EXPLORING IMDB MOVIES DATASET:

EDA AND ML MODELING IN PYTHON

MSDS422B GROUP 9:

ANAMIKA KUMARI

KARAN RASTOGI

SHRIMATHY SRINIVASAN

GANESH SUBRAMANIAN AUDHIKESAVAN

Northwestern
SCHOOL OF
PROFESSIONAL STUDIES

OUTLINE

- Introduction
- Dataset Walkthrough
- Tools and Libraries Used
- Exploratory Data Analysis
- Data Visualization
- Machine Learning Modeling
- Challenges
- Conclusion & Future Scope

INTRODUCTION

The magic of movies transcends borders and languages, touching hearts and sparking imaginations across the globe. From timeless classics to groundbreaking blockbusters, every film has a story to tell, a message to convey, and an impact on its audience.

With IMDb, the world's most comprehensive and trusted source for movie information, we have the power to unlock the secrets behind these cinematic wonders.

With this project, we want to delve deeper into the following:

1. **Discovering Hidden Gems of Movies**
2. **Genre and Audience Analysis**
3. **Rating and Popularity Patterns**
4. **Directors and the Revenue Budget Analysis**
5. **NLP Analysis on Keywords and Genres**

Northwestern
SCHOOL OF
PROFESSIONAL STUDIES

DATASET WALKTHROUGH

- Dataset : imdb-movies.csv (Movies information data set from year 1960 to 2015)
- Link: https://www.kaggle.com/datasets/schoolofaitvm/imdbdataset
- This dataset has 10867 records and 21 columns.
- Features/Columns:

id	imdb_id	popularity	budget	revenue	original_ title	cast	homepage	director	tagline	keywords	overview
runtime	genres	production_c ompanies	release_date		vote_count	vote_avera ge	release_ year	budget_a dj	revenue_ adj		

- Features Type:

S.No	Feature Type	Columns
1.	Int	Id, imdb_id, budget, revenue, runtime, vote_count
2.	Float	Popularity, vote_average, budget_adj, revenue_adj
3.	String	Original_title, cast, homepage, director, tagline, keywords, overview, genres, Production companies,
4.	Date	Release_date, release_year

Northwestern
SCHOOL OF PROFESSIONAL STUDIES

TOOLS AND LIBRARIES USED:

- Anaconda IDE
- Jupyter Notebook
- Python
- Google Collab

Libraries and Packages Used:

- Sklearn
- Matplotlib
- Seaborn
- Numpy
- Pandas
- Tensorflow & Keras

DATA PREPROCESSING

IMDb

Data Duplication

1. Some entries were duplicated across rows
2. Duplicated entries were removed

Null Data Values

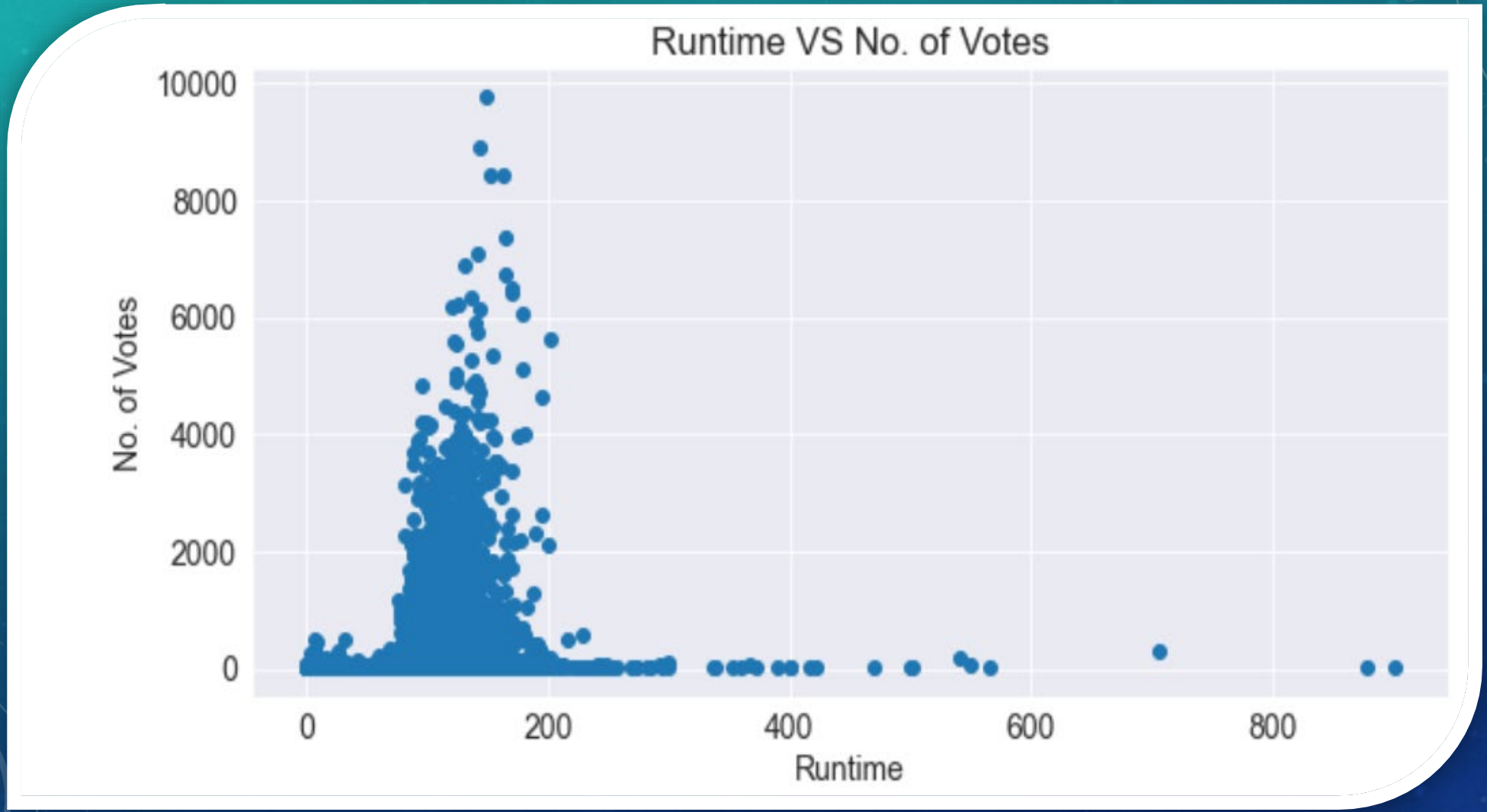
1. Columns such as cast(76), homepage (7930), director(44), tagline(2824), keywords (1493), production companies (1030), genres(23) have null values.
2. Columns which have low null values have their respective records dropped.

Dropping Unnecessary Columns

1. ‘homepage',' tagline’ were dropped as they contain non relevant information
2. NaN values in keywords and production_companies were replaced with Not Known string for analysis purpose.

EXPLORATORY DATA ANALYSIS

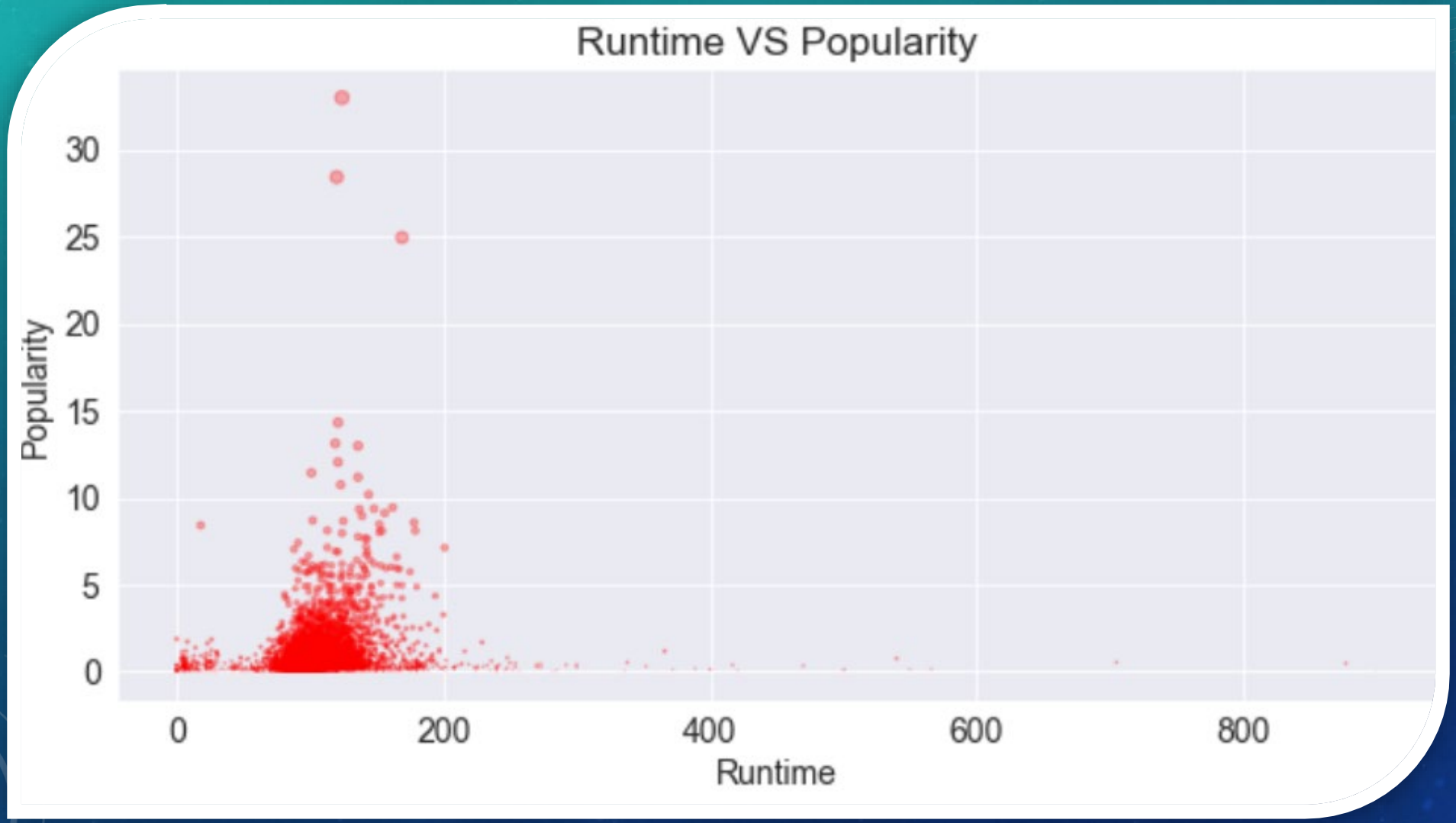
- Graph of Runtime vs Number of Votes



Northwestern
SCHOOL OF
PROFESSIONAL STUDIES

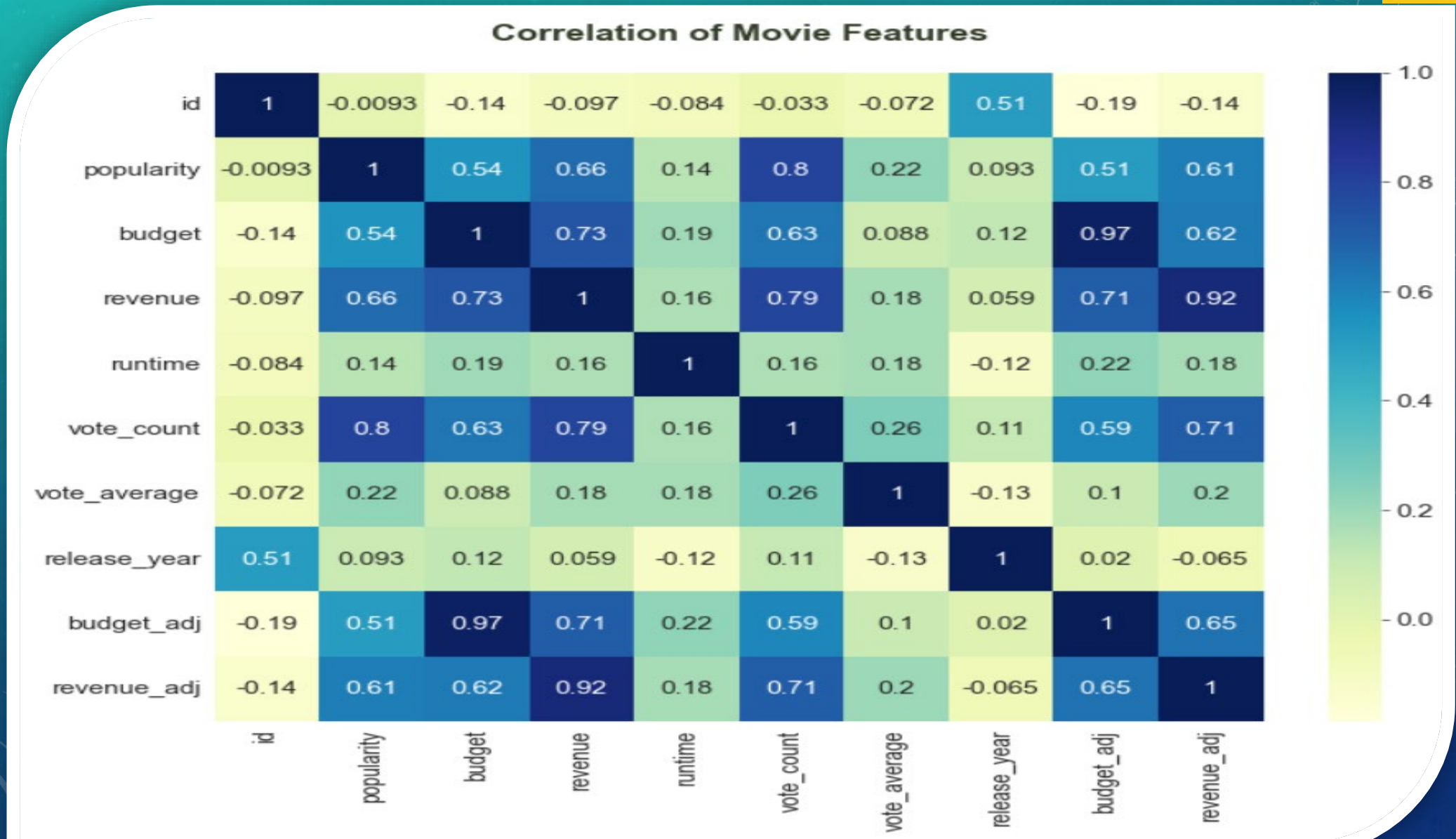
EXPLORATORY DATA ANALYSIS

- Graph of Runtime vs Popularity



CORRELATION HEAT MAP

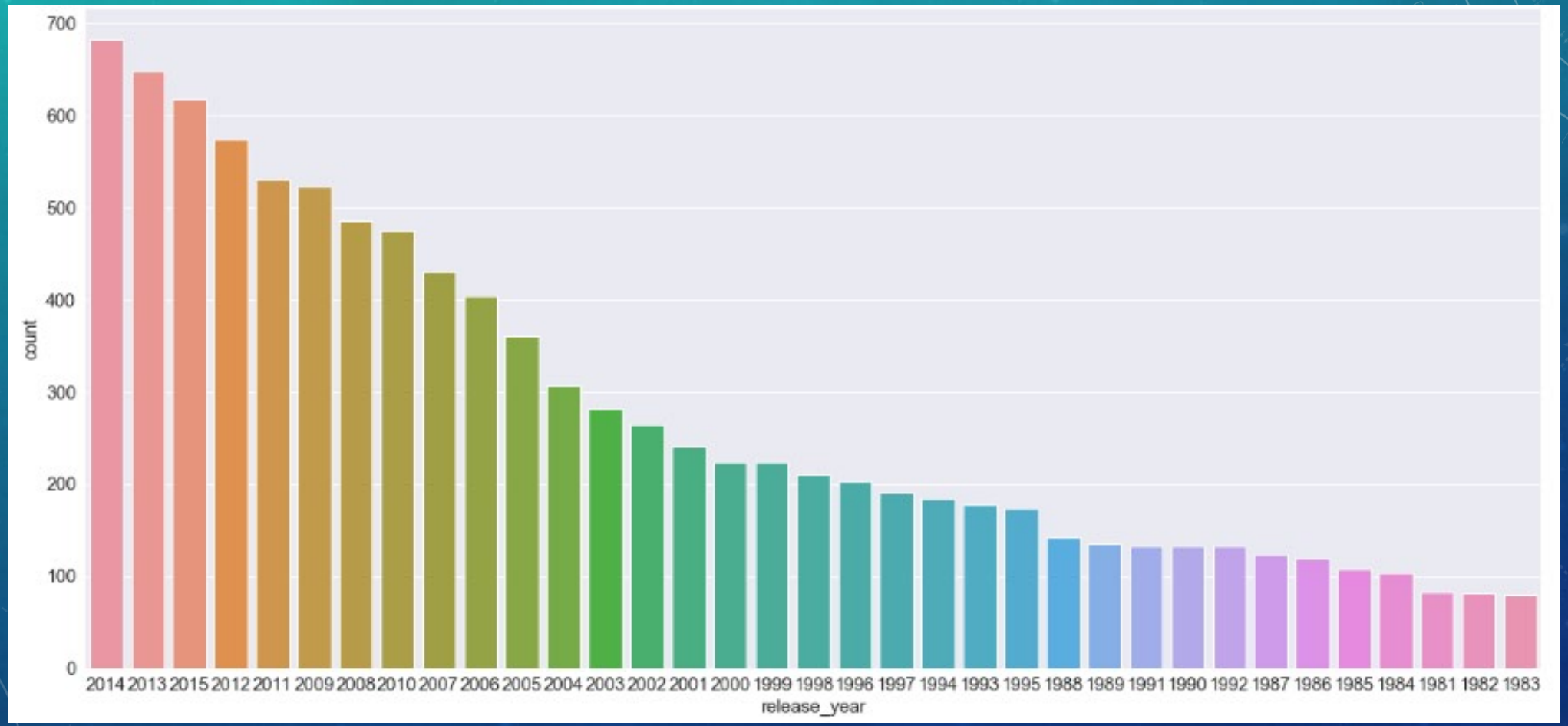
IMDb



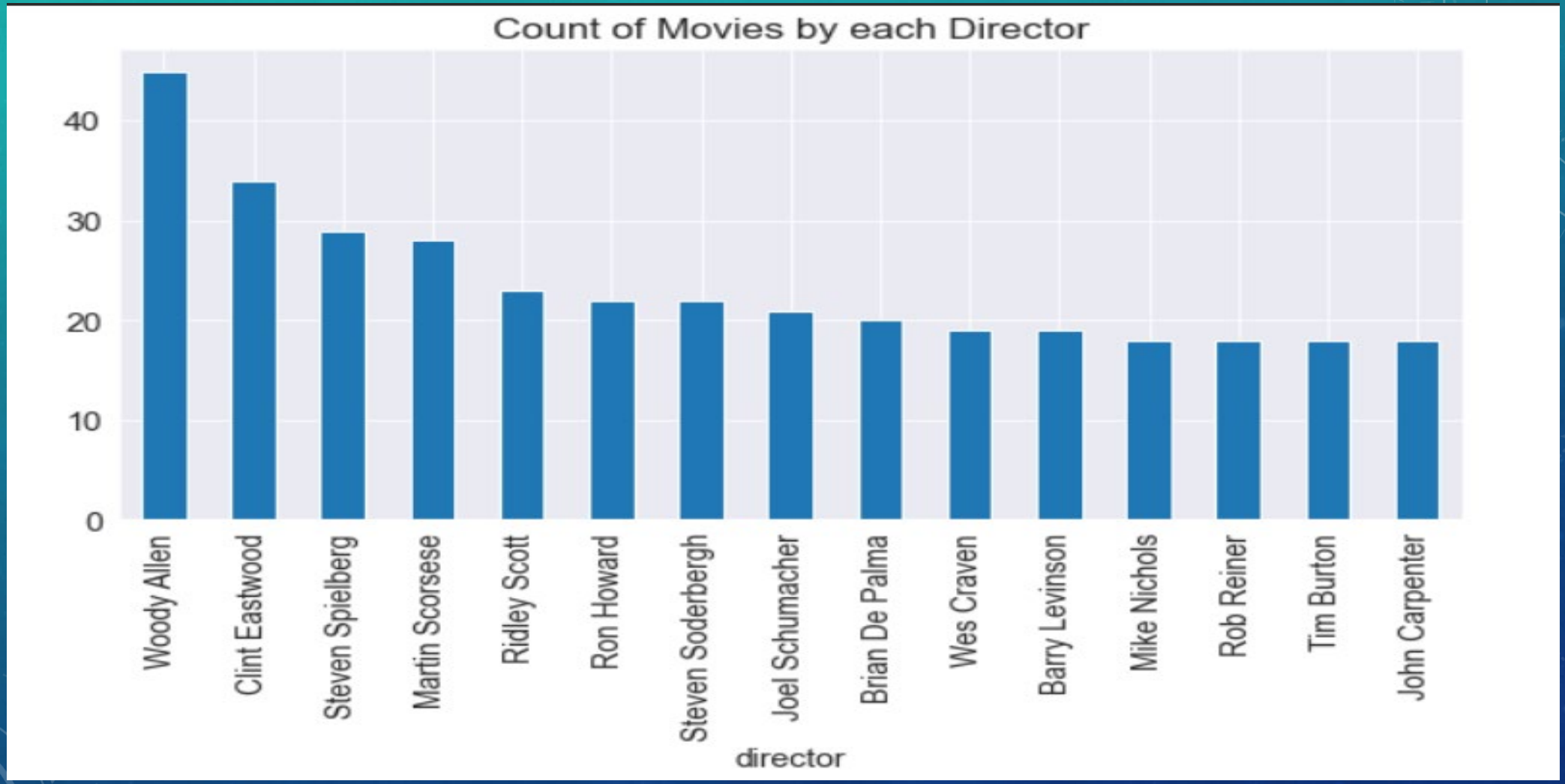
Northwestern
SCHOOL OF
PROFESSIONAL STUDIES

DATA VISUALIZATION

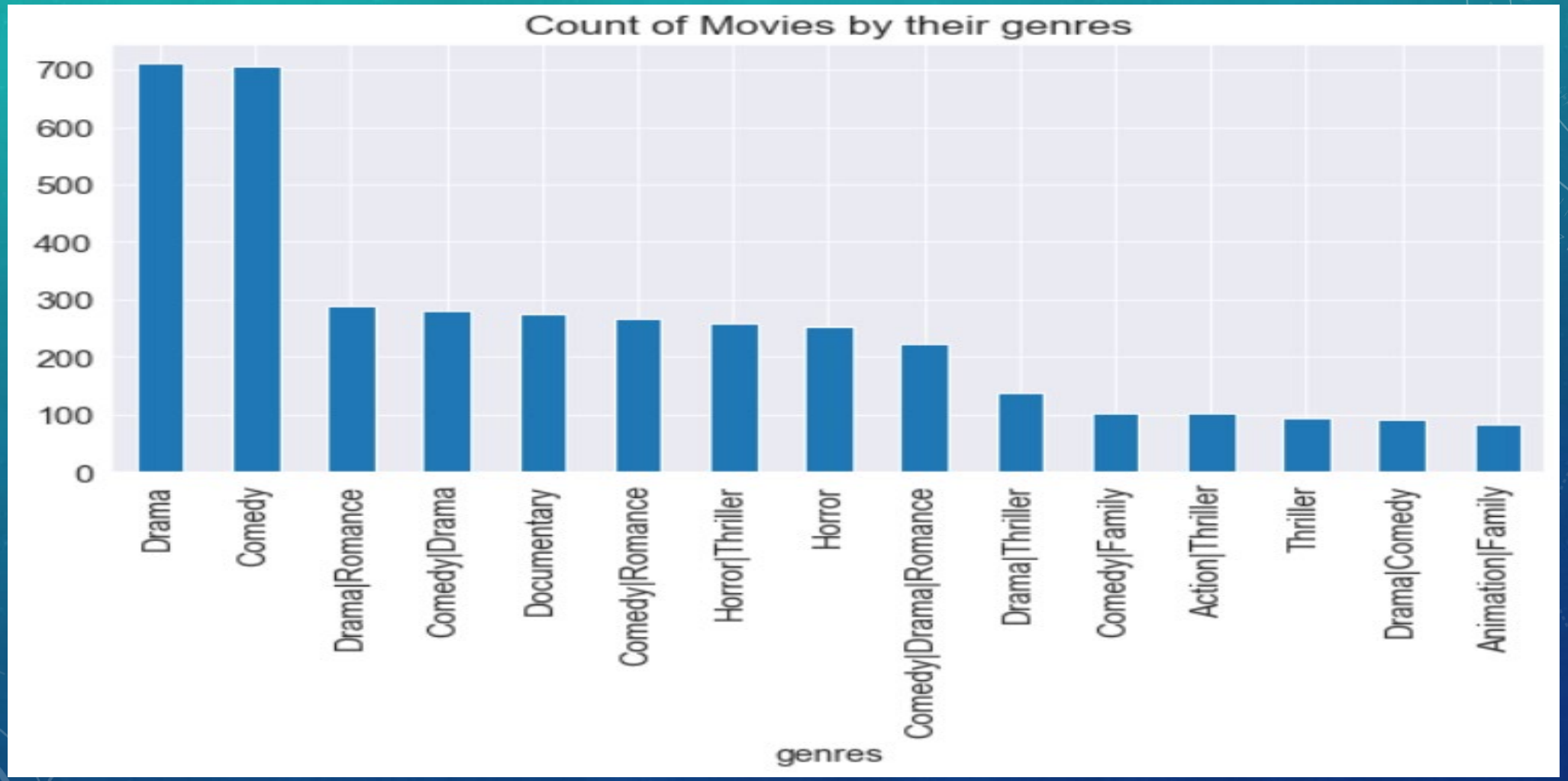
- Year wise movie count by graph



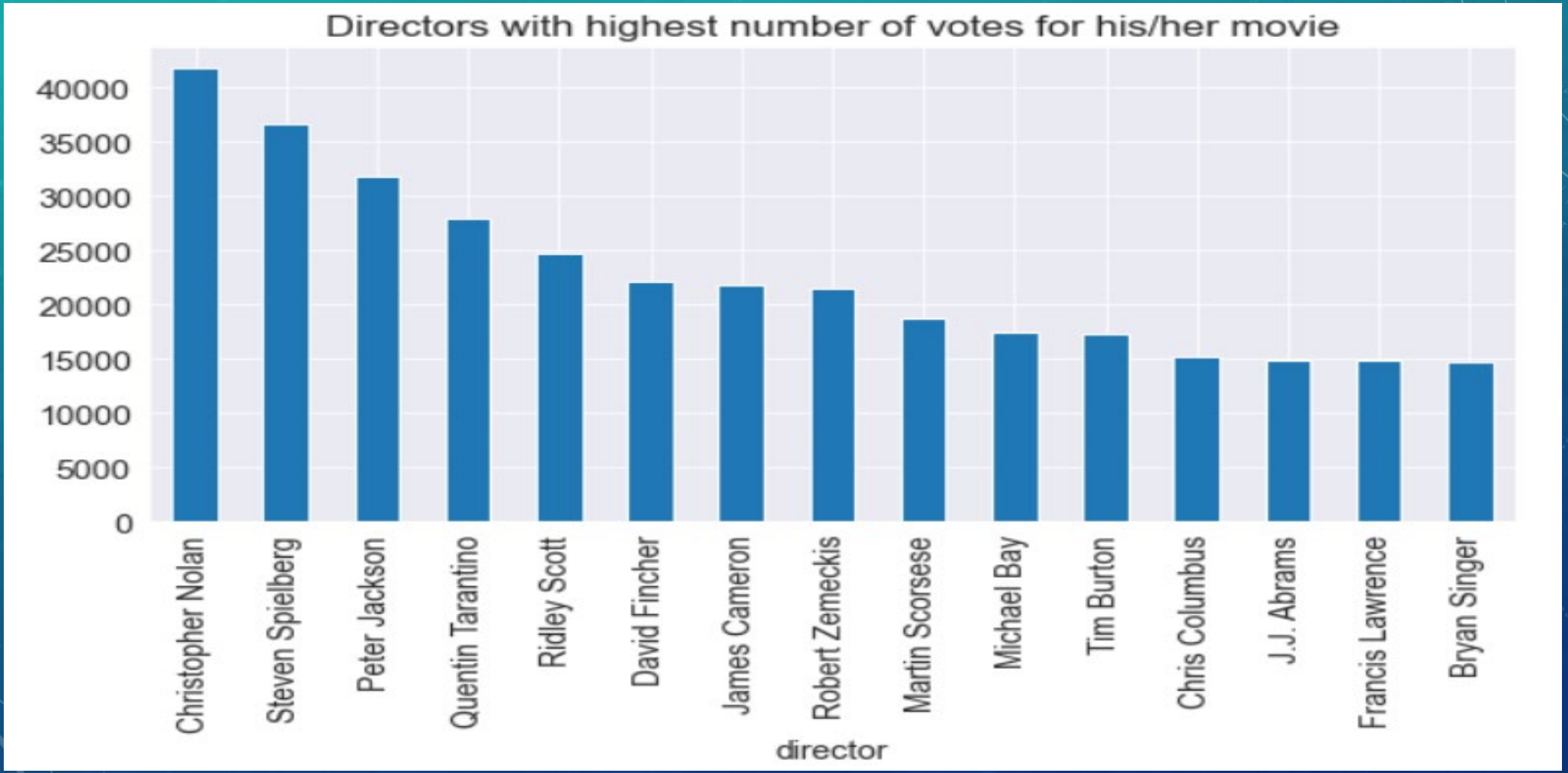
NUMBER OF MOVIES CREATED BY EACH DIRECTOR



NUMBER OF MOVIES BY THEIR GENRES

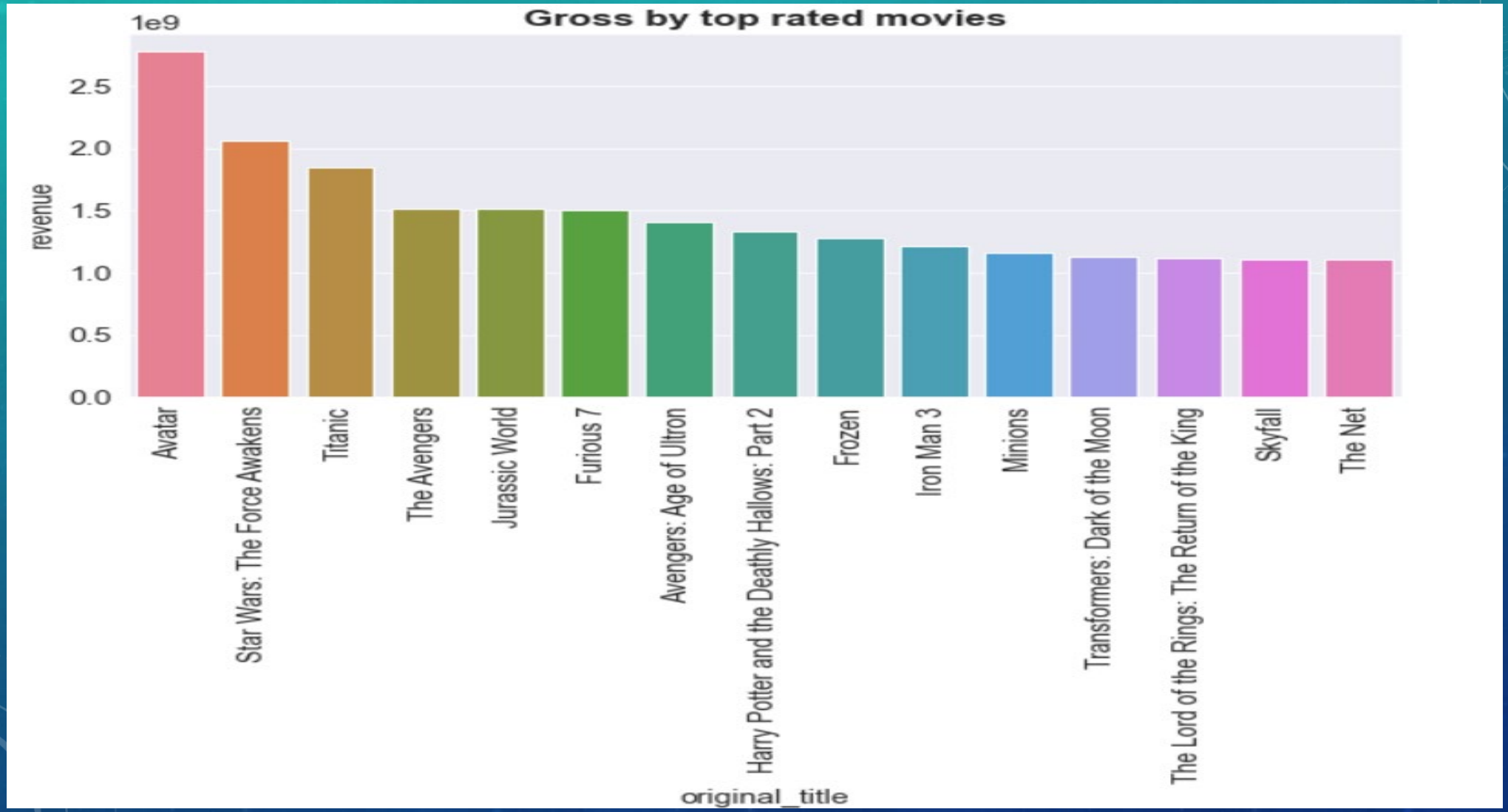


DIRECTORS WITH HIGHEST NUMBER OF VOTES FOR HIS/HER MOVIE

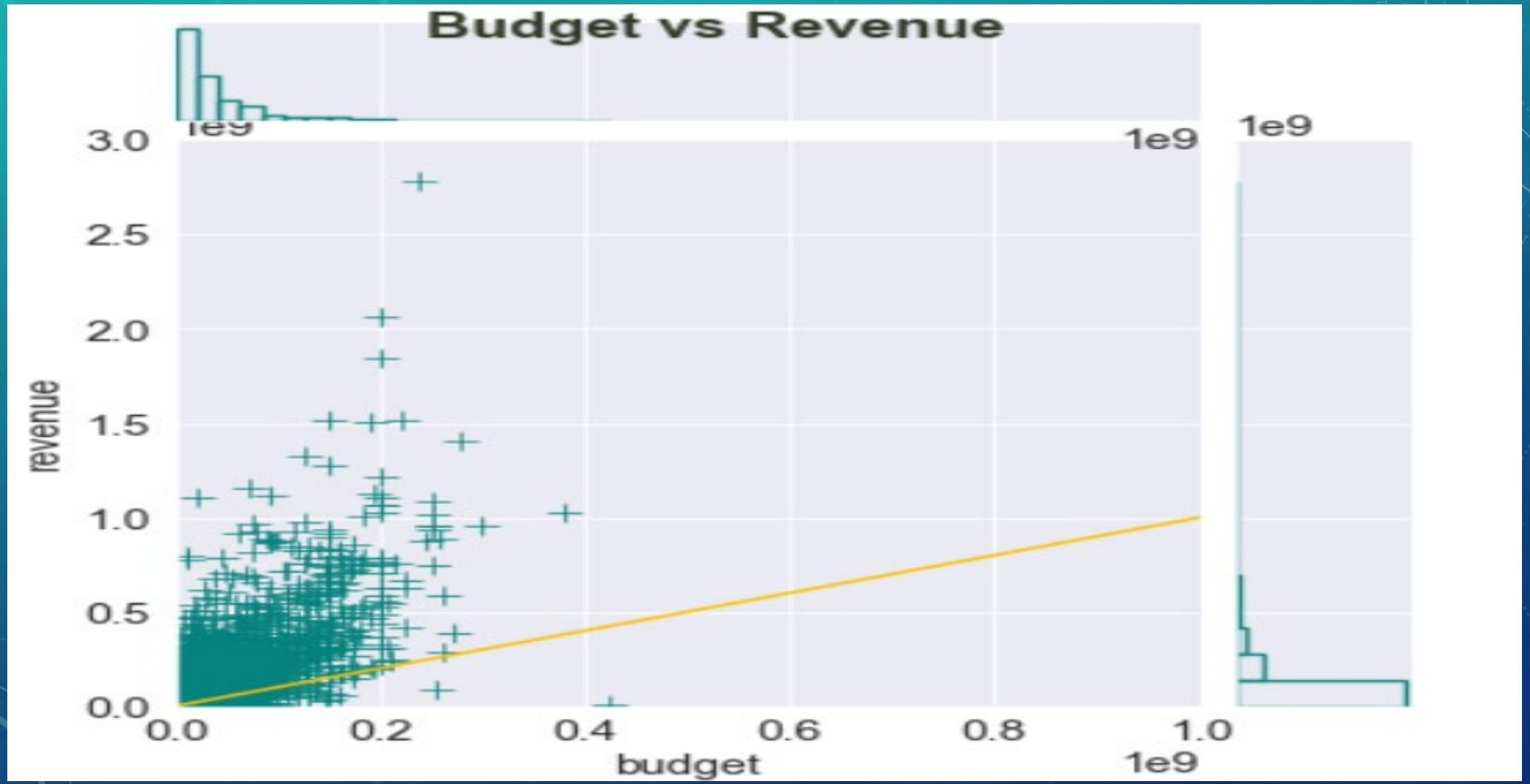


Northwestern
SCHOOL OF
PROFESSIONAL STUDIES

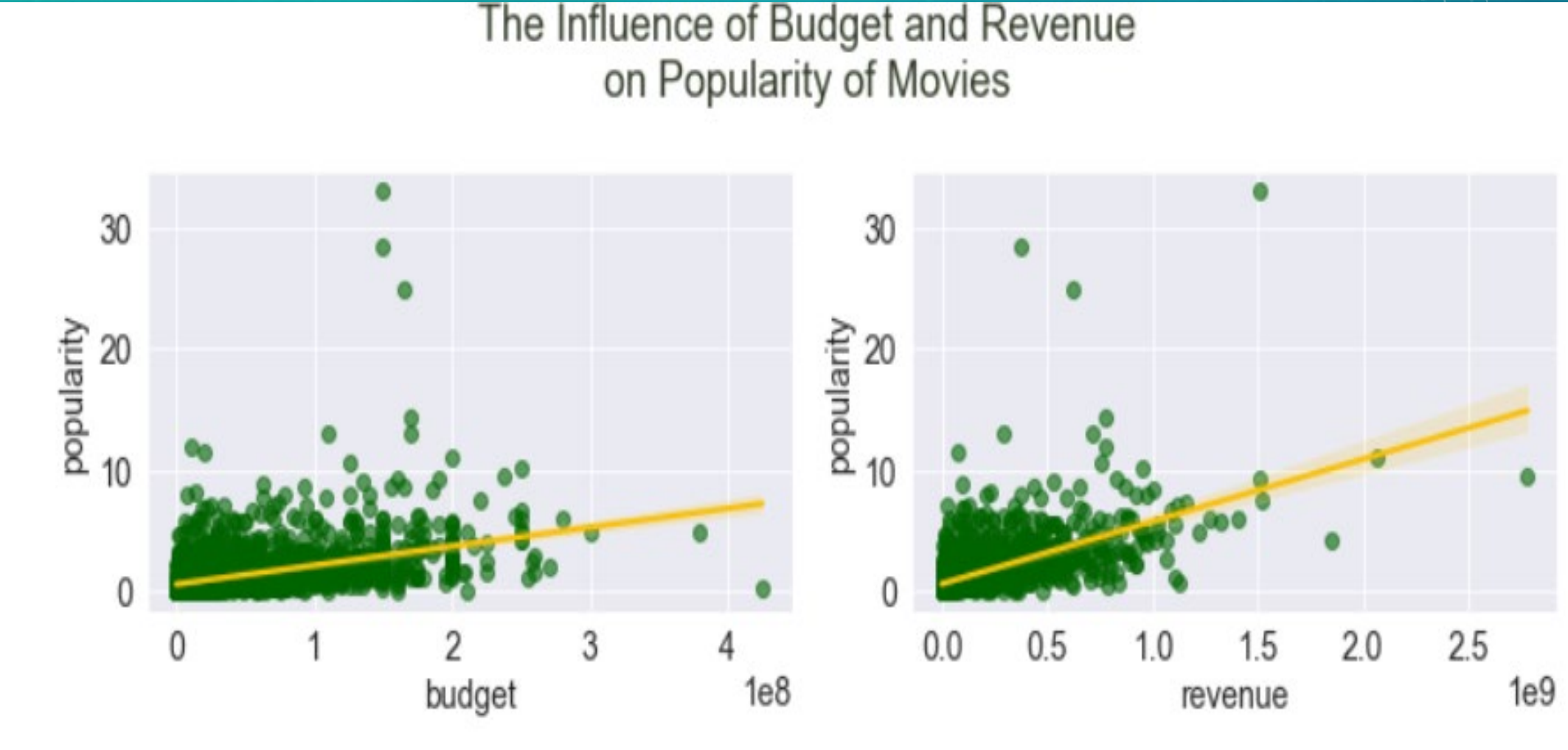
BAR PLOT TO SHOW TOP 15 MOVIES BY DESCENDING ORDER OF GROSS REVENUE



SCATTERPLOT: BUDGET VS REVENUE



INFLUENCE OF BUDGET AND REVENUE ON POPULARITY RATING



ML MODELING: INTENT & USE CASES

Revenue: Estimation and Prediction

- Linear Regression with Cross Validation
- Linear Regression with Regularization such as Lasso and Ridge
- XGBoost Regressor
- XGBoost Regressor with Cross Validation and Hyperparameter Tuning

Sentiment Analysis

- Natural Language Processing Multinomial NB for Keywords, Genres and overview features
- LSTM (Long Short-Term Memory)

Influence Factors

- Neural Networks
- Gradient Boost

ML MODELS OBSERVATIONS AND RESULTS

- **Regression Analysis**

1. Dependent Variable: Revenue

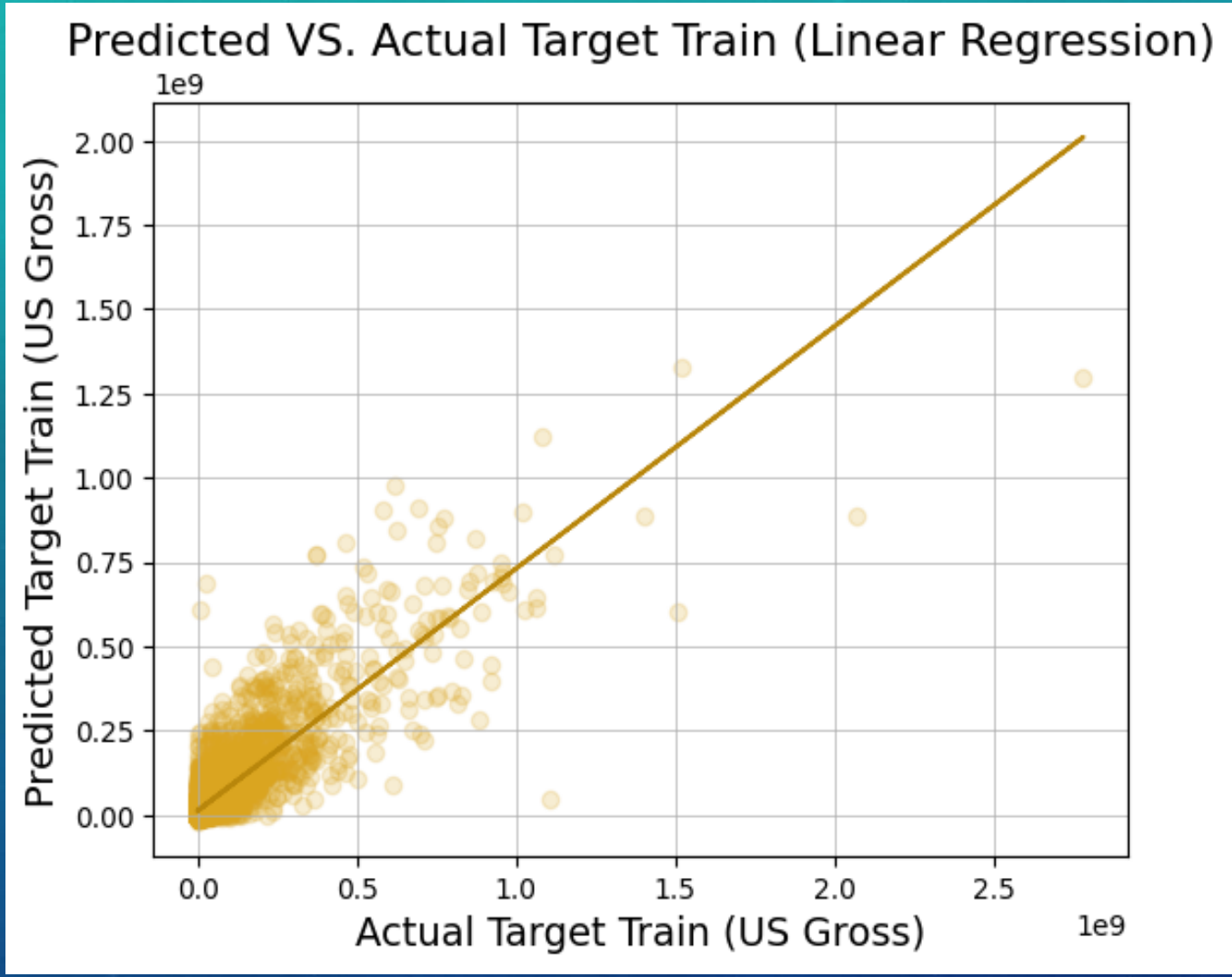
2. Independent Variable: Popularity, Budget and Vote_Count

The models and their performance are as follows:

S.No.	Model Name	Training Score	Validation/Test Score
1.	Linear Regression	0.718	0.67
2.	Linear Regression K fold CV	0.718	0.71
3.	Polynomial Regression (degree=2)	0.768	0.69
4.	Ridge Regression	0.718	0.67
5.	Ridge Regression with CV	0.718	0.71
6.	Lasso Regression with CV	0.718	0.713

*** Model highlighted with purple cells are best performing model/s amongst all.

Predicted vs Actual Values with Linear Regression



Northwestern
SCHOOL OF
PROFESSIONAL STUDIES

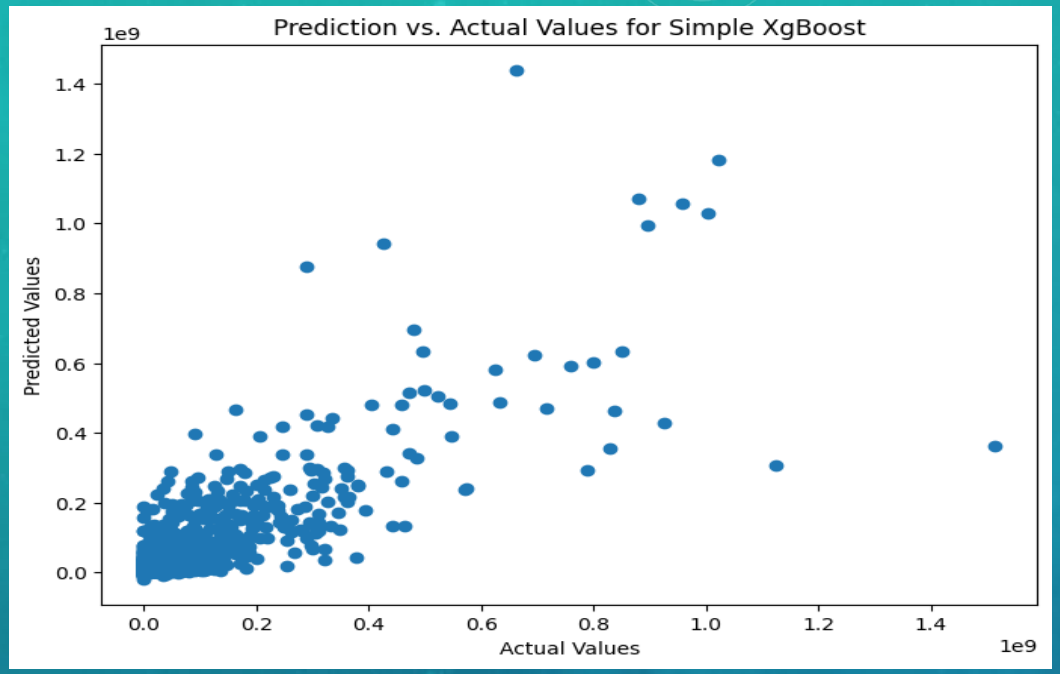
REGRESSION ANALYSIS (CONTD)

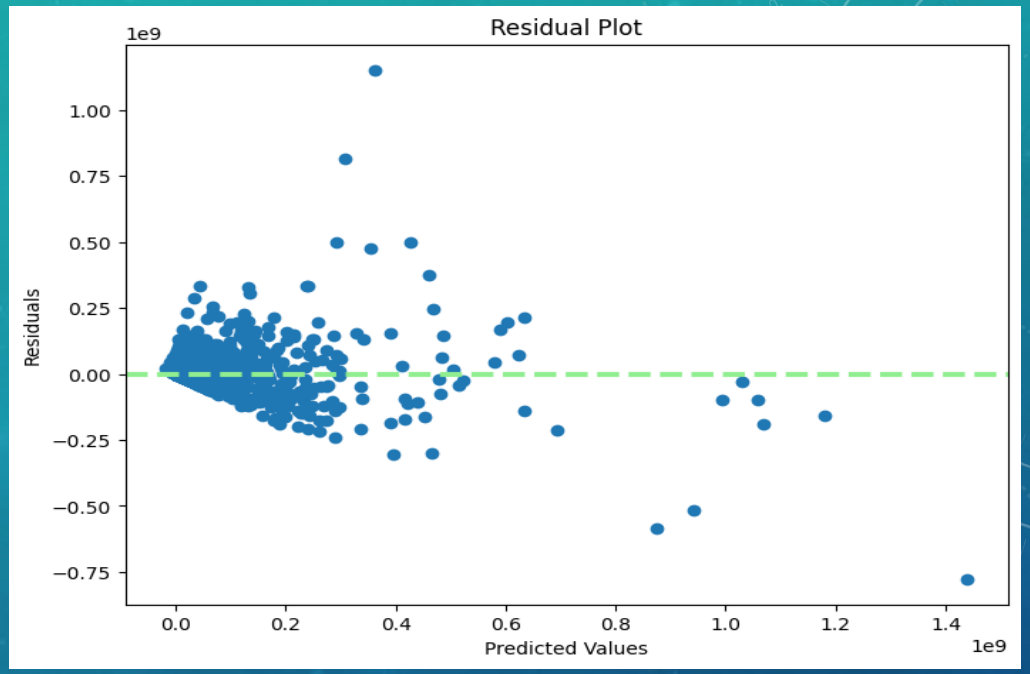
- Further we also tried XGBoost Regressor with RFE and cross validation and hyperparameter Tuning.

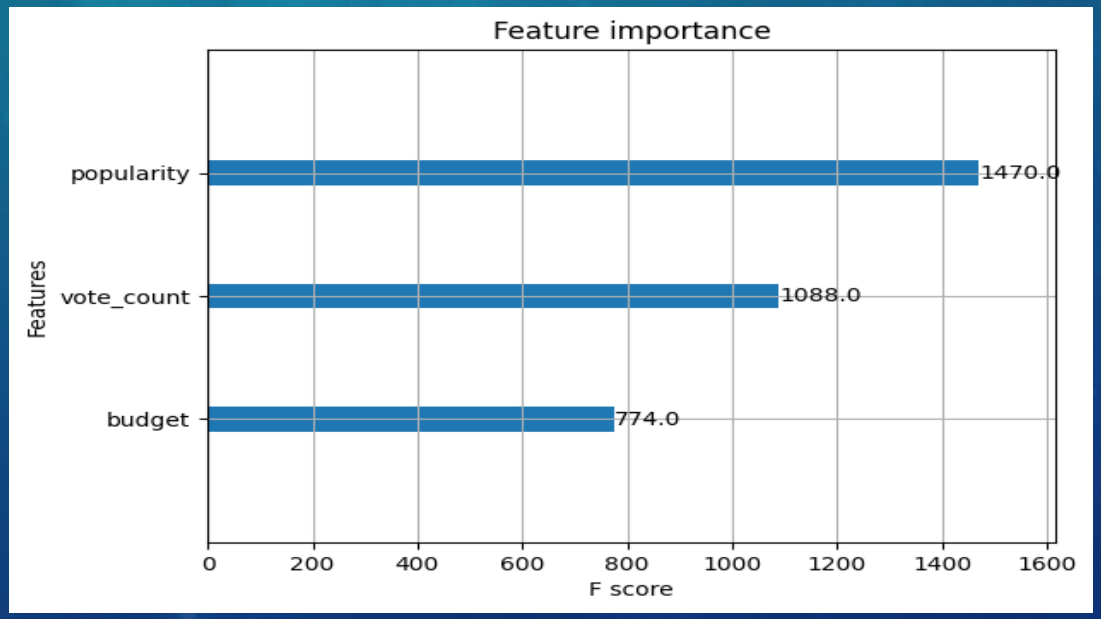
S.No	Regression Model	MAE	RMSE	R2
1.	XGBoost Regressor	26256665.263	75964208.27	0.64
2.	XGBoost Regressor with Cross Validation and Hyper parameter Tuning	22622848.21	71043604.95	0.7139
3.	XGBoost Regressor with CV and Hyperparameter tuning and RFE	26922083.78	67367538.39	0.668

***** Model highlighted with purple cells is best performing model amongst all.**

Northwestern
SCHOOL OF
PROFESSIONAL STUDIES







Northwestern
SCHOOL OF
PROFESSIONAL STUDIES

SENTIMENT ANALYSIS

- Aim is to use NLP for Sentiment analysis on ‘Keywords’ and ‘Genres’ columns to get the Positive, Neutral and Negative Sentiment. Metrics Results are as follows:

Sentiment Analysis for Keywords:

	precision	recall	f1-score	support
negative	0.92	0.47	0.62	285
neutral	0.85	0.98	0.91	1298
positive	0.88	0.68	0.77	291
accuracy			0.86	1874
macro avg	0.88	0.71	0.77	1874
weighted avg	0.87	0.86	0.85	1874

Sentiment Analysis for Genres:

	precision	recall	f1-score	support
negative	1.00	0.97	0.98	29
neutral	1.00	0.99	0.99	1444
positive	0.97	0.99	0.98	401
accuracy			0.99	1874
macro avg	0.99	0.98	0.98	1874
weighted avg	0.99	0.99	0.99	1874

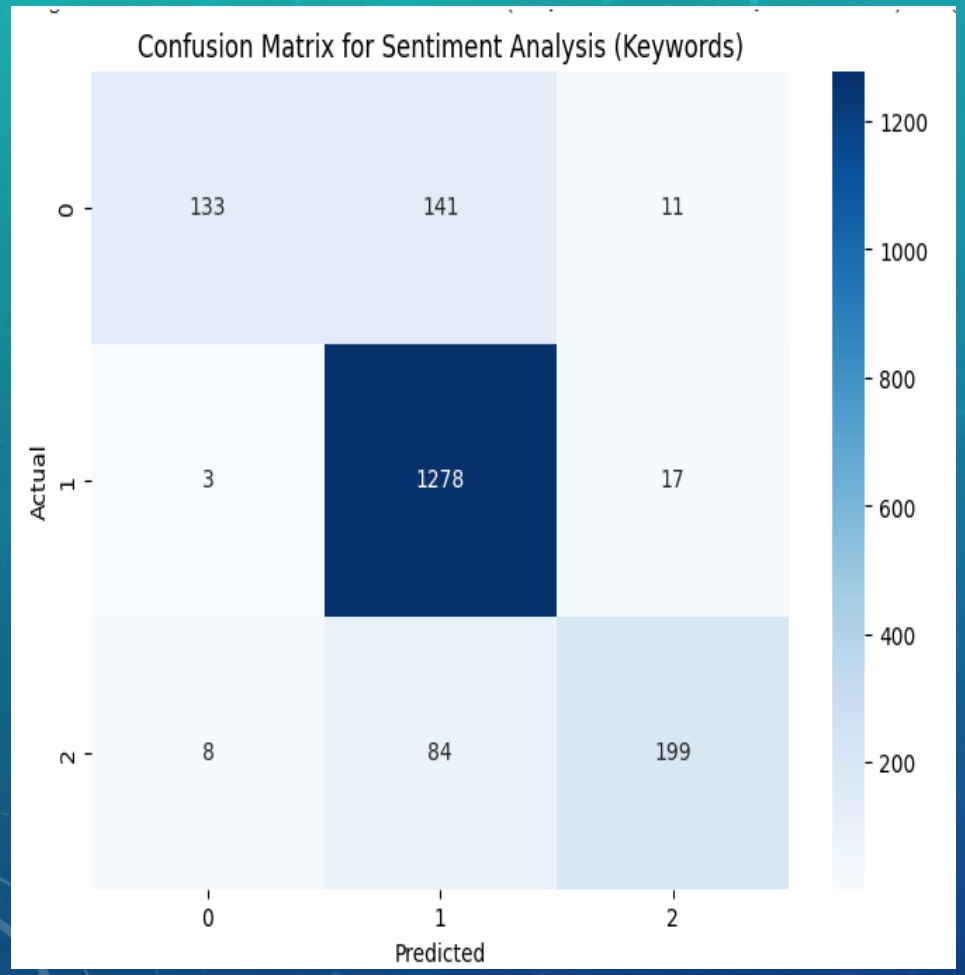
Northwestern
SCHOOL OF
PROFESSIONAL STUDIES

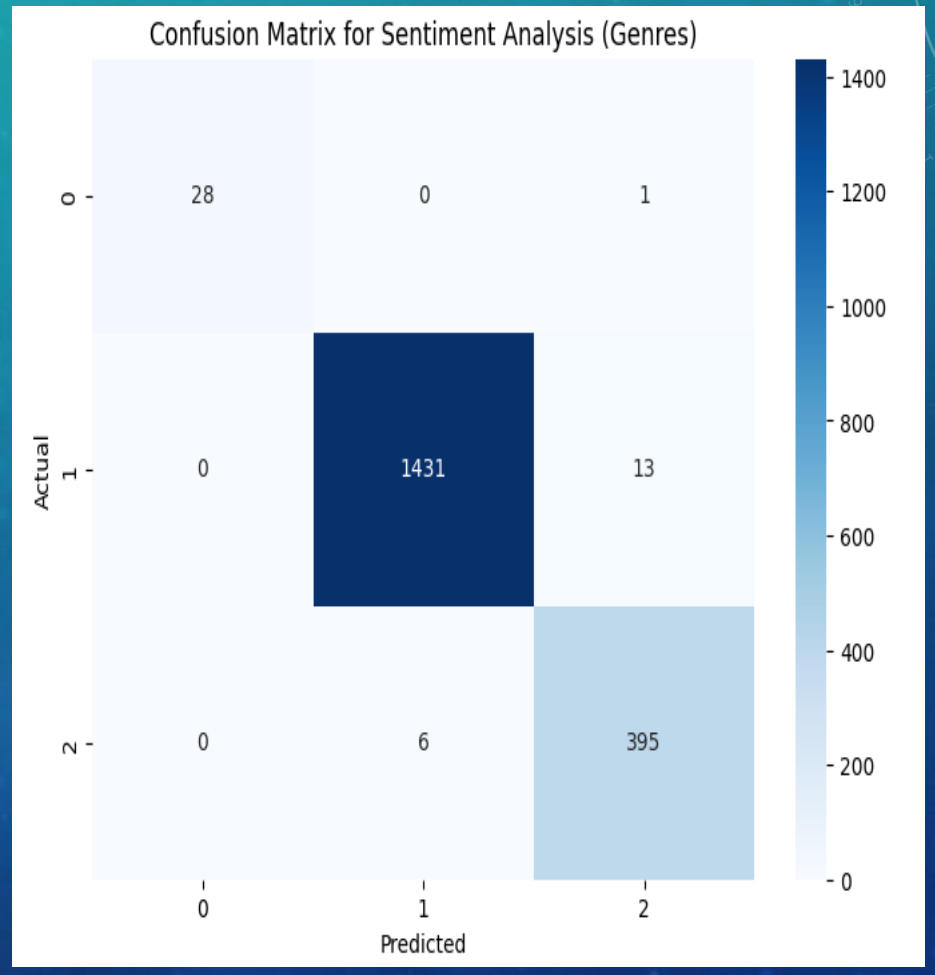
SENTIMENT ANALYSIS (CONTD...)

Sentiment Analysis Models Used Summary

	Keywords	Genres	Overview
Model	MultinomialNB	MultinomialNB	MultinomialNB
Description	Naive Bayes classifier	Naive Bayes classifier	Naive Bayes classifier
Accuracy	0.86	0.99	0.66
Precision	0.92, 0.85, 0.88	1.00, 1.00, 0.97	0.74, 0.00, 0.64
Recall	0.47, 0.98, 0.68	0.97, 0.99, 0.99	0.44, 0.00, 0.94
F1-Score	0.62, 0.91, 0.77	0.98, 0.99, 0.98	0.55, 0.00, 0.76

CONFUSION MATRIX FOR KEYWORDS & GENRES





Northwestern
SCHOOL OF
PROFESSIONAL STUDIES

LONG SHORT TERM MODEL(LSTM FOR SENTIMENT ANALYSIS)

•The LSTM model's performance in sentiment analysis on the combined text data is relatively poor, as indicated by the accuracy of around 54.27%.

•The consistent validation accuracy throughout the training suggests that the model's architecture or data representation might need further refinement to better capture sentiment nuances from the combined text.

•Additional feature engineering, hyperparameter tuning, and exploring more sophisticated model architectures could potentially enhance the model's performance.

```
Epoch 1/5
235/235 [==============================] - 18s 65ms/step - loss: 0.9930 - accuracy: 0.5148 - val_loss: 0.9891 - val_accuracy: 0.5427
Epoch 2/5
235/235 [==============================] - 15s 63ms/step - loss: 0.9901 - accuracy: 0.5159 - val_loss: 0.9906 - val_accuracy: 0.5427
Epoch 3/5
235/235 [==============================] - 16s 68ms/step - loss: 0.9899 - accuracy: 0.5159 - val_loss: 0.9971 - val_accuracy: 0.5427
Epoch 4/5
235/235 [==============================] - 20s 83ms/step - loss: 0.9903 - accuracy: 0.5159 - val_loss: 0.9920 - val_accuracy: 0.5427
Epoch 5/5
235/235 [==============================] - 15s 64ms/step - loss: 0.9903 - accuracy: 0.5159 - val_loss: 0.9925 - val_accuracy: 0.5427
```

Northwestern
SCHOOL OF PROFESSIONAL STUDIES

COMPARISON BETWEEN DIFFERENT MODELS FOR SENTIMENT ANALYSIS

Model/Scenario	Multinomial NB (Keywords)	Multinomial NB (Genres)	LSTM (Combined Keywords & Genres)
Accuracy	0.86	0.99	0.54

Northwestern
SCHOOL OF
PROFESSIONAL STUDIES

INFLUENCE FACTORS ON MOVIE RATINGS

The primary objective of this analysis is to uncover the influence of production companies, directors, genres, and cast members on movie vote ratings. By leveraging machine learning models, we aim to:

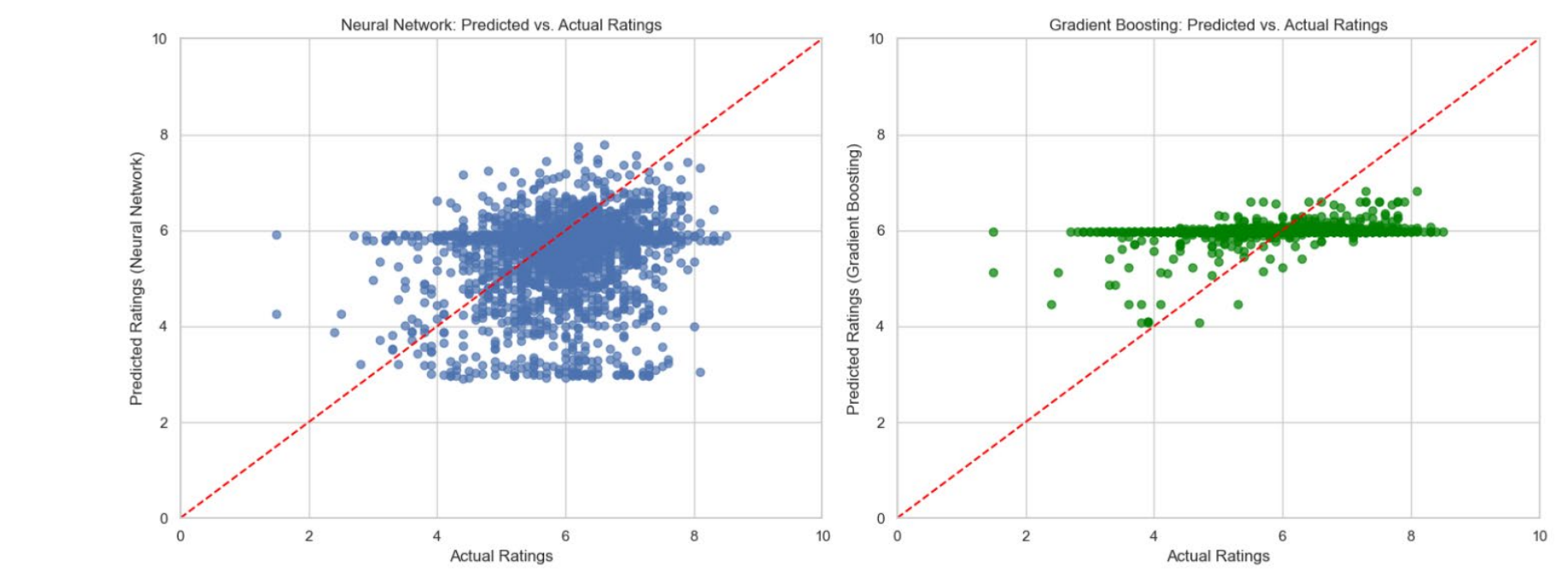
- ✓ Investigate how different factors contribute to movie ratings.
- ✓ Provide actionable insights for the film industry to enhance decision-making processes.
- ✓ Offer a comprehensive analysis of the relationships between key components and ratings.

MODEL SELECTION

Models used for determining influence of various features on the IMDB Ratings:

- **Gradient Boosting Regressor:** Robust ensemble model
- **Neural Network:** Complex relationships captured by multi-layer architecture;

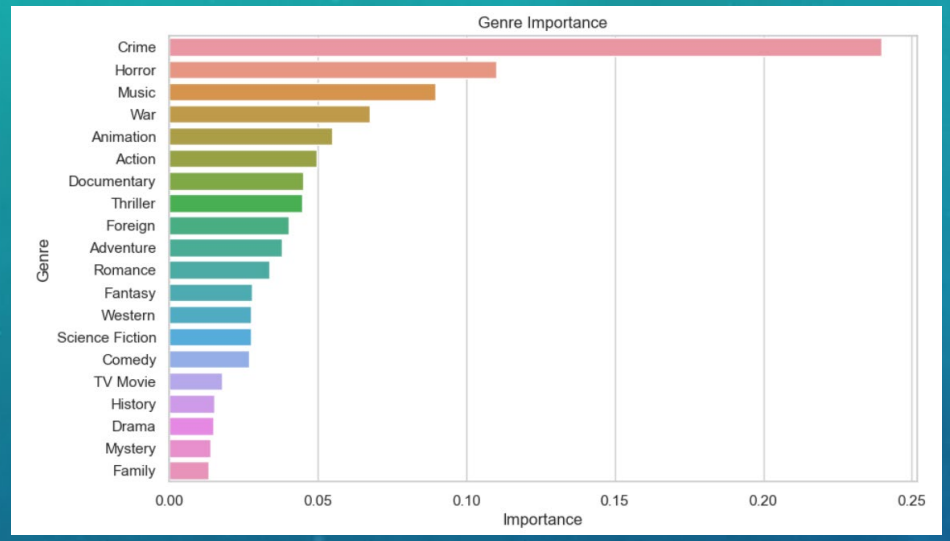
Feature Importance: Both models reveal the impact of production companies, genre, cast, and director collaborations on movie ratings.

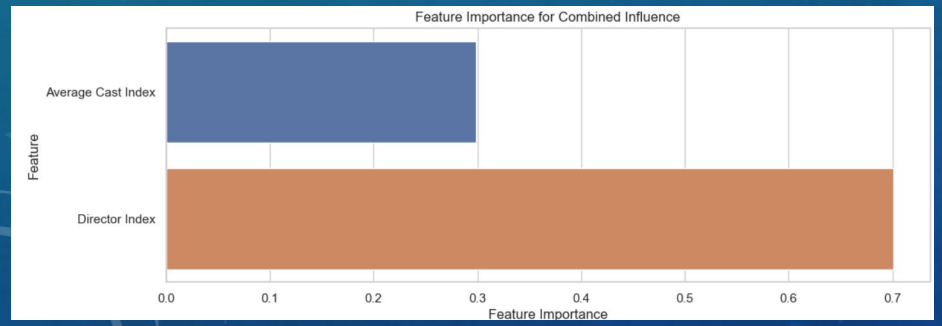


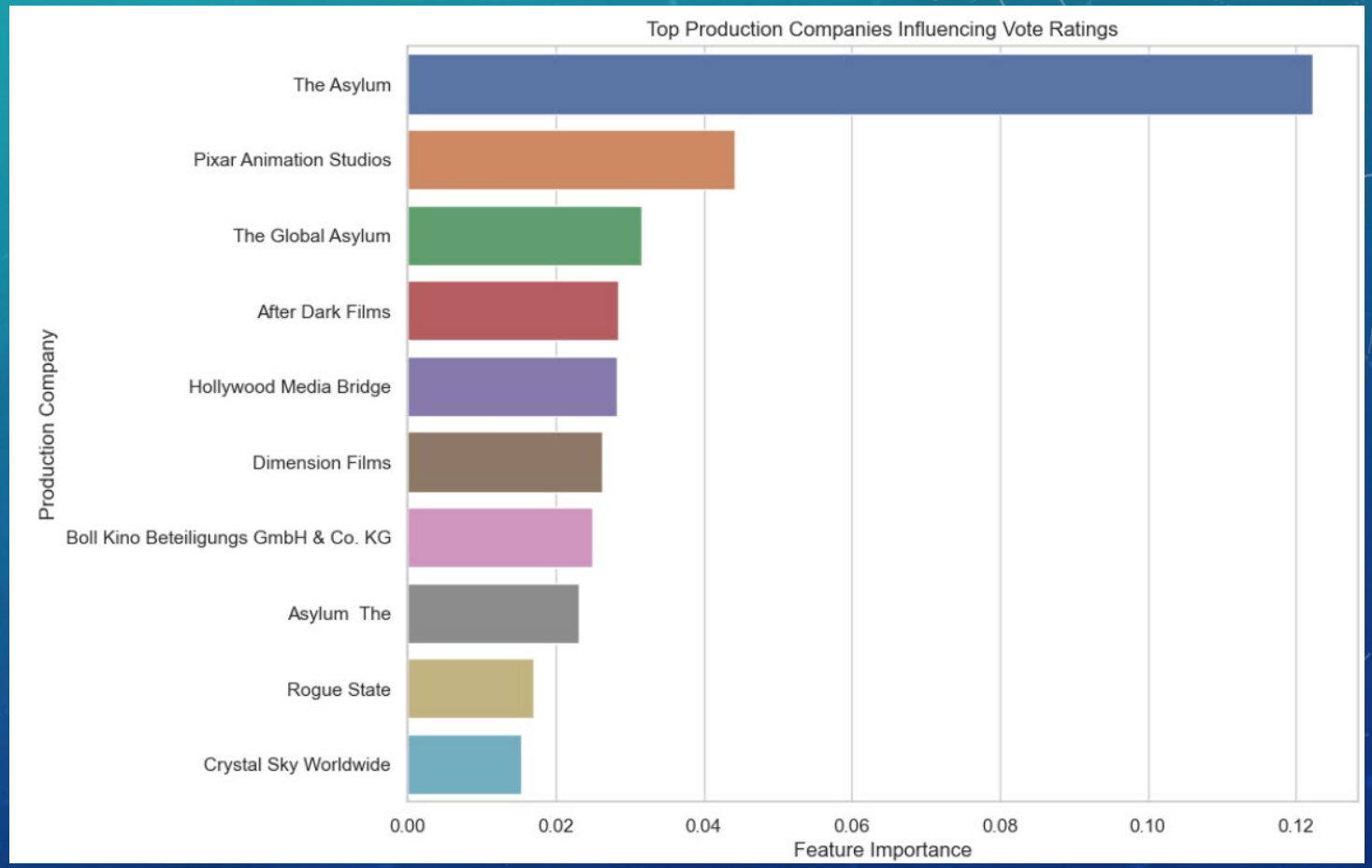
PERFORMANCE AND INSIGHTS

MODELS ->	GRADIENT BOOSTING REGRESSOR	NEURAL NETWORK
INFLUENCE FACTORS	MEAN SQUARED ERROR (MSE)	
Production Companies	0.801031	1.526187
Director And Cast Index	0.894547	0.873106
Genre	0.744598	0.736392

FEATURE IMPORTANCE







CHALLENGES

Elusive Optimal Accuracy

The XGBoost regressor plateaued at a 71% accuracy, signalling complex feature correlations and multicollinearity. Extensive hyperparameter tuning and cross-validation was needed

Hurdles in Heteroscedasticity

The Residual Plot unveiled heteroscedasticity's presence, posing a barrier to effective transformations. The quest for innovative features emerged to enhance model performance.

Navigating Sentiment Analysis

Sentiment Analysis faced intricacies in capturing positive sentiments accurately.

These challenges unlocked innovative avenues and highlighted the importance and need of additional relevant features for enhanced accuracy.

CONCLUSION

Decoding Cinematic Magic: Our analysis pierced the silver screen's veil, unraveling the enigma of movie influence factors—production companies, directors, genres, and more.

From Data to Decision: Armed with insights, the film industry can chart informed paths, leveraging historical data to guide budget decisions, create compelling narratives, and captivate audiences.

Sentiments Whispered and Heard: Our sentiment analysis decoded audience sentiments, giving filmmakers the power to comprehend and respond to viewer emotions.

Challenges as Stepping-Stones: Through challenges, we found stepping-stones toward innovation, inspiring us to forge ahead with inventive features and techniques.

PAVING THE WAY AHEAD…

Explore- Sky Is The Limit:

Our findings offer a compass for production companies and directors, guiding budget decisions through historical data's lens—votes, popularity scores, revenue, and profit intertwining.

We look forward to discovering more factors that will help us enhance our accuracy and gain better insights

Decoding User Sentiment:

Merging user reviews with our dataset unveils sentiment nuances, providing filmmakers with a magnifying glass to understand audience perspectives. (IMDB DATASET with 50k records)

Beyond Horizons:

The voyage doesn't culminate here. The journey forward encompasses a broader horizon, where we expand features and employ advanced techniques to conquer challenges and unveil deeper revelations.

Northwestern
SCHOOL OF
PROFESSIONAL STUDIES

Thank You!