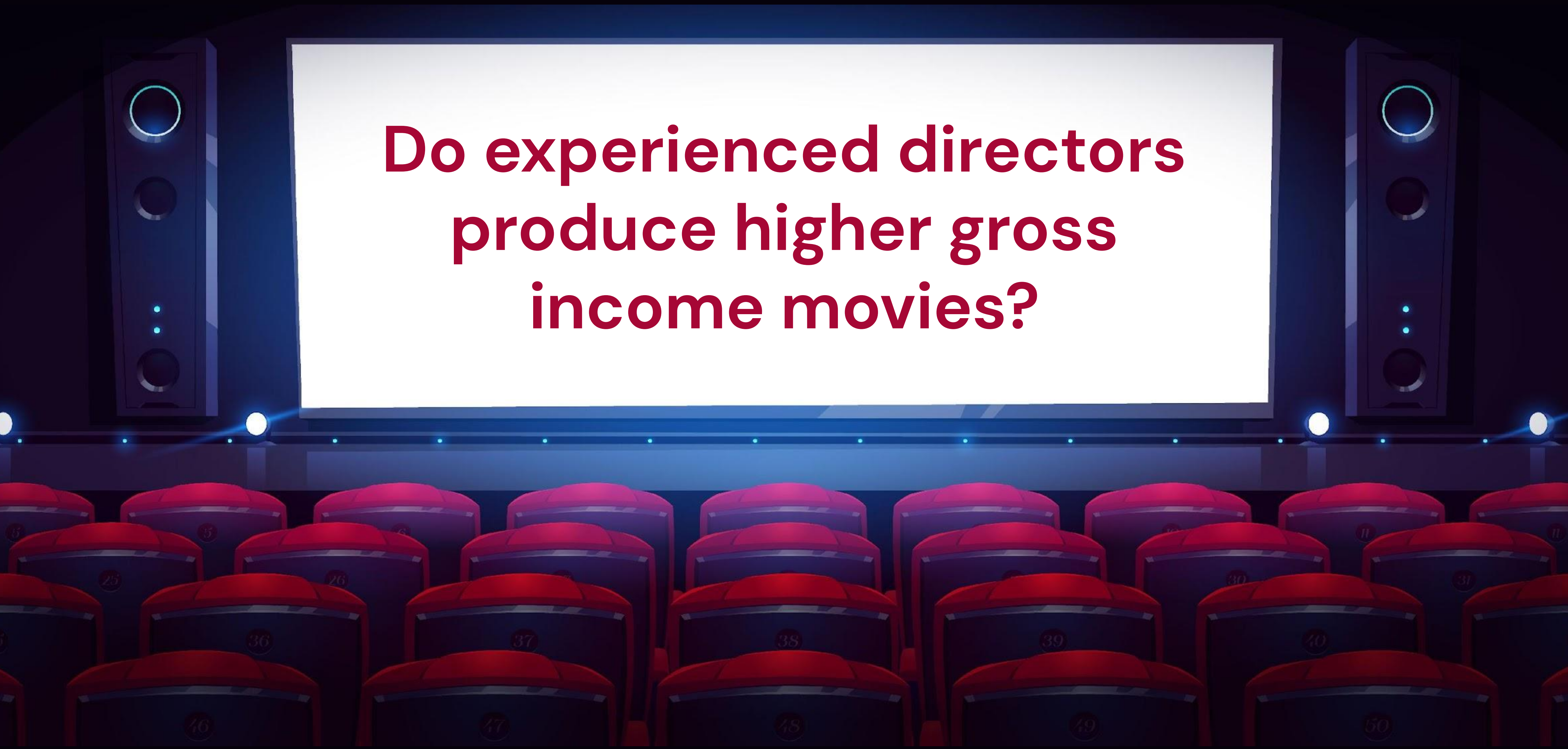
Final project, UofT Data Analytics Bootcamp 2021

Film Industry Box Office Analysis

Instructor
Daniel de R.

Prepared by
Callistus Ikeata
Rojin Shahba



Do experienced directors produce higher gross income movies?

According to our analysis,

YES!

Project Overview

Our Goal

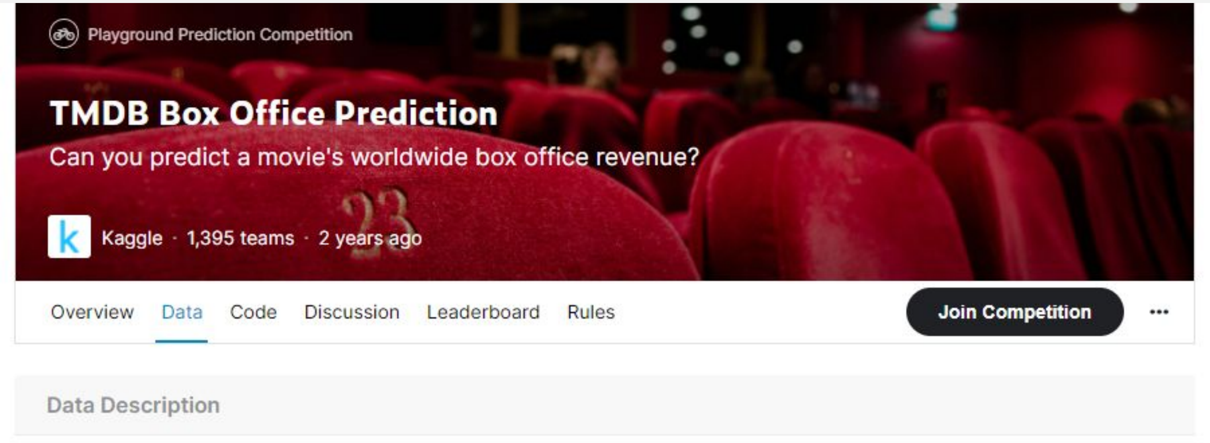
Implement a machine learning model to determine key factors in the film production industry and forecast the gross income of future productions.

Why Film Industry Box Office?

$41.7 B
Industry
in 2018

$9.85 M
Academy Awards Viewers
in 2021

Data Source



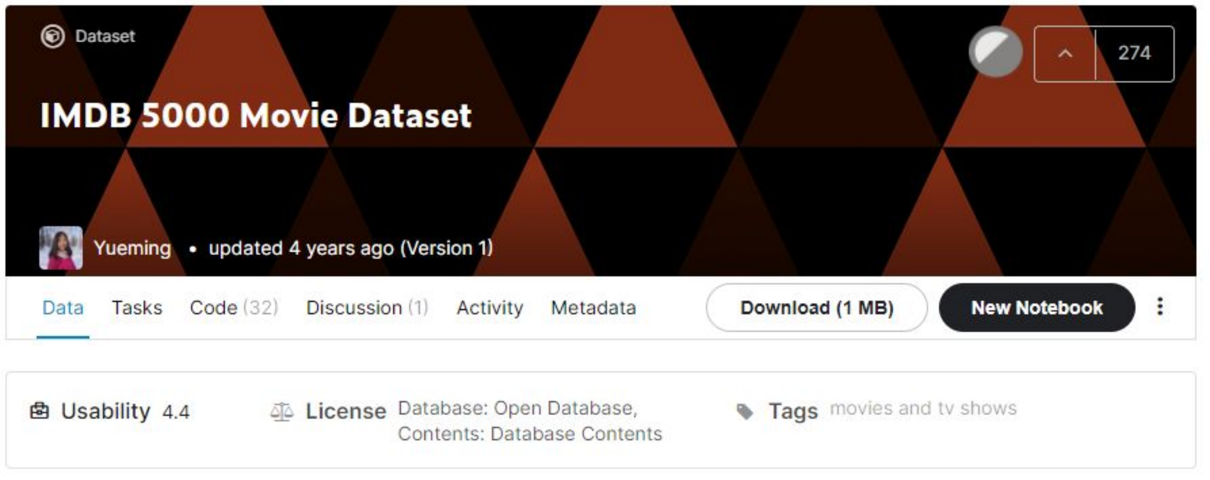
Playground Prediction Competition

TMDB Box Office Prediction

Can you predict a movie's worldwide box office revenue?

Kaggle · 1,395 teams · 2 years ago

Data Description



Dataset

IMDB 5000 Movie Dataset

Yueming • updated 4 years ago (Version 1)

Usability 4.4

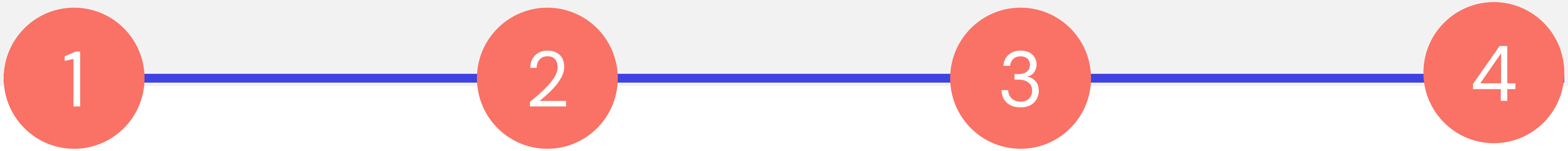
License Database: Open Database, Contents: Database Contents

Tags movies and tv shows

Insights We Are Looking to Gain

- How much does the director's portfolio play a difference?
- Do projects with higher budgets promise higher revenues?
- Are certain genres more profitable than others?

Data Exploration



1 Finalized the topic

2 Searched for data sources

3 Cleaned the data

4 Organized the data

Data Analysis

1. Created the schema
2. Build the database on internal server
3. Created new tables using the JOIN command
4. Migrated the database on Amazon RDS
5. Utilized Tableau to visualize the data

Tools & Technologies

python™

jupyter

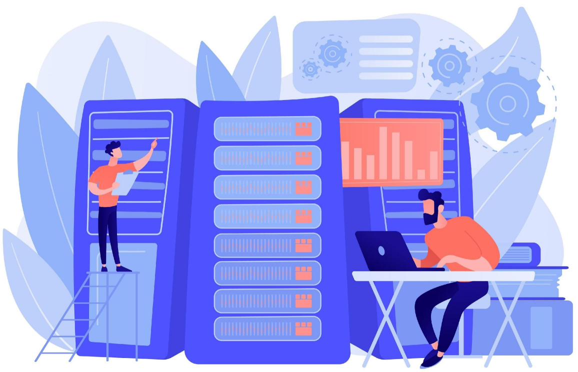
QUICK DBD

PostgreSQL

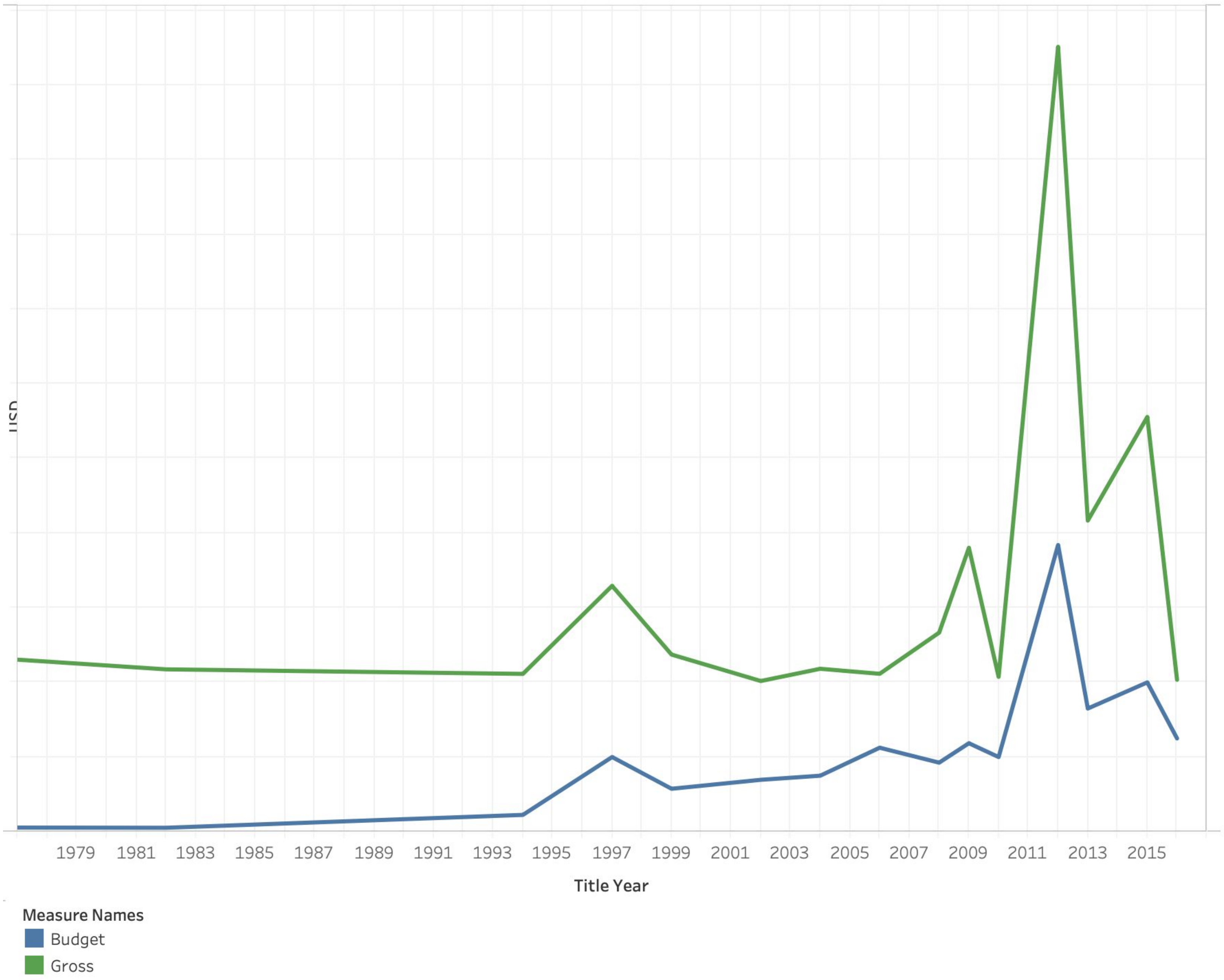
amazon RDS

tableau®

Result of Analysis

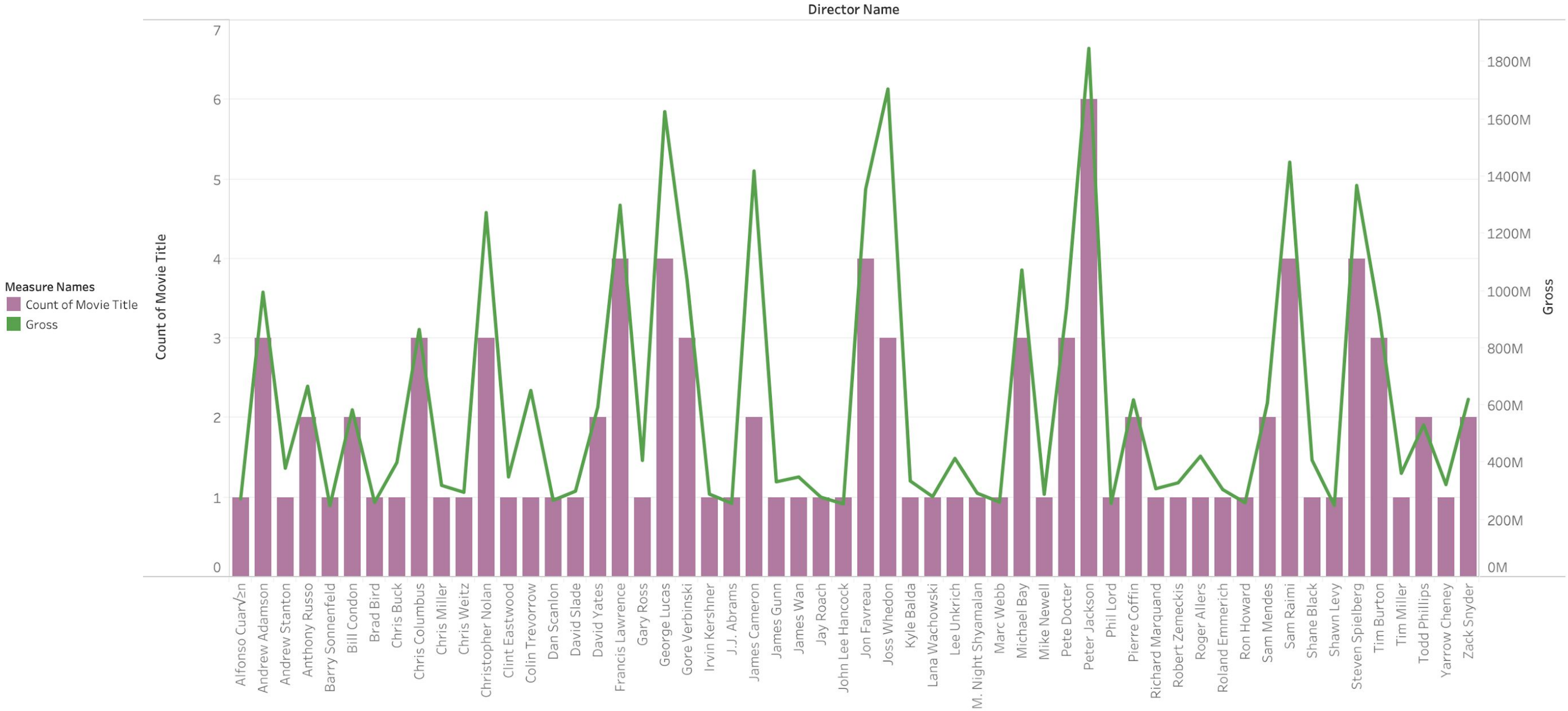


Top 20 Highest Gross Movies - Budget vs Gross



Result of Analysis

Top 100 Highest Gross Movies Directors



Preprocessing & Features

Preprocessing

Preprocessed data for machine learning training using:

- Numerical & textual data
- Textual features as categorical feature (e.g., director's name)

Features

Features used for gross income prediction:

- ➢ Numerical Features:
  - Director's Facebook likes
  - Movie's budget
  - Count of critics review for the movie
  - Cast's total Facebook likes
  - Movie's IMDB score
  - Duration of the movie
- ➢ Text Features:
  - Director's name

Preprocessing & Features

Data Cleaning

- Rows with NaN, missing gross value, and missing major feature values were removed from dataset.

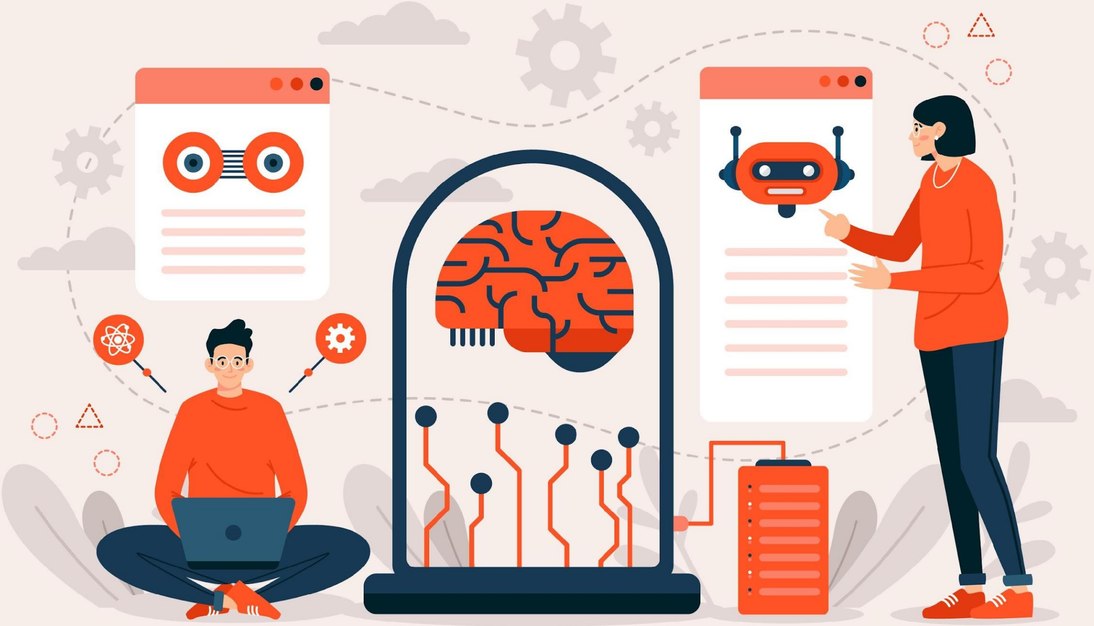
Numerical Features Preprocessing

- To test and train the dataset, the numerical data (movie's budget) was split into two groups.
  1. High budget movies
  2. Low budget movies

Motivation for using Textual Data as Categorical Data

- The focus for the textual data was the director's name column. Our goal was to draw conclusions on gross predictions based on the portfolio of the director and the count of their total movies produced.

Machine Learning Model



Random Forest & Neural Network

1. Prepared the input data and created a model
2. Trained and fit training data to the model

Why Neural Network model?

- Effective at detecting complex
- Nonlinear relationships
- Greater tolerance to messy data

Model Evaluation

- ✔ Both the Random Forest and Deep Learning models were able to predict correctly whether a **director's influence can significantly predict the gross income of a movie** by over **85%** of the time.

- ✔ Implementation and training times between the models varied:
  - Random Forest classifier was able to train on the large dataset and predict values **faster**, while the deep learning model required more time to train the data points.

Random Forest and Deep Neural Network Performance Evaluation

Evaluation Metrics	Deep Neural Network	Random Forest
Accuracy	0.8893	0.888

Further Research Opportunities

Does social media popularity of the cast impact film’s box office numbers?

Join Our Conversation!

Q & A Session

Thank You!