COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH

Information Visualization

CS5764

LAB # 1

In this LAB, you will learn how to use Python packages to make a simple line plot. You will learn how to add title, legend x-label and y-label to your graph. You will also practice finding the Pearson's correlation coefficient for two random variables. [All numbers must be displayed with [2-digit precision]. Set np.random.seed(5764).

1- Develop a python program that asks a user to input two numerical numbers [use input() function]: mean, variance & number of observations. Then Using the NumPy package in python create a normally distributed random variable x with the zero mean [default], variance [1]. Create another random variable y normally distributed with the mean [5] and variance [2]. Number of samples for both x and y = 1000. Display the first three values of x and y on the console. [10pts]

2- Write a python program that calculates the Pearson's correlation coefficient between two random variables x and y in question 1 with default values. Hint: You need to implement the following formula [take a screen shot of the developed function and place it into your report]. Calculate the sample Pearson's correlation coefficient between x & y [previous question] using the developed function and display the number on the console. Does the correlation coefficient between x & y make sense? Explain your answer. [10pts]

$$r = \frac{\sum(x_t - \overline{x})(y_t - \overline{y})}{\sqrt{\sum(x_t - \overline{x})^2}\sqrt{\sum(y_t - \overline{y})^2}}$$

3- Display a message on the console that shows the following information: [10pts]
   a. *The sample mean of random variable x is :*
   b. *The sample mean of random variable y is :*
   c. *The sample variance of random variable x is :*
   d. *The sample variance of random variable y is:*

4- Using the matplotlib package in python display the line plot of the random variable x and y in one figure differentiating x and y with legend. Add an appropriate x-label, y-label, title, and legend to each graph. Hint: You need to us the plt.plot() if [10pts]

```
import matplotlib.pyplot as plt
```

5- Using the matplotlib package in python display the histogram plot of the random variable x and y in one figure differentiating x and y with legend. Add an appropriate x-label, y-label, title, and legend to each graph. Hint: You need to us the plt.hist() if [10pts]

```
import matplotlib.pyplot as plt
```

In this section of the LAB, you will work with the real dataset and Panda's package. The Pandas package will be explained in detail in later lectures. For this section of the LAB, you will need to use the following package:

```
import pandas as pd
```

6- Using pandas' package in python read the 'tute1.csv' dataset [GitHub]. Read the dataset provided by the URL. The set is the timeseries dataset with Sales, AdBudget and GDP column. [5pts]

7- Using the developed function in question 2, find the Pearson's correlation coefficient between features and display a message on the console that shows the followings: [10pts]

   a. *The sample Pearson's correlation coefficient between Sales & AdBudget is:*
   b. *The sample Pearson's correlation coefficient between Sales & GDP is:*
   c. *The sample Pearson's correlation coefficient between AdBudget & GDP is:*

8- Display the scatter plot between Sales & AdBudget. Include the correlation coefficient for these two features in the title. Add the x & y label and grid to the plot. Does the graph make sense considering the correlation coefficient? Explain your answer. [5pts]

9- Display the scatter plot between Sales & GDP. Include the correlation coefficient for these two features in the title. Add the x & y label and grid to the plot. Does the graph make sense considering the correlation coefficient? Explain your answer. [5pts]

10- Display the scatter plot between *GDP* & AdBudget. Include the correlation coefficient for these two features in the title. Add the x & y label and grid to the plot. Does the graph make sense considering the correlation coefficient? Explain your answer. [5pts]

11- Display the line plot of Sales, AdBudget and GDP in one graph versus time. Add an appropriate x-label, y-label, title, and legend to each graph. [10pts]

12- Plot the histogram plot of Sales, AdBudget and GDP in one graph. Add an appropriate x-label, y-label, title, and legend to each graph. [10pts]

Submission guidelines:

- The softcopy of the developed Python code .py must also be submitted separately.
- Please make sure the developed python code runs without any error by evaluating it through PyCharm software. The developed python code with any error will subject to 50% points penalty.
- Add an appropriate x-label, y-label, legend, and title to each graph.
- Submission: report (pdf format) + .py . The python file is a supporting file and will not replace the solution.
- Only a report will be graded not the python file.
- The python file must regenerate the provided results inside the report.
- Report submission without python file, receives zero.
- Python file without report, receives zero.