



ASSESSMENT 1 PART A

PRACTICAL 5

Weighting: 6% (this is the fifth of five practical assessments totalling 30%).

Due Date: Sunday 11:59 pm, end of Week 5.

Introduction: Plotting Panda Data Frames, manipulating data and plotting

Purpose

The purpose of this assessment is for you to demonstrate your ability to apply what you have learnt each week in the creation of a program to solve a problem.

Your tasks

Q1. Plotting Panda Data Frames with Date indexes

Demonstrate your understanding Date indexes and plots by completing the following in a Jupyter Notebook:

- 1) Create six time series stored in a Pandas data frame:
 - a) The data frame must have an index that is a range of dates from 2016-01-01 until today.
 - b) Each column contains one set of random numbers in a range 0 to 1 (each column will have one random number for each date).
- 2) Calculate and print the number of days from 2016-01-01 until today.
- 3) Show the first 30 rows of your data (using a Pandas data frame function).
- 4) Calculate the cumulative sum of the Data frame (using the `<DataFrame>.cumsum()` function) and store it in a new data frame called: `df_accumulate`. Display the first five rows of your data to verify the values.
- 5) Plot the six sets of row values of `df_accumulate` in a graph
- 6) Plot the six sets of row values of `df_accumulate` in separate graphs using subplots then plot the accumulate sum of each of the columns (random numbers).

Your creativity and style will be evaluated.

Q2. Manipulating mental health data

Download the `mental-health-in-tech-survey.csv` dataset. This dataset is from Open Sourcing Mental Illness and the data was taken from a 2014 survey that measures mental health and frequency of mental health disorders of workers in the information technology workplace. This dataset has 27 columns. Go to the [Kaggle website](#) and look at the column categories for this data.

a) In your Jupyter Notebook read this file data into a Pandas data frame. Be sure to parse the Timestamp as a date time rather than a string.

Print the first five rows to show that the data has loaded and print the total number of rows using the shape property.

Write Python code to return the following data:

- Select all rows where the Age is less than 1
- Select all rows where the Age is greater than 99

b) Look at the Pandas data frame `drop()` function in the [Pandas documentation](#) Drop can be used to remove data from the data set.

Using your solutions from part a, drop the data values from your data set where the age is less than 1 or greater than 99.

- Print the shape of your new data frame using the shape property.
- Has the number of columns changed?
- Has the number of rows changed?
- Explain why each has or has not changed.

Hints:

- The `drop()` function takes the indexes you want to remove, which can be rows or columns. In this practical you need to remove rows. To get the index of the rows matching the criteria, you would use `.index` i.e. `<select rows>.index`
`<DataFrame name>.drop(<DataFrame>[criteria or condition].index,inplace=True)`

- As a mutable object, remember that you will not actually change the original data frame, but instead, `drop()` will return a new data frame with the data dropped. To make `drop()` change the original you can add the parameter: `inplace=True` to your drop call.
- For example, suppose that your data frame is named as `df_data` and you want to remove by the column 'salary', only the rows where the salary is lower than 0. This is the code remove for that criteria.
`df_data.drop(df_data[df_data.salary<0].index,inplace=True)`

c) Print the maximum, mean and minimum age values in your data.

Q3. Plotting mental health data

For all of the plots in this section, you should select appropriate titles, axes labels, and colours to make the plots easy to understand.

- Create a box plot showing the distribution of ages for 1) people with no family history of mental illness, and 2) people with a family history of mental illness.
- Create a horizontal bar plot showing the number of respondents for each country who said their employer discussed mental health as part of an employee wellness program.

Q4. Cleaning data

- Have a look at the survey responses for gender. You will see that there are a wide variety of ways that respondents identify their gender. Some of these are clearly different ways of indicating the same gender, for example: 'female', 'F', 'f' and 'Female'.

For the Gender data, change all instances that match any of the following to 'Male'

'male'

'M'

'm'

‘Make’

‘Man’

‘Mail’

‘Malr’

‘msle’

‘Mal’

‘maile’

‘Male ‘

‘Cis Male’

‘Male (CIS)’

‘cis male’

‘Cis Man’

‘Guy (-ish) ^_^’

Change all instances that match any of the following to ‘Female’

‘female’

‘F’

‘f’

‘Female ‘

‘Femake’

‘woman’

‘femail’

‘Woman’

'Cis Female'

'cis-female/femme'

'Female (cis)'

b) plot using two pie plots:

- the proportion of Males that have sought/not sought treatment for a mental health condition
- proportion of Females that have sought/not sought treatment for a mental health condition

Directions

Submit your assessment via online submission, as a Jupyter Notebook, providing your responses to questions in this Assessment brief. This assessment will be written in Python.

Save your notebook and upload it through the assignment link. Your tutor will assess your work based on the rubric. Make sure you have a look at the rubric so you understand what your tutor will be looking for.

Requirements

You must submit your Jupyter notebook file (containing your Python code) through the Canvas submission link.

Rubric

| Assessment Criteria | Level of performance | | | | | |
|---------------------|--|--|---|---|--|-----------|
| Q1 | High Distinction (5/5) All six correct. Plots have appropriate titles, axes labels, legends and colours. | Distinction (4.2/5) All six correct. Some aspects of plot incomplete. | Credit (3.7/5) Five correct. | Pass (3.2/5) Four correct. | Fail 0/5) Less than 4 correct. | /5 |
| Q2 | High Distinction (5/5) Correct answers for all questions. | Distinction (4.2/5) Correct answers highlighted parts for a and c and most of b (including code parts) but may have minor error. | Credit (3.7/5) Correct answers highlighted parts for a and c and non-code parts of b. | Pass (3.2/5) Correct answers highlighted parts of at least a and c. | Fail 0/5) Does not correctly answer highlighted parts of both a and c. | /5 |

| | | | | | | |
|----|---|---|---|---|---|----|
| Q3 | High Distinction (5/5) Boxplot and bar plot meet specification and have appropriate titles, axes labels and colours. | Distinction (4.2/5) Boxplot and bar plot meet specification and have mostly appropriate titles, axes labels, and colours (some aspect missing). | Credit (3.7/5) Boxplot and bar plot meet specification. | Pass (3.2/5) Boxplot and bar plot included with at least one showing correct selection of data. | Fail 0/5) Boxplot or bar plot missing. | /5 |
| Q4 | High Distinction (5/5) All instances of indicated data is changed. Pie plots correctly reflect the data and are labelled and include percentages in wedges. | Distinction (4.2/5) Most data cleaned. Pie plots are labelled and include percentages in wedges. | Credit (3.7/5) Evidence of at least some data cleaned. Pie plots are labelled and include percentages in wedges. | Pass (3.2/5) Evidence of at least some data cleaned and at least one pie plot of data is shown. | Fail 0/5) Data uncleaned or pie plot missing. | /5 |

| | | | | | | |
|------------------|-----|--|---|---|---|------------|
| Jupyter Notebook | N/A | Distinction (5/5) All problems include a header related to what data is being generated and an explanation of what is shown in the plot and what conclusions can be drawn from the plot. | Credit (3.7/5) All problems include a header related to what data is being generated and a description of the data in the plot. | Pass (3.2/5) All problems include a header and a description. | Fail 0/5) Notebook contains only the code solution. | /5 |
| TOTAL | | | | | | /25 |