Jal Jeevan Mission Water Quality
Data Completion

July 27, 2024

Ensuring no Missing Files

The data collection is automated with Python and is detailed in the python code reports, which explain how the source websites are scraped to retrieve all the required files. Due to the large volume of data, minor discrepancies, such as missing files for certain districts, can occur. To address this issue, all downloaded files are reviewed to track the number of district files obtained for each state and each year. This data is then compared with official records of states and districts in India to identify any missing districts in the downloaded dataset. The reported missing districts are manually checked on the source website, and if available, they are downloaded. This ensures the completeness of the dataset and eliminates any possibility of data incompleteness.

Following steps are done.

- First, we create a comprehensive dictionary from all the CSV files scraped from 2009 to 2024. This dictionary forms a comprehensive key-value pair of states and all the districts within each state. This helps include differently spelled variations of the same districts in different years in the list.

  For example, variations like Mysore, Mysuru, etc.

```
comprehensive_dict = {
    "State1": ["District1", "District1Variant", "District2"],
    "State2": ["District3", "District4"],
    ...
}
```

- Next, we create a new dictionary for a particular year. For example, if we want to find the missing files in our data for the year 2014-15, we make the key-value pair of states and all the districts within each state from that year's files.

```
year_specific_dict_2014_15 = {
    "State1": ["District1", "District2"],
    "State2": ["District3"],
    ...
}
```

- We then subtract the dictionary made for a particular year from the comprehensive dictionary to identify all the districts missing for that year.

```
missing_districts_dict_2014_15 = {
    "State1": ["District1Variant"],
    "State2": ["District4"],
    ...
}
```

  But this list will also include districts that might not be present on the website for that year.

- To remove these districts, we now form a dictionary of key-value pairs of states and all the districts within each state from that year's files using the data from the website. We then find the intersection of the subtracted dictionary and the dictionary from the website to identify which district files are actually missing.

```
website_data_dict_2014_15 = {
    "State1": ["District1", "District2"],
    "State2": ["District3"],
    ...
}
```

- Finally, we create the actual missing districts dictionary by finding the intersection of the missing districts dictionary and the website data dictionary to identify which district files are actually missing.

```
actual_missing_districts_dict_2014_15 = {
    "State1": ["District1Variant"],
    ...
}
```