Practice Assignment 02

Create a GitHub repository called “st2195_assignment_2” that includes:

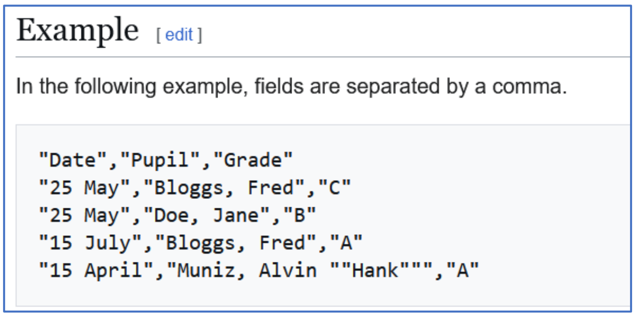
1. a README.md file with a short markdown description of this assignment [1 point]
2. a folder called “r_csv” with a R code for scraping the CSV example in the Wikipedia page https://en.wikipedia.org/wiki/Delimiter-separated_values and saving the resulting output in the local folder (in CSV) [4.5 points]
3. a folder called “python_csv” with a Python version of the code in point 2 [4.5 points]

Note that it is advised to use the packages rvest (R) and BeautifulSoup (Python) for scraping operations. RSelenium (R) and Selenium (Python) can also be used, but they are generally more complicated to setup.

Hint: For more information on rvest, you may want to have a look at https://rvest.tidyverse.org/articles/harvesting-the-web.html

Additional Notes:

- Task clarifications
    - Look for example data in CSV format from Wikipedia URL (see below)



Example [edit]

In the following example, fields are separated by a comma.

```
"Date","Pupil","Grade"
"25 May","Bloggs, Fred","C"
"25 May","Doe, Jane","B"
"15 July","Bloggs, Fred","A"
"15 April","Muniz, Alvin ""Hank""","A"
```

    -
    - Scrape the data and write to a file in CSV format
    - Read the CSV file to a data frame to verify that it was correctly saved

Hints:

- Research on how to use rvest package (in R) and BeautifulSoup (in Python).
- R - Some resources on how to use rvest to scrape data:
    - https://cran.r-project.org/web/packages/rvest/vignettes/rvest.html
    - https://www.r-bloggers.com/2023/01/web-scraping-in-r-2/
    - How to find the XPath to scrape tables in rvest - YouTube
    - You may use the following code to load the rvest package:
      install.packages("rvest") # if package not installed, run this once
      library(rvest)
- Python -- Some resources on how to use BeautifulSoup to scrape data:
    - https://realpython.com/beautiful-soup-web-scraper-python/
    - https://understandingdata.com/posts/web-scraping-with-beautifulsoup/

- You may use the following code to load the packages:

```
from bs4 import BeautifulSoup
import requests
import pandas as pd
```