SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

ANL252
Python for Data Analytics

Group-based Assignment

July 2021 Presentation

GROUP-BASED ASSIGNMENT

This assignment is worth 20% of the final mark for ANL252 Python for Data Analytics.

The cut-off date for this assignment is **22 August 2021, 2355hrs**.

This is a group-based assignment. You should form a group of **4 members** from your seminar group. Each group is required to upload a single report via your respective seminar group site in Canvas. Please elect a group leader. The responsibility of the group leader is to upload the report on behalf of the group. Those submitting individually will be given a 10 marks deduction.

It is important for each group member to contribute substantially to the final submitted work. All group members are equally responsible for the entire submitted assignment. If you feel that the work distribution is inequitable to either yourself or your group mates, please highlight this to your instructor as soon as possible. Your instructor will then investigate and decide on any action that needs to be taken. It is not necessary for all group members to be awarded the same mark.

Up to 25 marks of penalties will be imposed for inappropriate or poor paraphrasing. For serious cases, they will be investigated by the examination department. More information on effective paraphrasing strategies can be found on https://academicguides.waldenu.edu/writingcenter/evidence/paraphrase/effective.

Note to Students:

You are to include the following particulars in your submission: Course Code, Title of the GBA, SUSS PI No., Your Name, and Submission Date.

Question 1

Given the following data which contain 20 rows and 3 columns: X1, X2, and Y.

X1	X2	Y
4	0.2	1.16
6	0.1	0.06
8	0.3	-1.79
4	0.6	1.55
10	0.1	-4.88
1	0.4	1.37
9	0.6	-1.25
5	0.3	-1.1
2	0.5	3.23
7	0.5	-2.71
8	0.1	-0.99
2	0.9	3.23
2	0.8	4.55
8	1	2.7
7	0.9	-1.13
9	0.1	-0.88
1	0.2	2.08
4	0.2	1.62
6	0.7	-0.9
9	0.7	0.46

Note: Include your Python program code in the answers and show them in the "Consolas" or "Courier New" fonts (size 12). Make a screenshot of the program output if required.

(a) Construct a Python program to store the above data in a NumPy array.

(5 marks)

(b) Suppose a linear regression was fitted on these data. The estimated model is

$$\hat{Y} = 2 - 0.5X_1 + 2.5X_2,$$

where $\hat{Y}$ is the predicted (or expected) value of Y, X_1 and X_2 are the observed values of the columns X_1 and X_2. Design a Python program to compute $\hat{Y}$ for every row of the array and store the results in a separate NumPy array as well.

(5 marks)

(c) The residuals of the model ê are calculated by:

$$\hat{e} = Y - \hat{Y}$$

where Y is the actual value stored in the original NumPy array and $\hat{Y}$ is the predicted value of Y computed in (b). Use a Python program to compute ê for every row of the array and store the results in a separate NumPy array.

(5 marks)

(d) One of the main assumptions for linear regression is that the residuals must be normally distributed with zero mean and constant variance. Create a histogram of the residuals calculated in (c) by using the matplotlib package. Adjust the parameters of the chart so that the ticks on the x-axis can be read clearly, a title is given to the chart, and both the axes are labelled. Eventually, discuss whether you agree that the normality assumption with zero mean (the checking of constant variance is not required here) is valid based on this histogram.

(10 marks)

(e) The constant variance assumption can be checked by a scatter plot in which the x-axis represents the values of the predicted values $\hat{Y}$ and the y-axis represents the residuals $\hat{e}$. If the scatter plot does not show any pattern and the values of all the data points are more or less on the same level. Write a Python program to create such a scatter plot for checking the constant variance assumption. Adjust the parameters of the chart so that the ticks on both axes can be read clearly, a title is given to the chart, and both the axes are labelled. Eventually, discuss whether you agree that the constant variance assumption is valid based on this scatter plot.

(5 marks)

Question 2

The data of 19 students in a secondary school class are stored in a .csv data file named "class.csv". The gender, age, height, and weight are the features of the students that have been recorded. Employ your Python programming skills to carry out the tasks below.

Include your Python program code in the answers and show them in the "Consolas" or "Courier New" fonts (size 12). Make a screenshot of the program output if required.

(a) Prepare a Python program to read in and to convert the data from a .csv text file into a pandas DataFrame. Check the existing missing data in the dataset and adjust the reader accordingly.

(6 marks)

(b) The data should be sorted by the age of the students in the descending order and then by their gender in the ascending order. Employ the corresponding Python syntax to carry out this task.

(8 marks)

(c) Identify the location of the missing values in the DataFrame. Report the rows and columns where the missing data are found.

(5 marks)

(d) If missing values are detected in the DataFrame, they have to be treated according to the columns they belong to. Here are the instructions of how we should deal with the missing data in each column:

Gender – replace missing values by the gender with the highest frequency
Age – replace missing values by the median age

Height – replace missing values by the mean height
Weight – replace missing values by the mean weight

Design your own Python program to determine the corresponding statistics for each column to replace the missing values in it.

(16 marks)

(e) Use Python code to detect outliers in the DataFrame and delete the corresponding rows if they exist.

(15 marks)

Question 3

Explain some differences between inner and outer join when merging two or more DataFrames and how they should be carried out using the pandas package (max. 200 words).

(20 marks)

---- END OF ASSIGNMENT ----