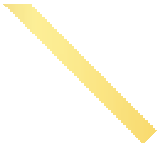
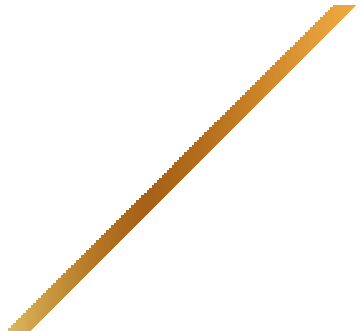
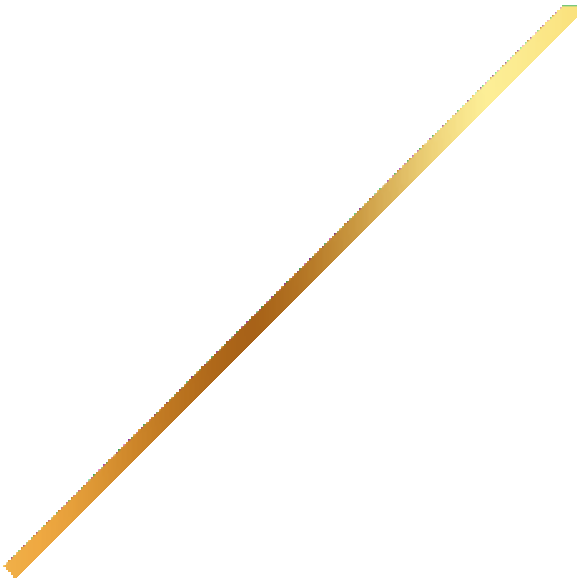
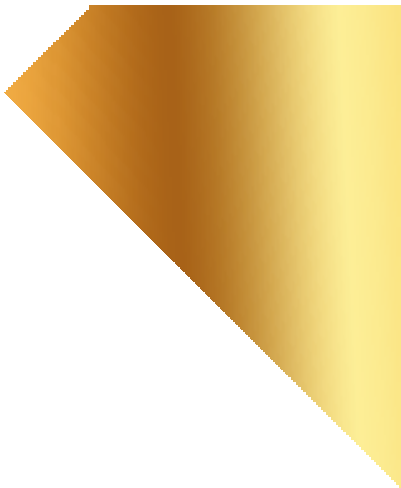
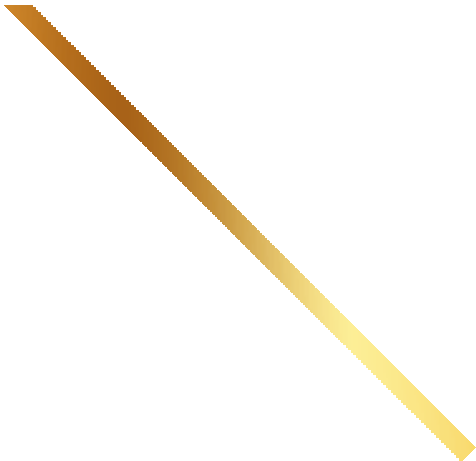
**2023**



DATA

SCIENCE

**TOP 10,000 POPULAR MOVIES**

***AHMED MUSTAFA & CO***

***42010147***

**Problem Definition**

The problem addressed in this report is the analysis and prediction of movie revenue. The goal is to develop a model that can accurately predict the revenue of a movie based on its budget and popularity. By understanding the factors that contribute to a movie's revenue and building a predictive model, we can provide insights to movie industry professionals and aid in decision-making processes.

Method

Data Collection: The dataset used for this analysis is the "Top 10,000 Popular Movies TMDB" dataset obtained from Kaggle. It contains information about the budget, revenue, and popularity of various movies

.

Data Cleaning: The dataset was preprocessed to handle missing values, remove irrelevant columns, and ensure data quality and consistency.

Exploratory Data Analysis (EDA): An exploratory analysis was performed to gain insights into the relationships between the variables, identify any patterns or trends, and understand the distribution of the data.

Feature Selection: Based on the EDA, the features selected for prediction were the budget and popularity of the movies. These features were deemed most relevant in determining the movie's revenue.

Model Selection: Linear regression was chosen as the predictive modeling technique due to its simplicity and ability to capture linear relationships between variables. The scikit-learn library in Python was used to implement the linear regression model.

Model Training: The dataset was split into training and testing sets. The training set was used to fit the linear regression model to the data.

Model Evaluation: The trained model was evaluated using the mean squared error (MSE) metric, which measures the average squared difference between the predicted and actual revenue values.

# Experiment

Data Cleaning: Missing values were handled by imputation or removal. Irrelevant columns were dropped, and data quality checks were performed to ensure accuracy.

Exploratory Data Analysis: Visualizations such as scatter plots, histograms, and correlation matrices were created to explore the relationships between variables and understand the data distribution.

Model Training and Evaluation: The linear regression model was trained on the training set and evaluated using the MSE metric. The performance of the model was assessed by comparing the predicted revenue with the actual revenue values in the testing set.

# References

Kaggle :

"Top 10,000 Popular Movies TMDB" dataset. Retrieved from

https://[www.kaggle.com/datasets/ursmaheshj/top-10000-popular-movies-](http://www.kaggle.com/datasets/ursmaheshj/top-10000-popular-movies-) tmdb-05-2023

This report provides an analysis of movie revenue prediction using linear regression. The results and findings can be utilized by movie industry professionals, analysts, or researchers to gain insights into the factors influencing movie revenue and to make informed decisions related to budget allocation and marketing strategies.