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.

```
import pandas as pd
df = pd.read_html("https://www.w3schools.com/html/html_tables.asp")
df[0].to_html("first_table.html")
df[1].to_html("second_table.html")
print("Two HTML files have been created...")
```

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.

```
#load the XML file first
from google.colab import files
uploaded = files.upload()
```

```
#start loading and parsing the xml file
from xml.dom import minidom
doc = minidom.parse('employees.xml')

# doc.getElementsByTagName returns NodeList
#get the company's name
name = doc.getElementsByTagName("name")[0]
print(name.firstChild.nodeValue)

#get the employees details
staffs = doc.getElementsByTagName("staff")
for s in staffs:
    sid = s.getAttribute("id")
    nickname = s.getElementsByTagName("nickname")[0]
    salary = s.getElementsByTagName("salary")[0]
    print("id:%s nickname:%s, salary:%s" %
        (sid, nickname.firstChild.data, salary.firstChild.data))
```

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.

[ ]
```
# First find the dataset in GitHub
# copy the link and use it in the following command
! git clone https://github.com/maajdl/Yeh-concrete-data-sklearn.git
```

[ ]
```
# read the CSV file with using the correct path
import pandas as pd
cData= pd.read_csv("Yeh-concrete-data-sklearn/Concrete_Data_Yeh.csv")
```

[ ]
```
# a statistical information about the dataset variables
summary=cData.describe()

# export the summary to excel sheet
summary.to_excel("Stat_summary.xlsx")
print("The summary excel sheet has been created...")
```

[ ]
```
# the correlation and covariance
corr_matrix= cData.corr()
cov_matrix=cData.cov()

# to export to csv file
corr_matrix.to_csv("corr_matrix.csv")
cov_matrix.to_csv("cov_matrix.csv")

print("The corr and cov data have been exported...")
cData.dtypes
```

[ ]
```
# find the average of the “Age (day)” variable.
avg=cData.age.mean()

# find the samples with age > the avg

result = cData.loc[cData['age'] >= avg]
result.to_json("conData1.json")

result = cData.loc[cData['age'] < avg]
result.to_json("conData2.json")

print("two json objects have been created...")
```