DATA MINING ASSIGNMENNT 2

DATA ANALYSIS OF INDIAN SPOKEN LANGUAGES

Requirements:

- Python version 3.8 or 3.9
- **Python Libraries needed:**
    - Pandas 1.3.2
    - Numpy 1.21.1
    - Csv
    - Scipy 1.6.2

Submission Structure:

1) Datasets Original: Contains mostly files downloaded from Census India website and Census1.csv which is taken from previous assignment i.e., assignment 1

    a) C17 Folder – Contains individual Csv files for India and each state and union territory containing data about no of people who speak a particular language as either mother tongue or 2nd Language or 3rd Language.
    b) Census1.csv – Taken from Assignment 1 and contains census data at district level for each state and union territory which is aggregated to form Census2.csv
    c) DDW-0000C-14.xls – Contains data pertaining to No of males and females and Total population for India and each state and union territory age group wise and separately for urban and rural parts.
    d) DDW-0000C-14.csv – Derived from DDW-0000C-14.xls with certain columns like District Code, Table name and etc dropped and even some rows dropped containing file info.
    e) DDW-C08-0000.xlsx – Contains data pertaining to No of males and females and Total population for India and each state and union territory age group wise and literacy group wise. Used to create Literacy Census.csv
    f) DDW-C18-0000.xlsx – Contains data pertaining to Bilingual and Trilingual No of males and females and Total population for India and each state and union territory age group wise. Used to create Bi and Tri Age.csv

g) DDW-C19-0000.xlsx – Contains data pertaining to Bilingual and Trilingual No of males and females and Total population for India and each state and union territory literacy group wise. Used to create Bi and Tri Litercay.csv

h) Literacy Census.csv – Contains data pertaining to total no of people, no of males and females belonging to a literacy group for India and each state and union territory. Derived from DDW-C08-0000.xlsx

2) Dataset

a) Agewise Census.csv – Gives Total, Male and Female Population for India and each state and union territory for each age group separately and for all age group too.

b) Bi and Tri Age.csv – Gives Number of Males, Females and Total who are Bilingual and Trilingual for India and each state and union territory for each age group. It was observed that column for Bilingual Data contains data of people who speak at least 2 languages. Derived from C18 file by dropping some columns done using Microsoft Excel.

c) Bi and Tri Age2.csv – Gives Number of Males, Females and Total who are Bilingual and Trilingual for India and each state and union territory for each age group. Derived from Bi and Tri Age.csv and contains data in Bilingual Columns as people who speak exactly 2 languages.

d) Bi and Tri Literacy.csv – Gives Number of Males, Females and Total who are Bilingual and Trilingual for India and each state and union territory for each literacy group. It was observed that column for Bilingual Data contains data of people who speak at least 2 languages. Derived from C19 file by dropping some columns done using Microsoft Excel.

e) Bi and Tri Literacy2.csv – Gives Number of Males, Females and Total who are Bilingual and Trilingual for India and each state and union territory for each age group. Derived from Bi and Tri Literacy.csv and contains data in Bilingual Columns as people who speak exactly 2 languages.

f) Census2.csv – It is derived from Census1.csv. Census1.csv contains census data at district level and we need data at state level, so all districts of each state were aggregated in Census1.csv to form Census2.csv with state code as required for this assignment.

g) Census3.csv – It is derived from DDW-C08-0000.XLSX. C08 contains census data for each state and union territory for each age group and each literacy group and we aggregate this data to get census data for each state and union territory for its urban and rural area. Therefore, Census3.csv contains census for India and each state and union territory for urban and rural region separately.

h) Language Total Speakers Statewise.csv – It gives data for India and each state and union territory, no of people speaking a language as either mother tongue or 2nd language or 3rd language.

i) Literacy Total.csv – Contains census data for India and each state and union territory for each literacy group separately.

j) Mono Gender.csv – Contains data for no of male, female and total people who speak one language for India and for each state and union territory separately.

k) Mother Tongue Statewise.csv – It gives data for India and each state and union territory, no of people speaking a language as thier mother tongue.

l) regions.csv – Contains state name and corresponding region as mentioned in Q7

m) TEMP_STATECODE.csv – Contains state name, state code from previous assignment and state code for this assignment. Used to build Census2.csv from Census1.csv. File was made manually using Microsoft Excel.

3) Output – Contains csv files generated as output in Q1 - Q9. In total 19 csv files are generated as output.

a) Q1 – percent-india.csv

b) Q2 – gender-india-a.csv (Monolingual or Only 1 language)
gender-india-b.csv (Bilingual or Exactly 2 languages)
gender-india-c.csv (Trilingual or 3 and more languages)

c) Q3 – geography-india-a.csv (Monolingual or Only 1 language)
geography-india-b.csv (Bilingual or Exactly 2 languages)
geography-india-c.csv (Trilingual or 3 and more languages)

d) Q4 – 3-to-2-ratio.csv
2-to-1-ratio.csv

e) Q5 – age-india.csv

f) Q6 – literacy-india.csv

g) Q7 – region-india-a.csv (Mother Tongue)
region-india-b.csv (Mother Tongue, 2nd language, 3rd language)

h) Q8 – age-gender-a.csv (Trilingual or 3 and more languages)
age-gender-b.csv (Bilingual or Exactly 2 languages)
age-gender-c.csv (Monolingual or Only 1 language)

i) Q9 – literacy-gender-a.csv (Trilingual or 3 and more languages)
litercay-gender-b.csv (Bilingual or Exactly 2 languages)

literacy-gender-c.csv (Monolingual or Only 1 language)

4) Python – Contains python scripts for each question and to generate some Input Datasets

a) Q1 – percent-india.py

Input: Census2.csv, Mono Gender.csv, Bi and Tri Age2.csv

b) Q2 – gender-india.py

Input: Census2.csv, Bi and Tri Age2.csv

c) Q3 – geography-india.py

Input: Census3.csv, Bi and Tri Age2.csv

d) Q4 – 3-to-2-ratio.py

Input: Bi and Tri Age2.csv

2-to-1-ratio.py

Input: Bi and Tri Age2.csv, Mono Gender.csv

e) Q5 – age-india.py

Input: Literacy Total.csv, Bi and Tri Literacy2.csv

f) Q6 – literacy-india.py

Input: vaccine.csv

g) Q7 – region-india.py

Input: Mother Tongue Statewise.csv, regions.csv, Language Total Speakers Statewise.csv

h) Q8 – age-gender.py

Input: Agewise Census.csv, Bi and Tri Age2.csv

i) Q9 – literacy-gender.py

Input: Literacy Total.csv, Bi and Tri Literacy2.csv

j) Agewise_Converter.py

Input: DDW-0000C-14.csv

Output: Agewise Census.csv

k) C17 to Total Speakers.py

Input: C17 files

Output: Language Total Speakers Statewise.csv

l) C17_to _Mother_Tongue.py

Input: C17 files

Output: Mother Tongue Statewise.csv

m) Census3gen.py

Input: DDW-C08-0000.XLSX

Output: Census3.csv

n) Census_1_to_2.py
   Input: Census1.csv, TEMP_STATECODE.csv
   Output: Census2.csv

o) Exactly 2 finder.py
   Input: Bi and Tri Age.csv, Bi and Tri Literacy.csv
   Output: Bi and Tri Age2.csv, Bi and Tri Literacy2.csv

p) Literacy Converter.py
   Input: Literacy Census.csv
   Output: Literacy Total.csv

q) Monolingual Data.py
   Input: Census2.csv, Bi and Tri Age2.csv
   Output: Mono Gender.csv

5) Script – Contains .sh files to execute python scripts

a) Complete Assignment – assign2.sh
b) Q1 – percent-india.sh
c) Q2 – gender-india.sh
d) Q3 – geography-india.sh
e) Q4 – 3-to-2-ratio.sh
   2-to-1-ratio.sh
f) Q5 – age-india.sh
g) Q6 – literacy-india.sh
h) Q7 – region-india.sh
i) Q8 – age-gender.sh
j) Q9 – literacy-gender.sh

<u>Steps To Execute Assignment:</u>

1. Move into the ‘Scripts’ folder and open Terminal/Command Prompt etc.
2. To run complete assignment, type ‘bash assign2.sh’, all the generated csv files will be stored in the ‘Output’ folder.

Note: You may delete one or more csv files from the ‘Output’ folder but do not delete ‘Output’ folder itself.

3. You may even run particular question script individually by typing ‘bash filename.sh’ and generated csv is stored in ‘Output’ folder.
4. There are no scripts to run python programs which modify or create files present in Dataset folder. To run these python script type “python ‘file path’ “. Remember to put filename in quotes as there are spaces in some python file name

NOTE:

1. **Do not delete ‘Output’ folder or else scripts will start generating error as all the csv files generated are stored in the ‘Output’ folder.**
2. **Some Files in Dataset folder are pre-processed and formatted for use in analysis using Microsoft Excel, so do not replace or modify them. Others for which python script is available in Python folder can be deleted and generated again.**
3. **While opening output csv files containing age group, some spreadsheet software treats it as either date type column or numeric type. If using LibreOffice then when opening output file, select ‘formatted quoted field as text’ to see the age group column properly.**
4. All Numeric Data has been rounded to 4 decimal places.
5. Most of the data processing done using Microsoft Excel is dropping some columns and removing header rows. For some files row values were summed to get new row values.
6. We have used One Sample T-test for calculating p values in Q2 and Q3.
7. For Q2, to calculate p-value we have taken sample has array of for Trilingual male: female ratio, Bilingual male: female ratio and Monolingual male: female ratio and population mean as male: female ratio of India/state/ut.
8. For Q3, to calculate p-value we have taken sample has array of for Trilingual urban population: rural population ratio, Bilingual urban population: rural population ratio and Monolingual urban population: rural population ratio and population mean as urban population : rural population ratio of India/state/ut.
9. For Q2 and Q3, there are 3 csv output files each, but p-value is same across the 3 files. The 3 files differ in male-percentage and female-percentage column.
10. For Q6 and Q9, to calculate total, male, and female population in each literacy group there was mismatch in column headers across file C-08 and C-19 so literacy group Matric/Sec but below grad contains sum of values of Matric/Sec, Higher Sec/Inter, Non-tech diploma, Tech diploma.