Data Collection and Preprocessing Phase

Date	16 June 2025
Team ID	SWTID1749705685
Project Title	Movie Box Office Gross Prediction using Machine Learning
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
tmdb_5000_movies.csv	Presence of null values in runtime	Moderate	Dropped records with missing runtime values using dropna(subset=['runtime'])
tmdb_5000_movies.csv	Null dates preventing extraction of month/day	Moderate	Parsed release_date with pd.to_datetime(); filled missing release_month and release_DOW with mode values.

tmdb_5000_movies.csv	Extremely skewed revenue and budget outliers	High	Removed records where budget > $500M or revenue > $3B to reduce distortion.
tmdb_5000_movies.csv	Nested JSON fields, not usable in raw form	High	Applied literal_eval() and extracted primary genre and keywords.
tmdb_5000_movies.csv	Zero or missing budgets and revenues	High	Filtered out records with budget < $1K and revenue < $100K
tmdb_5000_movies.csv	Duplicate or redundant columns after merging datasets	Low	Dropped unnecessary columns (title_x, title_y, id) to reduce noise.
tmdb_5000_credits.csv	Mismatch with id in movies dataset for some records	Low	Ignored cast for this project to reduce complexity (could be included in future models).
tmdb_5000_credits.csv	Missing director in some records (no "Director" role)	Moderate	Filled missing directors as "Unknown" to ensure no null values during encoding.