IC152: Assignment 5
Data statistics, Dictionary, Recursive Functions and Plots

Problems 1 and 2 require you to submit two kinds of files: 1. Python code 2. Output images of the plots. Plot images can be created using any technique, for instance, screen/window capture or use matplotlib.pyplot.imsave() instead of matplotlib.pyplot.imshow(). Nothing is to be submitted for Problem 0.

If you are solving this assignment in the A11’s PC Lab: Keep Fn + F9 pressed during the start of your machine (do not repeatedly press, keep it continuously pressed), and then select the second option with “ubuntu”. Please check if you are able to login to moodle, else change the machine.

Problem 0: Import and print the Data.csv file[1] using the following snippet of code. Download Data.csv first into your working directory on your laptop.

```
######### code starts ###########
import pandas as pd
# Use pandas to read the CSV
csvData = pd.read_csv(’Data.csv’,sep=’,’)
# Convert dataframe into a numpy array
csvDataNum = csvData[[’State’,’A’,’B’,’C’,’D’,’E’,’F’,’G’,’H’,’I’,’J’,’K’]].to_num
# Convert numpy array into a list (of lists)
data = csvDataNum.tolist()
# Access values as usual from data # For eg. data[0][3] is 25 print(data[0][3])
```

[1] For more background information and the source of the given data, see Climate Vulnerability Assessment for the Himalayan Region Using a Common Framework, https://dst.gov.in/sites/default/files/. For West Bengal, only the hilly region of West Bengal has been considered.

######### code ends ###########

The column names given as A, B, C, D, E, F, G, H, I, J, K in the Data.csv file have the following meaning:

A Percentage crop area insured under all Insurance Schemes (2013-15)
B Percentage farmers taking loans (2015-16)
C Average person days per household under MGNREGA (2006-2016)
D Average Percentage area with > 30% slope
E Road Density
F Population density (2011): person/sq. km
G Percentage of marginal farmers (2011-12)
H Livestock to human ratio (2017-18)
I Per Capita Income (2014-15)
J Number of Primary Health Centres per 100,000 households (2017-18)
K Percentage of women in the overall workforce (2011)

Problem 1: Use data from Problem 0 for this question.

a. Write functions working on data which provide information on i) population density, ii) percentage of marginal farmers and iii) percentage of women in the overall workforce, by giving:

    - Highest = the state with the highest value
    - Lowest = the state with the lowest value
    - Median = the median
    - Average = the average value
    - Mode = the value with highest frequency

    Don’t use predefined pandas methods for these functions. Print the information in a way you find best represented.

b. For the given order of states in the first column of the csv, create a single bar chart which takes as variables the:
    i. the percentage area with slope > 30%
    ii. the road density

   For each state the 2 values should be displayed next to each other.
c. Create a bar chart for the states ordered by increasing percentage area with slope > 30% showing the road density.

Problem 2: Consider the following functions:

$$f(n) = \begin{cases} 1 & \text{if} \quad n < 2 \\ 1.65\, f(n-1) & \text{if} \quad n \geq 2 \end{cases}$$

$$g(n) = \begin{cases} 1 & \text{if} \quad n < 2 \\ g(n-1) + g(n-2) & \text{if} \quad n \geq 2 \end{cases}$$

$$h(n) = \begin{cases} 2 & \text{if} \quad n < 2 \\ 2\, h(n-2) & \text{if} \quad n \geq 2 \end{cases}$$

$$k(n) = \begin{cases} 3 & \text{if} \quad n < 3 \\ k(n-1) + k(n-3) & \text{if} \quad n \geq 3 \end{cases}$$

a. Implement f, g, h and k as recursive Python functions.
b. Create a scatter plot for each of the functions. The values of the x-axis should range from 0 to 9.
c. Display the curves of f, g, h and k within one plot. Add the legend, that is, for every curve indicating which of the functions it represents. Again, for the range of the x-axis use the range from 0 to 9. After that try larger ranges as well. What are your

observations? Write down the observations using “print()” function calls.

Problem3: Write the following recursive functions:

a. prefRevCapStr( ), which prefixes a string with its capitalized reversal separated by an arrow (with a blank before and after the arrow):
For example, prefRevCapStr("Holi-to-come") should return the string "EMOC-OT-ILOH -> Holi-to-come"
Hint: Try thinking of a function of the string at n^{th} instance as the recursive sum (concatenation) of the strings at <= $(n-1)^{th}$ instances.

b. scatSubStr( ), which takes as input two strings, w and s and yields "yes" if w occurs in s as a scattered substring and "false" otherwise. A string w occurs scattered in a string s if it may be obtained by deleting some of the letters of s. For example, abb occurs scattered in cadbebb (delete c, d, e and one of the occurrences of b).

Problem 4: Run the script with the name “problem5.py”. Consider the equation: y = mx + c where we want to learn m and c, for the line that fits the given data. Now for each point (x_i, y_i) in the data, we can minimize the sum $(mx_1 + c - y_1)^2 + (mx_2 + c - y_2)^2 + ... + (mx_n + c - y_n)^2$ with respect to m and c, by taking the corresponding derivatives. Take pen and paper and solve for m and c (you will get interesting equations in terms of mean and standard deviation/variance of x_i’s and y_i’s). Write the code to find m and c, and add the code to plot the line that fits the given data (show data and line in the same plot).

Extra Problems:

1. In Problem 1: Do more analyses concerning the Indian Himalayan states with the data provided in the Data.csv file. For example, you might want to know whether a higher percentage of women in the overall workforce leads to a higher per capita income.
2. In Problem 2: Choose your own functions like the square function, exponential functions etc and plot them.
3. In Problem 3 b: Refine your solution by counting the number of scattered occurrences of w in s.
4. In Problem4: Find m and c which minimize the sum: $|mx_1 + c - y_1| + |mx_2 + c - y_2| + \ldots + |mx_n + c - y_n|$. Plot the line and print what you observe.

Bonus questions:

1. You have the data in file Data.csv of various indian states. Take the columns of ‘percentage of farmers taking loans’ and ‘per capita income’.
a. Calculate and print the correlation coefficient between these two columns. Determine if there is a relationship between these two variables.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

b. Create a scatter plot to visualize relationship between these two variables.

Create the folder having your python files, with name having your roll number followed by “_assignment5” (don’t use inverted commas in folder name), compress the folder with .zip extension and submit it on moodle.

Make sure that you delete all your files from the lab PC/Laptop, and shut it down before you leave.