Fourth Industrial Revolution (4IR) Summer School
Data Preparation – Day 2 Exercises

Question 1

Write a Pandas program to read the following HTML page which contains two tables.

https://www.w3schools.com/html/html_tables.asp

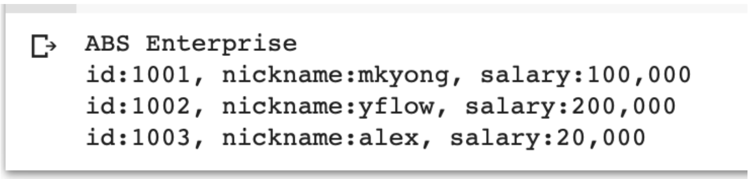
Read the tables to datafreams then write them to two separated HTML files.

Question 2

Write a Pandas program that uses the minidom class to read and parse the given XML file “employees.xml”.

Print the name of the company given in the xml file as well the details of the employees.

Sample of the expected output:



```
ABS Enterprise
id:1001, nickname:mkyong, salary:100,000
id:1002, nickname:yflow, salary:200,000
id:1003, nickname:alex, salary:20,000
```

Question 3

Write a Pandas program that read “Yeh-concrete-data-sklearn” dataset from GitHub repository. The dataset consists of 1030 samples. It has eight independent variables and one dependent variable “Concrete compressive strength(MPa, megapascals”.

- Use the describe() function to provide a statistical information about all the variables provided in the dataset. Then sore the results in a new excel sheet called “**summary**”.
- Calculate the correlation and covariance among the given variables and save the results into flat files.
- Find the average of the “Age (day)” variable. Then retrieve all the data points with Age (day) value greater than the average and store them in json object called “conData1.json” and similarly the data points with Age (day) value less than the average store them in another json object called “conData2.json.