

Week 2: Required Assignment 2.1

[Start Assignment](#)

Due Thursday by 10:29pm **Points** 10 **Submitting** a file upload **Attempts** 0 **Allowed Attempts** 1
Available after Apr 16 at 10:30pm


Learning Outcome Addressed

- Describe the importance of sources and quality of data
- Describe the various steps involved in cleaning and preparing data for analysis

Objective:

This assignment aims to enhance your understanding of data cleaning and preparation, vital steps in the machine learning process. You will work with the Titanic dataset to address common issues such as missing values, duplicates and transformations to prepare the data for effective modelling. Although not mandatory, you are encouraged to provide justifications for your decisions, even if you choose not to submit code or technical proof. No points will be deducted if workings or proof are not provided.

Assignment Instructions

Dataset: Download the Titanic dataset from Kaggle: [Kaggle Titanic Dataset](https://www.kaggle.com/competitions/titanic/data)  [_ \(https://www.kaggle.com/competitions/titanic/data\)](https://www.kaggle.com/competitions/titanic/data). Use the `train.csv` file for this assignment.

Note: You would need to register and create a free account for kaggle.com. We strongly recommend creating this account, as kaggle is a valuable resource that would be used for later assignments, as well for your own exploration beyond this programme.

Assignment Parts:

1. Understanding Raw Data:

- Review the provided dataset and describe its structure.
 - Find out the total number of records in the dataset.

- Find out the columns in the dataset. List the column names and describe them.
- Find the data types of each column.
- Identify common issues in raw data.
 - Find out the number of missing values in each column.
 - Find out if there are any duplicate records in the dataset.

2. Data Cleaning Techniques:

- Apply data cleaning methods such as handling missing values using mean, median and mode imputation.
 - Decide whether the columns with missing values should be retained or removed. Justify your decision.
 - If the column with missing values will be retained, what is your strategy to impute the missing values? Justify your decision.
- Remove non-essential columns from the dataset.
 - Identify the non-essential columns from the modelling point of view. Justify your decisions.
 - Justify the selection or removal of features based on their impact on model performance.

3. Data Transformation:

- Implement data transformation techniques for encoding categorical variables.
- Implement data transformation techniques for scaling (normalisation, standardisation).
- Discuss how data transformation ensures the data meets model requirements without altering its inherent meaning.

4. Reflection and Insights:

- Reflect on the challenges encountered during data cleaning and preparation.
- Highlight the importance of each step and its contribution to building a reliable and effective machine-learning model.

Submission Instructions:

- Submit a document detailing your approach and insights for each task.
- Code submission is optional but ensure your decisions are clearly justified.
- Work should be original and reflect your understanding of the concepts.
- Rename the dataset and assignment files as **Your_Name_Assignment name** and complete the assignment.
- Select the **Start Assignment** button at the top of this page.
- Upload the file containing your responses.
- Select the **Submit Assignment** button to submit your responses.

Suggested time: 45 minutes

This is a required assignment and counts towards programme completion.

Required Assignment 2.1 - Rubric

Criteria	Ratings			Pts
Understanding Raw Data	2 pts Full Marks Accurately identifies the dataset structure, number of records, and columns. Provides correct descriptions for the data types of each column. Thoroughly identifies missing values, duplicates, and any other issues. Clearly discusses common issues such as missing data, inconsistent data, or potential outliers.	1 pts Half Points Provides a general overview of the dataset structure, records, and columns, but with minor inaccuracies. Identifies most missing values or duplicates but misses some issues. Provides some analysis of common issues but lacks depth.	0 pts No Marks Fails to describe the dataset structure, columns, or record count. Misses key aspects like data types, missing values, and duplicates. Does not address any common data issues.	2 pts
Data Cleaning Techniques	3 pts Full Marks Properly applies appropriate methods (mean, median, mode imputation) for handling missing values and provides a strong justification for the approach. Correctly identifies and removes or retains non-essential columns with a clear and logical rationale for each decision. Clearly justifies how the selection or removal of features impacts model performance.	1.5 pts Half Points Attempts to handle missing values but may apply inappropriate techniques or lack strong justification. Identifies and removes or retains some non-essential columns but with limited or unclear reasoning. Provides some justification for feature selection but lacks detail on how it affects model performance.	0 pts No Marks Does not handle missing values or apply incorrect techniques with no justification. Fails to identify or justify the removal or retention of non-essential columns. Provides no reasoning behind feature selection.	3 pts
Data Transformation	3 pts Full Marks Accurately applies encoding techniques to categorical variables (e.g., one-hot encoding) and explains the method well. Implements	1.5 pts Half Points Applies encoding techniques but may not choose the most appropriate method or fails to explain the reasoning clearly.	0 pts No Marks Does not apply encoding techniques or applies them incorrectly. Fails to implement	3 pts

Criteria	Ratings			Pts
	appropriate scaling techniques (normalization or standardization) and justifies their use based on model requirements. Discusses how transformations affect model performance without altering the data's inherent meaning.	Applies scaling techniques but without a strong justification for the choice. Mentions transformations but with limited discussion on their effect on model performance.	any scaling techniques. Provides no discussion on the impact of transformations on model performance.	
Reflection and Insights	2 pts Full Marks Provides a clear, thoughtful reflection on the challenges encountered during the assignment. Explains how each step of the data cleaning and preparation process contributes to building a reliable machine learning model. Highlights the importance of the steps taken, demonstrating a solid understanding of the workflow.	1 pts Half Points Provides a general reflection on the challenges faced, but lacks depth or misses some key points. Mentions the importance of certain steps, but without strong reasoning or connection to the model-building process.	0 pts No Marks Does not provide any reflection on the challenges encountered. Fails to explain the importance of the data cleaning and preparation steps in the machine learning process.	2 pts
Total Points: 10				