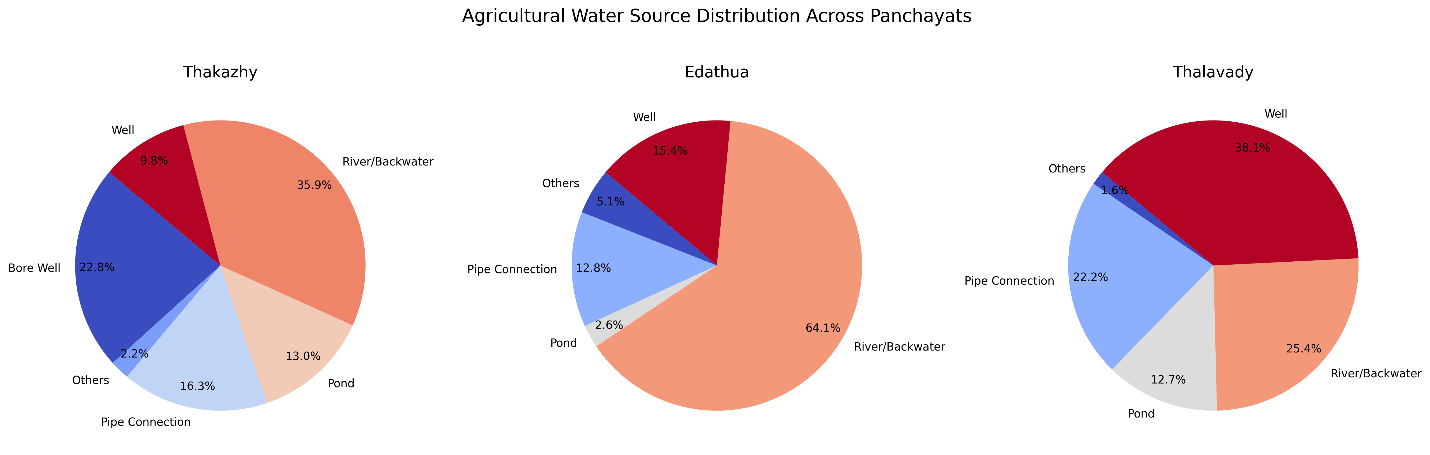
****

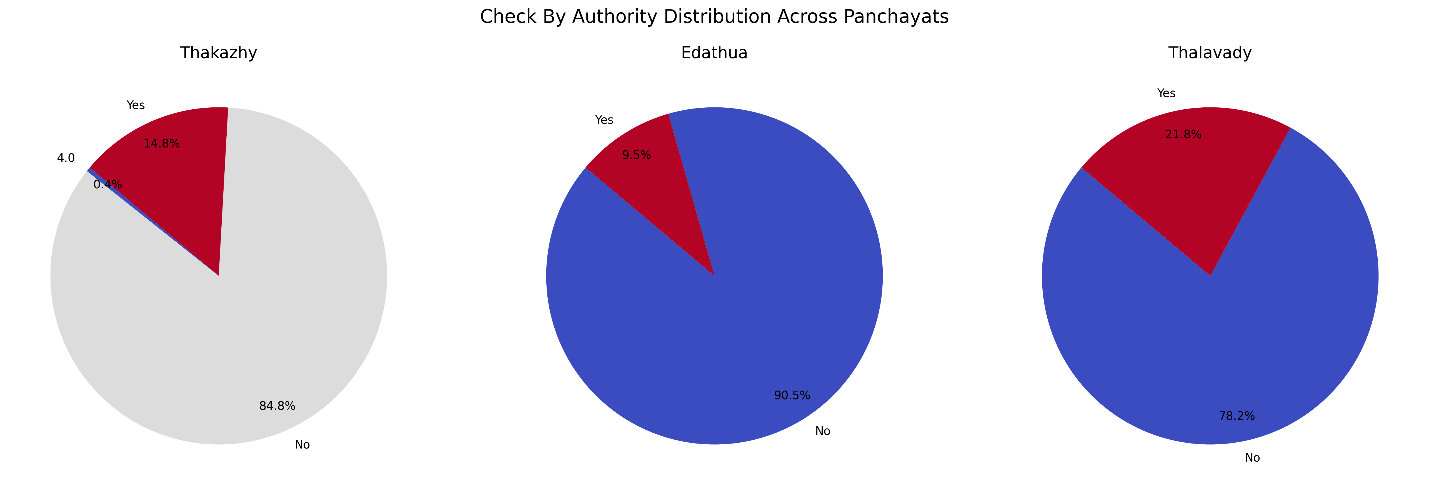
****

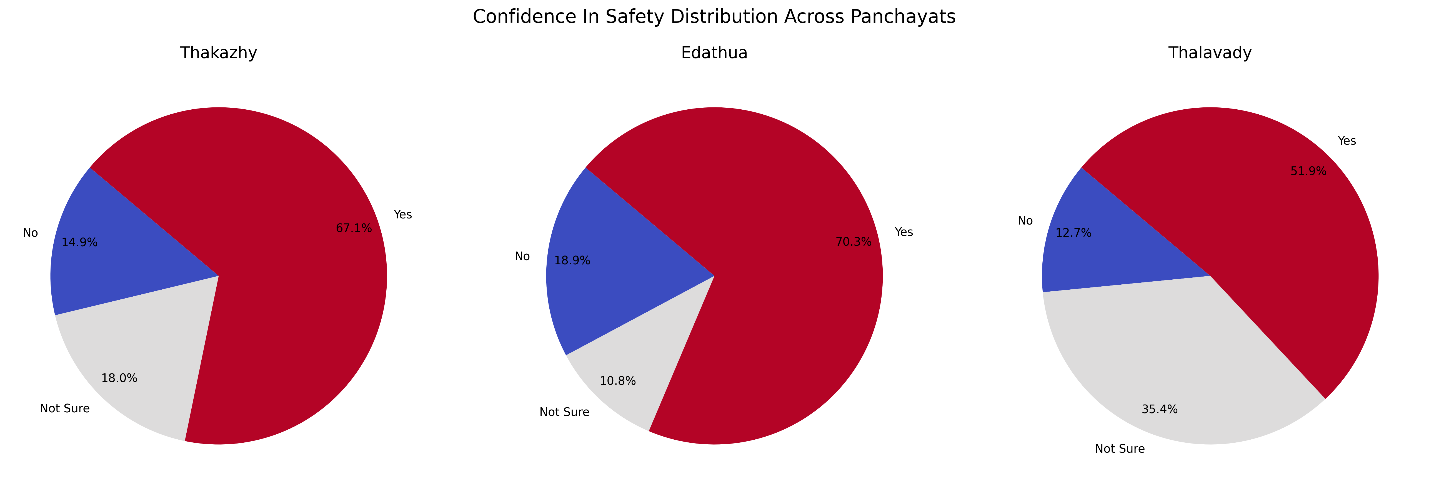
****

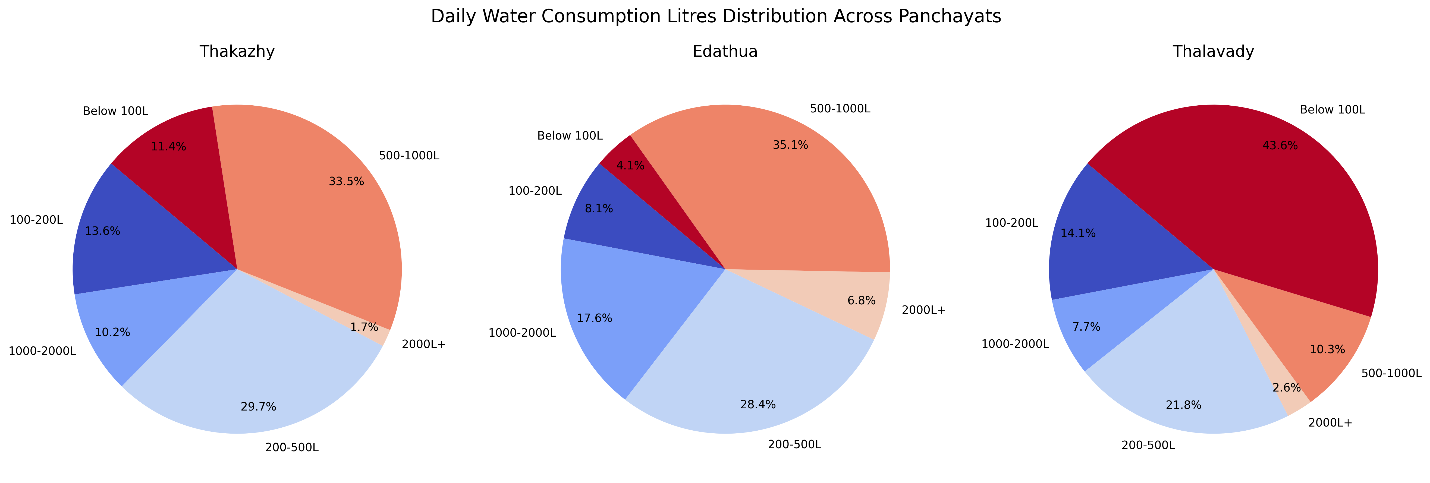
****

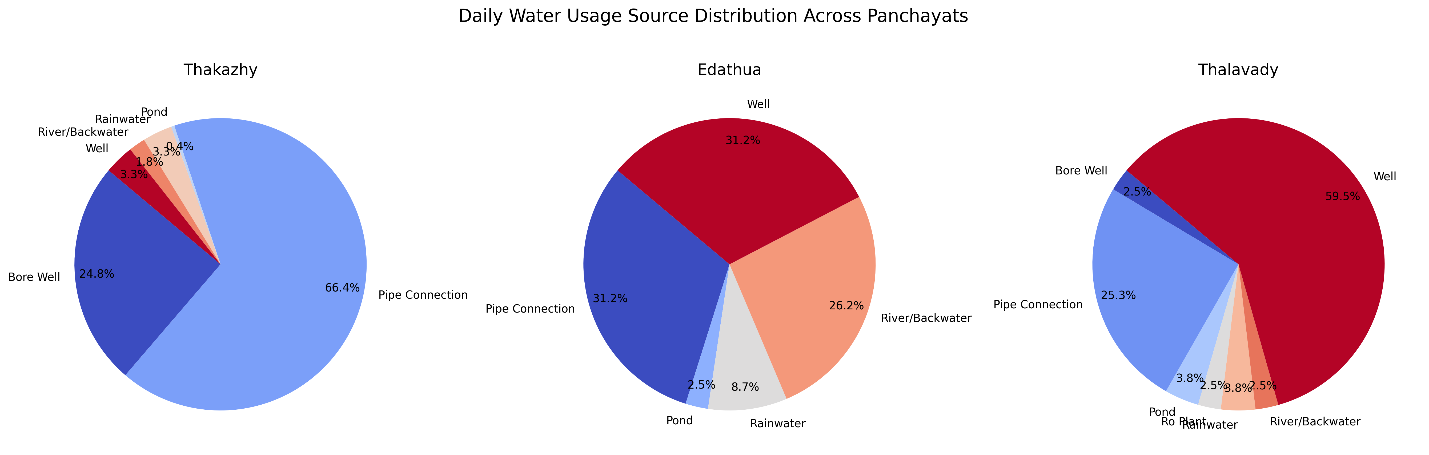
* + 1. **Percentage distribution**

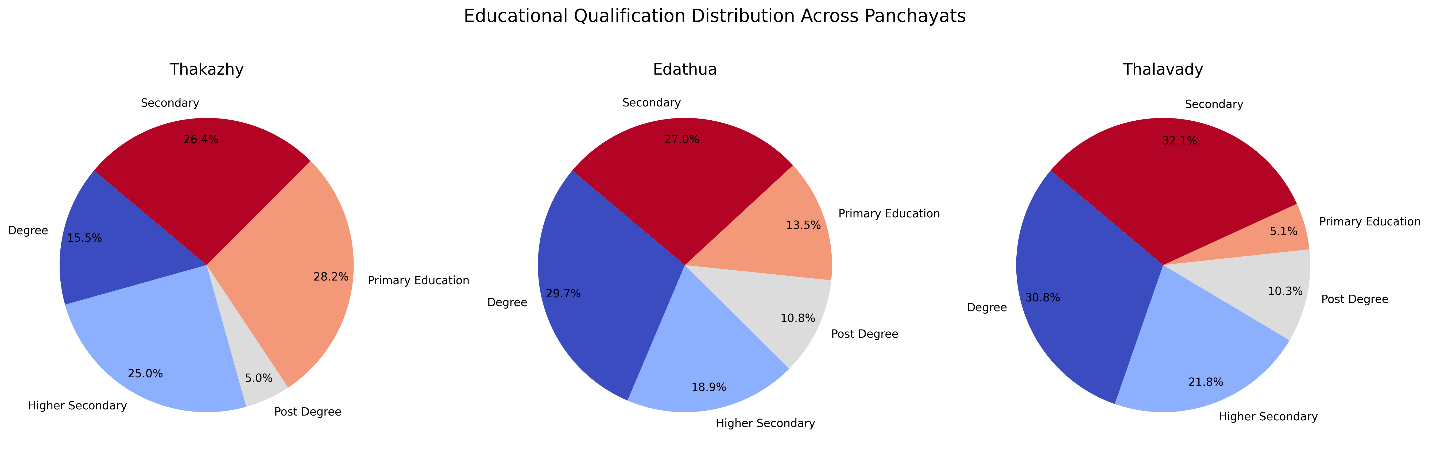
****

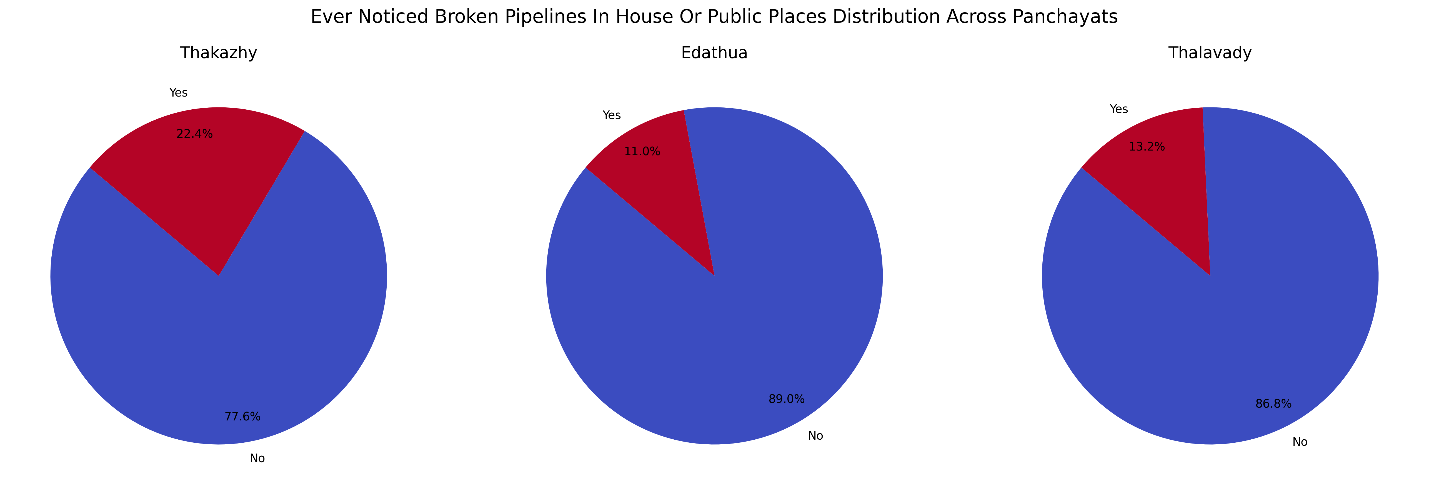
****

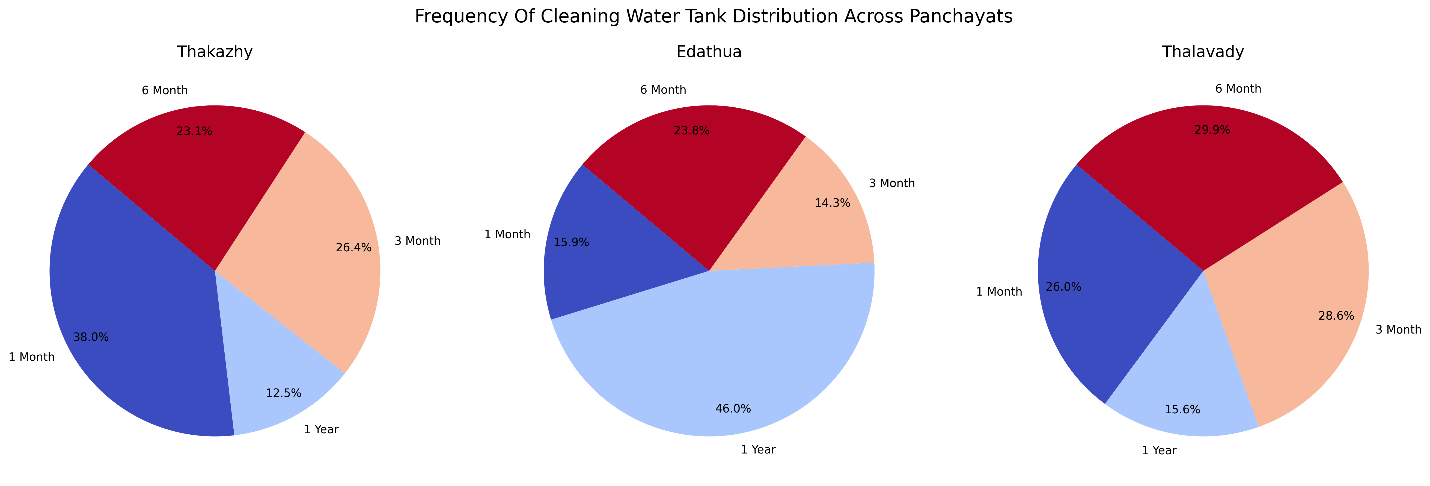
****

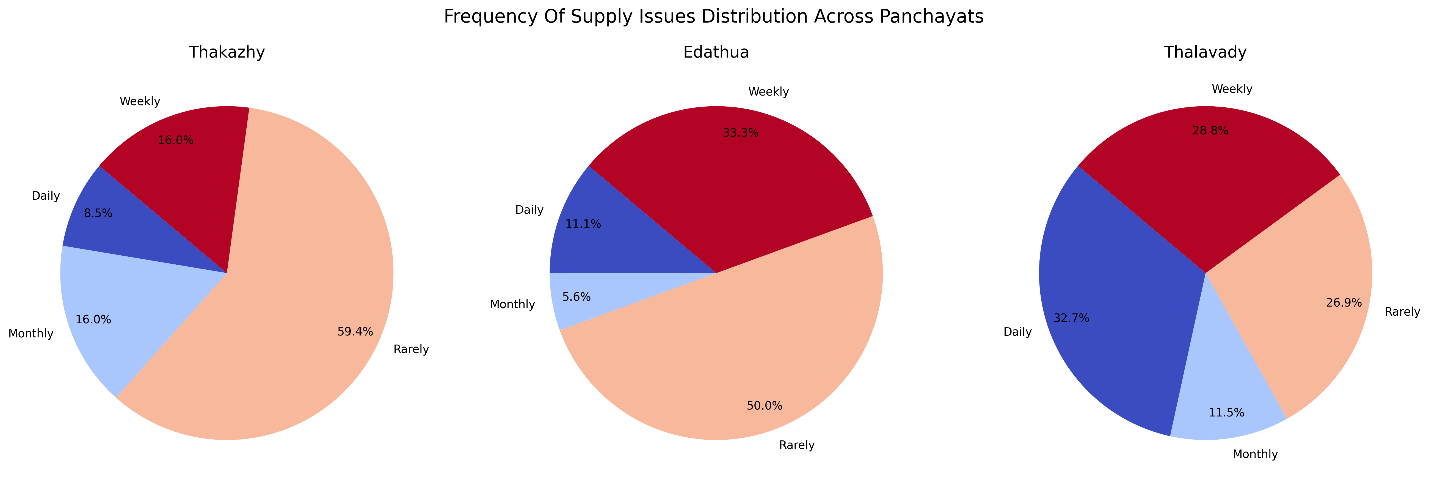
****

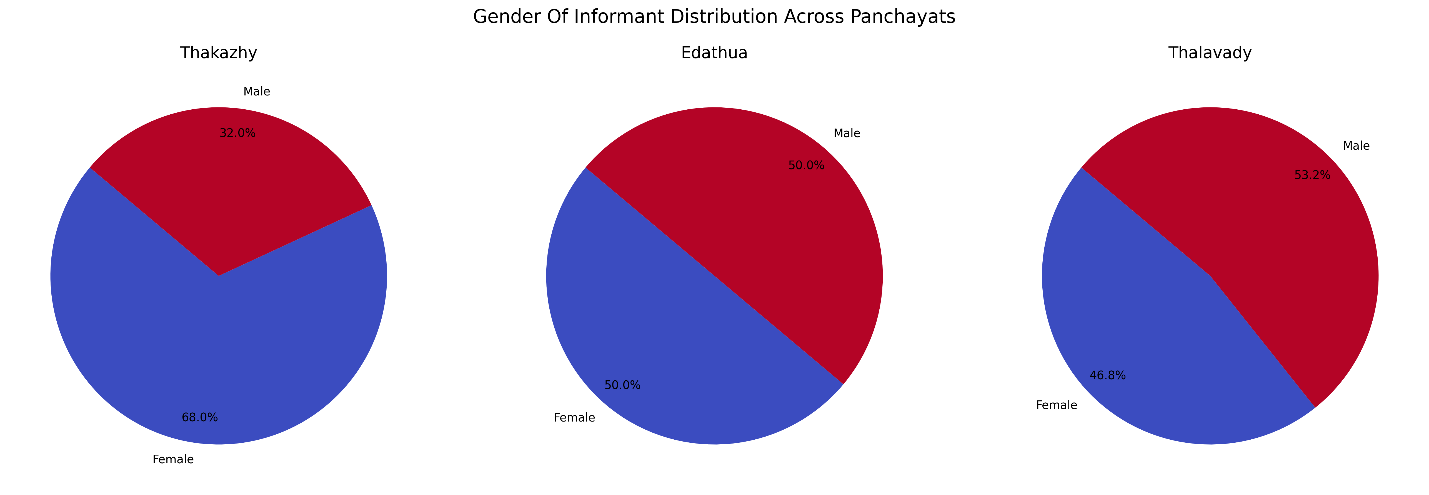
****

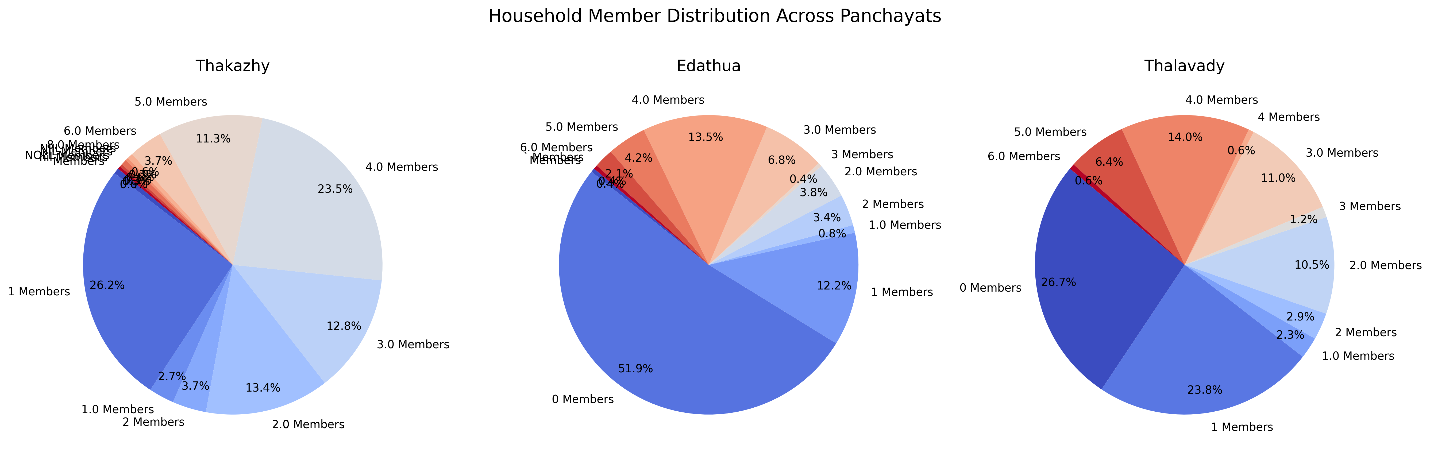
****

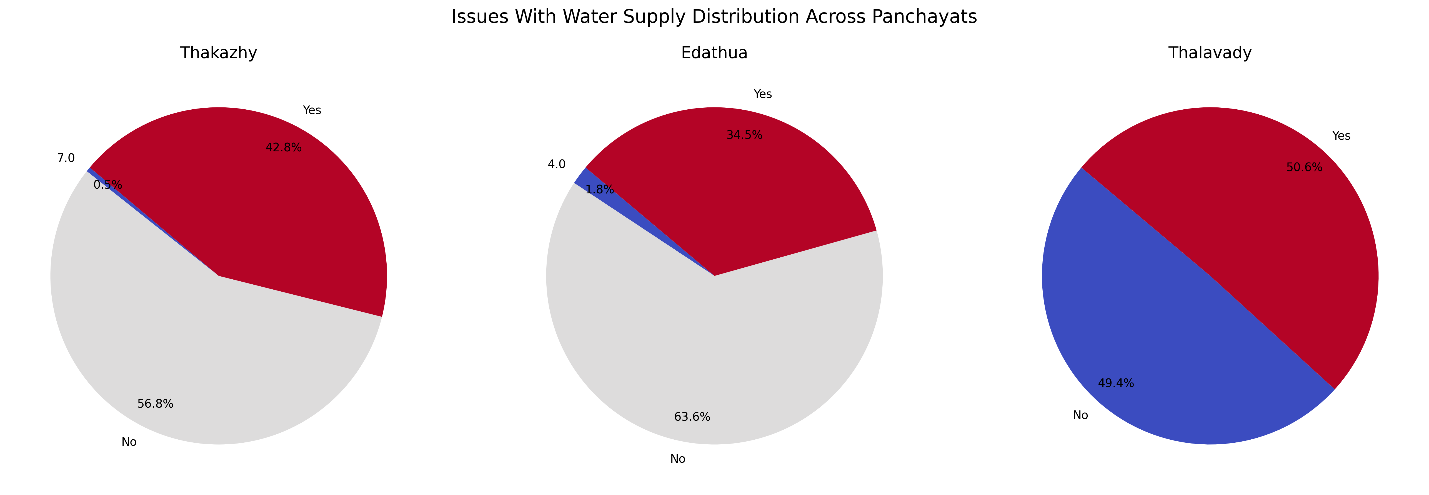
****

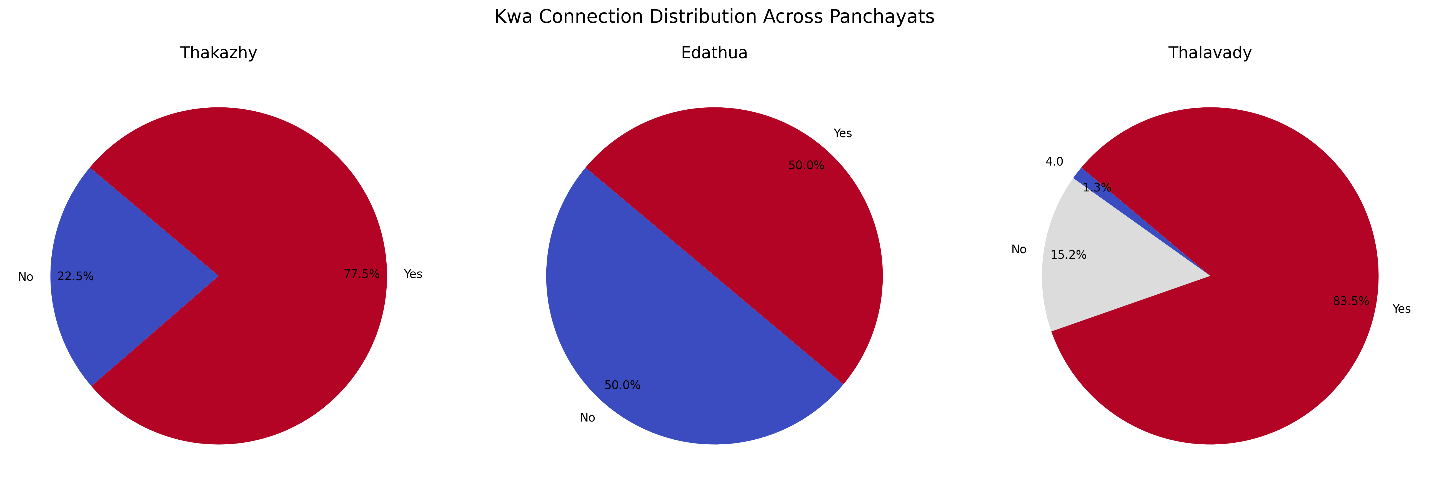
****

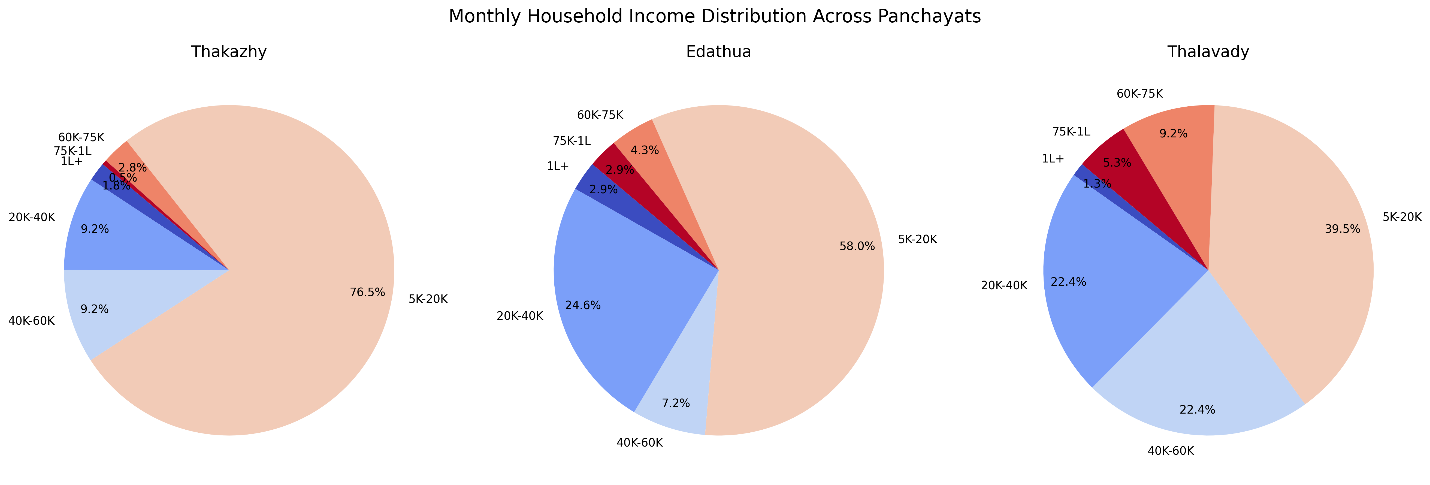
****

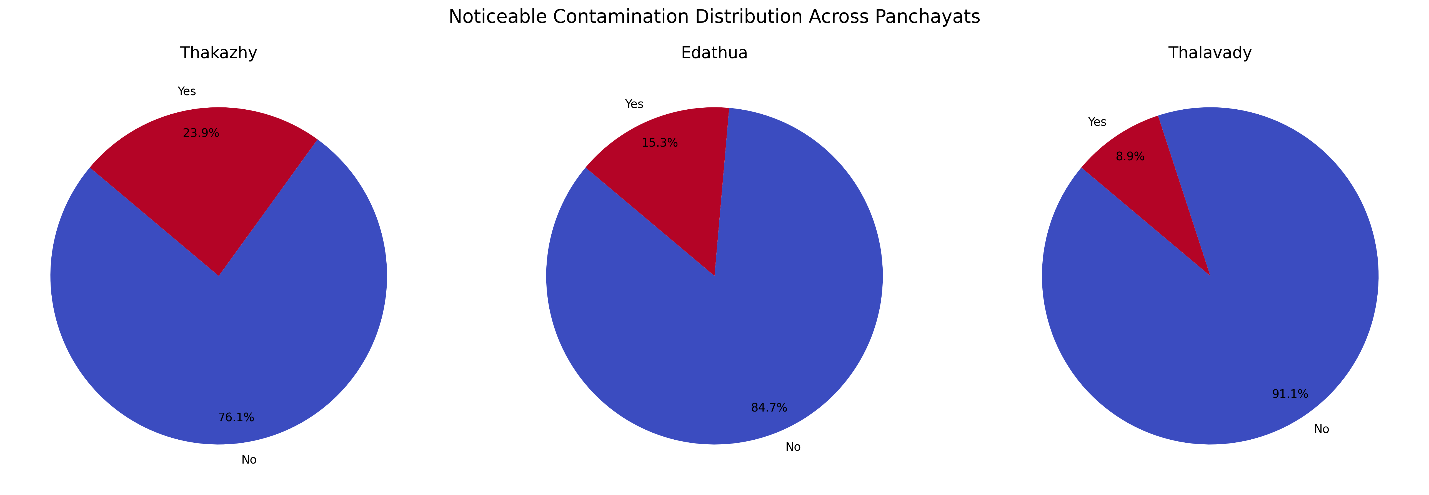
****

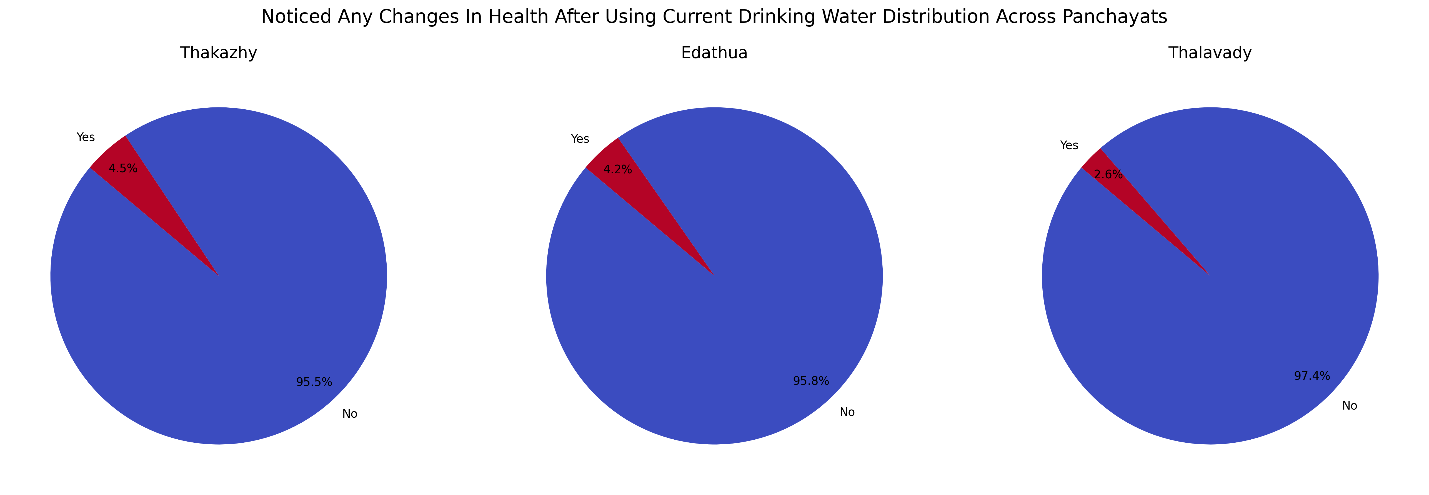
****

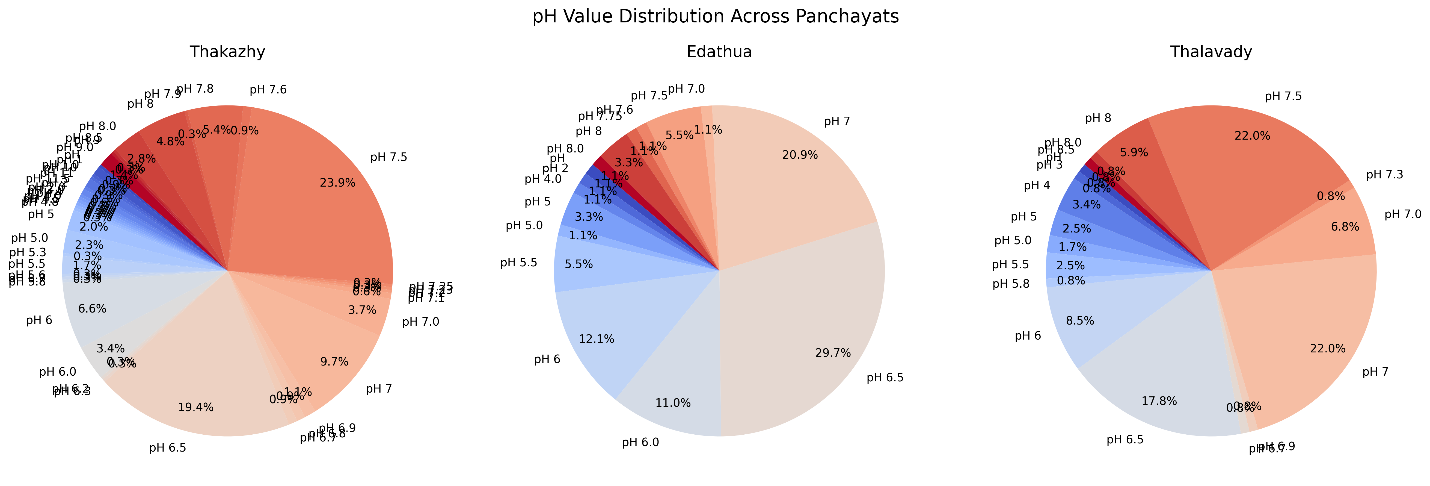
****

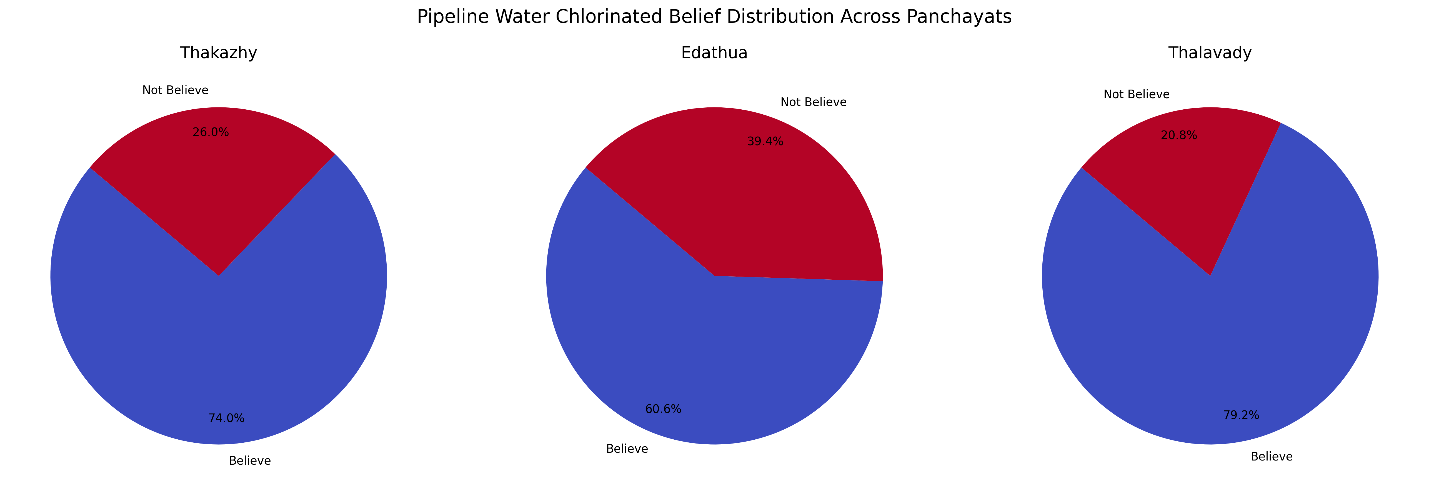
****

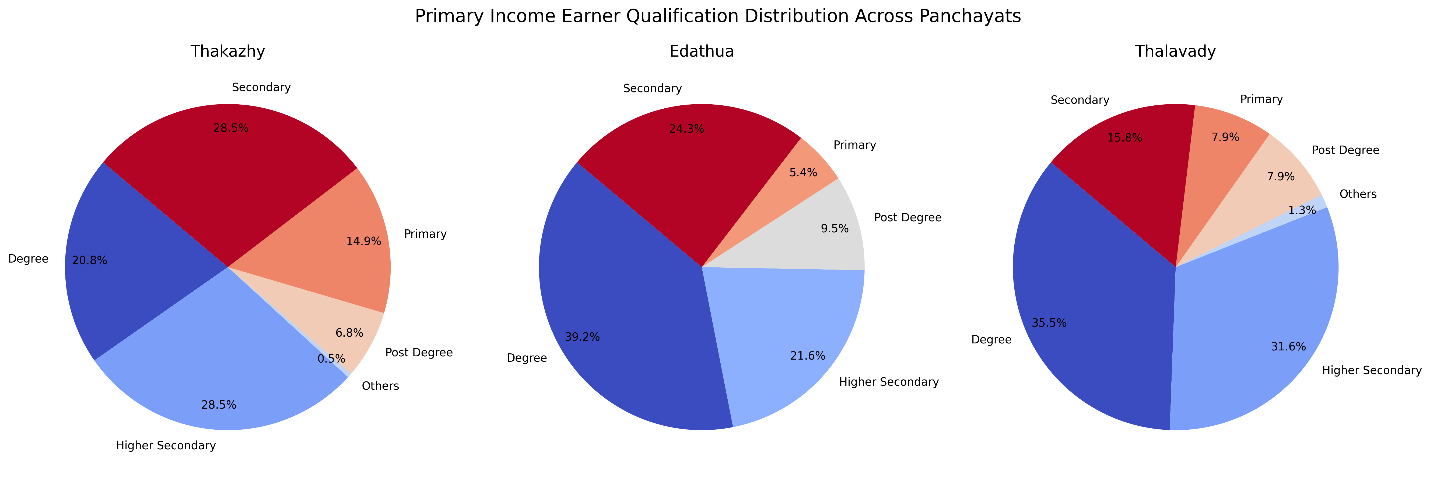
****

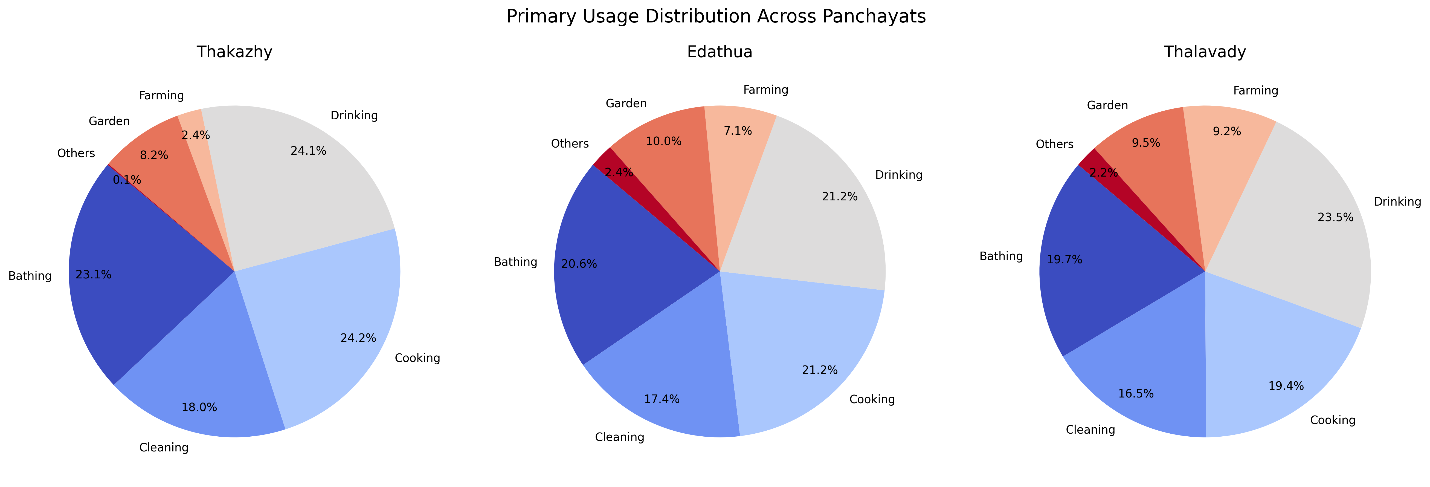
****

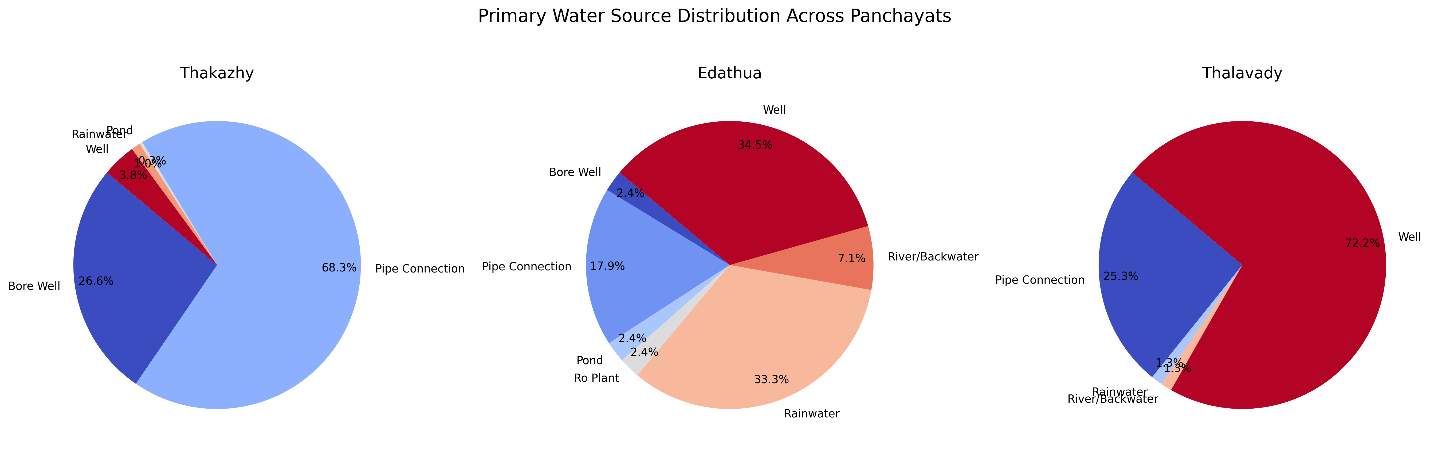
****

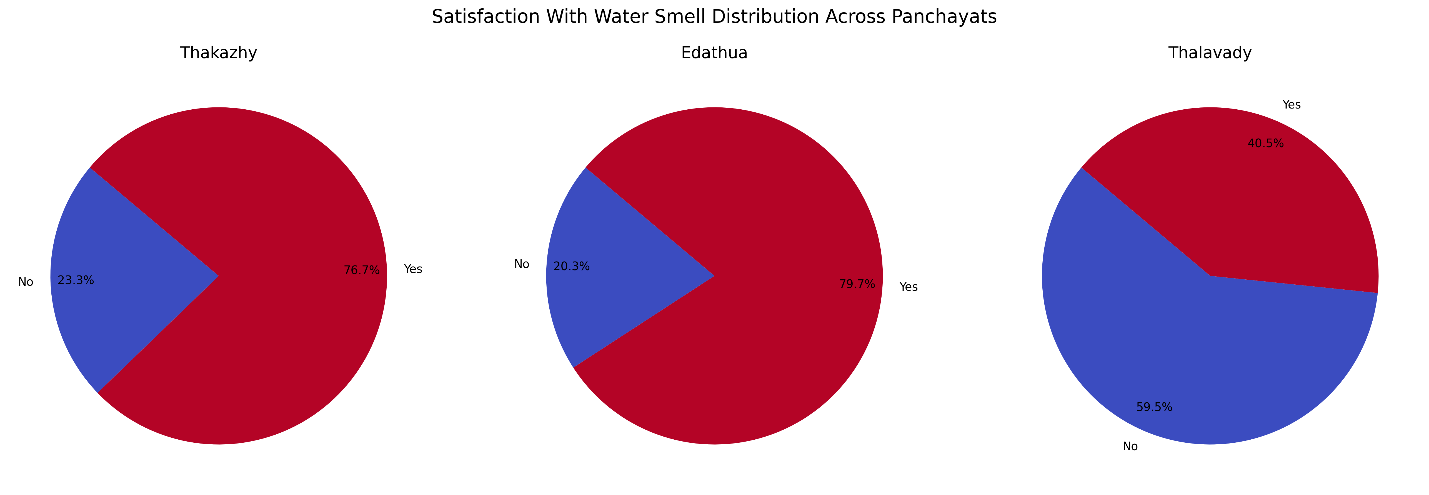
****

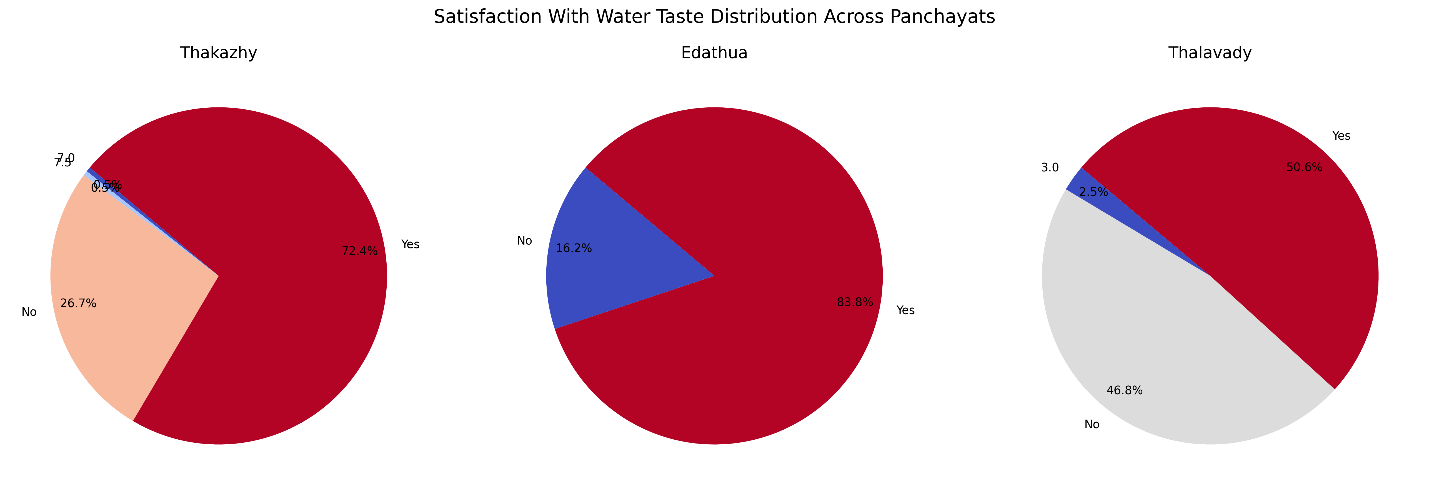
****

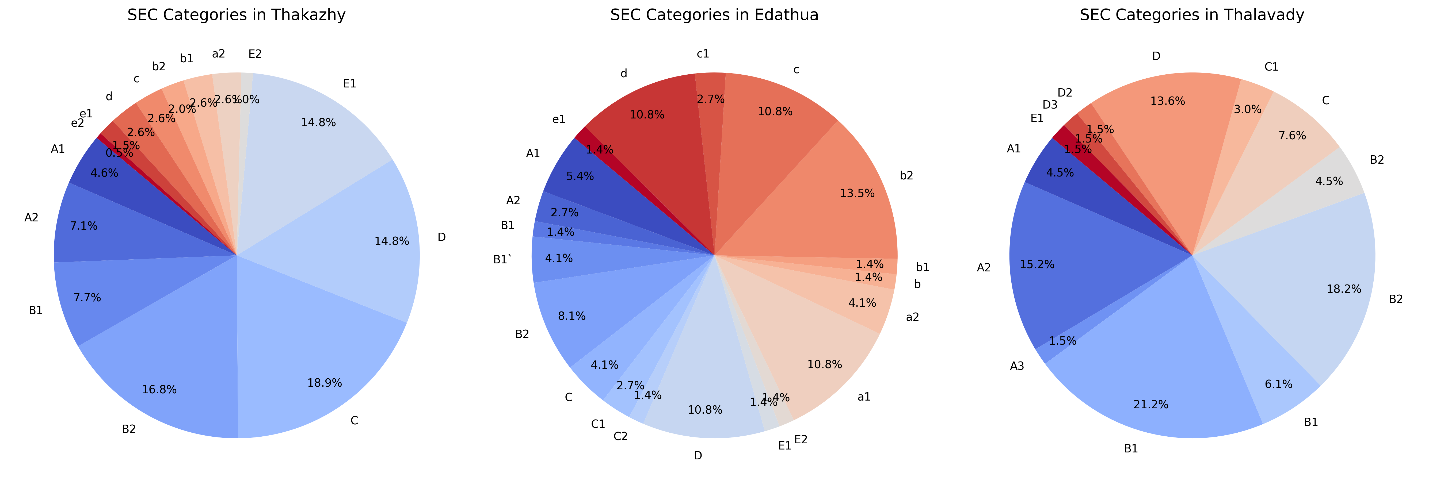
****

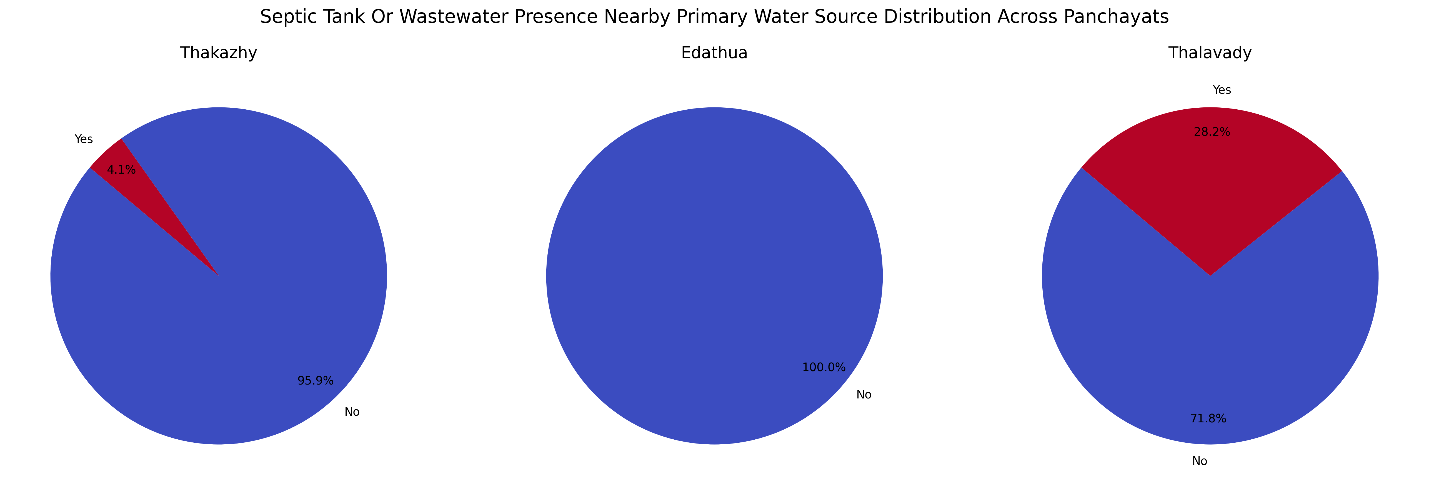
****

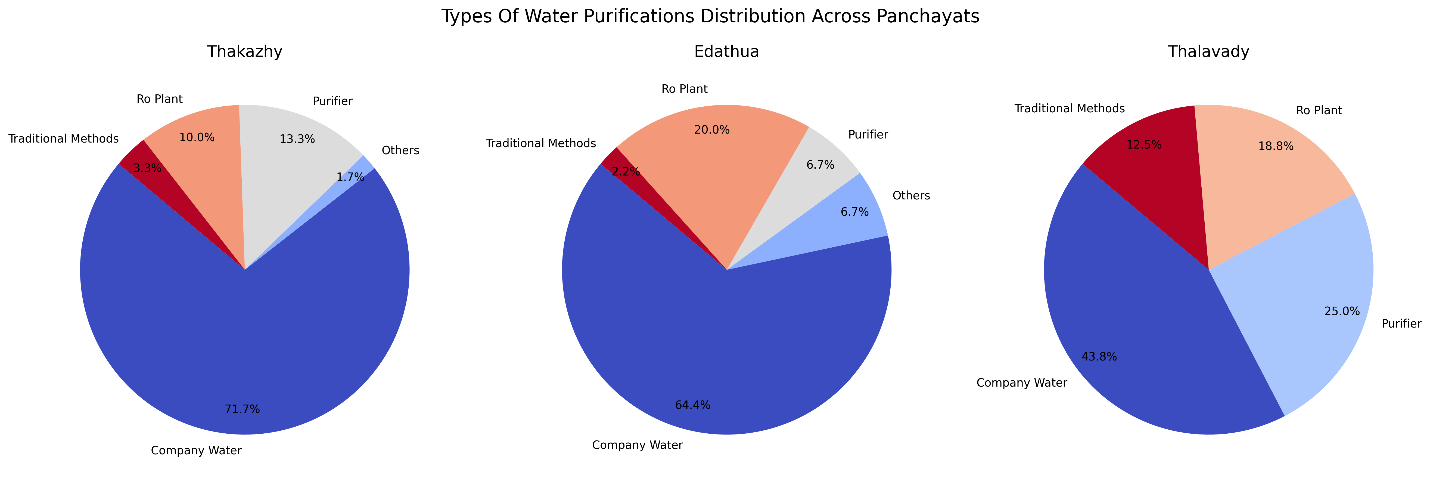
****

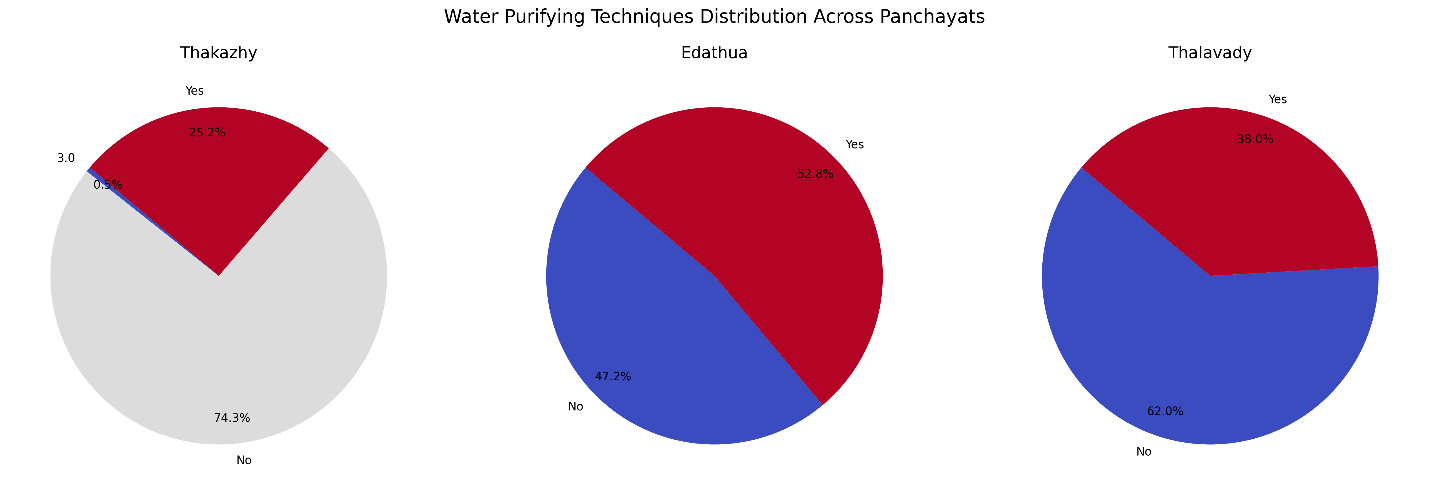
****

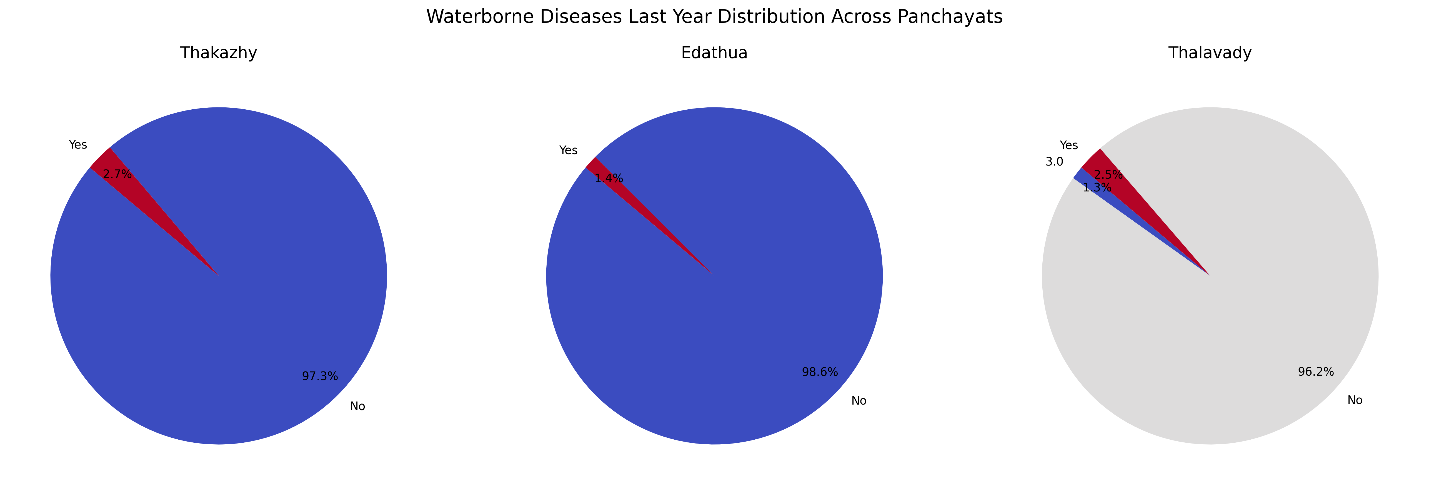
****

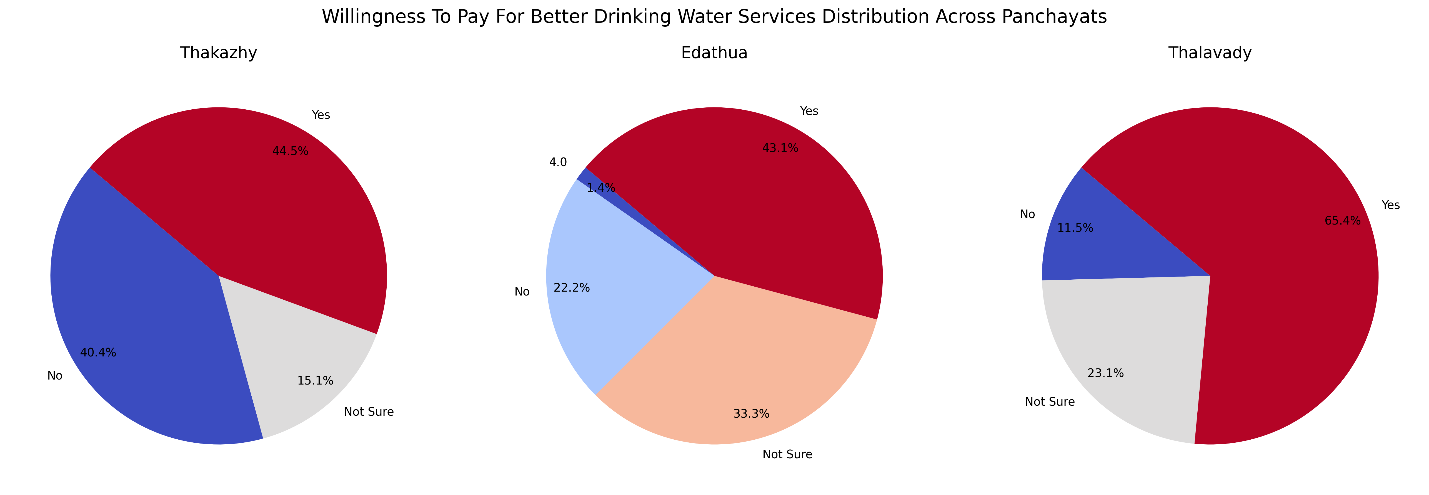
****

****

****

****

****

****

* 1. **Inferential Statistics**

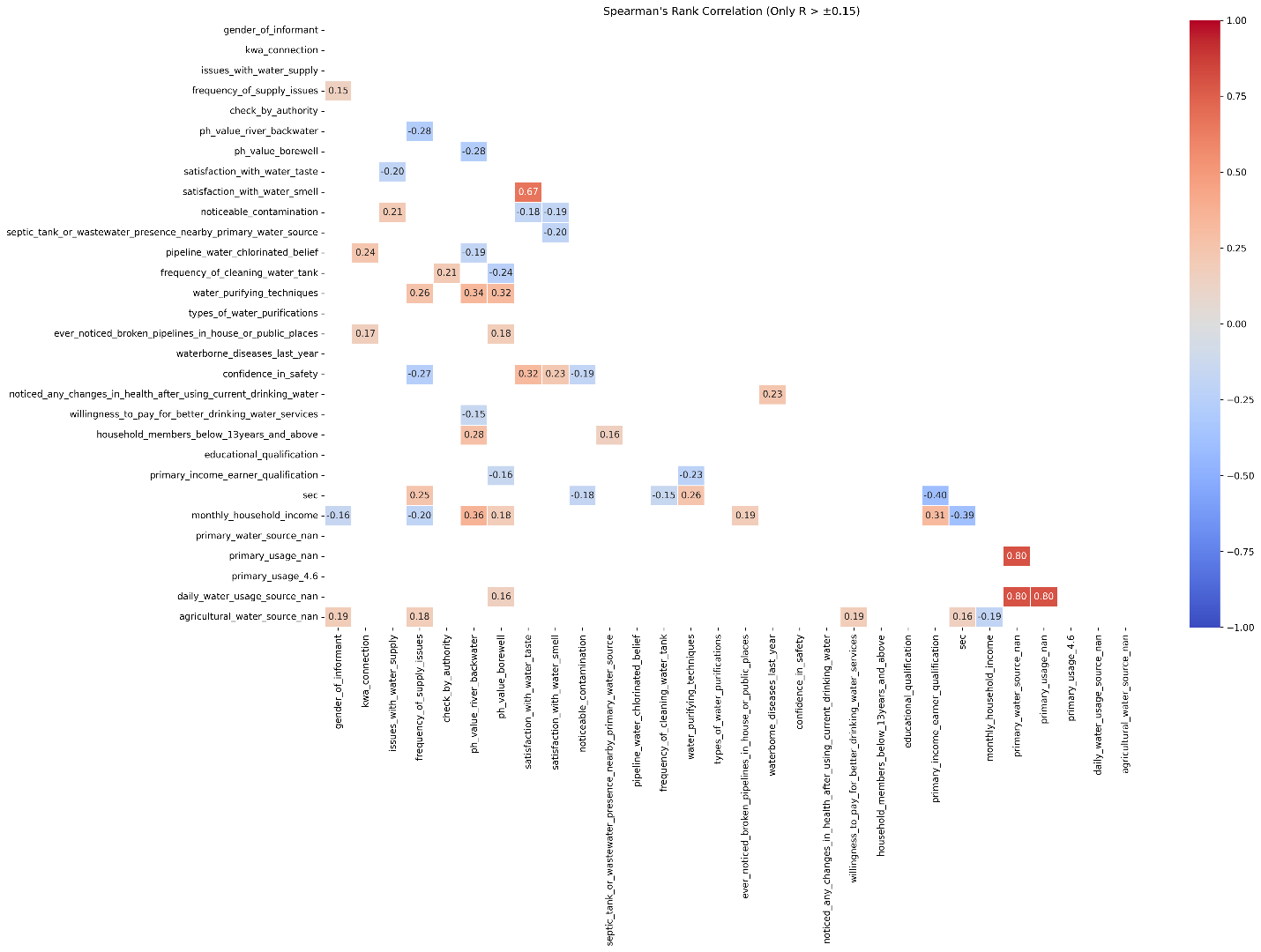
**5.2.1 Cramér’s** **V analysis**

|  |  |
| --- | --- |
|  | Cramér’s V is a statistical measure that shows how strongly two categorical variables are associated. It is based on the Chi-Square test, but instead of just saying whether two variables are related (p-value), it quantifies the strength of the relationship (ranges from 0 to 1).  0 → No Association (independent variables)  Between 0 and 1 → Partial Association  1 → Strong Association (fully dependent variables)  NOTE: Cramer's V only detect if there is an association, but not "how they related". To test "how they related" or direction of relationship do Spearman's rank correlation. |

|  |  |
| --- | --- |
| **Variable Pair** | **Cramér's V** |
| ph value well Vs satisfaction with water taste | 0.406111 |
| noticeable contamination Vs satisfaction with water taste | 0.215111 |
| pipeline water chlorinated belief Vs confidence in safety | 0.142605 |
| willingness to pay for better drinking water services Vs confidence in safety | 0.140102 |
| monthly household income Vs willingness to pay for better drinking water services | 0.155017 |
| monthly household income Vs sec | 0.402064 |
| sec Vs waterborne diseases last year | 0.169714 |
| sec Vs confidence in safety | 0.30765 |
| septic tank or wastewater presence nearby primary water source Vs confidence in safety | 0.023539 |
| water purifying techniques Vs waterborne diseases last year | 0.037346 |
| ever noticed broken pipelines in house or public places Vs waterborne diseases last year | 0.078982 |
| confidence in safety Vs noticed any changes in health after using current drinking water | 0.042265 |

Variable pairs with high Cramer’s V value (red highlighted in Table) indicates the likely association between those pairs.

**5.2.2 Spearman Correlation matrix**

****

Correlation heatmap plot above represent the correlations with possible strong relationships only. That means that pairs are inter-related significantly.

## Preliminary results conclusion

The figures and tables are the data analysis outputs. Analyst have not inferred outputs thoroughly to made a result and conclusion.

|  |
| --- |
| 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. |