**Lab 5: Analysis of “India Power Infrastructure Data (2004-2021)”**

Preamble: Having become familiar with Python programming and the use of various libraries such as Pandas, NumPy and Matplotlib, this exercise requires you to perform various data-scientific tasks to analyze the Indian Power Infrastructure data. The tasks include:

* Preliminary data cleaning (checking for duplicate entries/ missing entries/ incorrectly listed entries etc.)
* Understanding the nature of the data (through statistical analyses)
* Exploratory Data Analytics (using visualizations such as graphs and interactive plots)
* Drawing inferences and forecasting

Such tasks listed out in this assignment will have to be routinely conducted for all datasets hereafter, in this course.

You are required to prepare a complete report of all analyses conducted (e.g. what was the objective, what parameters/ features/ columns have been chosen, what is your interpretation of the results, overall conclusions) including the code and the outputs. You can use your Jupyter/colab or any other notebook as the report, ensuring that it is well commented and formatted (e.g. refer to the .ipynb files shared in the Python course).

**Step 1:** Read the description of the dataset provided below for this exercise.

**Context:** This Dataset is extracted from the RBI(Reserve Bank Of India) Annual Publication "HANDBOOK OF STATISTICS ON INDIAN STATES" and is open to be used for research and educational purpose. The particular dataset contains a CSV(Comma Separated Values) file on Indian Statewise Electricity Infrastructure from 2004 to 2021. To be used by researchers, analysts, professionals and students to carry out various projects and research especially for Time Series Analysis and Forecasting.

**Column description**

* State/Union Territory: States and Union Territories of India
* Year: Year from FY2004-2005 to FY2020-2021
* Power Requirement Net Crore Units: Power Requirement by each state in Net Crore Units
* Availability Of Power Net Crore Units: Power Availability for each state in Net Crore Units
* Availability Of Power Per Capita kiloWatt-Hour: Availability of PPC in kiloWatt-Hour
* Installed Power Capacity MegaWatt: Installed Power Capacity in MegaWatt

**Note**:

i) Per Capita Availability of Power is worked out based on Census Population 2011 and for the population for Andhra Pradesh and Telangana drawn from both Governments’ portals for the years 2014-15 and 2018-19.

ii) Figures of Telangana are w.e.f. June 2014 due to bifurcation of Andhra Pradesh into Andhra Pradesh and Telangana w.e.f. June 2, 2014.

**Acknowledgements**

Database on Indian Economy: https://dbie.rbi.org.in/DBIE/dbie.rbi?site=home

Central Electricity Authority, Ministry of Power, Government of India: <https://cea.nic.in/?lang=en>

**Step 2:** Perform the following tasks on the data

1. Load the .csv file into a Pandas dataframe and print the column headers of the dataframe to ensure that the file contents match the description provided in step 1.
2. Using inbuilt Pandas functions (or your own custom functions), check for missing values, duplicate entries, errors such as strings entered in numeric data fields, potential outliers.
3. Address the following questions (not necessary that all of them are through coding) –

**Sampling:**

* + 1. You have collected power data for several states. How would you perform simple random sampling to select a subset of states for a more detailed analysis?
    2. What are the advantages and disadvantages of using stratified sampling in this context?

**Descriptive Statistics:**

* + 1. Compute the mean, median, and standard deviation of the” Power Requirement Net Crore Units” and” Availability of Power Net Crore Units” columns.
    2. Create box plots to visualize the distribution of power requirements and availability across states.

**Random Variables:**

* + 1. Define a random variable *X* as the difference between” Power Requirement Net Crore Units” and” Availability Of Power Net Crore Units” for a randomly selected state. Calculate the empirical probability distribution of *X*.
    2. What is the probability that a randomly selected state will have a power requirement greater than its power availability (i.e., *P*(*X>* 0))?

**Mean and Expectation:**

* + 1. Calculate the expected power requirement for a randomly selected state based on the ”Power Requirement Net Crore Units” column.
    2. Compute the expected per capita power availability for all states.
    3. Determine the expected installed power capacity for a group of states that you randomly select.

**Sample Mean and Variance:**

* + 1. Calculate the sample mean and variance of the "Power Requirement Net Crore Units" for states with a population size greater than 1 million.
    2. Calculate the sample mean and variance of the "Power Requirement Net Crore Units" for states with a population size less than 1 million.

**Skewness:**

* + 1. Compute the skewness of the "Availability Of Power Net Crore Units" for all states.
    2. Calculate the skewness of the "Power Requirement Net Crore Units" for states with a population size greater than 500,000.

**Cluster Summary Statistics:**

* + 1. Stratify the states into three clusters based on their population size: small, medium, and large. Define the thresholds for each cluster.
    2. Calculate the mean, median, and standard deviation of "Power Requirement Net Crore Units" for each cluster.
    3. Calculate the mean, median, and standard deviation of "Availability Of Power Net Crore Units" for each cluster.
    4. Create box plots to visualize the distribution of power requirements and availability within each cluster.

1. Search online for projected population data for the next few years and use it to forecast the expected Power Requirements for each state. If you are making any assumptions for this calculation, state them clearly.