Data Mining - Assignment 2

File structure

1. The assignment is made using python 3.
2. Assignment folder contains
    A. Python code for all questions.
    B. Shell .sh file to run respective python code.
    C. 'assign2.sh' file to run shell script of all questions.
    D. Data folder - stores all data.
3. Data folder contains
    A. All initial data needed.
    B. For population - CENSUS DATA file given in assignment 1 is used.
    C. C17 folder - contains all state-wise c17 files.
    D. Output folder - All output files generated by the program are named as per naming conventions.

Library needed

Library needed to run assignment are.

1. openpyxl.
2. numpy
3. pandas
4. scipy.stats
5. collections

How to run code

'assign2.sh' is top-level script that runs the entire assignment. It run all the questions in sequence.

It can be run by `bash assign2.sh`

Other Details

Question 1

1. python file is named as `'percent_india.py'`
2. shell file is named as `'percent_india.sh'`
3. Apart from output file it genrate 'c18_modified.csv' which is dependency for other questions.

Question 2

1. python file is named as `'gender_india.py'`
2. shell file is named as `'gender_india.sh'`
3. To find p-value t_test is used by using 'scipy.stat' library's 'ttest_ind' function.

Question 3

1. python file is named as `'geography_india.py'`
2. shell file is named as `'geography_india.sh'`
3. To find p-value t_test is used by using 'scipy.stat' library's 'ttest_ind' function.

Question 4

1. python file is named as `'3_to_2_ratio.py'` and `'2_to_1_ratio.py'`
2. shell file is named as `'3_to_2_ratio.sh.'` and `'2_to_1_ratio.sh'`

Question 5

1. python file is named as `'age_india.py'`
2. shell file is named as `'age_india.sh.'`
3. For population data for diffrent age group c14 file is used.

Question 6

1. python file is named as `'literacy_india.py'`
2. shell file is named as `'literacy_india.sh.'`

Question 7

1. python file is named as `'region_india.py'`
2. shell file is named as `'region_india.sh.'`
3. c17 file for all states are c17 folder

Question 8

1. python file is named as `'age_gender.py'`
2. shell file is named as `'age_gender.sh'`
3. For population data for diffrent age group c14 file is used.

Question 9

1. python file is named as `'literacy_gender.py'`

2. shell file is named as `'literacy_gender.sh'`
3. For population data for diffrent age group c8 file is used.