BDM2 Project Instructions and Deliverables

Key Objectives:

The project should achieve the following objectives. The teams have freedom to choose problems and datasets as per their familiarity with the domain.

1. Choose a dataset with reasonable size with at least two data files / dataset.
2. One of the dataset should have at least 6 columns and 1000 records. The other one could be smaller.
3. Define the schema and load the data as per schema. Ensure handling of bad record should be designed in the pipeline.
4. Create a report of data quality e.g. number of null values etc. Drop the records with missing values if necessary.
5. Convert data types as necessary like timestamps etc.
5. Write/save the cleaned records to columnar tables in staging zone. Add partitions or clusters as required.
6. Read from the columnar tables to create pipelines to find insights.
7. The pipelines should contain few of the following operations.
   - selecting
   - filtering
   - grouping
   - sorting
   - joining
   - cubes or grouping sets

8. Store the pipelines results into curated tables.
9. Find at least 5 insights from results tables with charts (different types of charts)
9. Some insights should involving converting to pandas dataframe and using matplotlib or seaborn library and some can be drawn using the notebook inbuilt charts.
10. Examples of charts that can be used are histogram, distribution plot, bar plot, trend plot, scatter plot, heatmap etc.

Assessment Weightage:

Evaluation will not only be based on completing the above tasks, but also the following factors.

1. Code clarity
2. Processing Complexity
3. Creating valuable insights
4. Documentation (use markdown)

There may be penalty if code clarity and documentation is not proper.

Notebook Requirements:

The notebook should contain:

1. The project name, team members (with IDs) and description at the top.
2. Clear description of the problem and dataset.
3. Proper sections (loading, cleaning, processing, insights etc.) should be created for clarity.
4. Pipelines should be described with bulleted points.
5. Insights and inferences should be described for understanding.
6. There should be conclusion section with summary of accomplishment and few bulleted points of lessons learnt in developing the project.

Coding scheme for group project is 2N-c

Deliverables:

The deliverables should be the notebook with all outputs from the complete run of the notebook and the datasets.

Do NOT submit .zip files otherwise, the submission will not be considered.