

Assignment - 5

Part (A)

Consider this Dataset "Students Performance in Exams" and Write the code for the following basics questions. Download the dataset from link below ?

<https://www.kaggle.com/datasets/spscientist/students-performance-in-exams>

1. Load the dataset and display the first 10 rows.
2. Check for missing values in the dataset.
3. Rename column names to remove spaces and lowercase them and show the output by displaying the columns name after changings.
4. Create a new column average_score as the mean of math, reading, and writing scores.
5. A: 90+, B: 80–89, C: 70–79, D: 60–69, F: below 60 (based on average_score), create a column for grades as per given logic.
6. Display details of students who completed the test preparation course.
7. Find all female students who scored above 85 in math.
8. Group students based on parental level of education and calculate their average scores.
9. Group by gender and find the maximum average score.
10. Show top 5 students with the highest average_score.
11. Create a second DataFrame with student_id and sports_participation, merge it with the original using inner join (you can add data from yourself in new dataframe)

Part (B)

Consider this Dataset for this Part, (Download the dataset from Link Below) ?
"House Prices - Advanced Regression Techniques"

Kaggle Link: <https://www.kaggle.com/camnugent/california-housing-prices>

Dataset Cleaning and Handling Tasks:

1. Load the dataset and display the first 5 rows.
2. Find and display the number of missing values in each column.
3. Find and display the number of duplicate rows.
4. Drop duplicate rows (if any) from the dataset.
5. If missing values are less than 5% of data, drop the rows. otherwise, fill missing values using the mean of the column.
6. Show the dataset shape before and after cleaning.

(Submit all the questions before the deadline)