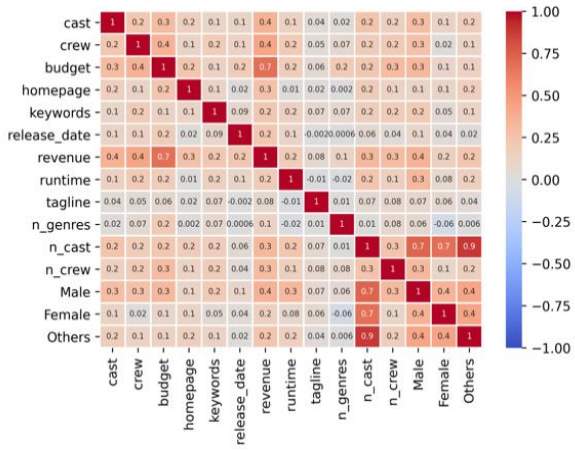
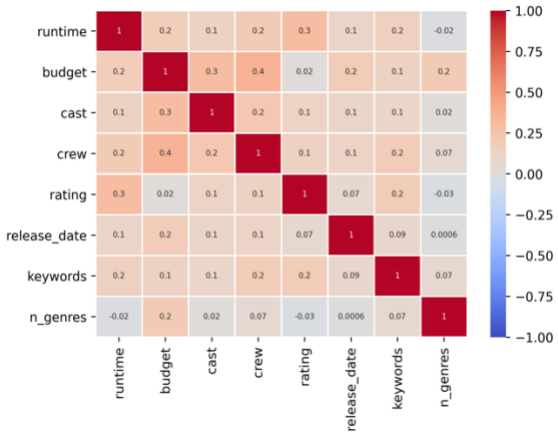
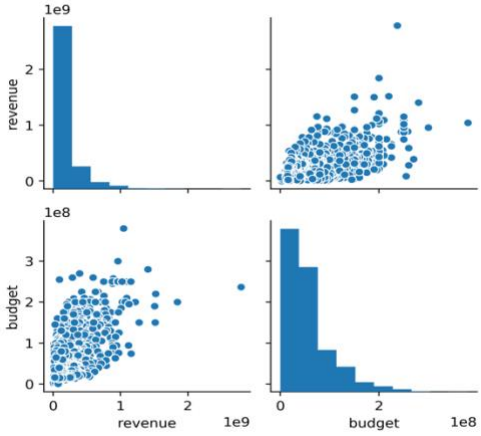
ASSIGNMENT 3

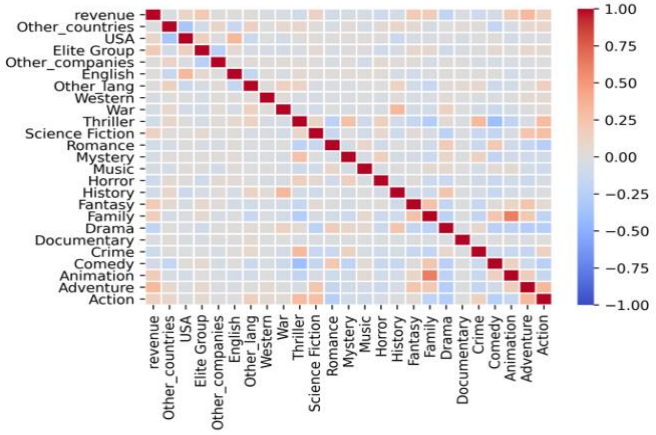
Data Pre-processing

After carefully examining the given training dataset, it is evident that preprocessing is crucial to standardize the data by removing outliers and transforming the non-numeric columns into categorical data.









Data Cleaning

Plotting multiple correlation heatmaps and pair-plotting among the columns gives the relevant feature columns that correlate with the target column. Hence filtering out less correlated columns and selecting only the relevant ones. JSON formatted columns such as cast/crew are parsed intensively to filter out noisy data and extract only useful information like names and gender. A similar approach is used to clean other JSON formatted columns. Also, crucial data such as holiday month is extracted from the given release date.

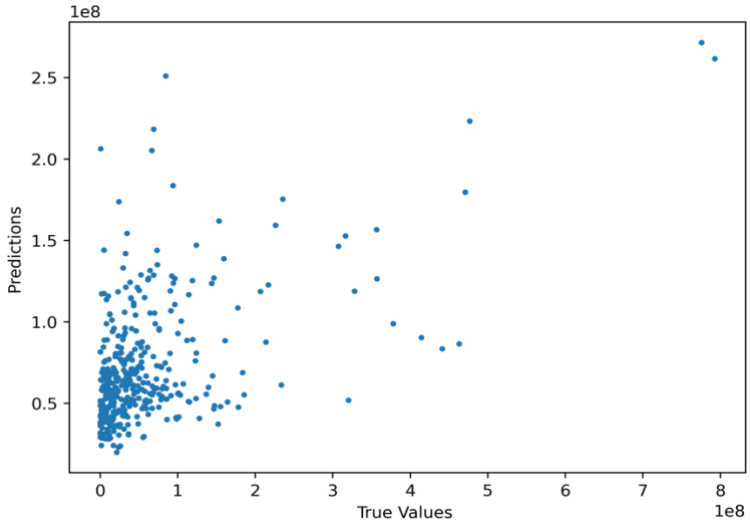
Data Transformation

A list containing an elite group of production companies, bankable cast and crew is downloaded from the internet and curated to transform comma-separated values into numeric form. Dummies of genres, production countries are thereby formed to ensure smooth one-hot encoding. Budget and revenue columns are scaled down in the range of 1-100 by using MinMaxScaler() to ensure uniformity throughout the dataset.

Part 1 – Regression

Linear Regression Algorithm is used with the normalization parameter set as True. The results are MSR as 8494608961541101.00 and Pearson Coefficient of Correlation as 0.40.

To further enhance the stats, **Random Forest Regressor** with random_state parameter set to 0 [zero] is used. It is noticeable that it is only the range of one feature that is split at each stage, which means no need to scale large values of any particular column. Overfit is significantly reduced by building a large number of weak/shallow decision trees and using a sort of majority vote to result in complex decision boundaries. Following results are obtained:

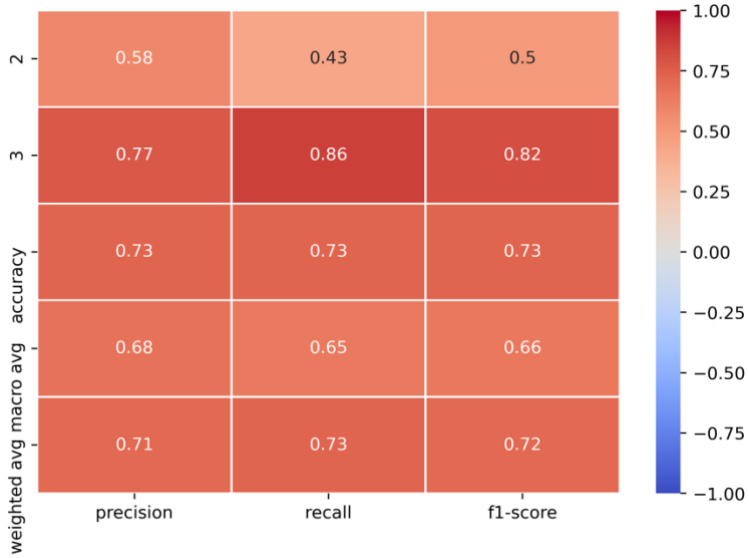


MSR	6200837367991657
Pearson Coefficient Correlation	0.53

Part 2 – Classification

KNeighbours Classifier is used with all default parameters. The results are Average Precision as 0.65, Average Recall as 0.54 and Accuracy as 0.70.

To further enhance the stats, **Gradient Boosting Classifier** with default parameters is used as it works without careful data cleaning. It considers each feature separately for splitting, the comparative scales of the features do not matter in boosting classifiers and it does not require any special treatment for outliers. Following results are obtained:



Average Precision	0.68
Average Recall	0.65
Accuracy	0.73