**Project 2: Unit 2 (Deadline 23 March 2024 11:59pm)**

This project has 3 parts and you need to do the following in R Studio and submit the R codes, R markdown and PDF report from the R markdown file.

**A link of video explaining your work using R markdown must also be submitted. DO NOT SUBMIT THE VIDEO FILE IN THE GOOGLE CLASSROOM, ONLY THE LINK IS NEEDED!**

**Part 1**: Replicate the following table using

1. Import "covnep_252days.csv" data in R, get summary of "totalCases" variable to get:

   - Min. 1st Qu. Median Mean 3rd Qu. Max.
   - 0 2 963 13376 19340 77816

Fix the problem with minimum value using base R code and get the summary of the same variable again to show that the minimum number of "totalCases" is 1 between 1st and 2nd COVID-19 cases in Nepal.

2. Import "SAQ8.sav" data in R Studio and get frequencies of q01, q03, q06 & q08 variables as per this table. You must show codes to compute frequencies, percentage, valid percentage and cumulative percentage in R script file **i.e. DO NOT COPY THE VALUES given in the tables below!**

| Statistics makes me cry | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Strongly agree | 270 | 10.5 | 10.5 | 10.5 |
| | Agree | 1338 | 52.0 | 52.0 | 62.5 |
| | Neither | 735 | 28.6 | 28.6 | 91.1 |
| | Disagree | 187 | 7.3 | 7.3 | 98.4 |
| | Strongly disagree | 41 | 1.6 | 1.6 | 100.0 |
| | Total | 2571 | 100.0 | 100.0 | |

| Statistics makes me cry | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Strongly agree | 270 | 10.5 | 10.5 | 10.5 |
| | Agree | 1338 | 52.0 | 52.0 | 62.5 |
| | Neither | 735 | 28.6 | 28.6 | 91.1 |
| | Disagree | 187 | 7.3 | 7.3 | 98.4 |
| | Strongly disagree | 41 | 1.6 | 1.6 | 100.0 |
| | Total | 2571 | 100.0 | 100.0 | |

| I have little experience of computers | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Strongly agree | 702 | 27.3 | 27.3 | 27.3 |
| | Agree | 1127 | 43.8 | 43.8 | 71.1 |
| | Neither | 344 | 13.4 | 13.4 | 84.5 |
| | Disagree | 252 | 9.8 | 9.8 | 94.3 |
| | Strongly disagree | 146 | 5.7 | 5.7 | 100.0 |
| | Total | 2571 | 100.0 | 100.0 | |

| I have never been good at mathematics | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Strongly agree | 383 | 14.9 | 14.9 | 14.9 |
| | Agree | 1487 | 57.8 | 57.8 | 72.7 |

| | | | | | |
|---|---|---|---|---|---|
| | Neither | 482 | 18.7 | 18.7 | 91.5 |
| | Disagree | 147 | 5.7 | 5.7 | 97.2 |
| | Strongly disagree | 72 | 2.8 | 2.8 | 100.0 |
| | Total | 2571 | 100.0 | 100.0 | |

3. Import “MR_drugs.xlsx” file in R Studio and get the following table using inco1, inco2, inco3, inco4, inco5, inco6 and inco7 variables. These variables are multiple response variables and each of them coded as 0=No and 1=Yes.

| $Income Frequencies | | | | |
|---|---|---|---|---|
| | | Responses | | Percent of Cases |
| | | N | Percent | |
| Income - Multiple Response[a] | inco1 | 226 | 12.8% | 23.5% |
| | inco2 | 607 | 34.5% | 63.0% |
| | inco3 | 293 | 16.6% | 30.4% |
| | inco4 | 50 | 2.8% | 5.2% |
| | inco5 | 82 | 4.7% | 8.5% |
| | inco6 | 151 | 8.6% | 15.7% |
| | inco7 | 352 | 20.0% | 36.6% |
| Total | | 1761 | 100.0% | 182.9% |

a. Dichotomy group tabulated at value 1.

**Part 2:** Do the web scraping and data wrangling of the following websites and show the final cleaned data in a single file:

1. https://data.covid19india.org (two JSON files)
2. https://aqicn.org/city/kathmandu (aqi forecast table)

**Part 3:** •You must search and download the **first 10 free pdf files of "Data Mining"** topic using **Google Scholar (https://scholar.google.com/) (Remove duplicate file/s, if required)**

• Place these 10 pdf files in a folder /directory named "MDS503P2"

• Set your working directory as "MDS503P2" in R using code to work with these files

• Install and use the "pdftools" package to read these ten pdf files from MDS503P2 in R Studio

• Once you read them in R studio then create a "corpus" and perform text mining using step-by-step process i.e. pre-processings (with or without stemming), TDM creation, frequency of the most frequent terms, associated terms of the most frequent term, word cloud using a sensible minimum frequency and topic models with interpretations as per session 11 slides

Note: Do not forget to submit the R Script file, R markdown file and knitted pdf report file of the Project work (Project 2:Unit 2) in Google classroom along with the link of the recorded video explaining each step (**DO NOT UPLOAD THE VIDEO FILE!**)