
📊 Dataset: Online Course Enrollments

📁 Sample Data (save as `course_enrollments.csv`)

```
EnrollmentID,StudentName,CourseName,Category,EnrollDate,ProgressPercent,Rating,Status
ENR001,Aditya,Python for Beginners,Programming,2024-05-10,80,4.5,Active
ENR002,Simran,Data Analysis with Excel,Analytics,2024-05-12,100,4.7,Completed
ENR003,Aakash,Power BI Essentials,Analytics,2024-05-13,30,3.8,Active
ENR004,Neha,Java Basics,Programming,2024-05-15,0,,Inactive
ENR005,Zara,Machine Learning 101,AI,2024-05-17,60,4.2,Active
ENR006,Ibrahim,Python for Beginners,Programming,2024-05-18,90,4.6,Completed
```

🧪 Exercise Set – Online Course Use Case

🔹 Data Loading

1. Load the data with schema inference enabled.
2. Manually define schema and compare both approaches.

🔹 Filtering and Transformation

3. Filter records where `ProgressPercent < 50`.
4. Replace null ratings with average rating.
5. Add column `IsActive` → 1 if Status is Active, else 0.

🔹 Aggregations & Metrics

6. Find average progress by course.
7. Get count of students in each course category.
8. Identify the most enrolled course.

🔹 Joins

9. Create second CSV: `course_details.csv`

```
CourseName,DurationWeeks,Instructor
Python for Beginners,4,Rakesh
Data Analysis with Excel,3,Anjali
Power BI Essentials,5,Rekha
Java Basics,6,Manoj
Machine Learning 101,8,Samir
```

10. Join `course_enrollments` with `course_details` to include duration and instructor.

🔹 Window Functions

11. Rank students in each course based on `ProgressPercent`.
12. Get lead and lag of `EnrollDate` by Category.

🔹 Pivoting & Formatting

13. Pivot data to show total enrollments by Category and Status.
14. Extract year and month from `EnrollDate`.

Cleaning and Deduplication

15. Drop rows where `Status` is null or empty.
16. Remove duplicate enrollments using `dropDuplicates()`.

Export

17. Write the final cleaned DataFrame to:
    - CSV (overwrite mode)
    - JSON (overwrite mode)
    - Parquet (snappy compression)
