Assessment Task 5: Advanced Data Analysis

Aim

The goal of this assignment is to build on the programming skills you have learned in this subject. This assignment focusses on more advanced data analysis and visualisations, and some aspects will require you to independently learn more about the libraries introduced in the subject. This assignment will require the use of the pandas library specifically, and will ensure you are able to use some key elements of that library.

Task

You are to read through the task description below and plan and implement a solution in Python 3 (any submission in Python 2 or any other language is unacceptable). You are required to submit an IPO chart for each function (see Assessment One for examples) but you do not need to provide detailed pseudocode or flowcharts for this assignment.

Detailed instructions

These are the same instructions as for Assessment Three, with minor clarifications on implementation.

We have been asked to create a program that will allow users to load sets of data from CSV files and then view, manipulate and perform simple statistical analysis on the data.

When the program is loaded the user will see the following menu:

Welcome to The DataFrame Statistician!

Please choose from the following options:

1 – Load data from a CSV file

2 – Analyse

3 – Visualise

4 – Quit

Option 4 will exit the program, every other option will do some task and then display the menu again until the user chooses 4 at this menu.

If the user enters anything other than a value between 1 and 4 an error message is to be displayed.

This program will begin with an empty DataFrame in memory, and then perform a series of operations on this data.

Option 1 – load data from a csv file

This option will ask the user for the name of a CSV file. This file will be loaded into a DataFrame. If the file does not exist then an error message will be displayed before returning to the main menu.

Your program should be able to handle any file in a format like the following:

day,min_temp,max_temp,rainfall,humidity

1,11,23,3,55

2,13,25,0,60

3,9,19,17,80

4,9,18,36,85

5,15,25,0,50

6,12,22,0,60

7,13,23,0,6

So the first row should be the names of the columns, and the following rows should consist of the data. **Your program should not be hardcoded to deal with the weather format above**, it should work with any file with this format.

Option 2 – analyse

This option will use the methods of the pandas library to produce a statistical report as shown below:

Rainfall

Number of values (n): 20

Mean: 35.81

Standard Deviation: 16.12

Std.Err of Mean: 3.52

Option 3 – visualise

When this option is chosen a new menu will be displayed to the user, asking them which type of plot to generate. They are to be given the choice between a line, bar, or box plot. After choosing their visualisation they will be asked if each column of the DataFrame should be plotted on a single plot, or if they should be generated as subplots.

Important note on visualisation

If you are using Jupyter Notebook to write your code then you need to add the following lines at the start of your program:

%matplotlib notebook

If you are using PyCharm (or another editor) then you need to do 2 things:

1 – import the matplotlib library at the start of your program

2 – use the pyplot.show() method of the matplotlib library at the end of your program

This may look like this:

import matplotlib as plt

.

.

.

plt.pyplot.show()

Submission

Your completed solution should be submitted to JCU Online as a single compressed zip file (*.zip). This file needs to contain one Python 3 (*.py) file and one Word (*.docx) or PDF (*.pdf) file. Please name your file LastnameFirstname.zip (for example John Smith's file would be named SmithJohn.zip). You should follow the same filename protocol for the planning and code files.