CS685A: Assignment 2

Analysis of Indian Census Data of 2011

Problem Description

- Can be found in the Assignment file.

Requirements:

- Python3 should be present on your system and should be accessible using the command `python3`
- Install all libraries mentioned in `Requirements.txt` by running the following command:

```
python3 -m pip install -r Requirements.txt
```

OR

```
pip3 install -r Requirements.txt
```

Running the Project:

- To run the complete project in one go, execute `assign2.sh`

```
./assign2.sh
```

- To run a certain segment, use the following scripts:
    1. `percent-india.sh` : To solve Q1
    2. `gender-india.sh` : To solve Q2
    3. `geography-india.sh` : To solve Q3
    4. `3-to-2-ratio.sh` : To solve Q4(a)
    5. `2-to-1-ratio.sh` : To solve Q4(b)
    6. `age-india.sh` : To solve Q5
    7. `literacy-india.sh` : To solve Q6
    8. `region-india.sh` : To solve Q7
    9. `age-gender.sh` : To solve Q8
    10. `literacy-gender.sh` : To solve Q9

- The individual scripts ***need to be executed*** in the order in which they are mentioned, as the output generated by one script is needed for the execution of the further scripts.

Directories:

- out/ : Stores the required output json and csv files.
- data/ : Data files used in the project.
- meta/ : Stores the meta data (a set of json files) generated by the execution of scripts. The contents inside this folder ***should not*** be modified when the scripts are being executed. The folder can be deleted before beginning the execution of `assign2.sh`. These files are essentially a transformation of dataset into a more usable format.

Data Files:

- `C17` : Population by Bilingualism & Trilingualism
- `DDW_PCA0000_2011_Indiastatedist` : Overall Population Data
- `DDW-0000C-08` : Educational Level By Age And Sex
- `DDW-0000C-13` : Single Year Age
- `DDW-C18-0000` : Population By Bilingualism, Trilingualism, Age And Sex
- `DDW-C19-0000` : Population By Bilingualism, Trilingualism, Education Level And Sex

Python Scripts:

- meta.py : Converts the csv files to json format for use by the scripts later on.
- q1.py : Solves Q1 and generates `percent-india.csv`.
- q2.py : Solves Q2 and generates `gender-india.csv`.
- q3.py : Solves Q3 and generates `geography-india.csv`.
- q4.py : Solves Q4 and generates `2-to-1-ratio.csv` and `3-to-2-ratio.csv`.
- q5.py : Solves Q5 and generates `age-india.csv`.
- q6.py : Solves Q6 and generates `literacy-india.csv`.
- q7.py : Solves Q7 and generates `region-india-a.csv` and `region-india-b.csv`.
- q8.py : Solves Q8 and generates `age-gender-1.csv`, `age-gender-2.csv` and `age-gender-3.csv`.
- q9.py : Solves Q9 and generates `literacy-gender-1.csv`, `literacy-gender-2.csv` and `literacy-gender-3.csv`.

Additional Notes:

- All Output files except those from Q4: Have the first line as header.
- In order to calculate p-value, `Welch's t-test` has been used.