ST 228: Data Analysis, ML and AI

Assignment # 1 (10 marks)

(Data Handling)

17th of Jan 2025

Due on: 22nd **of Jan 2025 before 5 PM**

Instructions :

- Clearly state all the assumptions made.
- Clearly quote the source of data used.
- Clearly show your work
- Reports exceeding with the word limit will not be considered.

1. Analysing 125 Years of Bangalore Monthly Rainfall Data (Excel 1)

i) Data Exploration
- Display the structure, size, and basic statistics of the data
- Perform a summary analysis
- Calculate the total, average, and variance of monthly and yearly rainfall.
- Identify the top 5 wettest and driest years.

ii) Trend Analysis
- Visualize the long-term trend of annual rainfall over 125 years.
- Compute a 10-year rolling average of annual rainfall and overlay it on the trend plot.

iii) Seasonal Analysis
- Group the data by month and calculate the monthly average rainfall over 125 years , show graphically.
- Identify the most and least variable months using the coefficient of variation (CV).

iv) Summary Report: Write a short (200–300-word maximum) report summarizing the key findings from your analysis.

v) Do you see a change in rainfall pattern over the last 125 years, if yes, demonstrate, if no demonstrate. **(5 Marks)**

2: Analysing Modified Bangalore Monthly Rainfall Data (Excel 2 - modified)

i) Identify and report the locations of missing values in the dataset, assign missing values using different methods and compare the impact:
- Mean of the corresponding month across all years.
- Median of the corresponding month.
- Seasonal average (mean of the month ± 2 months), and

- Few other ways that you think is more appropriate

ii) Detect and Visualize the outliers , choose a treatment strategy for the outliers (e.g., removal, capping, or transformation) and justify your decision.

iii) Re-analyze the dataset after cleaning:

- Calculate the 10-year rolling average of annual rainfall.
- Re-plot the seasonal average rainfall and variability (as in Question 1).
- Compare findings with the original dataset. Highlight key differences due to data cleaning.

iv) Extreme Events and Drought Analysis

v) Define and identify:

- Extreme Rainfall Events: Months with rainfall in the top 1% of the dataset.
- Drought Years: Years in the bottom 5% of total annual rainfall.

vi) Analyze the frequency of extreme events and droughts over decades.

vii) Summary Report: Write a 300–400-word report discussing the following:

- How missing values and outliers influenced the analysis.
- Comparison of results with the original dataset.
- Recommendations for ensuring data integrity in real-world rainfall datasets.

(5 Marks)