Data Collection and Preprocessing Phase

Date	17 June 2025
Team ID	SWTID1749705685
Project Title	Movie Box Office Gross Prediction using Machine Learning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

This data collection plan outlines the sources and structure of raw movie data used for predicting box office revenue. It ensures well-documented, clean, and relevant data to support accurate machine learning analysis and reliable results

Data Collection Plan Template

Section	Description
Project Overview	This project aims to predict the box office gross revenue of movies using machine learning models based on features like budget, genre, popularity, runtime, and more.
Data Collection Plan	The dataset was collected from The Movie Database (TMDB) and obtained from Kaggle, which includes two CSV files: `tmdb_5000_movies.csv` and `tmdb_5000_credits.csv`.
Raw Data Sources Identified	The datasets contain metadata about movies such as cast, crew, budget, revenue, genres, and more. These serve as the foundation for feature engineering and model training.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
tmdb_5000_movies.csv	TMDB 5000 Movies dataset with budget, revenue, genres, popularity, etc.	https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata	CSV	~5 MB	Public
tmdb_5000_credits.csv	TMDB 5000 Credits dataset with cast and crew information.	https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata	CSV	~3 MB	Public