DS3030: Data Analytics Lab

Assignment 4

Date: Aug 25, 2025
Timing: 2:00 to 4:00 PM
Max Marks: 12

Instructions

- Submit one `.ipynb` file containing all answers named as: **[student name]_assignment4.ipynb**
- Write the questions in **separate text blocks** before the answers.
- Write **justifications/comments** as required.

Part 1

Dataset: Industrial Characteristics Dataset

The dataset contains **27 key industrial characteristics** across **36 states/UTs**, including:

- Number of Factories
- Fixed Capital & Working Capital
- Number of Workers & Total Persons Engaged
- Wages to Workers & Total Emoluments
- Materials Consumed & Products Manufactured
- Net Value Added & Net Income
- Capital Formation metrics

Use the file:

`Table 4 Estimate of important characteristics by State 2022 2023.csv`

Q1.

Calculate z-scores for Fixed Capital across all states. Identify states with z-scores ≥ ±2.5 (outliers). Which states have exceptionally high or low industrial fixed capital?
(3 Marks)

Q2.

Apply 3-point moving average to smooth wage related data that is Wages to Workers, Total Emoluments, and Net Income across states.
(1 Mark)

Part 2

(The marks for the plots are given according to the aesthetic value, such as sufficient spacing between axes labels, readability of the labels via font sizes, and required titles)

Q3.

Load the dataset IHHL.csv and compute the total achieved IHHL (column *"IHHLTotalAch"*) of all the districts in each state. Print the results and verify.
(2 Marks)

Q4.

Generate a scatter plot from the results above, with x-axis labels as state names and y-axis as the total achieved IHHL. Ensure that the outlier values computed using Z-score are marked in red colour.
(2 Marks)

Part 3

Dataset: Indian Food Dataset

The dataset contains information about Indian dishes with attributes:

- `name` → Name of dish (e.g., Balu shahi, Gajar ka halwa)
- `ingredients` → Main ingredients used

- `diet` → Vegetarian/Non-vegetarian
- `prep_time` → Preparation time (minutes)
- `cook_time` → Cooking time (minutes)
- `flavor_profile` → Sweet, spicy, sour, etc.
- `course` → Snack, dessert, main course, etc.
- `state` → Indian state where the dish is popular
- `region` → North, South, East, West

Tasks:

Q5.

Using the `indian_food.csv` dataset, create a new column called **`total_time`** by adding the preparation time (`prep_time`) and cooking time (`cook_time`) for each dish.
Then, compute the **average total time** required to make food for each state and display the results in descending order.
(2 Marks)

Q6.

From the results above, select the **top 8 states** with the highest average total cooking time.
Plot a **pie chart** showing the distribution of average total cooking time across these states. Add percentages to the chart for clarity.
(2 Marks)