Introduction to data science - Assignment #2

Due date: Wednesday, Oct 21 at 9pm
Submit your solutions in the same way as with Assignment #1

The 'student.csv' and 'crime.csv' data files attached to the assignment file (available at Sakai) is taken from the UCI repository [1, 2]. The Student Data reports the student achievements at the secondary education of two Portuguese schools. This data includes the following 5 variables:

- 'grade': Final grade of Portuguese course subject (numeric: from 0 to 20)
- 'studytime': Weekly study time (numeric: 1 for less than 2 hours, 2 for 2 to 5 hours, 3 for 5 to 10 hours, or 4 for more than 10 hours)
- 'Internet': Internet access at home (binary: 1 for Yes or 0 for No)
- 'activities': Extra-curricular activities (binary: 1 for Yes or 0 for No)
- 'absences': Number of school absences (numeric: from 0 to 93)

The Crime Data file reports the number of violent crimes per 100,000 population for the communities within the United States. It also includes some socio-economic factors. The variables are as follows:

- 'PctPopUnderPov': Percentage of people under the poverty level (numeric: from 0 to 1)
- 'PctUnemployed': Percentage of unemployed people (numeric: from 0 to 1)
- 'PolicPerPop': Ratio of police officers to the population (numeric: from 0 to 1)
- 'Pcthomeless': Percentage of homeless people (numeric: from 0 to 1)
- 'PctBSorMore': Percentage of people with a bachelors degree or higher education (numeric: from 0 to 1)

- 'ViolentCrimesPerPop': Ratio of violent crimes to the population (numeric: from 0 to 1)

The 'sample-file.txt' file is a simple text file with 5697 lines which is taken from the learning container website [3].

Note. You must put the 'student.csv' and 'sample-file.txt' files in the same folder as your code file. If you use Jupyter notebook it may be at the address 'C:/Users/YOUR-USER-NAME'. You can also read the file with its address, for example:

```
f = open('C:/files/sample-file.txt')
```

Question 1

Fasting blood sugar (FBS) test is normal if it is lower or equal to 99 mg/dl. Write a code to get the FBS test result of the user using stdin, and print a message if it is out of the normal range.

Question 2

The following code each time randomly selects 2 numbers from the list of 1, 2, and 3 and adds each to the end of one of the two text files. Complete the code to compare the numbers added to the files, line by line, and if numbers are the same print 'Success!'.

```
from random import sample

for i in range(0,100):

    a = sample(['1','2','3'],1)
    b = sample(['1','2','3'],1)

    file_a = open(file='file1.txt', mode='a')
    file_b = open(file='file2.txt', mode='a')

    file_a.write('\n'+a[0])
    file_b.write('\n'+b[0])
```

```

file_a.close()
file_b.close()

################## your code here ##################
## you need to
## 1. write a for loop to read the files line by line
## and compare the values


#######################################################
```

Question 3

Write a code that counts the number of lines in 'sample-file.txt' file that include the 'q' character.

Question 4

The following code first reads the student csv file as a data frame, and then obtains the mean value of 'grade' for the students with or without access to the internet ('internet'=0 or 1) as a new data frame. Complete the code to create a bar plot of the average grade of these 2 groups of students.

```
import pandas as pd
import matplotlib.pyplot as plt

data = pd.read_csv('student.csv')

data = data[['internet', 'grade']]

data2 = data.groupby('internet').mean().reset_index()

print(data2)


################## your code here ##################
## you need to
## 1. Get the list of access to the internet and the
## list of corresponding mean grades from data2
## 2. Use these lists to create a plot


#######################################################
```

Question 5

Use a similar code to that for Question 4 to obtain a data frame including the average grade for students with different levels of study time ('studytime'= 1, 2, 3, or 4). Then create a line plot of the average grade for each level of 'studytime'.

Question 6

Use crime data to create a scatter plot of 'PctPopUnderPov' and 'ViolentCrimesPerPop'. Then scale these feature values and create a scatter plot with the new values.

Note. You can get a list of values for each feature and call `scaler.fit_transform()` on it separately. But you need to first convert it to an array and reshape it as below:

```
import numpy as np

crime_list = np.array(data['ViolentCrimesPerPop']).reshape(-1,1)
```

Question 7

Write a code that uses the `requests` and `bs4` packages to get all the paragraphs of the 'data science' page at wikipedia and then counts the number of paragraphs that contain the word 'learning'.

Question 8

Divide the data samples into two categories of students with 'studytime' (i) less than 3 or (ii) equal to or more than 3. Then create a box plot of 'grade' for each category and compare the two plots.

Question 9

Write a code to get the correlation of features in the crime data with the ratio of violent crimes.

Question 10

One definition for an outlier is the value that falls outside $(\mu - 3\sigma\ ,\ \mu + 3\sigma)$, where μ is the average of the values and σ is the standard deviation. Write a function that takes a list of values and a test value and checks if the test value is an outlier and prints a message.

References

[1] Student performance data set. https://archive.ics.uci.edu/ml/datasets/Student+Performance.

[2] Communities and crime data set. https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime.

[3] Large sample text file for download. https://www.learningcontainer.com/sample-text-file/.