DS3030: Data Analytics Lab
Assignment 5

Date: Sep 01, 2025

Timing: 2:00 to 3:45 PM — Max marks: 12

Instructions

- Submit one .ipynb file containing all answers named as [**student name**]**_assignment**[**number**]**.ipynb**
- Write the **Part number** in **separate text blocks** before the answers.
- Write **justifications/comments** as required.

1 Part I

Dataset: Indian Food Dataset (contains information about Indian dishes collected from different parts of the country)

Columns:

- **name** → the name of the dish (e.g., Balu shahi, Gajar ka halwa)
- **ingredients** → main ingredients used
- **diet** → whether the dish is vegetarian or non-vegetarian
- **prep_time** → preparation time (in minutes)
- **cook_time** → cooking time (in minutes)
- **flavor_profile** → taste category like sweet, spicy, sour
- **course** → type of course (e.g., snack, dessert, main course)
- **state** → Indian state where the dish is popular
- **region** → geographical region (North, South, East, West)

1. (4 points) **Task: Impute missing values in `prep_time` using KNN Imputer.**

   Workflow:

   1. Replace all values `-1` in `prep_time` with `NaN`.

2. Select the following columns for KNN similarity:
   - `diet`
   - `state`
3. Convert the categorical columns (`diet`, `state`) into numeric form using One-Hot Encoding.
4. Apply **KNN Imputer** with $k = 3$:
   - Use neighbors based on similarity of `diet` and `state`.
   - Fill missing values only in `prep_time`.
5. Update the dataset:
   - Ensure `prep_time` has no `-1` or `NaN`.
6. Display the final cleaned dataset.

2 Part II

Dataset: PGI.csv (contains missing values in some numeric columns). This dataset presents the Performance Grading Index (PGI) scores of various districts across Indian States and Union Territories for the academic year 2022–2023. It evaluates educational performance across multiple predefined categories and provides an overall score along with a qualitative Grade (e.g., "Prachesta-1").

Each district is assessed annually based on its performance in six categories, and an overall PGI score is calculated. Some districts may have missing data.

Task: Impute missing values using the following strategies:

- Mean
- Median
- Mode (most frequent)

1. For each strategy, compute the average of the `Overall` column after imputation. **(3)**
2. Write a brief conclusion (2–3 sentences) on the most suitable strategy and why. **Expected Result:** The average of the `Overall` column should be strictly greater than 270. **(1)**

3 Part III

Load **"Punjab_Girls_Enrollment_Class11_12_By_Year.csv"**.
This education catalog contains the data about the district-wise number of girls in Classes XI to XII of the state of Punjab from 2009 to 2022. The key features are "District" and the years. The dataset has been obtained from the Open Government Data (OGD) Platform Punjab. More details are available here.
Generate the **heatmap** of the **correlation matrix** between **every pair of years** in each of the following scenarios:

1. After dropping the rows with at least one missing value from the original dataframe. **(1)**
2. Without dropping the rows with missing values, impute them in the original dataframe using the KNN approach (with your choice of number of neighbours) **(2)**
3. Is there any significant difference in correlations (loss of information) between the cases of the previous two questions? Justify your answer by generating the heatmap of the absolute difference between the respective correlations. **(1)**