Dsf-pt09-grp2

Presented by:

- Humphrey Matagaro
- Charity Nguru
- Jael Kirwa
- Evelyn Mwangi
- Okech Maximillan
- Samuel Maina

Presentation Structure

Beginning

- Overview
- Business Understanding

Middle

- Data Understanding
- Data Analysis

End

- Recommendations
- Next Steps
- Thank You

Beginning.

- Overview
- Business Understanding

Project Overview

This project explores current movie trends to help guide the business decisions of a newly launched movie studio. Our project goal is to analyze various data sets from reputable movie sources such as Box Office Mojo(BOM), IMDB, Rotten tomatoes and The Numbers , to identify actionable insights that the movie studio can use to produce films with high box-office potential.

Business understanding

Stakeholder;

- The head of the new movie studio.

Business Problem;

- The new company wants to understand the factors that influence box-office performance and how they can apply this knowledge to create films that resonate with audiences and make profit.

Key business questions:

- Which are the best performing movies categories to produce?
- Which demographics watch movies ?
- What each demographic prefer to watch ?

MIDDLE

- Data Understanding
- Data Analysis

DATA UNDERSTANDING

Source of data

The analysis is based on five datasets from various movie industry sources;

- IMBD database - IMDB data source analysis
- Box Office Mojo-BOM movie company analysis
- Rotten Tomatoes
- The Movie DB
- The Numbers

Description of data

- **bom_movie_gross.csv** - Our data has a total of 5 columns and 3387 rows. foreign_gross has 1350 missing values. domestic_gross has 28 missing values. studio has 5 missing values. foreign_gross (object) column needs to be converted to numeric.
- **tn_movie_budgets.csv** -the dataset has 5782 rows and 6 columns. In the production_budget, domestic_gross and worldwide_gross columns, remove currency sign and commas to make them fully numeric.combine both the domestic gross income and worldwide gross income into a new column called total gross income and this will form the basis of our analysis later.

DATA ANALYSIS

1. BOM movie company analysis(Maina Ndirangu)
2. IMDB movie company analysis(Max & Jael)
3. TMDB data source movie analysis (Evelyn)
4. TN movie budgets dataset(Humphrey and Charity)

BOM MOVIE COMPANY ANALYSIS

(MAINA NDIRANGU)

Objectives:

1. To determine if foreign gross earnings are generally higher than domestic gross.
2. To investigate if the foreign gross has greater variability compared to domestic gross.

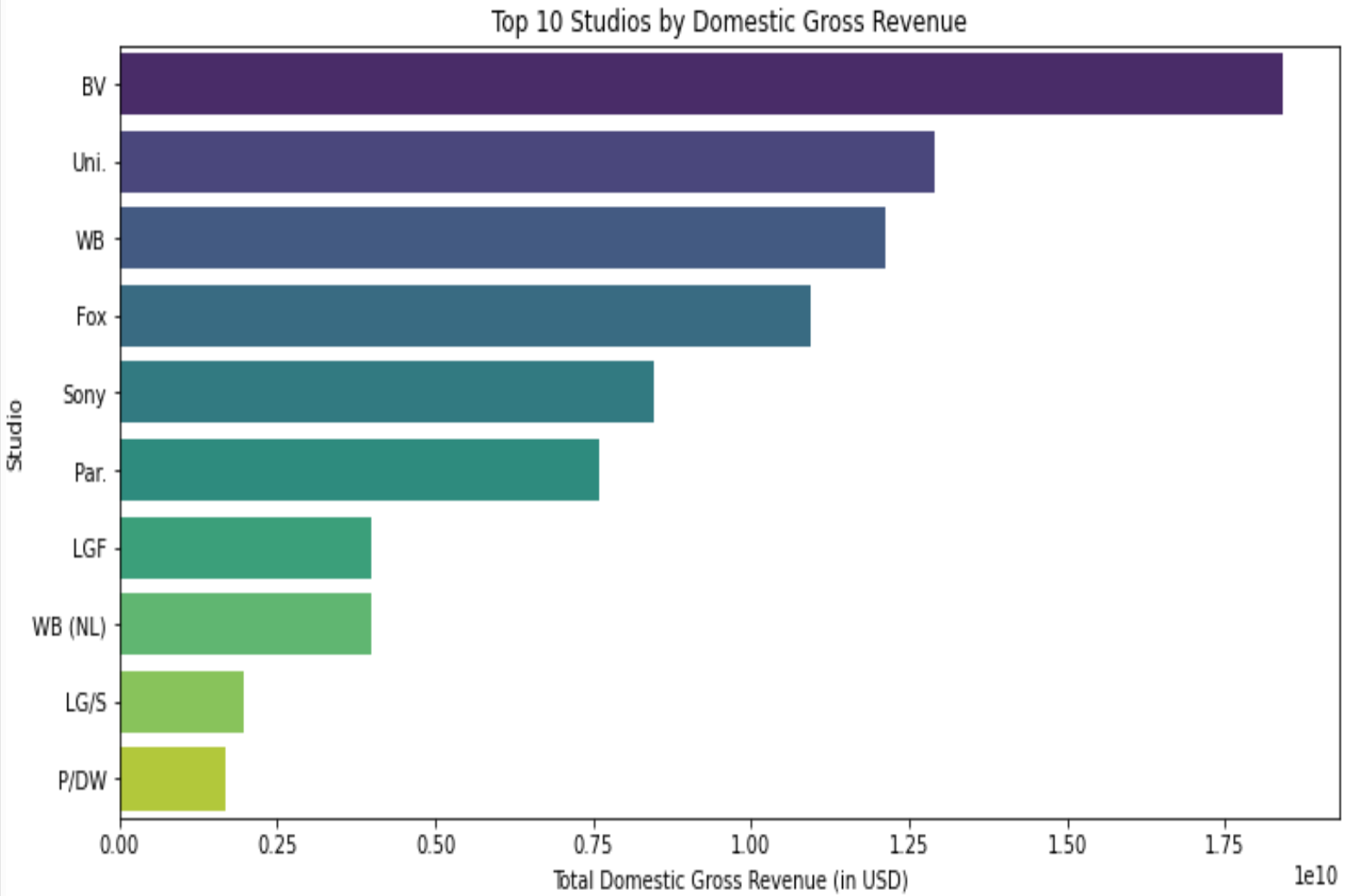
	domestic_gross	foreign_gross	year
count	2.007000e+03	2.007000e+03	2007.000000
mean	4.701984e+07	7.579038e+07	2013.506228
std	8.162689e+07	1.381796e+08	2.597997
min	4.000000e+02	6.000000e+02	2010.000000
25%	6.700000e+05	3.900000e+06	2011.000000
50%	1.670000e+07	1.940000e+07	2013.000000
75%	5.605000e+07	7.595000e+07	2016.000000
max	9.367000e+08	9.605000e+08	2018.000000

BOM MOVIE COMPANY STATISTICAL SUMMARY

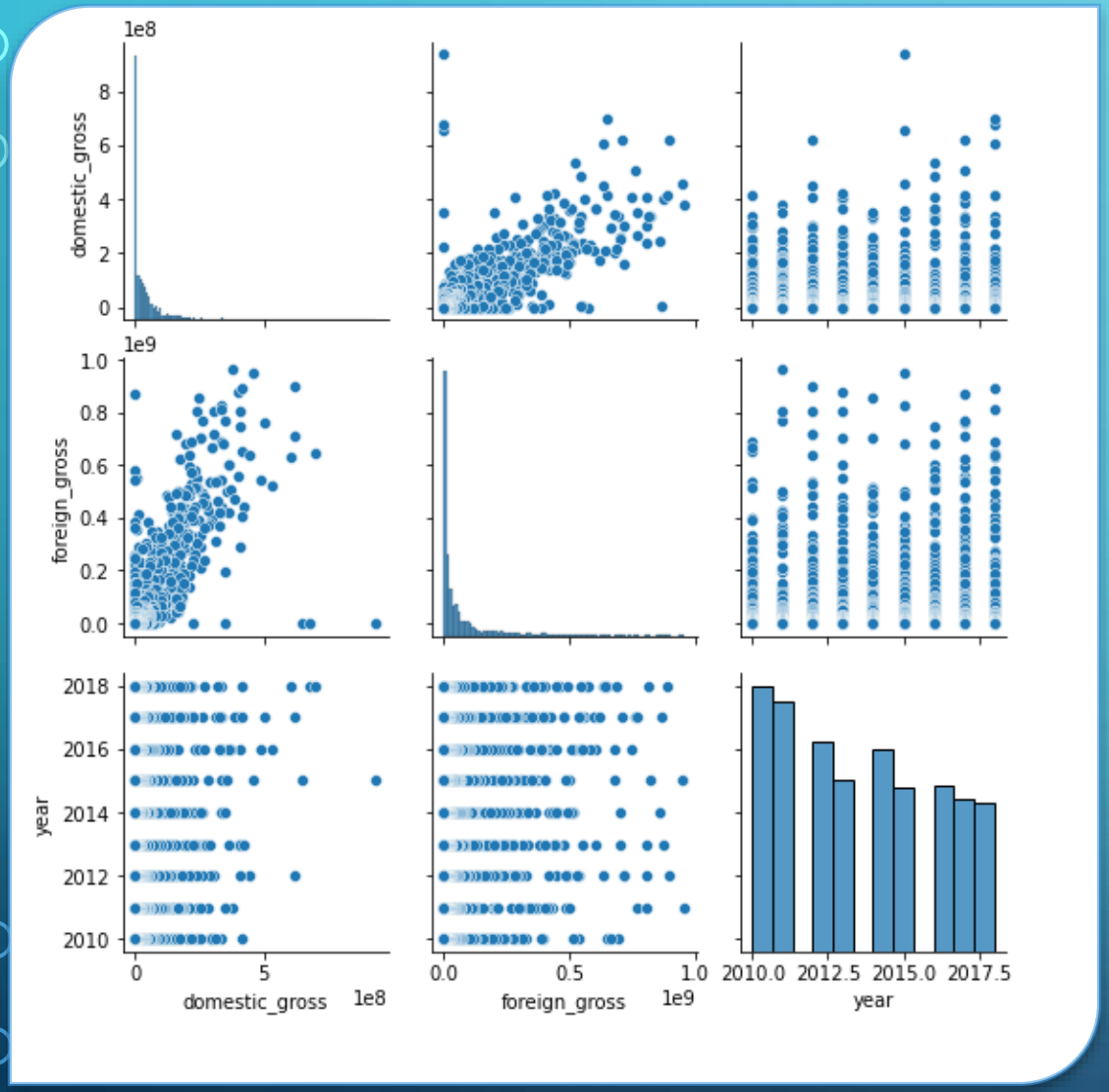
- The mean average domestic gross is $47.01 million$ and 75.79 million for foreign gross.
- 25% of movies were released before 2011, 75% released before 2016.

BOM MOVIE COMPANY VISUALIZATION 1 – BAR GRAPH.

- Total Domestic Gross Revenue vs Studio.
- We first used a bar chart that shows the top 10 movie studios ranked by their total domestic gross revenue (how much money the movies made in the United States).
- BV Studio topped in domestic revenue



BOM MOVIE COMPANY VISUALIZATION 2 - PAIRPLOT



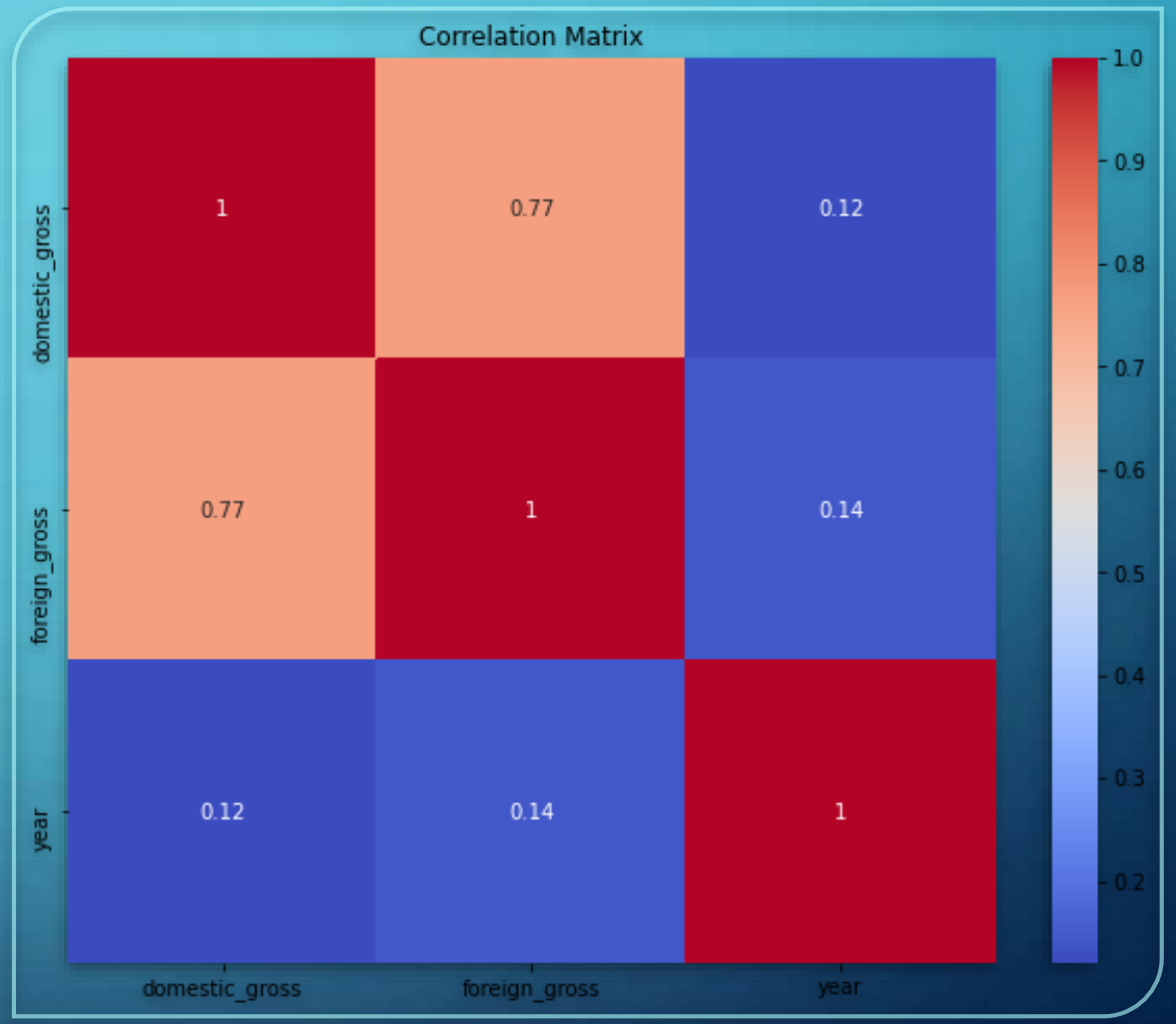
Checking on correlation of the various variables when plotted against each other and presence of outliers.

Scatter plots best visualize foreign and domestic gross revenues.

BOM MOVIE COMPANY VISUALIZATION 3 - HEATMAP

Relationships between different numerical variables in the dataset.

The foreign gross has greater variability compared to domestic gross (higher standard deviation i.e. $ 138.18 million.



BOM MOVIE COMPANY ANALYSIS

CONCLUSION

1. Foreign gross earnings are generally higher than domestic gross, as indicated by the higher mean and median values
2. The foreign gross has greater variability compared to domestic gross (higher standard deviation

IMDB MOVIE COMPANY ANALYSIS.

(MAX & JAEL)

Objectives:

1. Identify movie trends that Drive business success in the film industry.
2. Evaluate Movie performance to determine budget allocation and risk assessment.
3. Provide data insights that help in Movie creation strategies.
4. Help in Target identification.
5. Help in market research and competitor analysis.

IMDB MOVIE COMPANY ANALYSIS - LOADING THE DATASET

```
conn = sqlite3.connect('im.db')
cur = conn.cursor()
cur.execute("SELECT name FROM sqlite_master WHERE type='table';")
tables = cur.fetchall()
print(tables)
```

```
q1 = """SELECT *
        FROM movie_ratings
        """
pd.read_sql(q1,conn)
```

```
q2 = """SELECT *
        FROM movie_basics
        """
pd.read_sql(q2,conn)
```

```
q4 = """SELECT *
        FROM movie_akas
        """
pd.read_sql(q4,conn)
```

```
q5 = """SELECT *
        FROM principals
        """
pd.read_sql(q5,conn)
```

```
q6 = """SELECT *
        FROM known_for
        """
pd.read_sql(q6,conn)
```

```
q7 = """SELECT *
        FROM writers
        """
pd.read_sql(q7,conn)
```

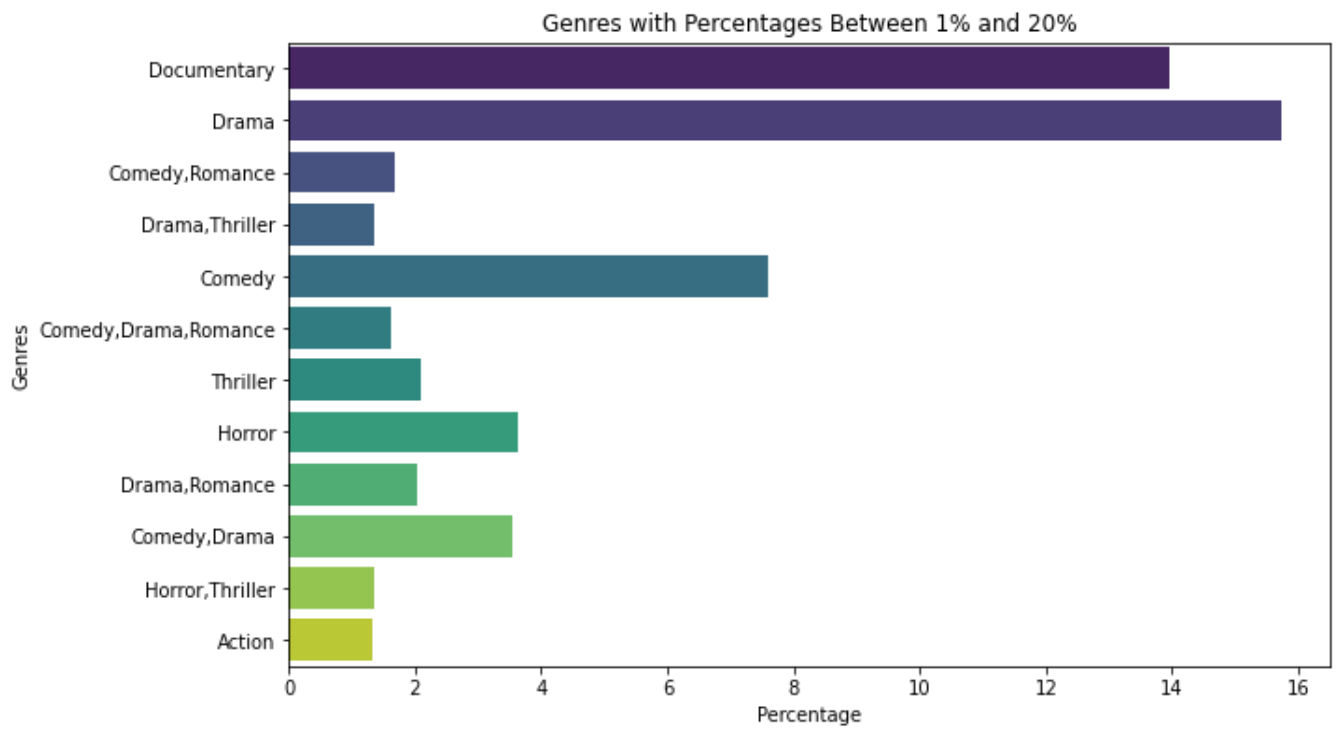
```
q7 = """SELECT *
        FROM persons
        """
pd.read_sql(q7,conn)
```

```
q = """SELECT *
        FROM directors
        """
pd.read_sql(q,conn)
```

```
q3 = """SELECT *
        FROM movie_ratings
        JOIN movie_basics
        ON movie_ratings.movie_id = movie_basics.movie_id
        """
im_ratings = pd.read_sql(q3,conn)
im_ratings.head()
```

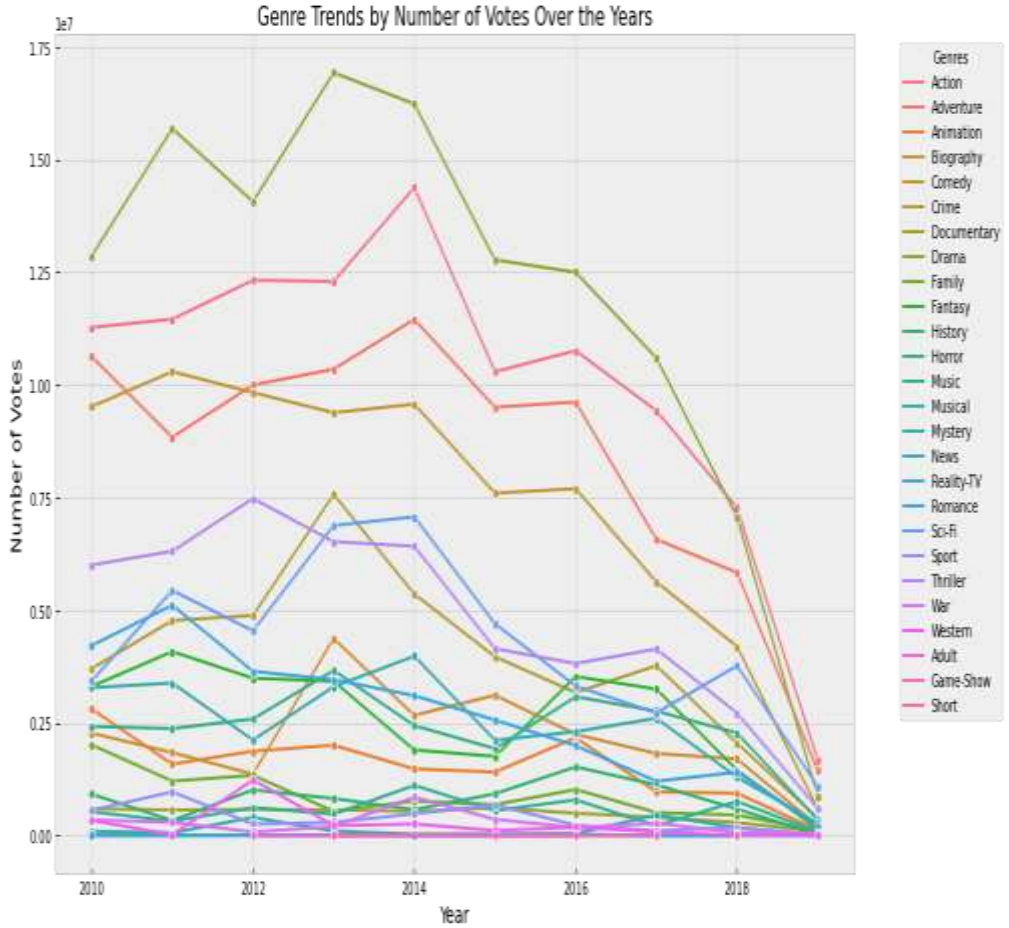
- Movie_ratings: 73856 rows × 3 columns
- Movie_basics: 146144 rows × 6 columns
- Im_ratings: (73856, 9)

IMDB MOVIE COMPANY VISUALIZATION 1 – BAR GRAPH.



- Genres with percentages between 1 and 20.
- Drama (Genre) leads at close to 16%

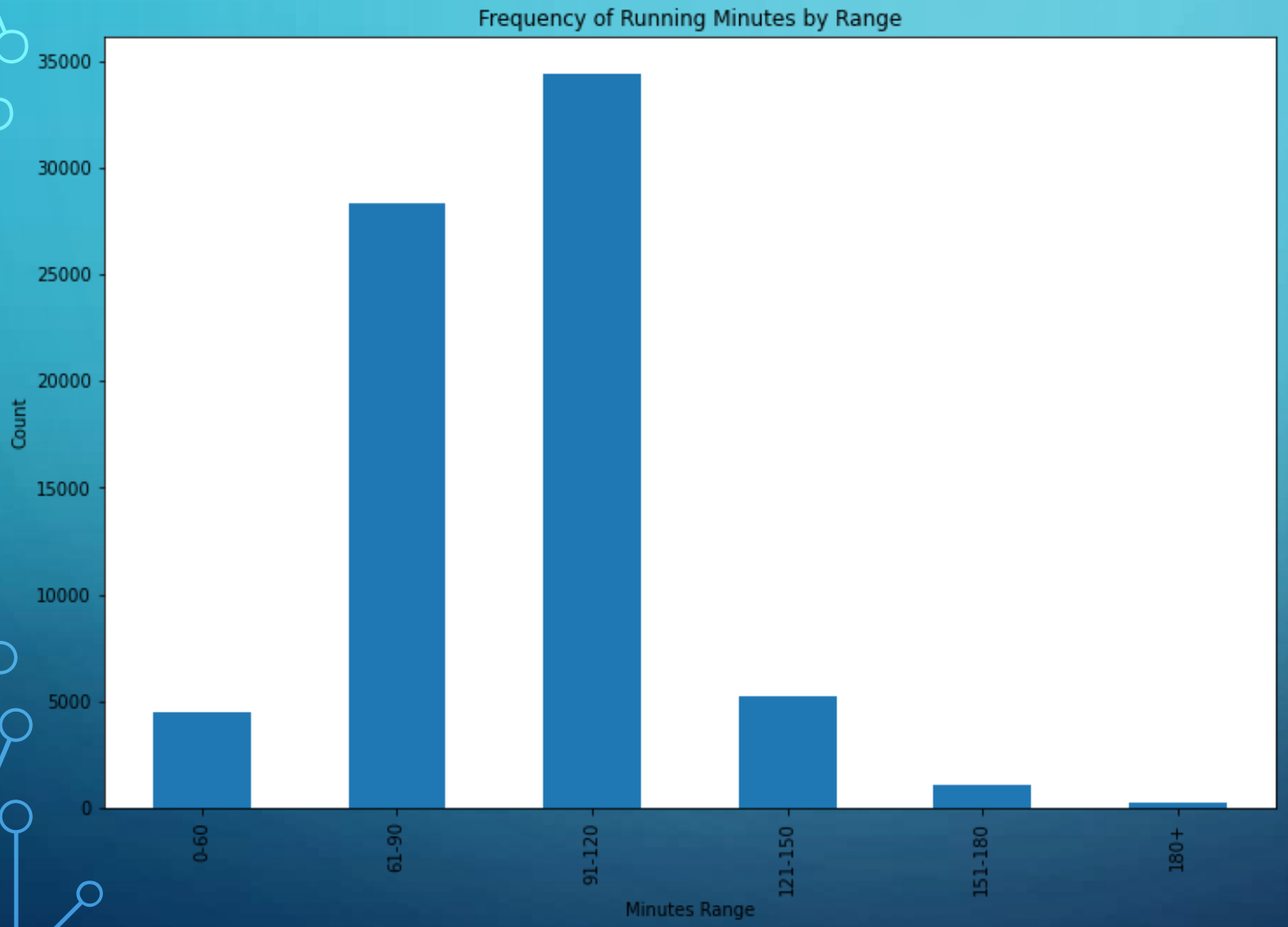
IMDB MOVIE COMPANY VISUALIZATION 2 – LINE GRAPH



Genre trend analysis from 2010 to 2018.

Drama (Genre) receiving higher votes in the period under review.

IMDB MOVIE COMPANY VISUALIZATION 3 – BARGRAPH.



Frequency of running minutes by range.

Most common range: 91-120 with 34429 occurrences

IMDB MOVIE COMPANY

CONCLUSION

From the analysis the DRAMA genre is the safest genre to produce based on the industry frequency of producing that genre. However, ACTIN-CRIME-DRAMA, despite having a small audience had the best quality score compared to the top 30 genres. This genre most-likely attracts enthusiasts and mite skew the ratings in favor of the production company. The industry go-to runtime minutes ranged between 91-120 minutes with longer movies having fewer audiences. Therefore, for a Blockbuster movie with a big budget, the action-crime-drama genre is most likely going to have better reviews and a bigger impact on the audience. For medium to low budget, however, the drama genre seems most appropriate due to its safe numbers.

TMDB DATA SOURCE MOVIE ANALYSIS

(EVELYN)

Main Objectives: Exploring Factors Influencing Movie Success.

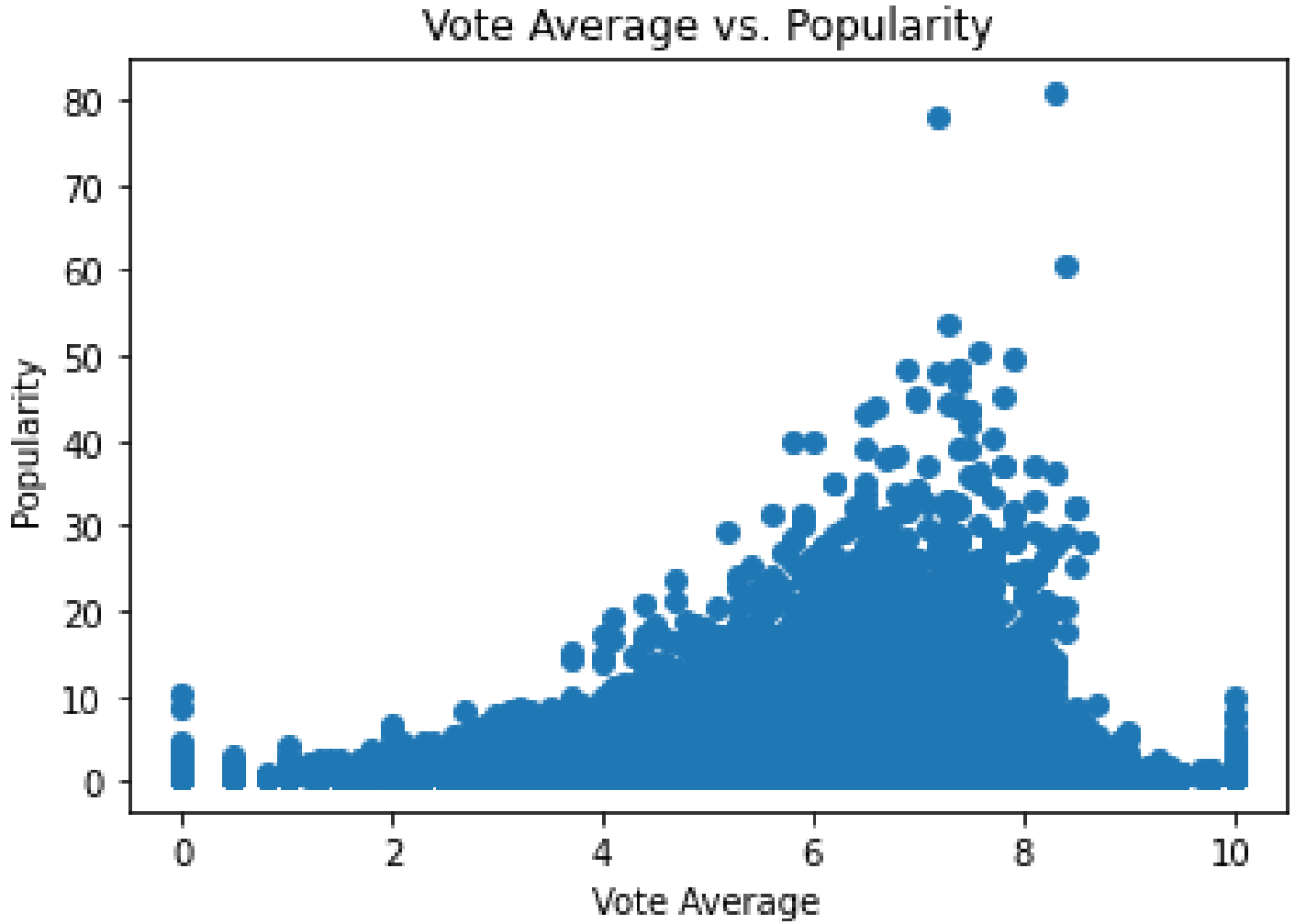
- To explore the relationship between movie popularity and other key metrics
- To investigate the factors that contribute to movie popularity
- To identify trends in movie popularity over time and analyze the impact of historical events and technological advancements on the film industry.

TMDB DATA SOURCE MOVIE SUMMARY STATISTICS.

	popularity	vote_average	vote_count
count	26517.00000	26517.00000	26517.000000
mean	3.130912	5.991281	194.224837
std	4.355229	1.852946	960.961095
min	0.600000	0.000000	1.000000
25%	0.600000	5.000000	2.000000
50%	1.374000	6.000000	5.000000
75%	3.694000	7.000000	28.000000
max	80.773000	10.000000	22186.000000

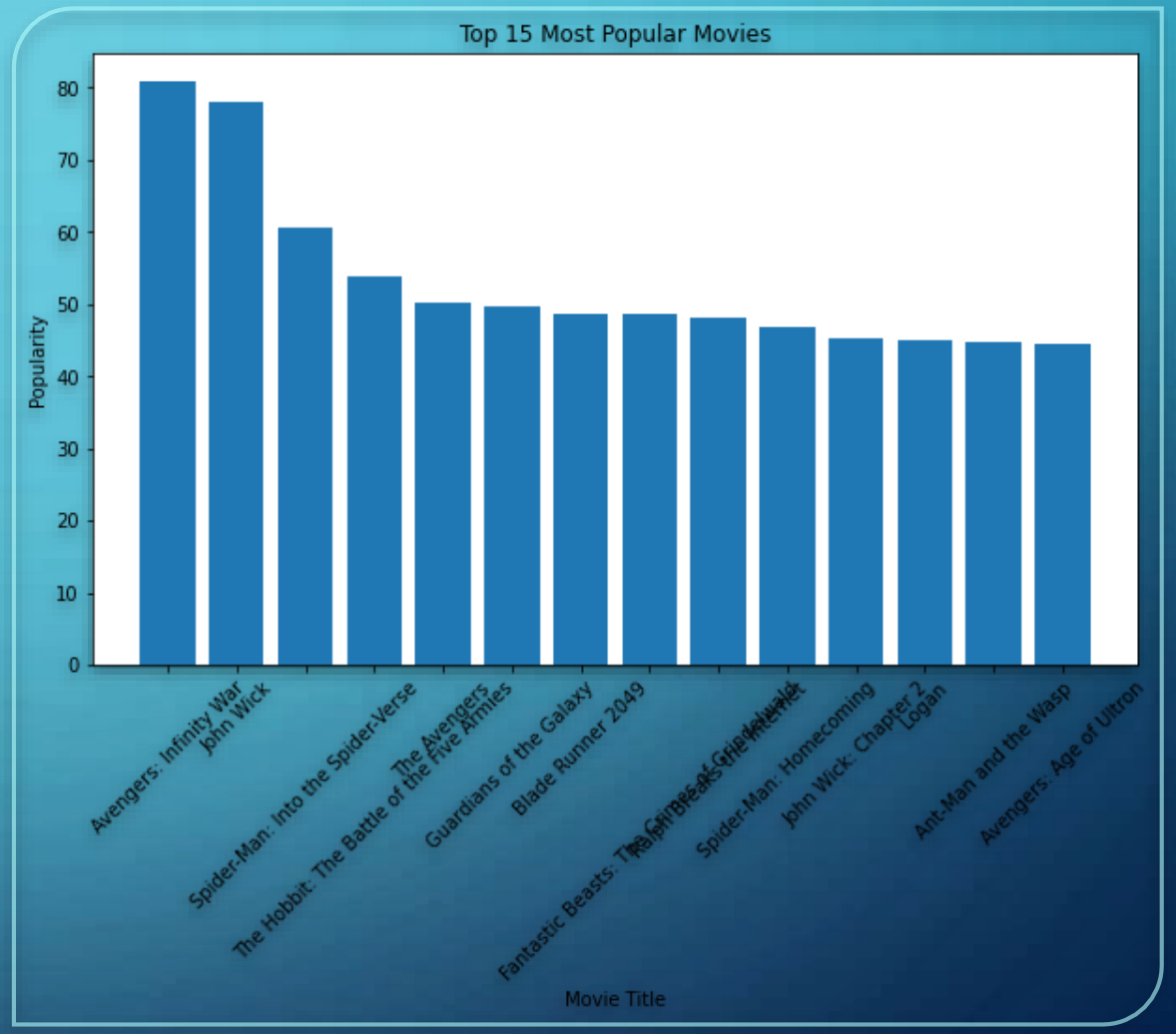
TMDB DATA SOURCE MOVIE VISUALIZATION 1 – SCATTERPLOT.

- Relationship between Vote Average and Popularity
- Movies with higher popularity tend to have better ratings.



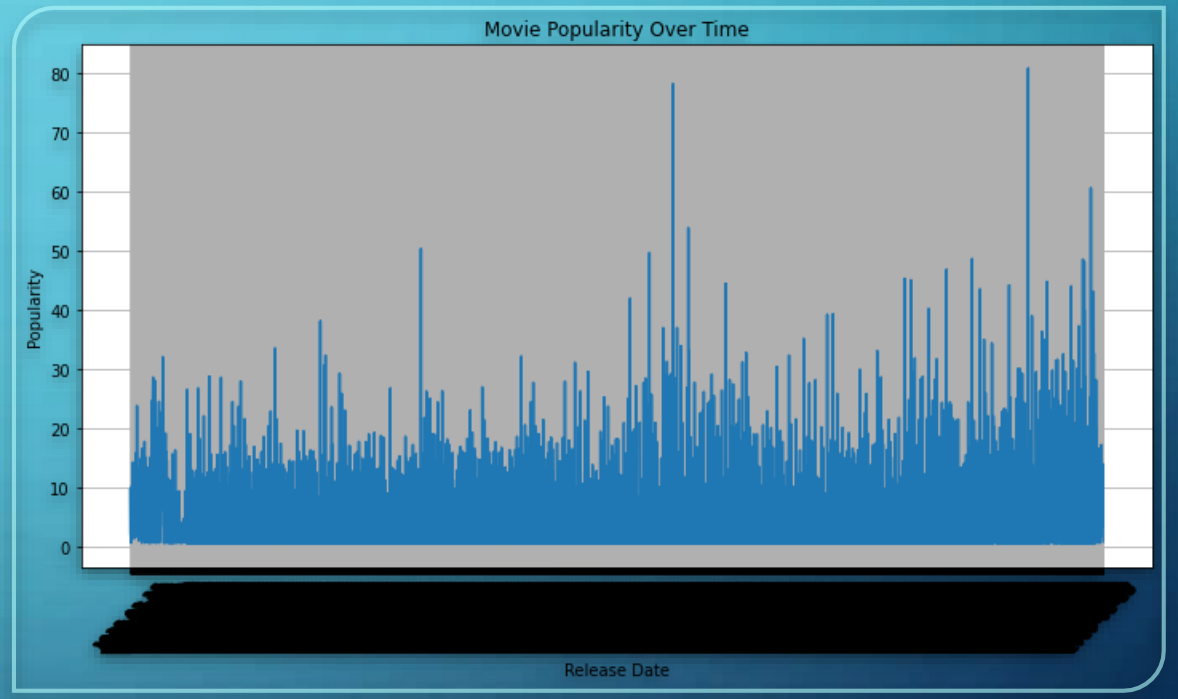
TMDB DATA SOURCE MOVIE VISUALIZATION 2 – BAR GRAPH.

Graph showing the top 15 most popular movies.

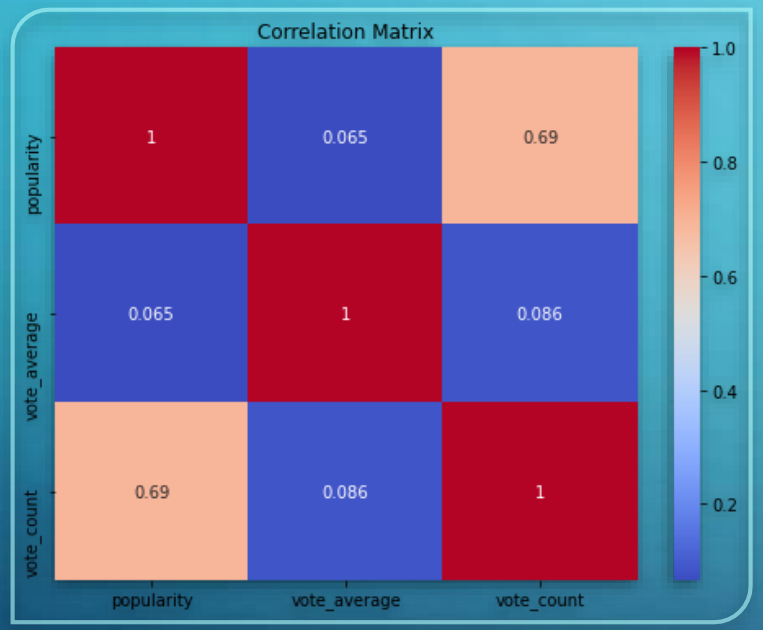


TMDB DATA SOURCE MOVIE VISUALIZATION 3 – TREND ANALYSIS.

High Volatility: The graph exhibits a high degree of volatility. Popularity levels fluctuate significantly over time with numerous peaks and troughs.



TMDB DATA SOURCE MOVIE VISUALIZATION 4 - HEATMAP



Indicating a strong positive relationship between the two variables. If popularity and vote average have a strong positive correlation, it suggests that movies with higher popularity tend to have higher average ratings.

Heatmap showing correlation matrix

TMDB DATA SOURCE MOVIE

CONCLUSION

- Language Influence: The analysis of vote average by original language might reveal potential biases or trends in how movies in different languages are rated. Further investigation is needed to draw definitive conclusions.
- The analysis provides a preliminary understanding of the relationship between popularity, vote average, release dates, language and other factors in the movie dataset.

TN MOVIE BUDGETS DATASET

(HUMPHREY AND CHARITY)

Null Hypotheis(H0)

The average total gross income does not increase with increase in production budget

Alternate Hypotheis(Ha)

The average total gross income increases with increase in production budget

Choosing the significance level(α) - Significance level=5%

The probability threshold for rejecting my null hypothesis is 5%

TN MOVIE BUDGETS DATASET STATISTICAL SUMMARY

- The value of R-squared being equal to 0.547 means that 54.7% of the data describes the linear model making it suitable for analysis
- The p-value is essentially zero(0.00), confirming that the relationship between total gross income and production budget is statistically significant.
- From the production budget coefficient of 4.2460, For every 1 dollar increase in the production budget, the total gross income increases by approximately 4.25 dollars, on average.

OLS Regression Results

Dep. Variable:	total_gross_income	R-squared:	0.547
Model:	OLS	Adj. R-squared:	0.547
Method:	Least Squares	F-statistic:	6993.
Date:	Fri, 24 Jan 2025	Prob (F-statistic):	0.00
Time:	09:01:18	Log-Likelihood:	-1.1748e+05
No. Observations:	5782	AIC:	2.350e+05
Df Residuals:	5780	BIC:	2.350e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.618e+05	2.66e+06	-0.286	0.775	-5.98e+06	4.45e+06
production_budget	4.2460	0.051	83.622	0.000	4.146	4.346

Omnibus:	3978.605	Durbin-Watson:	0.895
Prob(Omnibus):	0.000	Jarque-Bera (JB):	125596.505
Skew:	2.865	Prob(JB):	0.00
Kurtosis:	25.102	Cond. No.	6.57e+07

TN MOVIE BUDGETS DATASET

CONCLUSION

Spending more money on producing a movie leads to higher quality movie which further translates to more income.

END

- ✓Recommendations
- ✓Next Steps
- ✓Thank You

RECOMMENDATIONS

The Market share analysis per genre and popularity trend before weighted analysis show that there is a ready audience, through the number of votes, rating for Drama, Action and Documentaries, Mystery, Sci-Fi, Crime, Thriller, War, Action and Comedy in almost equal measure.

This therefore means;

a) Prioritize the action plan as per the market share per genre because it guarantees a ready target audience.

b) Scape for more data especially production budgets and revenue per genre to determine which genre to invest in.

c) The startup management should largely invest in marketing to increase visibility hence take advantage of popularity votes.

d) The management should also engage in portfolio diversification as a risk management remedy.

NEXT STEPS

Go ahead in venturing into this lucrative field with a focus on the below;

- High budget movies: Produce more of action, crime or drama genres.
- Medium to low budget movies: Drama genre will be the best bet due to the ratings received.
- In as much as movies in the English language are popular, for the foreign market try to offer translations into other diverse languages.
- Upon production of the movie focus should be directed towards intense marketing as it will raise the popularity scores.

THANK YOU