DS203 Final Project:

Preparation for the dataset:

- Students will be working together with their chosen pairs.
- Each team will be exploring datasets from the available sources based on their field of interest:
    - Here are a few data sources to start with:
    - UCI ML Repository
    - Kaggle
    - Environment Canada
    - Statistics Canada
    - Vancouver Open Data Portal
- The size of the chosen dataset should be almost 1GB and the underlying dataset should have multiple csv files within it.

Project Details:

- Import the dataset into your Spark environment and perform the initial analysis:
    a. Generate Dataframe from the csv data.
    b. Define and verify the schema of the dataset.
    c. Display the schema of the dataset.
    d. Print the number of rows and columns.
    e. Obtain the descriptive statistics.
- You can use **PySpark functions** or **Spark SQL** to perform the following operations:
    a. Filtering the data based on conditions. (Atleast 5 queries have to be generated. Perform filtering based on single condition and multiple conditions using and/or keywords)
    b. Create a new column based on the existing columns. - At least one column has to be created.
    c. Aggregate functions (sum, avg, min, max) - At least 2 queries have to be performed.
    d. Grouping the records based on a single or multiple columns. - At least 1 query has to be performed.

e. Sorting the records - At least 1 query has to be performed.
f. Join (Full, Left, Right, Self) - At least 1 join has to be performed.
g. Window functions (rank(), dense_rank(), lag(), lead()) - At least 2 queries have to be performed.
h. Aggregate window functions - At least 2 queries have to be performed.

Documentation:

The team has to document their final project with clear findings, observations, challenges and results. (Refer to the project template for ideas)

Submission Instructions:

- Please provide clear comments to the code.
- Please create a Github repository with your team members as collaborators with the following files:
  - Dataset files (.csv)
  - Code (.ipynb)
  - Project Documentation (.pdf)
- Upload your github repository to classrooms as a submission.

Presentation:

The team will be presenting their final project on Friday. The presentation slots will be provided shortly.

Grading Criteria:

- Dataset Selection (20%)
- PySpark queries/functions usage (50%)
- Presentation and engagement with the audience (20%)
- Code comments (10%)

Resources:

SQL Handy Guide