Movie Revenue Prediction - Minor Project Report

# 1. Project Overview

This project aims to predict movie revenue using a dataset of 10,000 movies and machine learning models. The dataset includes key features like popularity, vote average, vote count, runtime, and revenue. The goal is to build regression models that can estimate a movie's revenue based on these features.

# 2. Dataset Description

The dataset used is `Top\_10000\_Movies.csv` with 13 original columns and 10,014 rows. For this project, 5 columns were selected: popularity, vote\_average, vote\_count, runtime, and revenue. Rows with missing or zero revenue were removed.

# 3. Exploratory Data Analysis

Eight visualizations were created to understand the data distribution and relationships:

- Histogram of revenue  
- Boxplot of revenue  
- Scatter plot: popularity vs revenue  
- Scatter plot: vote count vs revenue  
- Bar plot: vote average vs revenue  
- Histogram of runtime  
- Plotly scatter: runtime vs revenue  
- Plotly histogram: vote average distribution

# 4. Model Building

The dataset was split into training (80%) and testing (20%) sets. Four regression models were trained:  
- Linear Regression  
- Decision Tree Regressor  
- Random Forest Regressor  
- K-Nearest Neighbors Regressor

# 5. Evaluation Metrics

Each model was evaluated using Mean Squared Error (MSE) and R-squared score. Random Forest showed the best performance. For Linear Regression, coefficients of features were printed. A comparison of actual vs predicted revenue was visualized for each model.

# 6. Conclusion

This project demonstrated that basic machine learning models can estimate movie revenue with reasonable accuracy using just a few numeric features. Data quality plays a critical role in model performance.

# 7. Future Improvements

- Add more features like genres, budget, and release month  
- Apply log transformation to revenue  
- Use hyperparameter tuning (GridSearchCV)  
- Try other models like XGBoost or SVR