Gladwellchebelyon /
GROUP7_BOX_OFFICE_MOVIES_ANALYSIS

Code Issues Pull requests Actions Projects Wiki Security Insights

GPL-2.0 license

3 stars 0 forks 1 watching 6 Branches 0 Tags Activity

Public repository

m… 6 Branches 0 Tags Go to file t Go to file + Add file Code ···

iamisaackn Rename Box_Office_Analysis.pdf to notebook.pdf a8be758 · 3 minutes ago

Data	Submission	2 days ago
Images	Updated folder	yesterday
Individual_Notebooks	Submission	2 days ago
.gitignore	Initial commit	last week
Group_7.ipynb	Submission	2 days ago
LICENSE	Initial commit	last week
README.md	Updated Links	yesterday
bom-analysis.jpg	Second draft	4 days ago
notebook.pdf	Rename Box_Office_Analysis.pdf to noteboo…	3 minutes ago

README GPL-2.0 license

Box Office Movies Analysis; EDA and Linear Regression Project

Table of Contents

Project Overview

As the entertainment industry surges and major corporations dive into original video content, a new company is poised to enter the competitive world of movie-making. Recognizing the complexities of the film business, especially for newcomers, the company seeks to establish a strong foundation by understanding current box office trends and transforming these insights into a strategic roadmap for their new studio.

To ensure a successful start, the company has enlisted our help as Group 7 members to identify which types of films are currently performing well at the box office and translate these findings into actionable recommendations.

Business Understanding

The Business objective is to identify which film genres will consistently bring in the most revenue for this new studio. There are many elements that contribute to a film's success and our goal is to analyze these factors through Exploratory Data Analysis (EDA) and linear regression. This will allow us to uncover trends and connections that will guide the studio's production choices.

Ultimately, we want to translate these insights into actionable recommendations that will help the studio create films that captivate audiences and turn a profit.

Objectives

Main Objectives

1.To determine which types of films are performing best at the box office.
2.To identify key factors that contribute to a film's success.

Specific Objectives

1.Investigate key variables such as production budgets, domestic and worldwide gross revenues, release years, and genres.
2.Develop predictive models to determine factors significantly impacting box office performance.
3.To investigate the relationship between production budget and box office revenue.
4.To examine the impact of release timing on a film's success.
5.To provide data-driven recommendations for film production and release strategies.

Data Understanding

This analysis, uses datasets from:

IMD Data Base
Box Office
Rotten Tomatoes
The Movie
The Numbers

The data contains information about various films, including their genres, budgets, box office revenues, movie ratings, and release dates.
Understanding the structure and contents of our data will be the first step in uncovering the insights needed to guide our new movie studio's strategy.

Exploratory Data Analysis (EDA)

The EDA section involves data cleaning, exploration, and visualization to uncover patterns and trends in the dataset. Key steps include handling missing values, encoding categorical variables, and visualizing distributions and relationships.

Univariate Analysis

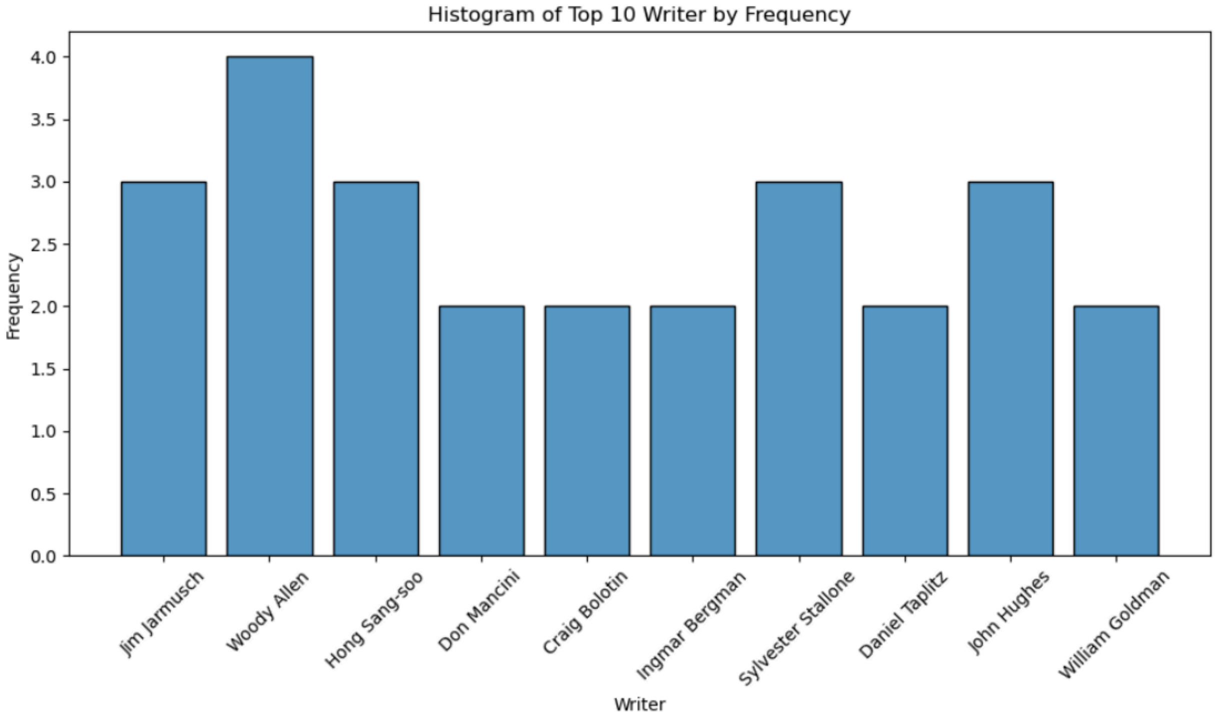
In the Univariate Analysis section, we focus on examining the statistical properties of individual variables in our dataset. By analyzing one variable at a time, we can identify patterns, detect outliers, and gain a clear understanding of each variable's behavior, which is essential for accurate data interpretation and further analysis.

Histogram of Top 10 Writers by Frequency of Movies Written

- The x-axis lists the writers, and the y-axis represents the frequency of movies, ranging from 0 to 10.

- **Woody Allen** has writen the most movies, with a frequency close to 8.
- Most other wrs, writers as **Don Mancin**, **Craig Bolotin**, and **Alfred Hitchcock**, have directed between 4 and 6 movies.
- The chart highlights the prominence of writers in terms of the number of movies they have directed.

```
# Get the top 10 directors, excluding 'Other'
top_10_writers = rtmdf['writer'].value_counts().head(11).index
top_10_writers = top_10_writers[top_10_writers != 'Other']

# Filter the DataFrame to include only the top 10 directors
top_10_df = rtmdf[rtmdf['writer'].isin(top_10_writers)]

# Plot the histogram
plt.figure(figsize=(10, 6))
sns.histplot(data=top_10_df, x='writer', shrink=.8)
plt.xticks(rotation=45)
plt.xlabel('Writer')
plt.ylabel('Frequency')
plt.title('Histogram of Top 10 Writer by Frequency')
plt.tight_layout()
plt.show();
```



Bivariate Analysis

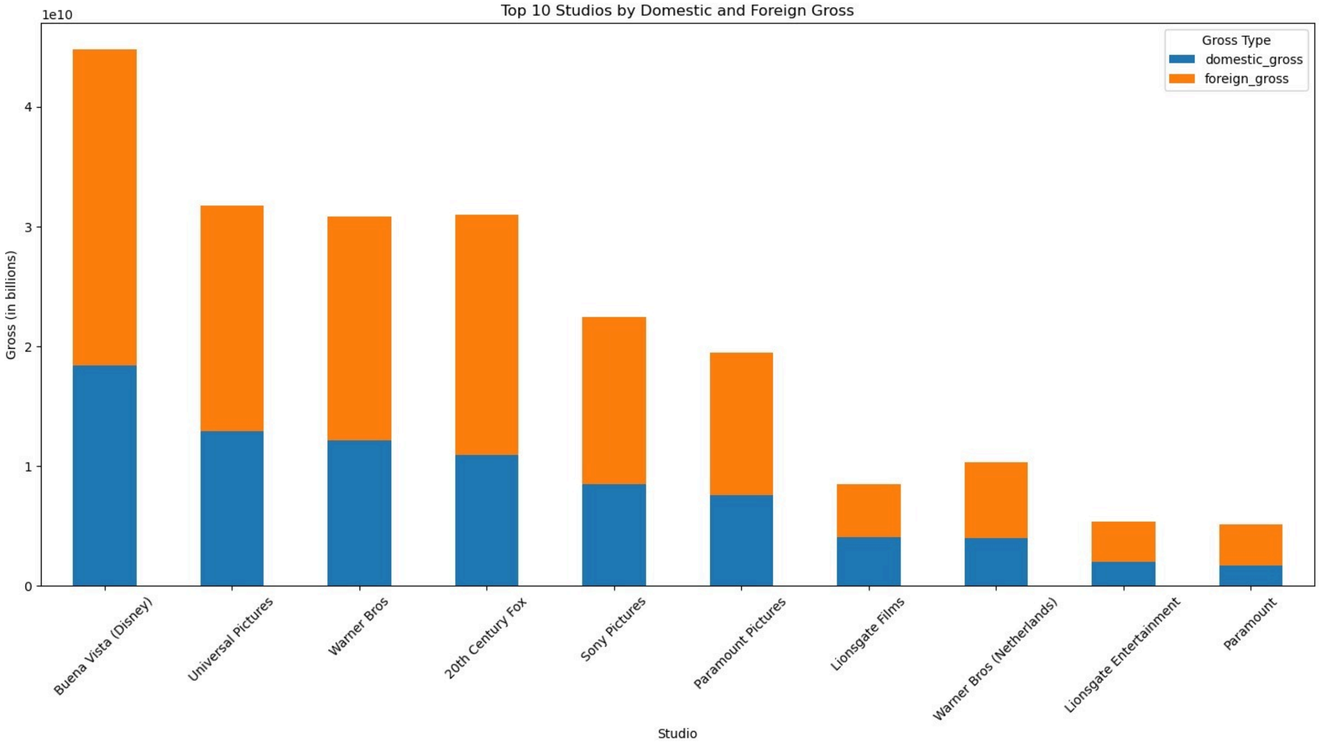
The Bivariate Analysis section investigates the relationships between pairs of variables exploring how the two variables interact with each other. This analysis helps us uncover associations, trends, and dependencies that might exist between variables.

Top 10 Studios by Domestic and Foreign Gross

- Comparing the gross earnings of the top 10 film studios in both domestic (blue bars) and foreign (orange bars) markets.
- The vertical axis represents the gross earnings in billions, ranging from 0 to 10.
- **Buena Vista (Disney)** has the highest combined gross, with significant earnings in both domestic and foreign markets.
- **Universal Pictures** and **Warner Bros** also show substantial earnings, with a notable portion coming from foreign markets.
- The chart highlights the global reach and financial performance of these major studios, indicating their success in both domestic and international markets.

```
# Filter out zeros in domestic and foreign gross
filtered_data = bomdf[(bomdf['domestic_gross'] != 0) & (bomdf['foreign_gross'] != 0)]

# Group by studio and calculate the sum of domestic and foreign gross
studio_gross = filtered_data.groupby('studio')[['domestic_gross', 'foreign_gross']].sum()

# Sort by domestic and foreign gross and select the top 10 studios
top_studios = studio_gross.sort_values(by=['domestic_gross', 'foreign_gross'], ascending=False).head(10)
```

```
# Plotting
fig, ax = plt.subplots(figsize=(18, 8))
top_studios.plot(kind='bar', stacked=True, ax=ax)

# Labeling
plt.title('Top 10 Studios by Domestic and Foreign Gross')
plt.xlabel('Studio')
plt.ylabel('Gross (in billions)')
plt.xticks(rotation=45)
plt.legend(title='Gross Type')
plt.show();
```



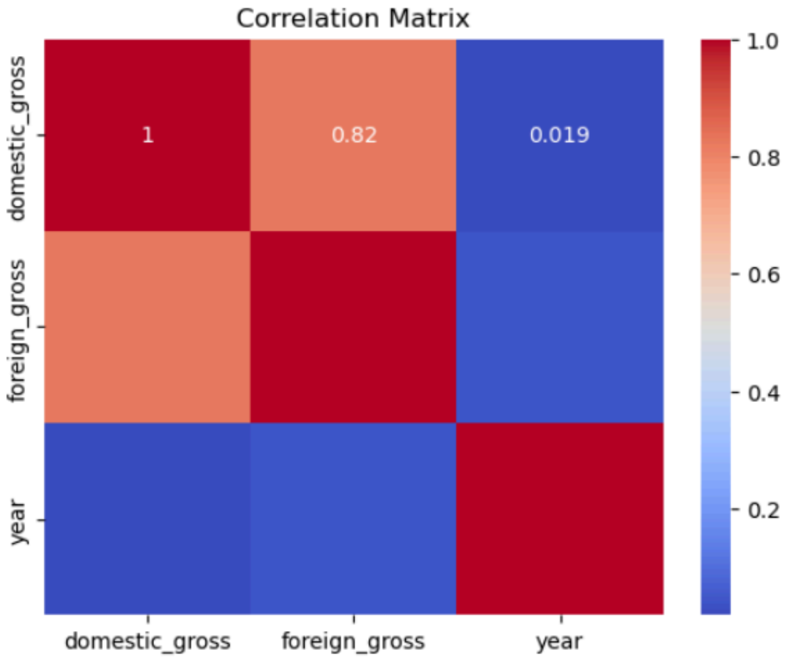
Multivariate Analysis

In the Multivariate Analysis section, we extend our examination to more than two variables simultaneously. This comprehensive approach provides deeper insights into the complex structure of our data, helping us identify patterns, correlations, and underlying factors that are crucial for building robust and accurate predictive models.

Correlation matrix of relationships between `domestic_gross`, `foreign_gross`, and `year`

- **Domestic Gross vs. Foreign Gross**: There is a strong positive correlation (0.82), indicating that movies with higher domestic earnings tend to also have higher foreign earnings.
- **Year vs. Domestic Gross**: There is a very weak positive correlation (0.019), suggesting almost no linear relationship between the year and domestic earnings.
- **Year vs. Foreign Gross**: The correlation is not explicitly shown, but it appears to be similarly weak.

```
#Correlation analysis
corr = bomdf[['domestic_gross', 'foreign_gross', 'year']].corr()
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



Statistical Data Analysis

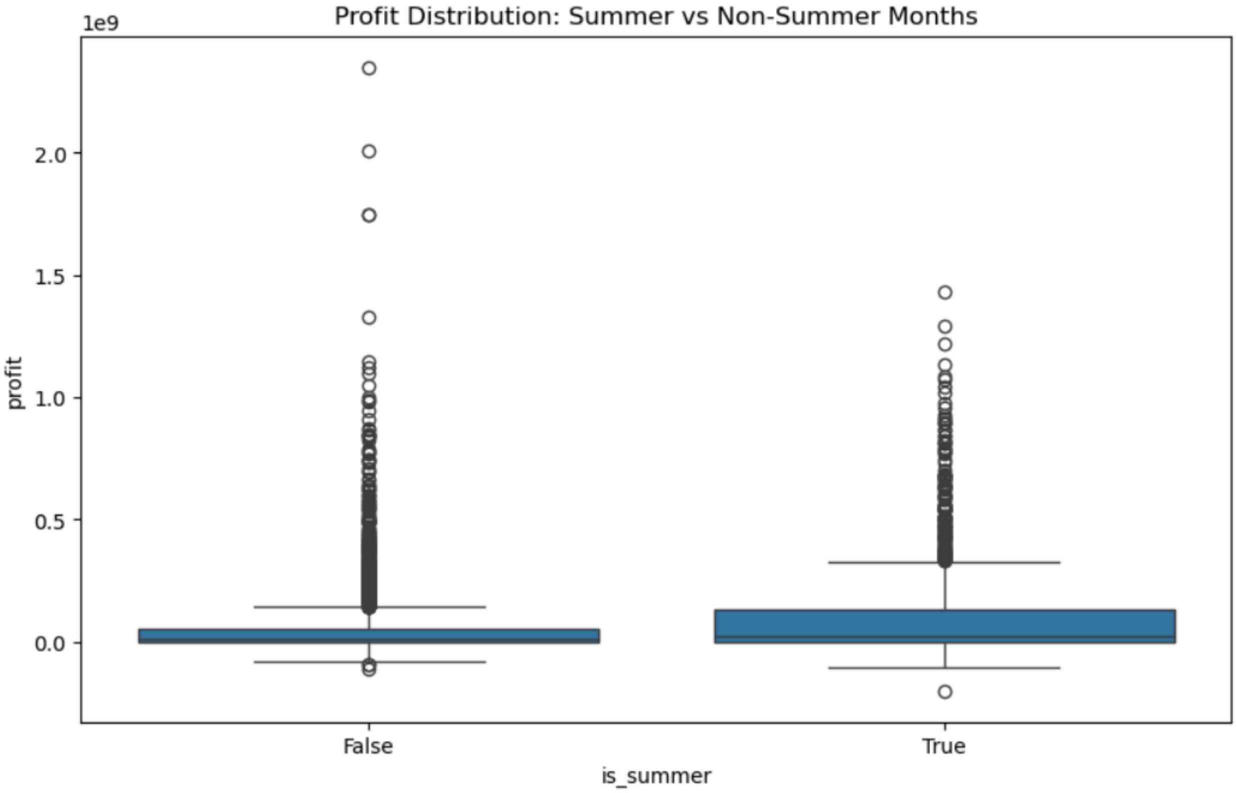
In this section, we apply statistical techniques to derive insights from our dataset. We use descriptive statistics like mean, median, variance, and standard deviation to summarize the data's central tendency and dispersion. Inferential statistics, including hypothesis testing, confidence intervals, and regression analysis, help us make predictions or generalizations about a population based on our sample. This analysis will help validate our findings, identify significant patterns, and supports data-driven decision-making.

Hypothesis Testing

Test of Movies Released in Summer and Non-Summer

Two Sample T-test
Null Hypothesis (H0): There is no significant difference in movie profits between movies released Summer and Non-Summer. Alternative Hypothesis (H1): There is a significant difference in movie profits between summer and non-summer months. ```bash # Splitting the dataframe into a subset for summer months TN_df['is_summer'] = TN_df['release_month'].isin([5, 6, 7])

```
# Creating a subset of non summer months from the remaing months and extracting profit values for both
subsets.

summer_profits = TN_df[TN_df['is_summer']]['profit']
non_summer_profits = TN_df[~TN_df['is_summer']]['profit']

# Performing a t-test to compare the mean profits between summer and non-summer movies.
t_stat, p_value = stats.ttest_ind(summer_profits, non_summer_profits)

print(f"T-statistic: {t_stat}")
print(f"P-value: {p_value}")

# Significance Level
a = 0.05
if p_value < a:
    print("Reject Null Hypothesis: There is no significant difference in movie profits between summer
months (May, June, July) and other months.")
else:
    print("Accept Null Hypothesis: There is no significant difference in movie profits between summer and
non-summer months.")
```
```
```



Based on the analysis the difference in movie profits between summer and non-summer months is statistically significant. From the box plot, we can see that the box for summer movies (True) is slightly higher than for non-summer movies (False). From both the t-test and the box plot visualisation we can infer that Movie profits tend to be high in the months of May, June and July (summer).

Linear Regression Model

A linear regression model is created to predict box office revenue based on factors such as budget, genre, and release timing. The model is evaluated using metrics such as mean squared error and R-squared.

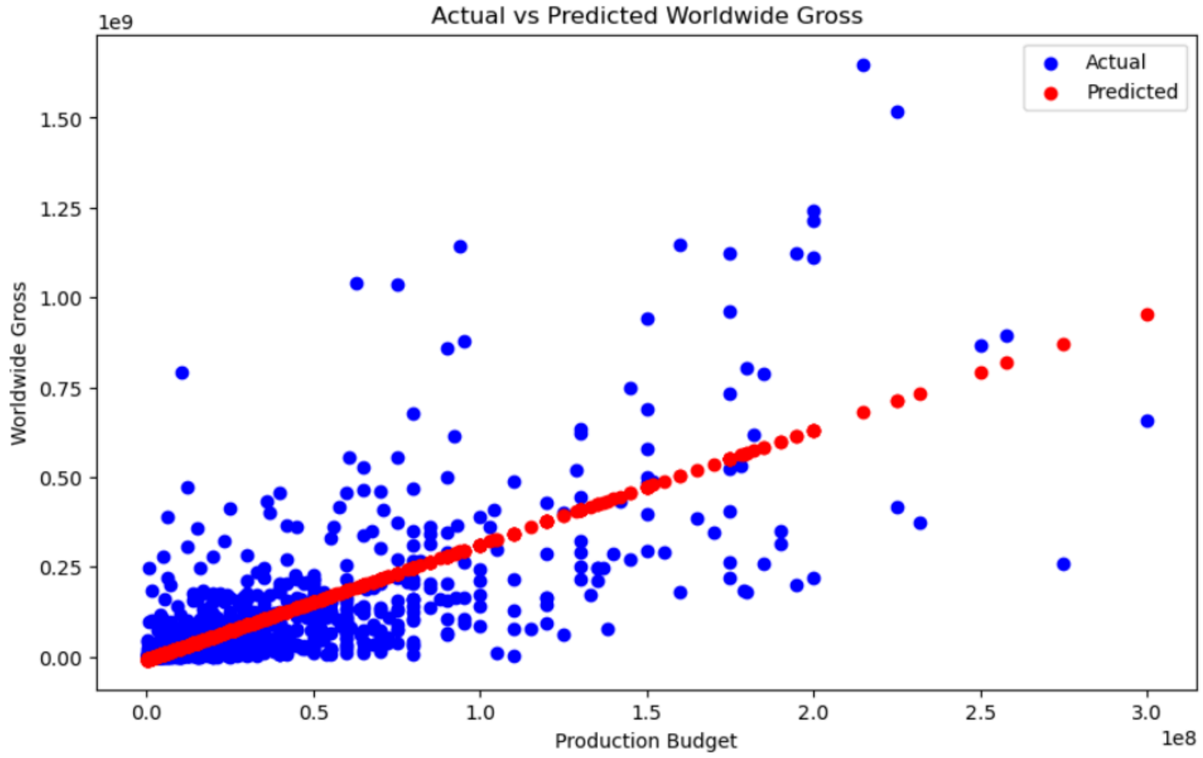
Predict the worldwide gross revenue of movies based on their production budget using a linear regression model

- The scatter plot compares actual and predicted worldwide gross revenues.
- The red regression line indicates the model's predictions.
- The model has a **Mean Squared Error (MSE)** of approximately (1.6 \times 10^{16}) and an **R-squared score** of about 0.50.
- An R-squared score of 0.50 suggests that the model explains about 50% of the variability in worldwide gross revenue based on production budget.
- The visualization helps assess how well the model predicts worldwide gross revenue from production budgets.

```
# Select the feature and target variable
X = TN_df[['production_budget']]
y = TN_df['worldwide_gross']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Scale the features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Train the model
model1 = LinearRegression()
model1.fit(X_train_scaled, y_train)

# Make predictions
y_pred = model1.predict(X_test_scaled)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

```
print("Model 1: Predicting worldwide gross based on production budget")
print(f"Mean squared error: {mse}")
print(f"R-squared score: {r2}")

# Visualize
plt.figure(figsize=(10, 6))
plt.scatter(X_test, y_test, color='blue', label='Actual')
plt.scatter(X_test, y_pred, color='red', label='Predicted')
plt.xlabel('Production Budget')
plt.ylabel('Worldwide Gross')
plt.title('Actual vs Predicted Worldwide Gross')
plt.legend()
plt.show()
```

```
Model 1: Predicting worldwide gross based on production budget
Mean squared error: 1.5991486331029718e+16
R-squared score: 0.5041078156339777
```



Installation

To run this project, follow these steps:

1. Clone the repository:

```
git clone https://github.com/Gladwellchebelyon/GROUP7_BOX_OFFICE_MOVIES_ANALYSIS.git
```

2. Navigate to the project directory:

```
cd GROUP7_BOX_OFFICE_MOVIES_ANALYSIS
```

3. Install the required dependencies:

```
pip install  requirements.txt
```

Usage

Open the Jupyter notebook and follow the instructions to reproduce the analysis and results:

Results and Insights

Findings

1. **Genre Ratings:** Documentary and Drama genres have the highest median ratings, indicating they are generally well-received.
   Horror and Action genres have lower median ratings, suggesting they might be less favorably reviewed on average.
2. **Production Budget vs. Profit:** As the production budget increases, the profit tends to increase as well.
   High-budget productions are much less common.
3. **Foreign vs. Domestic Gross:** Most studios have a higher foreign gross, indicating a larger international market presence.
   Buena Vista (Disney) leads in both domestic and foreign gross.
   Warner Bros and Universal show strong performances, with Warner Bros having a higher foreign gross compared to domestic.
   STX Entertainment, Focus Features, and Weinstein Company have significantly lower gross revenues, with both domestic and foreign gross below 1 billion.
4. **Director Popularity:** Steven Spielberg is the most popular director in the dataset.
   Barry Levinson and Ivan Reitman have the lowest frequencies among the top 10 directors.
5. **Writer Frequency:** Fyodor Dostoevsky and Jane Austen have the highest frequencies, indicating they are the most frequently mentioned or analyzed writers in the dataset.
   William Golding and Philip Roth have the lowest frequencies among the top 10 writers.
6. **Seasonal Profit Trends:** May, June, and July have the highest average profits.
   January and February have the lowest average profits.
   December shows relatively high average profits, likely due to the holiday season when people have more leisure time and are more likely to go to the movies.

Business Recommendations

1. **Focus on High-Earning Genres**: Prioritize producing films in genres that consistently show higher average ratings and box office returns, such as Drama, Comedy, and Documentary.
2. **Optimize Production Budgets**: Carefully balance production budgets to maximize profitability. Aim for a budget range that optimizes profitability without excessive spending, as higher budgets can lead to diminishing returns.
3. **Strategic Release Timing**: Schedule film releases during peak movie-going periods, such as summer and holiday seasons, to capitalize on higher average profits during these times.
4. **Invest in Proven Directors and Writers**: Collaborate with top directors and writers who have a track record of success. Their involvement can significantly impact a film's success.
5. **Leverage Popular Franchises**: Consider developing or acquiring established film franchises, which often have built-in audiences, reducing marketing costs and increasing box office returns.
6. **Maximize Foreign Markets**: Ensure strong international distribution and marketing strategies, as significant revenue is generated from foreign markets.
7. **Quality Over Quantity**: Focus on producing a smaller number of high-quality films rather than a large number of lower-quality releases, as quality films can outperform in profitability and audience reception.
8. **Effective Use of Marketing Budgets**: Allocate sufficient budget for marketing to ensure high visibility and audience awareness. Successful films often have robust marketing campaigns that drive initial box office performance.
9. **Moderate Budget Allocations**: Recognize that even moderate budget allocations can yield significant profits. Ensure that budget allocations are strategically planned to optimize profit margins.
10. **Data-Driven Decision Making**: Continuously gather and analyze data on film performance, audience preferences, and market trends to inform strategic decisions and optimize resource allocation.
11. **Utilize Audience Feedback**: Implement mechanisms to gather audience feedback on film concepts and trailers to refine production choices and marketing strategies.

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package

Contributors 3

iamisaackn Isaac Ngugi

MONISH254

Gladwellchebelyon Gladwell Chepkorir

Languages

- ● **Jupyter Notebook** 100.0%