CS 6103E Software Systems Laboratory

Monsoon Semester (2024-25)

1st Year MTech CSE/CS(IS)

Python Programming - Practise set 2

Submission date: on or before 30.09.2024 (11:59 pm)

Mode of submission: Create a folder named in the format P1_<ROLLNO> (e.g., *P1_saritha_m240064CS*). Place all the required files inside this folder, then compress it into a .zip file and upload the zipped folder.

Instructions/Information:

- The questions are designed mainly to focus on exception handling, strings, and data visualization topics.
- If you have any questions or need clarification, we have set up a Google document (clickhere) where you can post your queries by Saturday (i.e., 28/09/2024) 12 PM and we promptly provide responses in the same document.
- You can practice programs in a Jupyter notebook environment.
- Each machine in the NSL lab will be equipped with a Jupyter notebook environment for the Monday and Tuesday tests.
- You may refer the following useful youtube channel for learning more: Telusko - YouTube
- In addition to the three problems listed below, the Monday exam will also include a question based on a variation from the Pandas library.

Programming Questions:

Problem I: Define functions to catch and handle the following exceptions appropriately. Note: Refer "exception handling.ipynb" from eduserver to perform the tasks.

1. Take email id as input from the user and validate if the email contains an "@" symbol and any valid domain extension i.e. ending with ".xyz" where "xyz" can be of any length (e.g., "com", "in") and there should be at least one character between "@" and "." . Raise an exception for any email not following this format.

2. Take the phone number of the user as input and ensure that the phone number is exactly 10 digits long and contains only numeric characters. Raise an exception if it contains non-numeric characters or the length is not 10. If the length is incorrect, make a CustomError and invoke that and if the number is not exactly 10 digits long, invoke ValueError. If there are no errors, print "correct".

3. Create a dictionary specifically by taking input from the user. The dictionary should have 4 key-value pairs namely: {1: "Apple", 2: "Banana", 3: "Orange", 4: "Peach"}. Now take a key as an input and

validate if it is an integer and print the corresponding value for the key. Also validate whether the key exists in the dictionary. Print the dictionary in the end regardless of if there is any exception or not.

Problem II

1. **Wildcard Pattern Matching:** Given two strings, a source string *s* and a pattern string *p* that can include two special characters: '*?*' (matches any single character) and '*' (matches any sequence of characters (including an empty sequence)). Determine if the pattern *p* matches the entire source string *s*.

   Input: First input: string *s* (source string) and Second input: string p (pattern string)

   Output: True if the pattern matches the string, *False* otherwise.

   Example input: Enter the source string: abcdef

   Enter the pattern string: a*d*f

   Output: Pattern matches string: True

Problem III

The objective this problem is to learn the following:·

- Visualizations using discrete and continuous graphs: Bar plots and continuous plots
- Identifying outliers.
- Understanding basic data exploration: Data distributions.
- Understanding basic statistics: Confidence intervals and Standard errors.

1) Visualizing the data improves the understanding we have about the data, there are different types of visualizations, explanatory, exploratory, statistical, interactive, etc. In explanatory visualization we can use the data to plot discrete graphs such as histograms/bar graphs, another approach is to use continuous graphs such as line graphs. Using appropriate plots is key to visualizations. For example, the height of a person in each month can be represented using both discrete and continuous plots, whereas marks of a person in various subjects can be visualized only using discrete graphs.

   The *student.csv* file contains details of 300 undergraduate students who have enrolled for tuition. It recorded details such as roll number, name, age, and weights in the beginning of each trimester,

mid-test, final, and total marks for the subjects physics, chemistry, math, and english. The mid-tests are out of 20 and final tests are out of 80. You may read the dataset *student.csv* using pandas and perform the following tasks **(Hint**: Use matplotlib):

a) Plot the total marks secured for each subject by the student *"RN_15"* using bar graphs.
b) Plot the line graph for the student *"RN_15"* weights in the year.

2) Outliers/Anomalies are termed for unusual data observation values in a data field.
   a) Z-score is a statistical approach used to identify outliers, computed as:

$$Z\text{-}score_X = (X-\mu)/\sigma,$$

Here $Z\text{-}score_X$ is the Z-score value for observation X, X = current data observation value for the attribute, μ = mean of the attribute, and σ = standard deviation of the attribute.

Outliers of any attribute are identified by values with a *Z-score* outside the threshold ±3. This means that for data points with *Z-score* greater than +3 or less than -3 (note that *Z-score* is mostly applied to identify outliers in Gaussian/Normal distributed data). Now for the student data using *student.csv* file, identify and print the outliers/anomalies for the attribute ***Age*** using *Z-score*.

A. After discarding the outlier values identified using *Z-score* from the attribute ***Age***, plot a pie-chart representing the percentage of students enrolled for the tuition under different age groups as mentioned below:
   a. Students with age≤18, 18<age≤21, 21<age.

3) Data distribution refers to the way data points are spread out or arranged over a range of values. It is a description of the frequencies or probabilities of data values or intervals in a dataset. Depending on this observation they are categorized into Normal, Exponential, Poisson, Log-normal, Binomial, Uniform etc. Each distribution has its own set of properties based on which we can calculate or extract information. For example, normal distribution follows a bell shaped curve, where most of the data values observed concentrate around the mean value. If the spread of values (range) from the mean is symmetric then it is a normal distribution without skewness. If this spread is asymmetrically high on the right side (values greater than mean) the distribution is positively skewed (long right tail). If it is high on the left side (values lesser than mean) the distribution is negatively skewed (long left tail).

a) Read the *student.csv* data, plot the data distribution of the attribute ***Age*** with kernel density plot (seaborn.kdeplot). What is the type of data distribution obtained? (Normal/Uniform/Binomial/Exponential/Poisson/Log-normal). Is the data skewed?

b) *bike*.csv dataset tracks the number of bicycle rentals in a city on an hourly basis (where rentals for the hours 0-23 when combined will give the rentals on that day). The attributes of interest are ***Hour***: The hour of the day, ***Count***: The number of bike rentals in that hour. Additional features are also included to study the factors affecting bike rentals. Read the *bike*.csv, and plot the line plot with the number of bike rentals in a day. What is the type of data distribution obtained? (Normal/Uniform/Binomial/Exponential/Poisson/ Log-normal)

c) *heart_transplant*.csv dataset contains data about survival times after heart transplants. The attributes of interest are ***Age***: The age at which heart problem occurred, ***Age of death***: The age at which the patient died, ***Transplant***: 'control' refers to patients who didn't undergo a transplantation, 'treatment' refers to patients who are subjected to transplantation. Read the *heart_transplant.csv*, and convert the non-numeric attribute ***Transplan***t into numeric **(Hint: Replace 'treatment' with 1 and 'control' with 0).** Then using a bar plot, compare the mean survival times of patients who have undergone transplant with patients without transplant. Is there any advantage obtained through transplantation? (**Hint**: You may need to create a new attribute *survival time* by taking the difference between their age of death and their age of heart problem, and then filter the patients who have undergone transplant).