Last edited 3rd Jan 2026

Best Practices in Data Governance, Preparation and Analytics (CA6003)

Assignment submission Guidelines

1.Introduction

Data is the foundation of any successful AI system. The preparation and governance of data constitute a fundamental component in the development of reliable AI and machine learning solutions. This process typically begins with the systematic profiling and analytical exploration of dataset characteristics, followed by appropriate data cleaning, transformation, enrichment, and validation to ensure suitability for downstream learning tasks. To complete the data-to-insight workflow, a basic machine learning model is employed to evaluate and demonstrate how data preparation decisions influence model behaviour, performance, and robustness.

Effective data governance, preparation, and analytics represent a convergence of multiple disciplines spanning data engineering, statistics, ethics, and applied machine learning. The process is inherently technical, requiring the ability to interpret, validate, and transform raw data using appropriate tools and methodologies. At the same time, it demands critical analytical judgment to detect bias, outliers, and integrity issues, and to avoid common analytical fallacies during exploratory analysis. Evidence of mastery therefore lies in the ability to integrate these skills to justify data-centric decisions and communicate their impact clearly and rigorously. This group assignment (Group size 4- groups are pre assigned) provides students with the opportunity to apply the principles and best practices covered in the course by designing robust data preparation workflows and empirically demonstrating their importance within the AI lifecycle.

2. Assignment Objective

This assignment is explicitly designed to assess the following learning outcomes. Upon successful completion, students will demonstrate the ability to:

1. Prepare raw data for machine learning by performing systematic data profiling, cleaning, validation, transformation, and enrichment using appropriate tools and techniques (e.g., handling missing values, encoding, feature engineering, and standardisation).
2. Analyse datasets to detect and address bias, outliers, and data integrity issues, evaluate their impact on downstream model performance, and apply suitable mitigation strategies.
3. Design robust Exploratory Data Analysis (EDA) by interpreting summary statistics and visualisations, identifying analytical fallacies (e.g., correlation vs causation, Simpson's paradox), and applying strategies for accurate and meaningful data interpretation.
4. Demonstrating the Importance of Data Preparation using basic Machine Learning.

   Note that the key purpose is to identify that ML is evidence, not the goal. The assignment should reinforce data best practices. Only basic, interpretable models allowed (Linear / Logistic Regression, Decision Tree)

- Students must compare minimally processed data vs fully prepared data
- Marks are awarded for explanation, reasoning, insight and NOT for accuracy chasing

The emphasis of this assignment is data-centric AI practice, ethical reasoning, and analytical clarity rather than model complexity or predictive accuracy.

3. Video Presentation

Presentation Coverage: Your group is required to produce a 12-minute (maximum time) recorded video presentation, in which all group members must appear and present. The video should clearly communicate your investigative and analytical process, focusing on data governance, preparation, analytics, and their impact on basic machine learning outcomes.

The presentation should all the points mentioned in the assignment objectives along with the summary of the group's key contributions and reflection on what was novel or insightful about the data preparation.

Presentation Format and Requirements

- The presentation must be delivered as a single continuous video of no more than 12 minutes. Any extra time recorded will not be considered for evaluation.
- All group members must appear on screen and present during the video.
- Presentation slides may be used to support the presentation, but the spoken explanation and visual communication are the primary focus.
- Slides should be concise and visually clear; excessively wordy slides will not be scored favourably.

Originality and Use of Tools

All dataset analysis, data preparation, modelling, visualisation, and presentation design must be the group's original work. Plagiarism or submission of non-original content is considered a serious academic violation and may result in severe penalties.

Generative AI tools may be used only as assistance (e.g. code generation or debugging) and must be clearly acknowledged. Students are expected to demonstrate understanding and ownership of all submitted work.

Submission Instructions

On behalf of the group, the group leader (or any one group member) must submit:

1. A public YouTube link to the 12-minute presentation video
2. Jupyter Notebook (.ipynb format)
    1. Submit the notebook in. ipynb format only.
    2. The notebook must be well-structured and clearly written, including:
        1. Clear headings and sections
        2. Detailed comments and markdown explanations
        3. Properly explained data cleaning, preprocessing, and analysis steps
    3. Code should be readable, organized, and reproducible.

Submissions that are unclear, poorly documented, or incomplete may result in loss of marks.

Last edited 3rd Jan 2026

Both items must be submitted via the NTULearn Blackboard Assignment portal by the stated deadline. Multiple submissions are permitted (up to three), but only the final submission before the deadline will be assessed. Students are strongly encouraged to submit only when the group is satisfied that the submission is final.

4. Peer Evaluation

Peer Evaluation Rubrics – At the end of the assignment, it is necessary for all students to assess the contributions of each team member based on:

1. **Teamwork**: demonstrating proactiveness in collaborating with team members and respect for each other.
2. **Quantity of work**: demonstrating fair share in the overall workload throughout the team project.
3. **Quality of work**: contributing ideas and research efforts that enhance the overall quality of the team's output.

Scoring – Peer evaluation exercise is **confidential** and will be carried out after the assignment submission deadline is over. All students must complete their peer evaluation before stipulated deadline. Failing which, they will receive **zero marks** for their peer evaluation component, regardless of what scores they received from their peers. Each student will give a rating scale between 1 to 5 on the above three assessment criteria for each of his or her group members. Please carefully note the rating rubric below. Do note that a score of "1" is not the best rating, but the worst.

Rating Scale	
5	Strongly agree (*best rating*)
4	Agree
3	Neutral
2	Disagree
1	Strongly disagree (*lowest rating*)

Managing Your Team – Setup regular meeting (face to face or online) to monitor assignment progress. If team members are not contributing to their allocated assigned task and responsibilities, do provide timely feedback to them. Encourage them to contribute and remind them that their individual marks in the course can be impacted by a poor peer evaluation rating from the teammates.

5. Deliverables and Deadlines

The table below outlines the various deliverables for Video Group Assignment and their respective weightage.

Deliverables	Submission Mode	Weightage	Deadline
Video Presentation – Completed Video presentation (≤ 12 min) along with the Jupyter notebook of the work done. Note1: Publish the video on youtube and make it public. Include your group number while uploading the youtube link	Upload to NTULearn Assignment portal	90%	Before Sunday, **Week #5** (8 Feb@12 noon)

Note2: Any video beyond 12 min limit, we will discard the material after 12 min. Note 3: The Jupyter notebook(.ipynb format) should be well-structured and clearly written, including clear headings and sections, Detailed comments and markdown explanations, Properly explained data cleaning, preprocessing, and analysis steps. The code should be readable, organized, and reproducible			
Peer Evaluation – Completed peer assessment for each of your teammates using the peer review system on NTULearn, which should appear in the **Assignment** folder.	NTULean peer review system	10%	Before Saturday, **Week #6** (14 Feb@12 noon)

Last edited 3rd Jan 2026

Assessment Criteria for Video presentation

Presentation - Assessment Criteria	Weightage
Appropriateness –Appropriate selection and use of data analysis techniques and visualizations at each stage of the workflow, including dataset exploration, data profiling, cleaning, bias analysis, exploratory data analysis, and model evaluation.	25%
Correctness & Clarity - Technical correctness of work carried out across all stages of the project, including data preparation, transformation, bias mitigation, exploratory analysis, selection and application of basic machine learning models, and evaluation of results. Insights derived from the analysis should be clearly explained, logically reasoned, and accurately communicated. Assessment marks will be allocated according to the clarity with which each team member's contributions are distinguished and documented	25%
Data Interpretation and Analytics – Demonstration of sound analytical judgment in interpreting summary statistics, distributions, and relationships within the data. This includes the ability to identify redundancy, multicollinearity, skewness, and potential analytical fallacies, and to justify conclusions drawn from the analysis in a rigorous and evidence-based manner.	25%
Novelty and depth of data centric insights – Originality and depth in the choice of problem context, dataset, data preparation strategies, bias analysis, and analytical approach. Novel or insightful interpretations of the data, as well as thoughtful use of basic machine learning to justify data preparation decisions, will be credited.	25%
Total	100%

Some possible online sources for datasets

[1] Kaggle - huge repository of community published data & code - https://www.kaggle.com/datasets

[2] Data.world open datasets - https://data.world/datasets/open-data

[3] Singapore Statistics - https://www.singstat.gov.sg/

[4] Singapore Government Published Dataset - https://data.gov.sg/

[5] Singapore Geo Data dataset - https://data.gov.sg/dataset/national-map-line

[6] US COVID-19 Datasets - https://data.cdc.gov/browse?limitTo=datasets

[7] COVID-19 data - https://github.com/owid/covid-19-data/tree/master/public/data

[8] Our World in Data - Coronavirus Source Data - https://ourworldindata.org/coronavirus-source-data

[9] UCI Machine Learning Repository - https://archive.ics.uci.edu/

[10] Asian Development Bank (ADB) dataset - https://data.adb.org/search/content/type/dataset

[11] DataHub.io Stock Market Data - https://datahub.io/collections/stock-market-data

[12] Google's Dataset Search - https://datasetsearch.research.google.com/

Note: Some sites will require you to sign up as a member to gain access.