

Data Collection and Preprocessing Phase

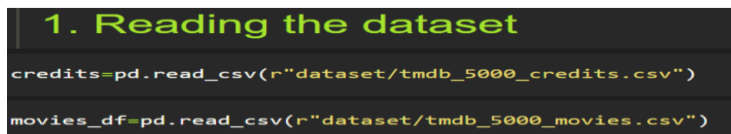
Date	10 JUNE 2024
Team ID	739711
Project Title	Movie Box Office Gross Prediction
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	Basic statistics, dimensions, and structure of the data.
Univariate Analysis	Exploration of individual variables (mean, median, mode, etc.).
Bivariate Analysis	Relationships between two variables (correlation, scatter plots).
Multivariate Analysis	Patterns and relationships involving multiple variables.
Outliers and Anomalies	Identification and treatment of outliers.

Data Preprocessing Code Screenshots

Loading Data	 <pre> 1. Reading the dataset credits=pd.read_csv(r"dataset/tmdb_5000_credits.csv") movies_df=pd.read_csv(r"dataset/tmdb_5000_movies.csv") </pre>
--------------	---

Handling Missing Data

```
movies.isnull().sum()
```

```
budget          0
genres           0
homepage        3091
id              0
keywords         0
original_language 0
original_title   0
overview         3
popularity       0
production_companies 0
production_countries 0
release_date     1
revenue          0
runtime          2
spoken_languages 0
status           0
tagline          844
title_x          0
vote_average     0
vote_count       0
title_y          0
cast             0
director         30
dtype: int64
```

Save Processed Data

```
import pickle
pickle.dump(mr,open("model_movies.pkl","wb"))
```