AYTS 5001 – Week 8 Lab: Data Preprocessing

Duration: 2 hours | Practice lab (not graded) | Dataset: malazan_recruit_records.csv

Submission: Notebook + cleaned CSV on GitHub repo

Learning Outcomes

• Identify and handle missing data

• Standardise inconsistent data (case, whitespace, formats)

• Handle incorrect data types

• Detect and remove duplicates

• Scale numeric data for analysis

Background

The Malazan Empire's recruitment office has compiled records from across the continent. Unfortunately, their data is inconsistent, incomplete, and riddled with formatting errors. Your task is to clean and preprocess this data so that it can be used for analysis.

Tasks

Task 1 – Load and Inspect (15 min)
Load the dataset, display first rows, and count missing values.

Task 2 – Clean Text Data (20 min)
Strip spaces, standardise capitalisation, and normalise Has_Magic_Potential values to True/False.

Task 3 – Handle Missing and Invalid Numeric Data (25 min)
Convert columns to numeric, replace missing values with mean, and drop duplicates.

Task 4 – Feature Scaling (20 min)
Scale Morale_Score to 0–1 using MinMaxScaler.

Task 5 – Save and Export (10 min)
Save your cleaned dataset as malazan_recruit_records_clean.csv and push it with your notebook to GitHub.

Checklist

☐ Dataset loaded

☐ Missing values handled

☐ Strings standardised

☐ Booleans normalised

☐ Numeric columns cleaned

☐ Morale scaled

☐ Duplicates removed

☐ File saved and uploaded