STAT 650 Assignment-02

Instructions :

- This assignment is based on materials coverved in Lectures 04, 05 and 06.
- We highly recommend that you write your solutions in **Jupyter Notebook** and convert them to a **PDF** file. However, you may write the solutions by hand, scan and upload it as **.pdf** file.
- The PDF file should be under 15MB in size. It must be uploaded as a single file and not separate files for separate pages. Do not take a photo of each page and then paste them into a document - this will make your file too big and the results will generally not be very readable anyway.
- Please make sure that the solutions are neat, legible and in order (even if you choose to solve them in different order).
- Include **STAT650--UIN** at the top of the first page.
- Name the file as **UIN_assign2.pdf** (For eg, if someone's UIN is 123456789, then the file should be named 123456789_assign2.pdf). Otherwise, your submission will not be graded.
- You should upload your file through Canvas, by 11:59PM U.S. Central time, on the due date. You can make multiple submissions within the due date, but only the latest submission will be considered for grading.
- You may take 6 hours extra after the due time, but 10% of your marks will be deducted.
- It is strictly prohibited to share or distribute the content in this document.

The aim of this assignment is to get familiar with Python operations and concepts in exploratory data analysis.

Problem 1:

Use the **dataset1.xlsx** to answer this question.

Data description:

This dataset contains worldwide internet usage of 39 countries in 2003 published by the United Nations. The aim is to assess the influence of four different factors - GDP, CO_2 , number of cellular phone subscribers and fertility on the internet usage.

The variables in the dataset are:

1. **Nation**: Name of different nations
2. **Internet**: The percentage of adult residents who used the Internent
3. **GDP**: Gross Domestic Product, per capita in thousands of U.S. dollars
4. **CO2**: Carbon dioxide emissions, per capita
5. **Cellular**: Percentate of adults who are cellular phone subscribers
6. **Fertility**: Mean number of children per adult women

Note: The variable $\mathbf{Nation}$ is given here as a supportive information only.

Question:

(a) Find the datatypes of the variables in the dataset. Are there any missing values?

(b) Compute the summary statistics for each of the variables: '**INTERNET**', '**GDP**', '**CO2**', '**CELLULAR**', '**FERTILITY**'.

(c) Describe the shape of the distribution of each variable using a suitable graphical method. Justify your choice.

(d) Check for the normality of each numeric variables.

(e) What is the target varible of this dataset ?

(f) By using a suitable graphical representation, visualize associations between the target variable and other variables.

(g) Compute the correlation between variables: '**INTERNET**', '**GDP**', '**CO2**', '**CELLULAR**', '**FERTILITY**' using the **Pearson's**, **Spearman's rank** and ϕ_k correlation coefficients. Create heat maps to visualize the correlation matrixes.

(h) How does the target variable associate with the other variables (linear and/or non-linear)? Justify your answer.

(i) Which variable has the highest association with the target variable? Plot the target variable against the variable you selected using `lmplot`.

[Marks: $5 \times 10 = 50$]

Problem 2:

The file **dataset2.pkl** contains a simulated dataset having 100 rows and 5 columns.

(a) Read the dataset **dataset2.pkl** as a dataframe and name it "**df1**". Set the column names to: **"A", "B", "C", "D", "E"**.

(b) Get the summary statistics (**only mean, median, and standard deviation**) of each column and save it as **"df1_sm"**.

(c) For column **"B"** in **"df1"**, find which rows have values **greater than 1.00**?

(d) Create a dataframe **df2** where the values greater than 1.00 in column "**B**" of **df1** are replaced with **NaN** (Not A Number). NaN is a particular data type. Do not replace with the string "NaN".

(e) Calculate the number of NaN values per column of **df2**.

(f) Create a new dataframe **"df2_imputed"**, where the NaN values in **df2** are replaced with the corresponding **column mean** (this procedure is called mean imputation). Compute its summary statistics (only mean, median, and standard deviation) and save them as **"df2_sm"**.

(h) Create a dataframe **"df3_imputed"** where the NaN values in each row of **df2** are replaced with the corresponding row mean. Compute summary statistics (only mean, median, and standard deviation) and save them as **"df3_sm"**.

(i) Creat a boxplot visualization to compare the columns in the three datasets "df1", "df2_imputed", and "df3_imputed".

(j) Which procedure do you think is better: column mean imputation or row mean imputation? Justify your answer.

[Marks : $5 \times 10 = 50$]