**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
   1. C17 Folder
   2. Census1.csv
   3. DDW-0000C-14.xls
   4. DDW-0000C-14.csv
   5. DDW-C08-0000.xlsx
   6. DDW-C18-0000.xlsx
   7. DDW-C18-0000.xlsx
   8. Literacy Census.csv
2. Dataset
   1. Agewise Census.csv
   2. Bi and Tri Age.csv
   3. Bi and Tri Age2.csv
   4. Bi and Tri Literacy.csv
   5. Bi and Tri Literacy2.csv
   6. Census2.csv
   7. Census3.csv
   8. Language Total Speakers Statewise.csv
   9. Literacy Total.csv
   10. Mono Gender.csv
   11. Mother Tongue Statewise.csv
   12. regions.csv –
   13. TEMP\_STATECODE.csv
3. Output – Contains csv files generated as output in Q1 - Q9. In total 19 csv files are generated as output.
   1. Q1 – percent-india.csv
   2. 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)

* 1. 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)

* 1. Q4 – 3-to-2-ratio.csv

2-to-1-ratio.csv

* 1. Q5 – age-india.csv
  2. Q6 – literacy-india.csv
  3. Q7 – region-india-a.csv (Mother Tongue)

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

* 1. 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)

* 1. 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)

1. Python – Contains python scripts for each question and to generate some Input Datasets
   1. Q1 – percent-india.py

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

* 1. Q2 – gender-india.py

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

* 1. Q3 – geography-india.py

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

* 1. 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

* 1. Q5 – age-india.py

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

* 1. Q6 – literacy-india.py

Input: vaccine.csv

* 1. Q7 – region-india.py

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

* 1. Q8 – age-gender.py

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

* 1. Q9 – literacy-gender.py

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

* 1. Agewise\_Converter.py

Input: DDW-0000C-14.csv

Output: Agewise Census.csv

* 1. C17 to Total Speakers.py

Input: C17 files

Output: Language Total Speakers Statewise.csv

* 1. C17\_to \_Mother\_Tongue.py

Input: C17 files

Output: Mother Tongue Statewise.csv

* 1. Census3gen.py

Input: DDW-C08-0000.XLSX

Output: Census3.csv

* 1. Census\_1\_to\_2.py

Input: Census1.csv, TEMP\_STATECODE.csv

Output: Census2.csv

* 1. 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

* 1. Literacy Converter.py

Input: Literacy Census.csv

Output: Literacy Total.csv

* 1. Monolingual Data.py

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

Output: Mono Gender.csv

1. Script – Contains .sh files to execute python scripts
   1. Complete Assignment – assign2.sh
   2. Q1 – percent-india.sh
   3. Q2 – gender-india.sh
   4. Q3 – geography-india.sh
   5. Q4 – 3-to-2-ratio.sh

2-to-1-ratio.sh

* 1. Q5 – age-india.sh
  2. Q6 – literacy-india.sh
  3. Q7 – region-india.sh
  4. Q8 – age-gender.sh
  5. Q9 – literacy-gender.sh