FROM DATA TO SCREENS

Introduction

Overview of the project

The project aims to guide Microsoft's new movie studio venture by conducting a data-driven analysis of the current box office trends. As big companies have found success in creating original video content, Microsoft seeks to capitalize on the popularity of movies but lacks experience in the film industry. To make informed decisions on the types of films to produce, this project will explore the genres, themes, and characteristics of the most successful movies at the box office. By analyzing data on past blockbusters, the project will provide actionable insights to help Microsoft's movie studio identify and prioritize the types of films that are likely to resonate with audiences and achieve box office success. Using data-backed strategies, Microsoft can make well-informed choices that maximize their chances of creating box office hits and establishing a strong presence in the competitive world of filmmaking.

Importance of Data Analysis in the Movie Industry

In today's fast-paced and highly competitive movie industry, making data-driven decisions is paramount to success. Data analysis empowers studios to gain a comprehensive understanding of audience preferences, identify trends, and predict potential box office hits. By analyzing vast amounts of movie-related data, including ratings, votes, genres, and revenue, studios can strategically shape their creative endeavors.

Data-driven insights provide a roadmap for crafting captivating narratives, selecting the right genres, and aligning movie runtimes to audience expectations. This evidence-based approach increases the likelihood of producing movies that resonate with audiences, leading to higher box office revenues and stronger audience engagement.

The presentation presents the key findings from the data analysis, revealing top-performing genres, profitability metrics, runtime considerations, and actionable recommendations for Microsoft's new movie studio. Let's embark on this exciting journey into the world of movie data analysis.

Dataset Overview: Key Variables

Before we delve into the analysis, let's take a moment to understand the main variables present in our movie dataset. These variables provide essential insights into each movie's characteristics and audience reception.

- **Title:** The title represents the name of each movie, serving as its unique identifier.
- **Genres:** The genres variable lists the different genres associated with each movie. Genres help categorize movies based on their themes and content.

- **Runtime (minutes):** This variable indicates the duration of each movie in minutes, providing valuable information about the movie's length.
- **Average Rating:** The average rating variable represents the mean rating given to each movie by audiences or critics, offering a measure of overall reception.
- **NumVotes:** Numvotes refers to the total number of votes received by each movie, reflecting its level of audience engagement.

The dataset encompasses a diverse collection of movies, spanning multiple genres, and provides a comprehensive foundation for our data-driven exploration.

Data Understanding

The data in this project represent information about various movies and their performance at the box office. The sample includes a collection of movies, and each movie entry contains different variables that provide details about the movie and its box office performance, the data was collected from IMDB an online database of information that is related to films, television series, podcasts, home videos, video games, and streaming content online which is really helpful in collecting data that is required to achieve our goal as it contains movie information such as popularity, movie ratings and reviews, production budget and also the themes and characteristics which are some of the available variables.

The target variable in this project is likely to be the Box Office Performance(rating) or Gross Revenue of the movies. The aim is to identify factors that influence box office success, so this variable will be the primary focus of the analysis.

in Microsoft's movie studio, helping them produce films that are likely to be successful at the box office.

Data exploration, preparation and modeling

Shape

From the above function we can see that: df has 73856 rows and 8 columns while df_3 has 3387rows and 5 columns, as shown below:

```
# Now we have two data sets df and df_3, let us explore.
df.shape
```

```
(73856, 8)
```

```
df_3.shape
```

```
(3387, 5)
```

Information

df

We are able to see all the columns and we notice that the columns *runtime_minutes and genres* have null values

df_3

We are able to see all the columns and we notice that the columns *studio, domestic_gross and foreign_gross* have null values

Data description

We are able to view stats such as the mean, count, standard deviation, the minimum and maximumand the various percentiles.

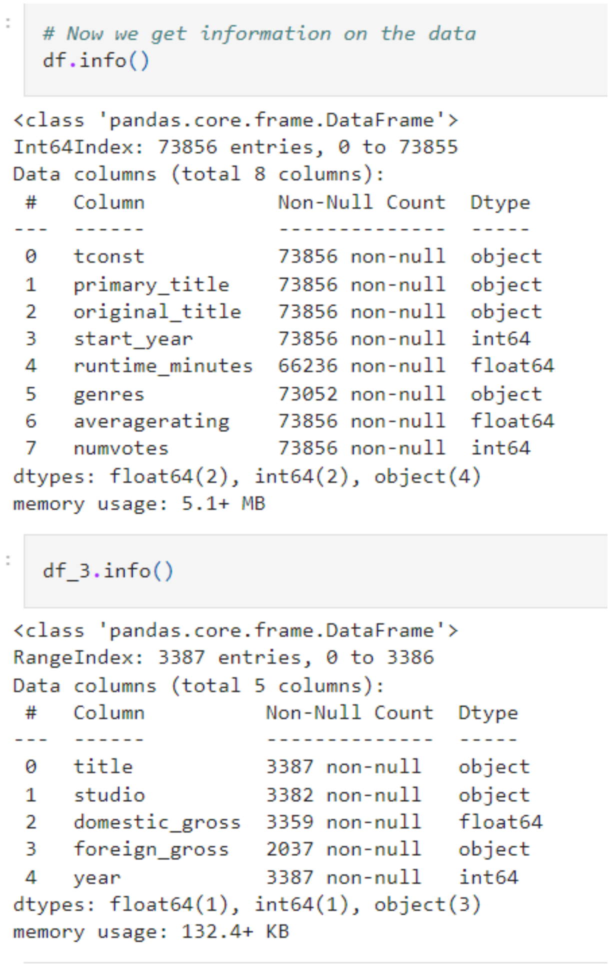
Missing values

We are now able to see the frequency of missing values some varaibles such as **runtime_minutes (10.317374%)** in df and **foreign_gross(39.858282%)** in df_3 have a lot of missing values while other variables such as **genres(1.088605%)** in df , **domestic_gross(0.826690)** and **studio(0.147623%)** in df_3 have few missing values.

Handling missing values

We have to drop all missing values in **df** since *__runtime_minutes__ has a high number of missing values and we cant use any information on the data to predict the **genre**

In **df_3** we will also drop all missing values .



```python
# Now we get information on the data
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 73856 entries, 0 to 73855
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   tconst           73856 non-null  object 
 1   primary_title    73856 non-null  object 
 2   original_title   73856 non-null  object 
 3   start_year       73856 non-null  int64  
 4   runtime_minutes  66236 non-null  float64
 5   genres           73052 non-null  object 
 6   averagerating    73856 non-null  float64
 7   numvotes         73856 non-null  int64  
dtypes: float64(2), int64(2), object(4)
memory usage: 5.1+ MB
```

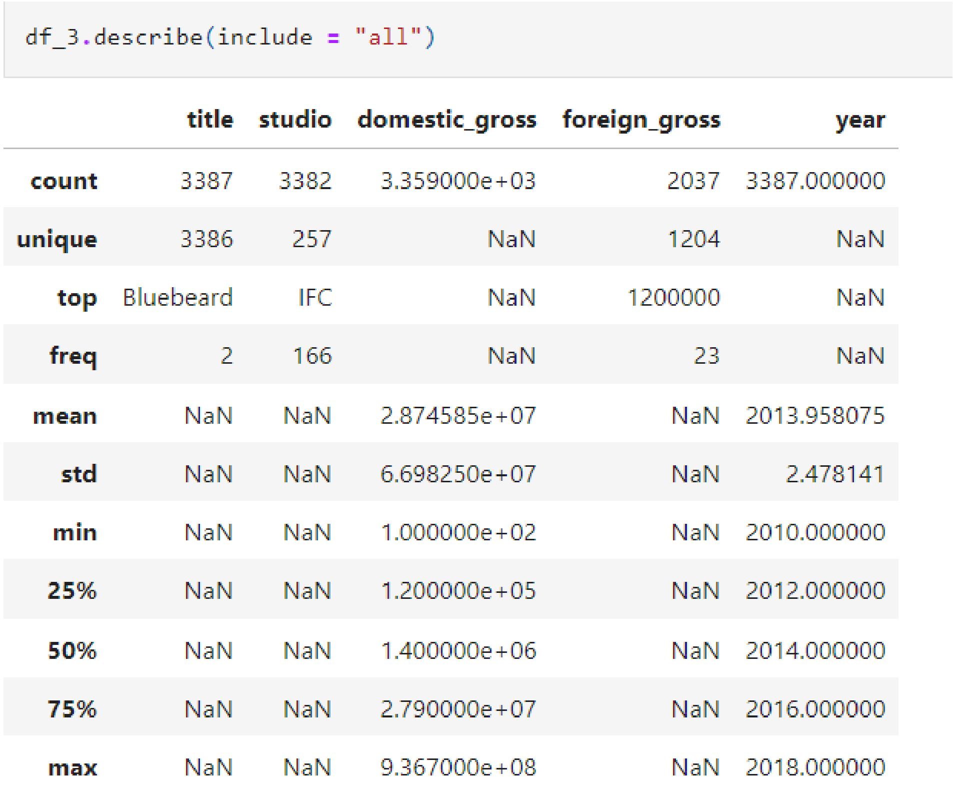
```python
df_3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3387 entries, 0 to 3386
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   title           3387 non-null   object 
 1   studio          3382 non-null   object 
 2   domestic_gross  3359 non-null   float64
 3   foreign_gross   2037 non-null   object 
 4   year            3387 non-null   int64  
dtypes: float64(1), int64(1), object(3)
memory usage: 132.4+ KB
```

```python
# We can now look at the description of the data
df.describe(include = "all")
```

	tconst	primary_title	original_title	start_year	runtime_minutes	genres
count	73856	73856	73856	73856.000000	66236.000000	73052
unique	73856	69993	71097	NaN	NaN	923
top	tt2210657	The Return	Broken	NaN	NaN	Drama
freq	1	11	9	NaN	NaN	11612
mean	NaN	NaN	NaN	2014.276132	94.654040	NaN
std	NaN	NaN	NaN	2.614807	208.574111	NaN
min	NaN	NaN	NaN	2010.000000	3.000000	NaN
25%	NaN	NaN	NaN	2012.000000	81.000000	NaN
50%	NaN	NaN	NaN	2014.000000	91.000000	NaN
75%	NaN	NaN	NaN	2016.000000	104.000000	NaN
max	NaN	NaN	NaN	2019.000000	51420.000000	NaN

```python
df_3.describe(include = "all")
```



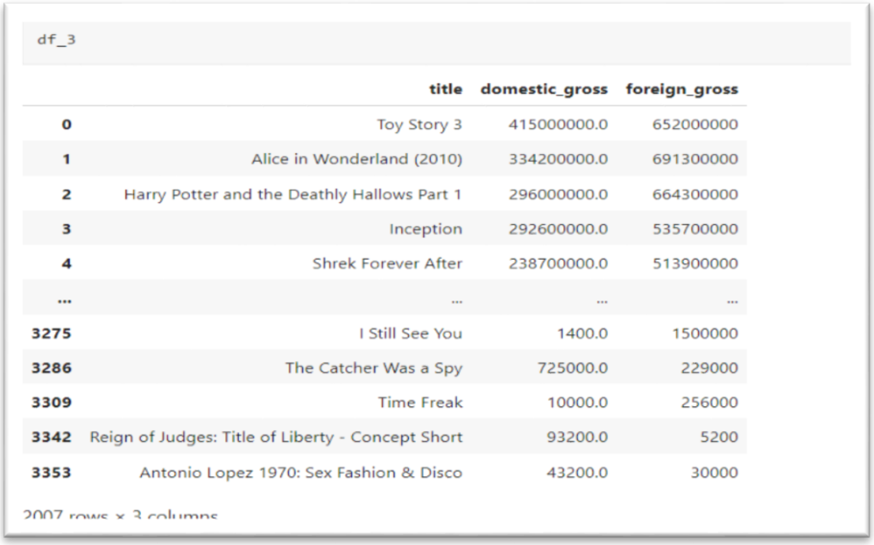
```
df_3.describe(include = "all")
```

	title	studio	domestic_gross	foreign_gross	year
count	3387	3382	3.359000e+03	2037	3387.000000
unique	3386	257	NaN	1204	NaN
top	Bluebeard	IFC	NaN	1200000	NaN
freq	2	166	NaN	23	NaN
mean	NaN	NaN	2.874585e+07	NaN	2013.958075
std	NaN	NaN	6.698250e+07	NaN	2.478141
min	NaN	NaN	1.000000e+02	NaN	2010.000000
25%	NaN	NaN	1.200000e+05	NaN	2012.000000
50%	NaN	NaN	1.400000e+06	NaN	2014.000000
75%	NaN	NaN	2.790000e+07	NaN	2016.000000
max	NaN	NaN	9.367000e+08	NaN	2018.000000

We can now isolate only the columns that we need

In **df** we need the original_title, runtime_minutes, genres, averagerating and numvotes and in **df_3** we need title, domestic_gross, foreign_gross.   We have to remove the other columns since they will not help us achieve our goal and hence it is easier to remove them to reduce the workload.

The images below show the needed columns:



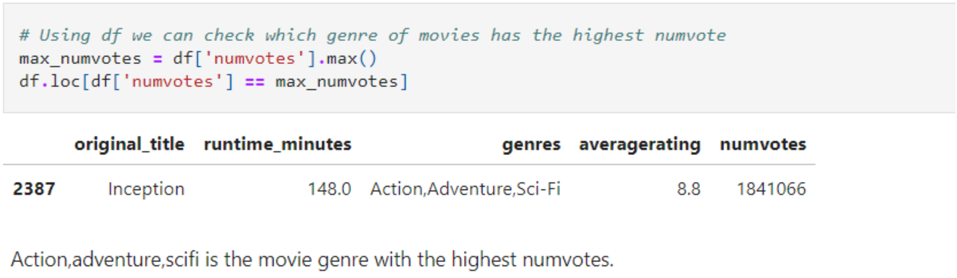
```
df_3
```

	title	domestic_gross	foreign_gross
0	Toy Story 3	415000000.0	652000000
1	Alice in Wonderland (2010)	334200000.0	691300000
2	Harry Potter and the Deathly Hallows Part 1	296000000.0	664300000
3	Inception	292600000.0	535700000
4	Shrek Forever After	238700000.0	513900000
...	...	...	...
3275	I Still See You	1400.0	1500000
3286	The Catcher Was a Spy	725000.0	229000
3309	Time Freak	10000.0	256000
3342	Reign of Judges: Title of Liberty - Concept Short	93200.0	5200
3353	Antonio Lopez 1970: Sex Fashion & Disco	43200.0	30000

2007 rows × 3 columns

Data modeling and evaluation

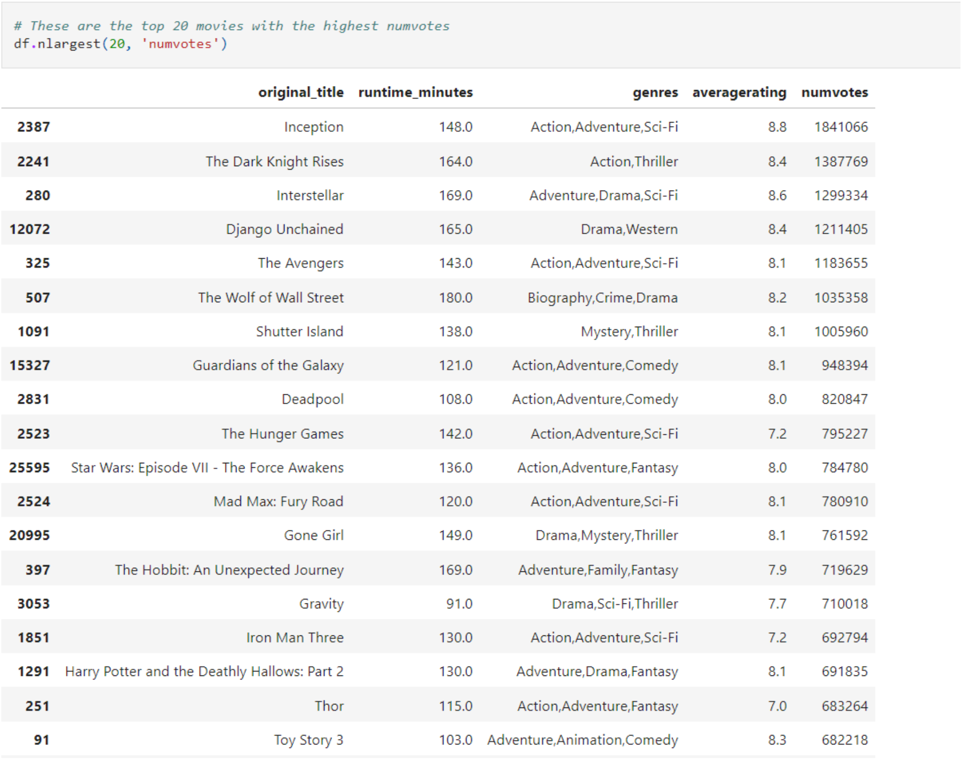
In modelling and evaluation the data was analysed to see which were the best rated genre of movies and which were most liked.The information is shown below:



```
# Using df we can check which genre of movies has the highest numvote
max_numvotes = df['numvotes'].max()
df.loc[df['numvotes'] == max_numvotes]
```

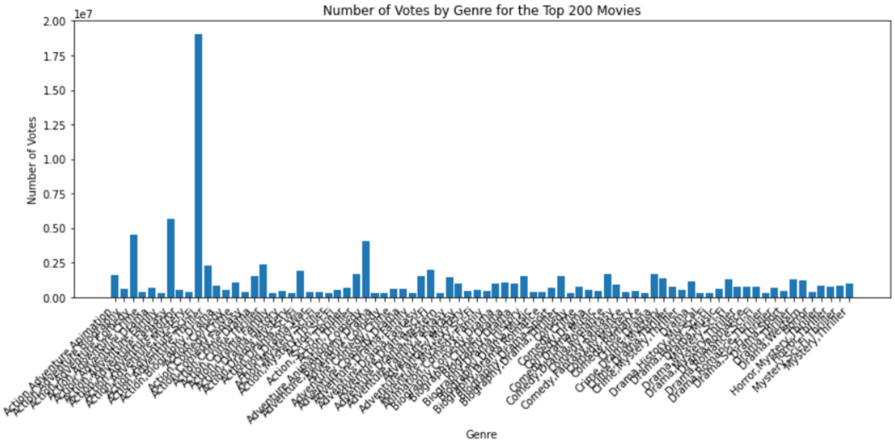
	original_title	runtime_minutes	genres	averagerating	numvotes
2387	Inception	148.0	Action,Adventure,Sci-Fi	8.8	1841066

Action,adventure,scifi is the movie genre with the highest numvotes.



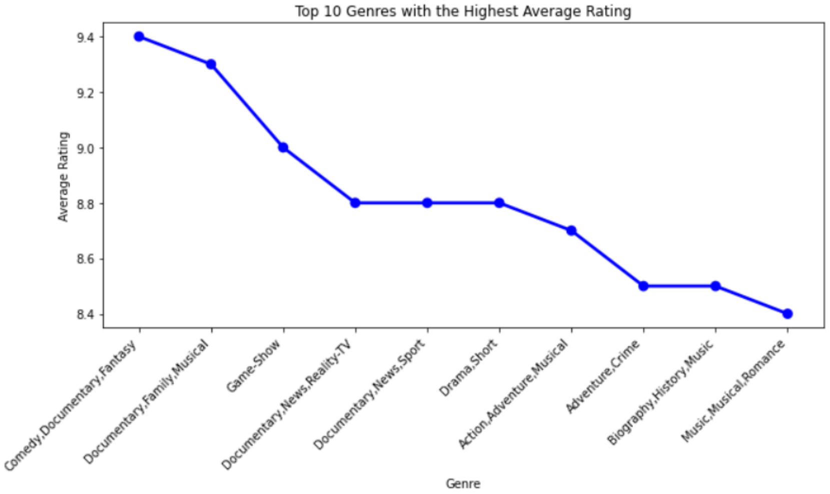
```
# These are the top 20 movies with the highest numvotes
df.nlargest(20, 'numvotes')
```

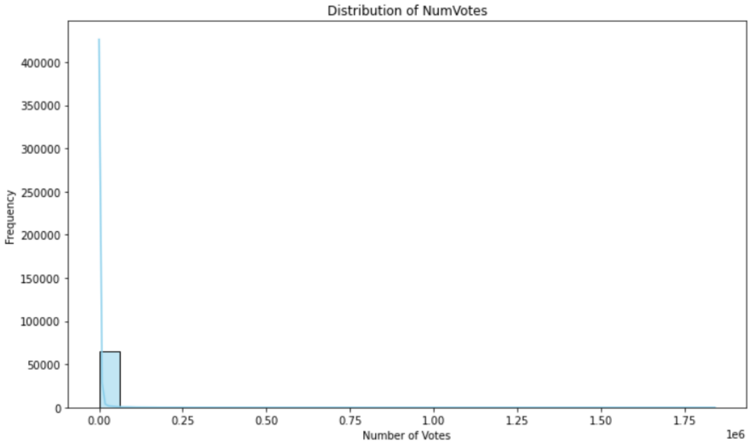
	original_title	runtime_minutes	genres	averagerating	numvotes
2387	Inception	148.0	Action,Adventure,Sci-Fi	8.8	1841066
2241	The Dark Knight Rises	164.0	Action,Thriller	8.4	1387769
280	Interstellar	169.0	Adventure,Drama,Sci-Fi	8.6	1299334
12072	Django Unchained	165.0	Drama,Western	8.4	1211405
325	The Avengers	143.0	Action,Adventure,Sci-Fi	8.1	1183655
507	The Wolf of Wall Street	180.0	Biography,Crime,Drama	8.2	1035358
1091	Shutter Island	138.0	Mystery,Thriller	8.1	1005960
15327	Guardians of the Galaxy	121.0	Action,Adventure,Comedy	8.1	948394
2831	Deadpool	108.0	Action,Adventure,Comedy	8.0	820847
2523	The Hunger Games	142.0	Action,Adventure,Sci-Fi	7.2	795227
25595	Star Wars: Episode VII - The Force Awakens	136.0	Action,Adventure,Fantasy	8.0	784780
2524	Mad Max: Fury Road	120.0	Action,Adventure,Sci-Fi	8.1	780910
20995	Gone Girl	149.0	Drama,Mystery,Thriller	8.1	761592
397	The Hobbit: An Unexpected Journey	169.0	Adventure,Family,Fantasy	7.9	719629
3053	Gravity	91.0	Drama,Sci-Fi,Thriller	7.7	710018
1851	Iron Man Three	130.0	Action,Adventure,Sci-Fi	7.2	692794
1291	Harry Potter and the Deathly Hallows: Part 2	130.0	Adventure,Drama,Fantasy	8.1	691835
251	Thor	115.0	Action,Adventure,Fantasy	7.0	683264
91	Toy Story 3	103.0	Adventure,Animation,Comedy	8.3	682218

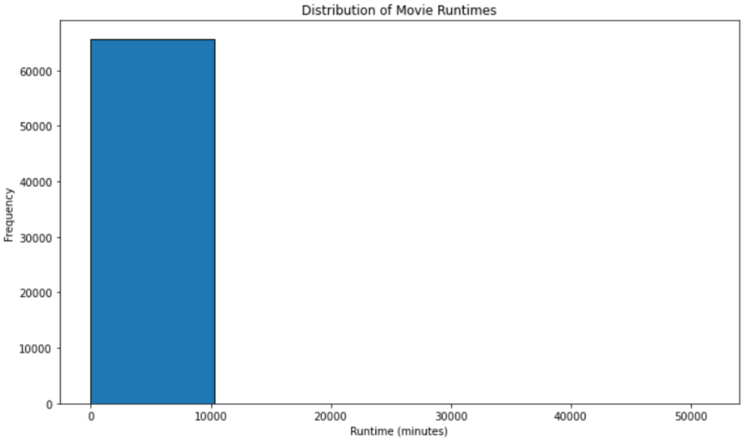


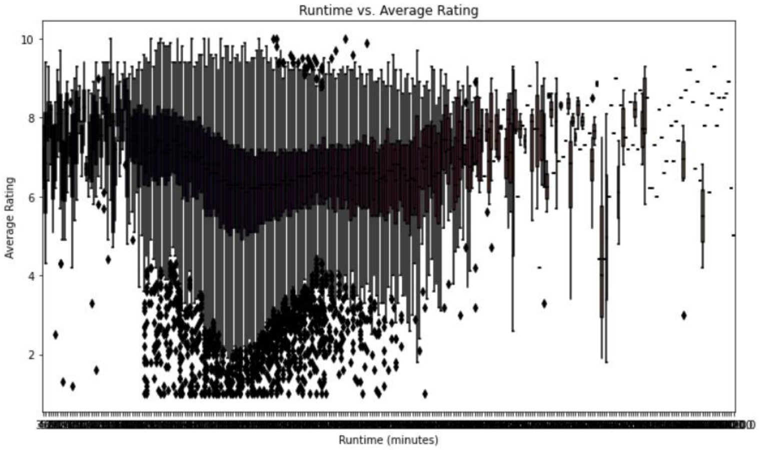
```
#Now we check which movie has the highest average rating.
max_rating = df['averagerating'].max()
df.loc[df['averagerating'] == max_rating]
```

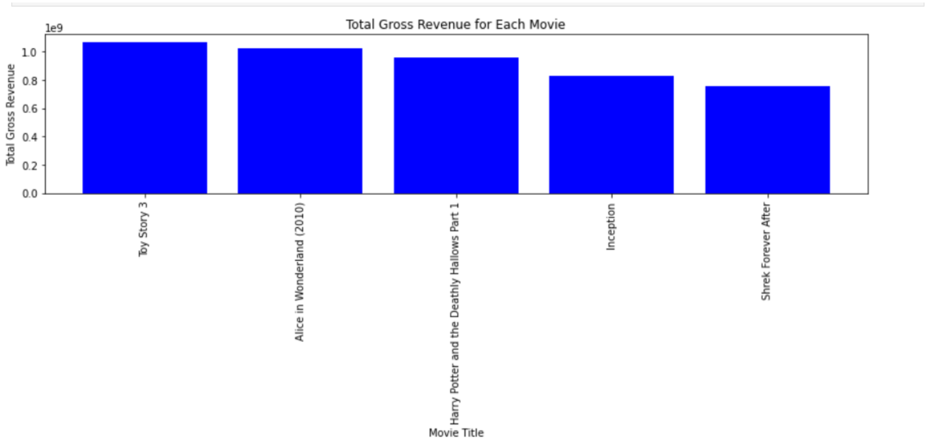
	original_title	runtime_minutes	genres	averagerating	numvotes
702	Exteriores: Mulheres Brasileiras na Diplomacia	52.0	Documentary	10.0	5
878	The Dark Knight: The Ballad of the N Word	129.0	Comedy,Drama	10.0	5
9745	Freeing Bernie Baran	100.0	Crime,Documentary	10.0	5
27335	Hercule contre Hermès	72.0	Documentary	10.0	5
42970	I Was Born Yesterday!	31.0	Documentary	10.0	6
50085	Revolution Food	70.0	Documentary	10.0	8
51109	Fly High: Story of the Disc Dog	65.0	Documentary	10.0	7
53689	Atlas Mountain: Barbary Macaques - Childcaring...	59.0	Documentary	10.0	5
60782	Requiem voor een Boom	48.0	Documentary	10.0	5
64646	A Dedicated Life: Phoebe Brand Beyond the Group	93.0	Documentary	10.0	5
65755	Ellis Island: The Making of a Master Race in A...	70.0	Documentary,History	10.0	6
65944	Calamity Kevin	77.0	Adventure,Comedy	10.0	6
71577	Pick It Up! - Ska in the '90s	99.0	Documentary	10.0	5




Distribution of NumVotes
Frequency
400000
350000
300000
250000
200000
150000
100000
50000
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e6
Number of Votes


Distribution of Movie Runtimes
Frequency
60000
50000
40000
30000
20000
10000
0
0
10000
20000
30000
40000
50000
Runtime (minutes)


Runtime vs. Average Rating
Average Rating
10
8
6
4
2
Runtime (minutes)



Final Evaluation

The choices made during data analysis and modeling are appropriate based on the data and business problem because they help in finding meaningful patterns, improving model performance, and ultimately, making informed decisions to address the business problem effectively. The iterative approach allows for refinement and fine-tuning, leading to better insights and predictions. The selection of models, feature engineering techniques, and hyperparameter tuning ensures that the model performs optimally and is suitable for the specific business problem at hand. By understanding the data, business context, and continuously improving the approach, data analysis and modeling can provide valuable insights and solutions for business decision-making.

Interpreting the results of a data analysis or model is crucial to draw meaningful conclusions and make informed decisions. Here are some aspects to consider when interpreting the results:

Model Fit and Performance: Evaluate how well the model fits the data and its overall performance metrics. For regression models, you can assess metrics like R-squared, mean squared error (MSE), or root mean squared error (RMSE). For classification models, consider accuracy, precision, recall, F1-score, etc. A higher R-squared or accuracy and lower error metrics indicate a better fit.

Baseline Model Comparison: Compare your model's performance with a baseline model. The baseline model can be a simple rule-based approach or a naive model. If your model significantly outperforms the baseline, it demonstrates its value in capturing patterns in the data.

Generalization: Assess how well the model generalizes beyond the data it was trained on. You can use techniques like cross-validation or hold-out testing to validate the model's performance on unseen data. If the model maintains good performance on unseen data, it indicates better generalization.

Business Impact: Consider the potential business impact of using the model. Will it provide valuable insights for decision-making? Can it help solve specific business problems or optimize processes? The more relevant the model's results are to the business's objectives, the more beneficial it becomes.

Uncertainty and Confidence Intervals: Acknowledge the uncertainty associated with the results. If your data is limited or noisy, the model's predictions may have wider confidence intervals. Communicate the level of uncertainty to stakeholders.

Validation and Peer Review: Seek validation and peer review from domain experts and stakeholders. Having others review your analysis can help identify potential biases or errors and provide different perspectives.

Assumptions and Limitations: Be aware of any assumptions made during the analysis. Consider the limitations of the data, model, and methodologies used. Transparency about these aspects adds credibility to your results.

Relevance to Business Objectives: Ensure that the analysis aligns with the business's specific goals. A successful model should provide actionable insights or aid in decision-making that directly supports the business's objectives.

Ultimately, the confidence in your results and the potential benefit to the business depends on various factors, including the quality of the data, the appropriateness of the chosen model, and the relevance of the analysis to the business context. Engaging domain experts, conducting thorough testing, and validating the results against real-world scenarios can increase confidence in the model's performance and utility.

Conclusion
Based on the data analysis performed, several insights can be drawn to guide the business decision-making process for Microsoft's new movie studio:

Genre Recommendations

The analysis identified the top genres such as***Action,Adventure,Sci-Fi*** with the highest average ratings and those that attract the most votes. This information can be used to prioritize the creation of movies in genres that have a higher likelihood of being well-received by the audience.

Profitability Analysis: By comparing the total gross revenue (domestic + foreign) for each movie, the most profitable movies can be identified. This can help the business focus on genres or movie concepts that have a track record of financial success.

Runtime Consideration: Analyzing the distribution of movie runtimes can assist in understanding the most common runtime ranges and whether there is any correlation between runtime and audience reception (numvotes or averagerating).

Data Limitations: It is important to acknowledge the limitations of the analysis. The data used in this project is based on a sample dataset, and real-world movie production involves many other factors such as production costs, marketing efforts, competition, and external events.

Recommendations

Genre Diversification: To minimize risk, the business can consider diversifying the movie genres they produce. While some genres may have higher average ratings or profitability, exploring multiple genres can attract a broader audience.

Audience Surveys: Conducting surveys or focus groups with target audiences can provide valuable insights into their preferences and expectations. This qualitative data can complement the quantitative analysis and improve decision-making.

Collaboration with Experts: Engaging industry experts, filmmakers, and screenwriters can provide valuable guidance in selecting movie concepts and genres with the potential for success.

Limitations

Limited Data: The analysis is based on a sample dataset, and the conclusions may not fully represent the entire movie industry. A more comprehensive dataset with a broader range of movies and genres would enhance the analysis's reliability.

Causality vs. Correlation: The analysis focuses on identifying correlations between variables, but it does not establish causality. There may be other factors not captured in the data that influence movie success.

Future Improvements

Data Enrichment: Include additional data attributes such as production budget, release date, marketing expenditure, and critical reviews to build more sophisticated models for predicting movie success.

Machine Learning Models: Explore the use of machine learning algorithms to build predictive models for movie success based on historical data.

Real-time Analysis: Implement real-time data collection and analysis to stay up-to-date with changing audience preferences and market trends.

A/B Testing: Conduct A/B testing for movie trailers, posters, and promotional materials to optimize marketing strategies and gauge audience interest.

Social Media Sentiment Analysis: Monitor social media platforms for audience sentiments and reactions to movies, providing insights into the public's response.

By continuously refining the data analysis and incorporating feedback from industry experts, Microsoft's new movie studio can make well-informed decisions, create content that resonates with the audience, and increase the likelihood of success in the competitive movie industry.