THE NUMBERS.COM

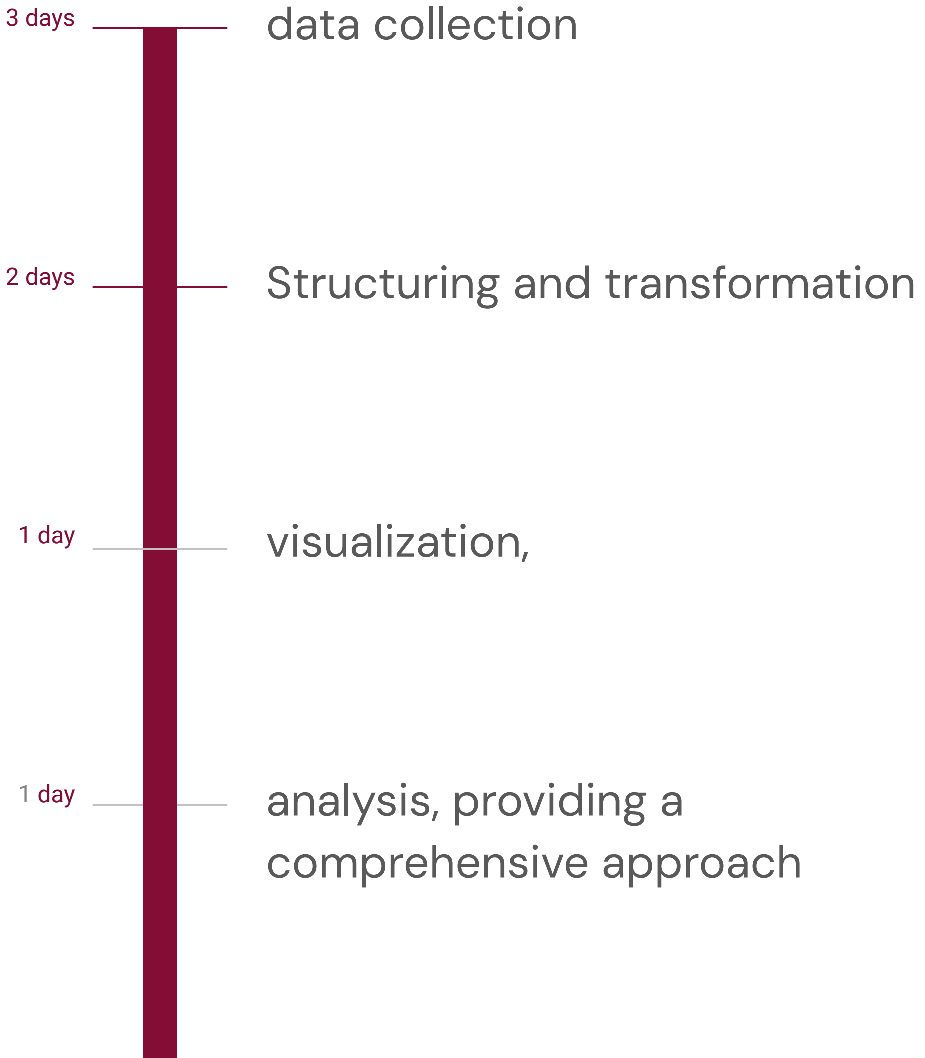
BOT DATA EXTRACTION, 1000 MOVIES ANALYSIS

BY:
DIVYAM BAJAJ

Objective of the Project

To analyze a dataset of over 1,000 movies based on IMDb ratings, worldwide box office performance, and profitability, examining various factors such as genre, profit margin, sources, production methods, production companies, and directors.

Workflow Overview



3 days data collection

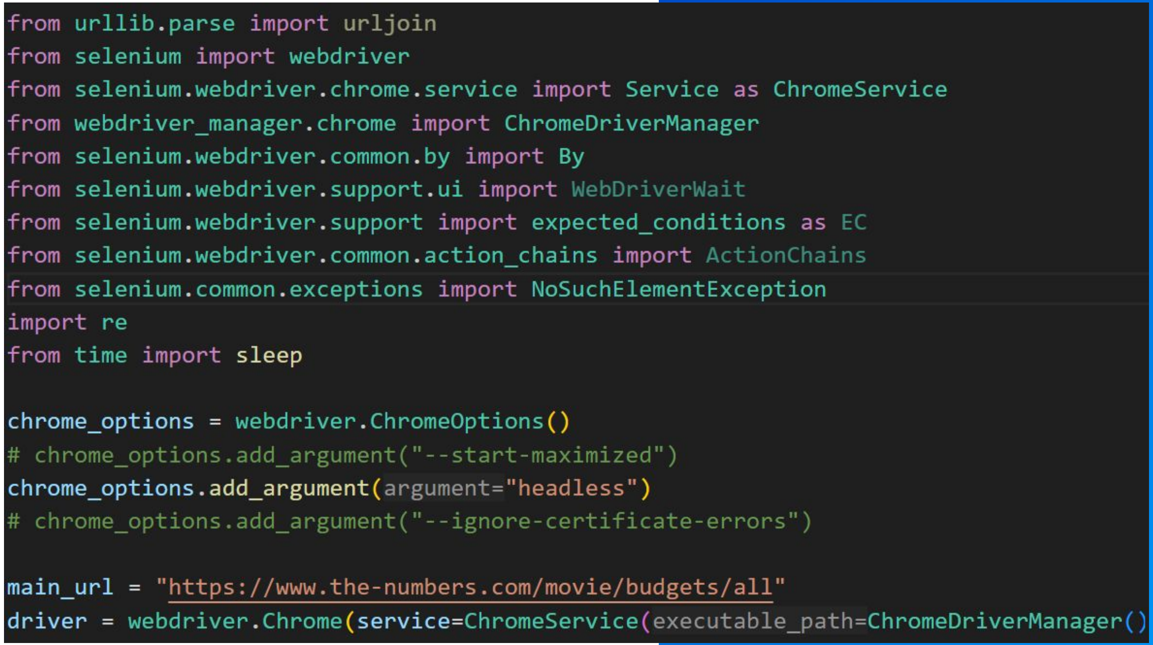
2 days Structuring and transformation

1 day visualization,

1 day analysis, providing a comprehensive approach

Data Collection Process

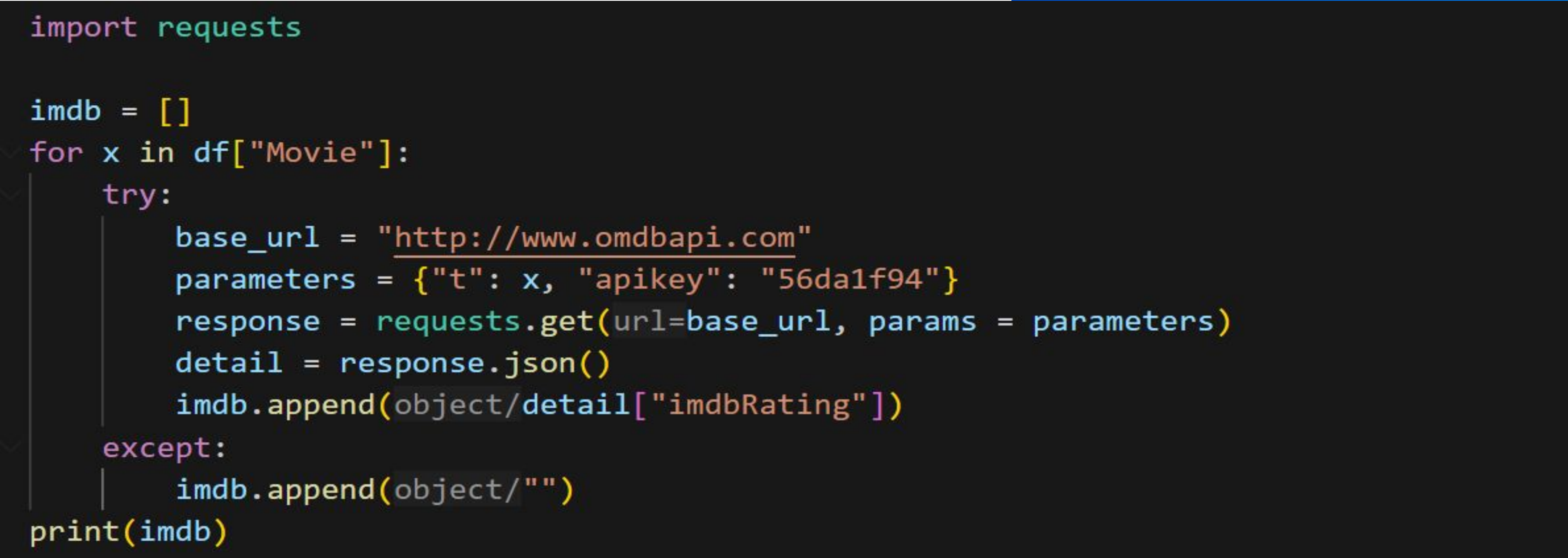
A Selenium bot was developed to extract data from The-Numbers.com, effectively navigating complex HTML structures across multiple pages to collect data on at least 1,000 movies.



```
from urllib.parse import urljoin
from selenium import webdriver
from selenium.webdriver.chrome.service import Service as ChromeService
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.action_chains import ActionChains
from selenium.common.exceptions import NoSuchElementException
import re
from time import sleep

chrome_options = webdriver.ChromeOptions()
# chrome_options.add_argument("--start-maximized")
chrome_options.add_argument(argument="headless")
# chrome_options.add_argument("--ignore-certificate-errors")

main_url = "https://www.the-numbers.com/movie/budgets/all"
driver = webdriver.Chrome(service=ChromeService(executable_path=ChromeDriverManager()
```

Retrieving IMDb Ratings

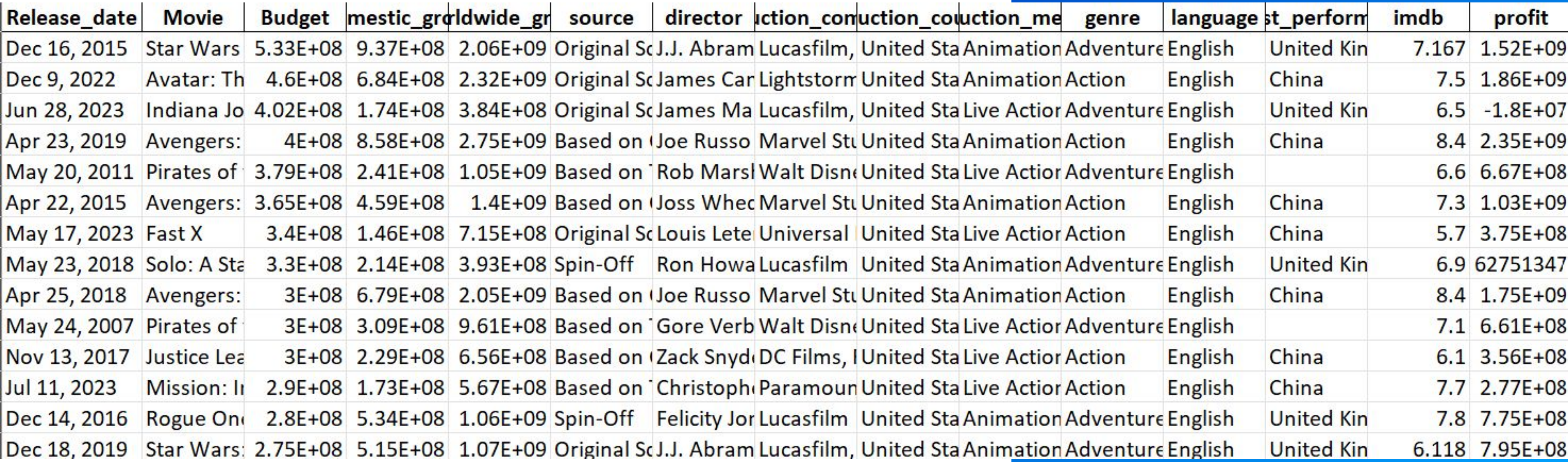
IMDb ratings were retrieved via API calls using a search method to match titles, which were then integrated into the main dataset for analysis.



```python
import requests

imdb = []
for x in df["Movie"]:
    try:
        base_url = "http://www.omdbapi.com"
        parameters = {"t": x, "apikey": "56da1f94"}
        response = requests.get(url=base_url, params = parameters)
        detail = response.json()
        imdb.append(object/detail["imdbRating"])
    except:
        imdb.append(object/"")
print(imdb)
```

Data Structuring Techniques

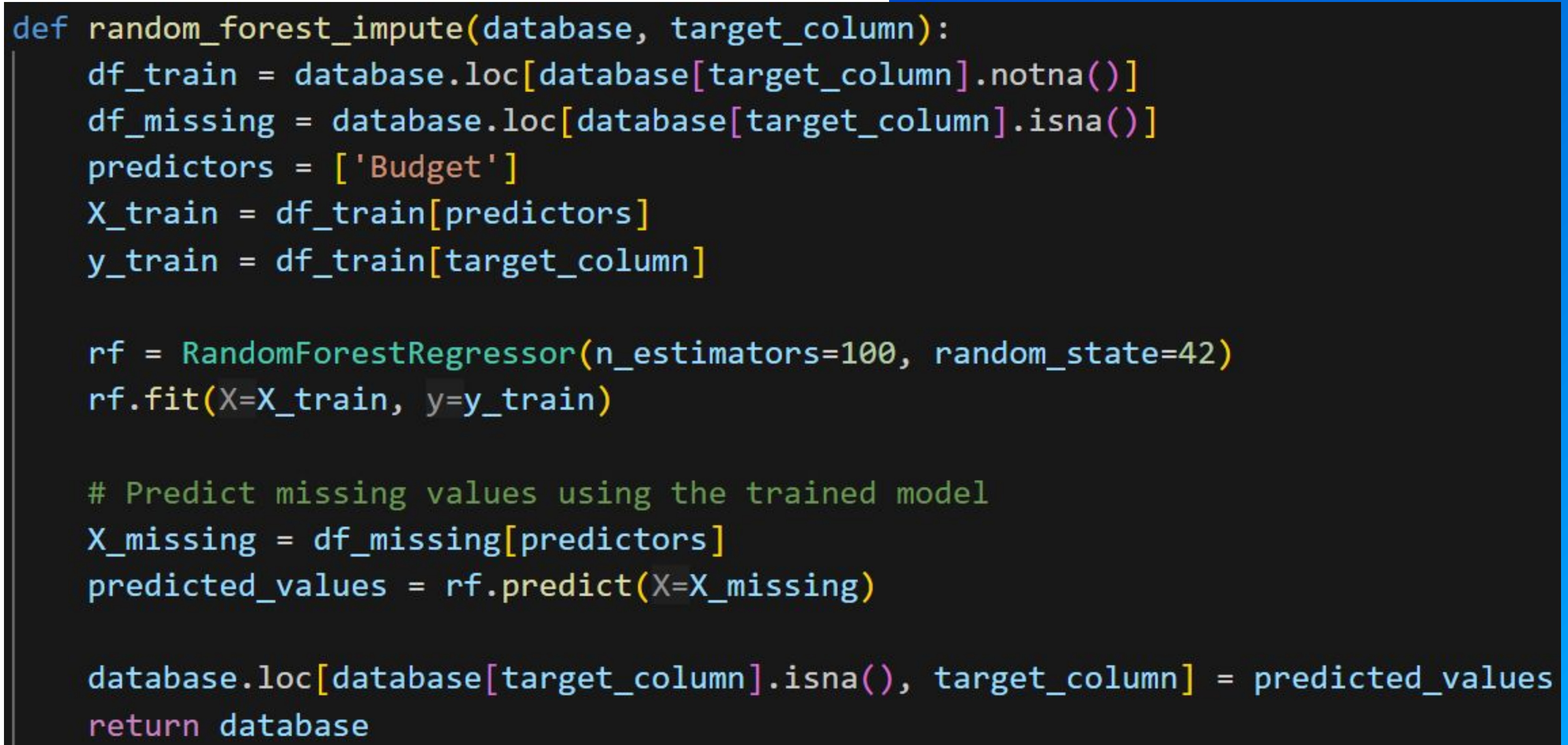
Data was organized into dataframes using Pandas and NumPy arrays, preparing the dataset for thorough analysis and visualization.



Release_date	Movie	Budget	mestic_gr	ldwide_gr	source	director	uction_con	uction_cou	uction_me	genre	language	st_perform	imdb	profit
Dec 16, 2015	Star Wars	5.33E+08	9.37E+08	2.06E+09	Original Sc	J.J. Abram	Lucasfilm,	United Sta	Animation	Adventure	English	United Kin	7.167	1.52E+09
Dec 9, 2022	Avatar: Th	4.6E+08	6.84E+08	2.32E+09	Original Sc	James Car	Lightstorm	United Sta	Animation	Action	English	China	7.5	1.86E+09
Jun 28, 2023	Indiana Jo	4.02E+08	1.74E+08	3.84E+08	Original Sc	James Ma	Lucasfilm,	United Sta	Live Actior	Adventure	English	United Kin	6.5	-1.8E+07
Apr 23, 2019	Avengers:	4E+08	8.58E+08	2.75E+09	Based on (	Joe Russo	Marvel Stu	United Sta	Animation	Action	English	China	8.4	2.35E+09
May 20, 2011	Pirates of	3.79E+08	2.41E+08	1.05E+09	Based on T	Rob Marsl	Walt Disne	United Sta	Live Actior	Adventure	English		6.6	6.67E+08
Apr 22, 2015	Avengers:	3.65E+08	4.59E+08	1.4E+09	Based on (	Joss Whec	Marvel Stu	United Sta	Animation	Action	English	China	7.3	1.03E+09
May 17, 2023	Fast X	3.4E+08	1.46E+08	7.15E+08	Original Sc	Louis Lete	Universal	United Sta	Live Actior	Action	English	China	5.7	3.75E+08
May 23, 2018	Solo: A Sta	3.3E+08	2.14E+08	3.93E+08	Spin-Off	Ron Howa	Lucasfilm	United Sta	Animation	Adventure	English	United Kin	6.9	62751347
Apr 25, 2018	Avengers:	3E+08	6.79E+08	2.05E+09	Based on (	Joe Russo	Marvel Stu	United Sta	Animation	Action	English	China	8.4	1.75E+09
May 24, 2007	Pirates of	3E+08	3.09E+08	9.61E+08	Based on T	Gore Verb	Walt Disne	United Sta	Live Actior	Adventure	English		7.1	6.61E+08
Nov 13, 2017	Justice Lea	3E+08	2.29E+08	6.56E+08	Based on (	Zack Snyd	DC Films, I	United Sta	Live Actior	Action	English	China	6.1	3.56E+08
Jul 11, 2023	Mission: I	2.9E+08	1.73E+08	5.67E+08	Based on T	Christoph	Paramoun	United Sta	Live Actior	Action	English	China	7.7	2.77E+08
Dec 14, 2016	Rogue On	2.8E+08	5.34E+08	1.06E+09	Spin-Off	Felicity Jor	Lucasfilm	United Sta	Animation	Adventure	English	United Kin	7.8	7.75E+08
Dec 18, 2019	Star Wars:	2.75E+08	5.15E+08	1.07E+09	Original Sc	J.J. Abram	Lucasfilm,	United Sta	Animation	Adventure	English	United Kin	6.118	7.95E+08

Missing Data Transformation

Missing values were addressed using Random Forest prediction, while regex was employed to convert special characters (e.g., "$") to numerical data, ensuring data integrity.



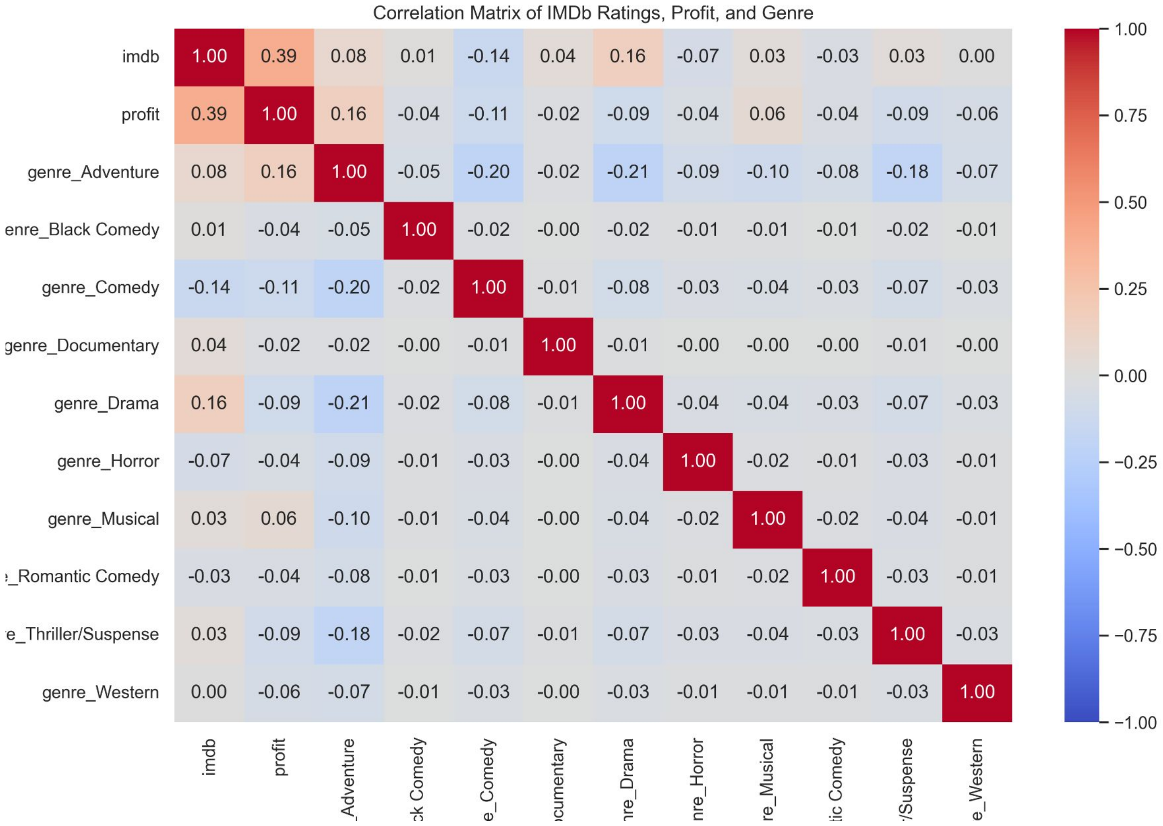
```python
def random_forest_impute(database, target_column):
    df_train = database.loc[database[target_column].notna()]
    df_missing = database.loc[database[target_column].isna()]
    predictors = ['Budget']
    X_train = df_train[predictors]
    y_train = df_train[target_column]

    rf = RandomForestRegressor(n_estimators=100, random_state=42)
    rf.fit(X=X_train, y=y_train)

    # Predict missing values using the trained model
    X_missing = df_missing[predictors]
    predicted_values = rf.predict(X=X_missing)

    database.loc[database[target_column].isna(), target_column] = predicted_values
    return database
```

Data Visualization Tools

Using Matplotlib, Seaborn, and Power BI, various visuals were created, including:

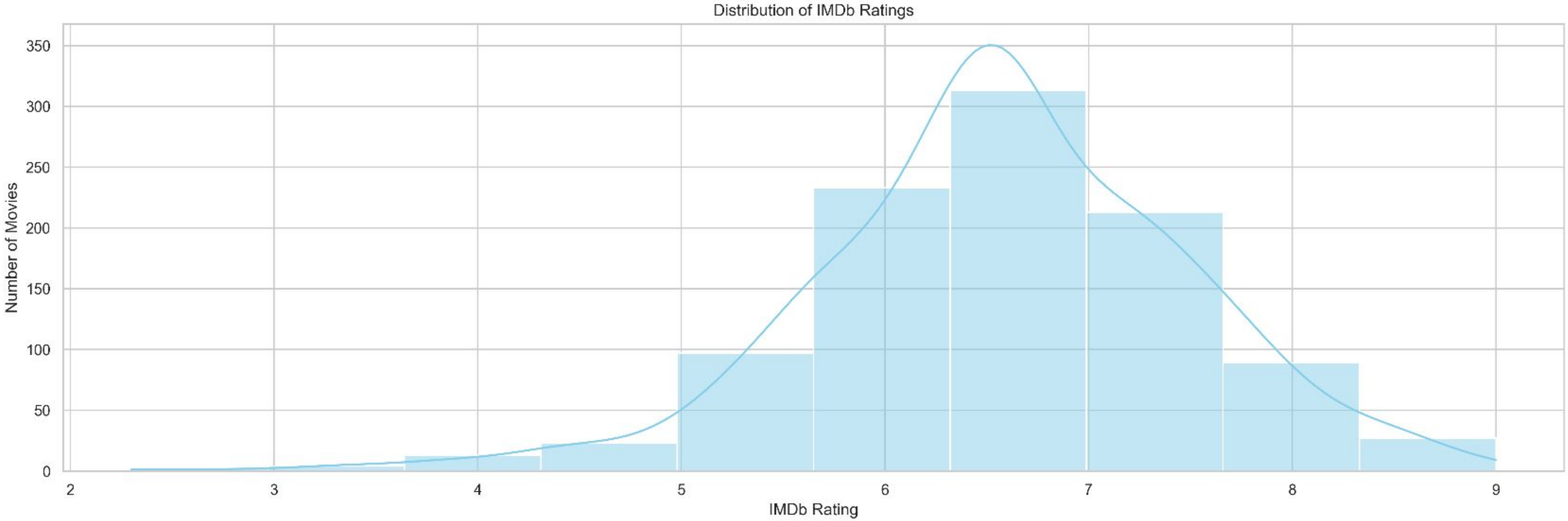
- Histogram: Distribution of IMDb ratings.
- Correlation Matrix: Analyzing IMDb and profit by genre.
- Bar Graphs: Worldwide gross by genre and performance by production method.
- Line Chart: Top-performing countries by earnings.
- Map: Geographic distribution of top-performing countries, genres, and production companies.
- Pie & Donut Charts: Relationship between genres and sources.
- Box Plot: IMDb ratings by genre and source.
- Bubble and Table Charts: Movie distribution by genre, source, and rating.

Correlation Matrix: Analyzing IMDb and profit by genre

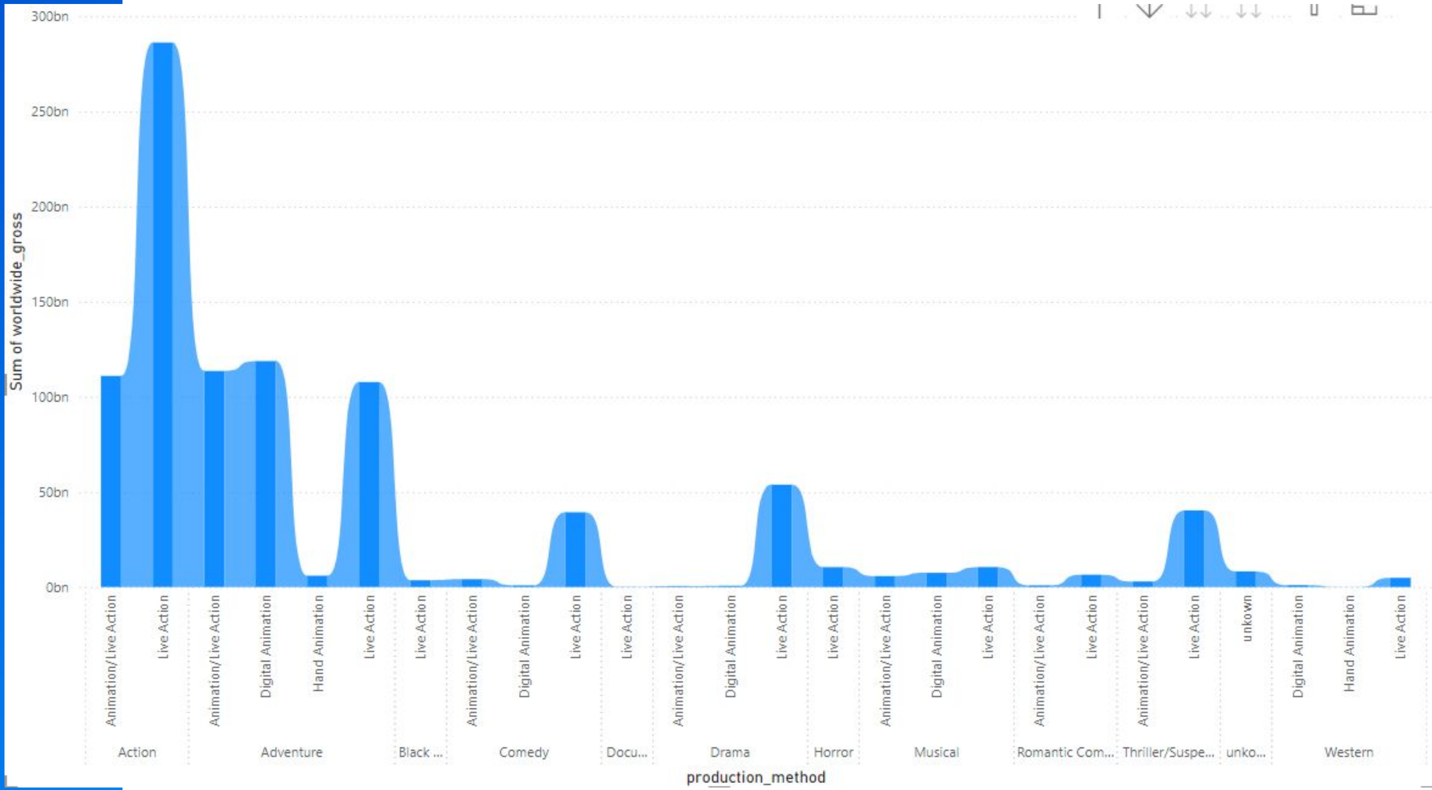

Correlation Matrix of IMDb Ratings, Profit, and Genre

	imdb	profit	_Adventure	ːk Comedy	e_Comedy	ɔcumentary	nre_Drama	nre_Horror	re_Musical	ːic Comedy	/Suspense	e_Western
imdb	1.00	0.39	0.08	0.01	-0.14	0.04	0.16	-0.07	0.03	-0.03	0.03	0.00
profit	0.39	1.00	0.16	-0.04	-0.11	-0.02	-0.09	-0.04	0.06	-0.04	-0.09	-0.06
genre_Adventure	0.08	0.16	1.00	-0.05	-0.20	-0.02	-0.21	-0.09	-0.10	-0.08	-0.18	-0.07
enre_Black Comedy	0.01	-0.04	-0.05	1.00	-0.02	-0.00	-0.02	-0.01	-0.01	-0.01	-0.02	-0.01
genre_Comedy	-0.14	-0.11	-0.20	-0.02	1.00	-0.01	-0.08	-0.03	-0.04	-0.03	-0.07	-0.03
genre_Documentary	0.04	-0.02	-0.02	-0.00	-0.01	1.00	-0.01	-0.00	-0.00	-0.00	-0.01	-0.00
genre_Drama	0.16	-0.09	-0.21	-0.02	-0.08	-0.01	1.00	-0.04	-0.04	-0.03	-0.07	-0.03
genre_Horror	-0.07	-0.04	-0.09	-0.01	-0.03	-0.00	-0.04	1.00	-0.02	-0.01	-0.03	-0.01
genre_Musical	0.03	0.06	-0.10	-0.01	-0.04	-0.00	-0.04	-0.02	1.00	-0.02	-0.04	-0.01
ː_Romantic Comedy	-0.03	-0.04	-0.08	-0.01	-0.03	-0.00	-0.03	-0.01	-0.02	1.00	-0.03	-0.01
'e_Thriller/Suspense	0.03	-0.09	-0.18	-0.02	-0.07	-0.01	-0.07	-0.03	-0.04	-0.03	1.00	-0.03
genre_Western	0.00	-0.06	-0.07	-0.01	-0.03	-0.00	-0.03	-0.01	-0.01	-0.01	-0.03	1.00

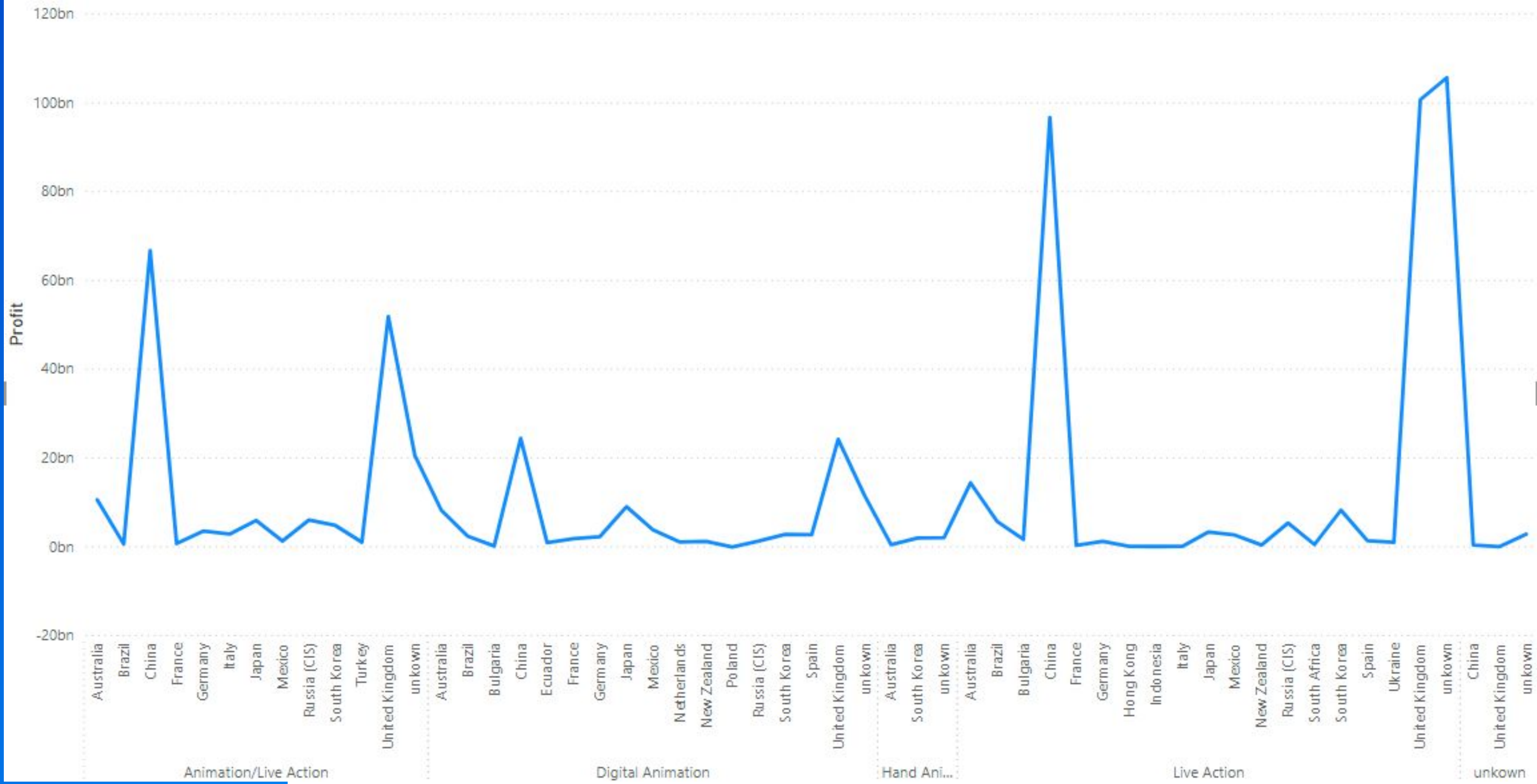
Histogram: Distribution of IMDb ratings.



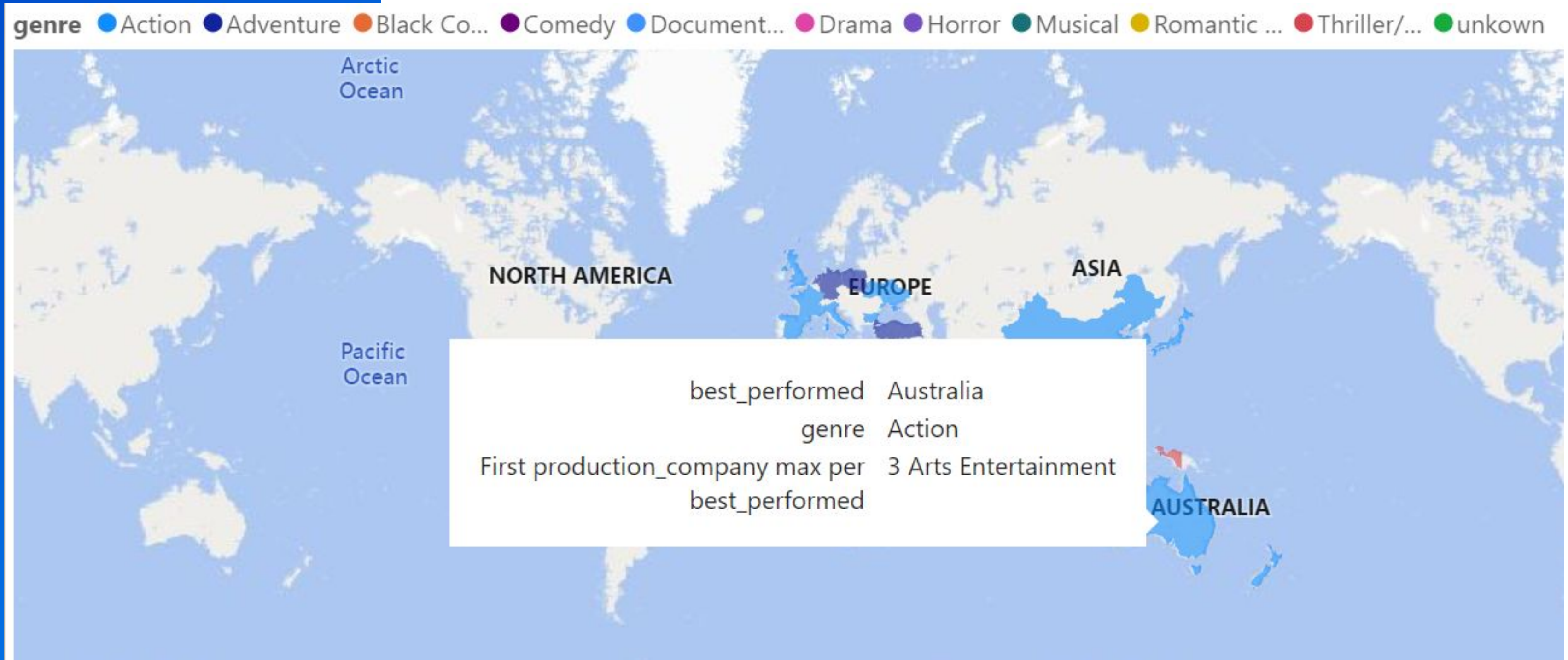
Bar Graphs: Worldwide gross by genre and performance by production method.



Line Chart: Top-performing countries by earnings.

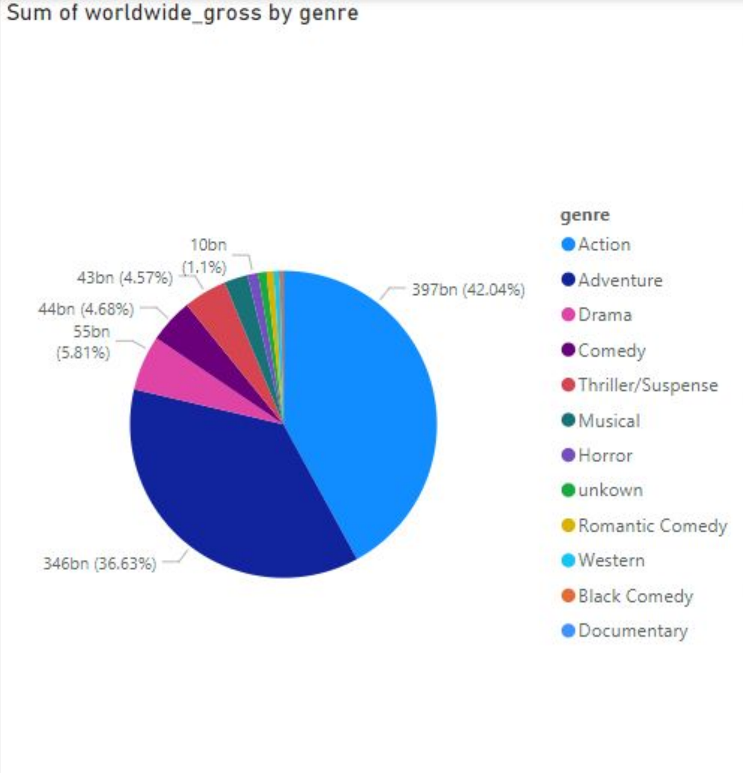


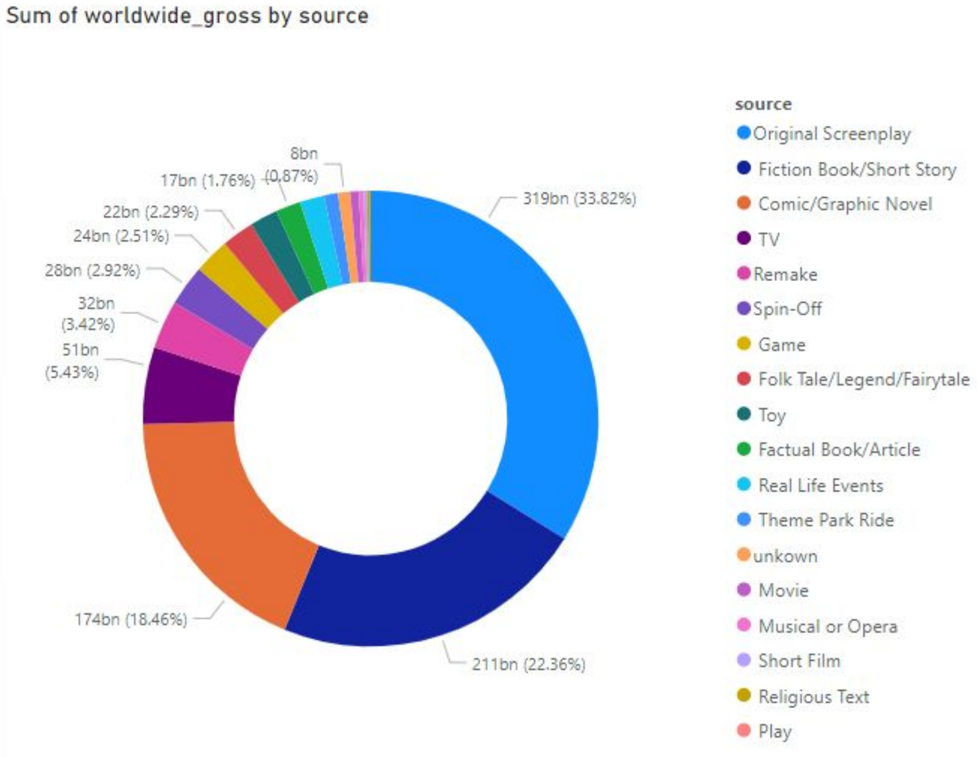
Map: Geographic distribution of top-performing countries, genres, and production companies.


genre ● Action ● Adventure ● Black Co... ● Comedy ● Document... ● Drama ● Horror ● Musical ● Romantic ... ● Thriller/... ● unkown
Arctic Ocean
NORTH AMERICA
EUROPE
ASIA
Pacific Ocean
AUSTRALIA

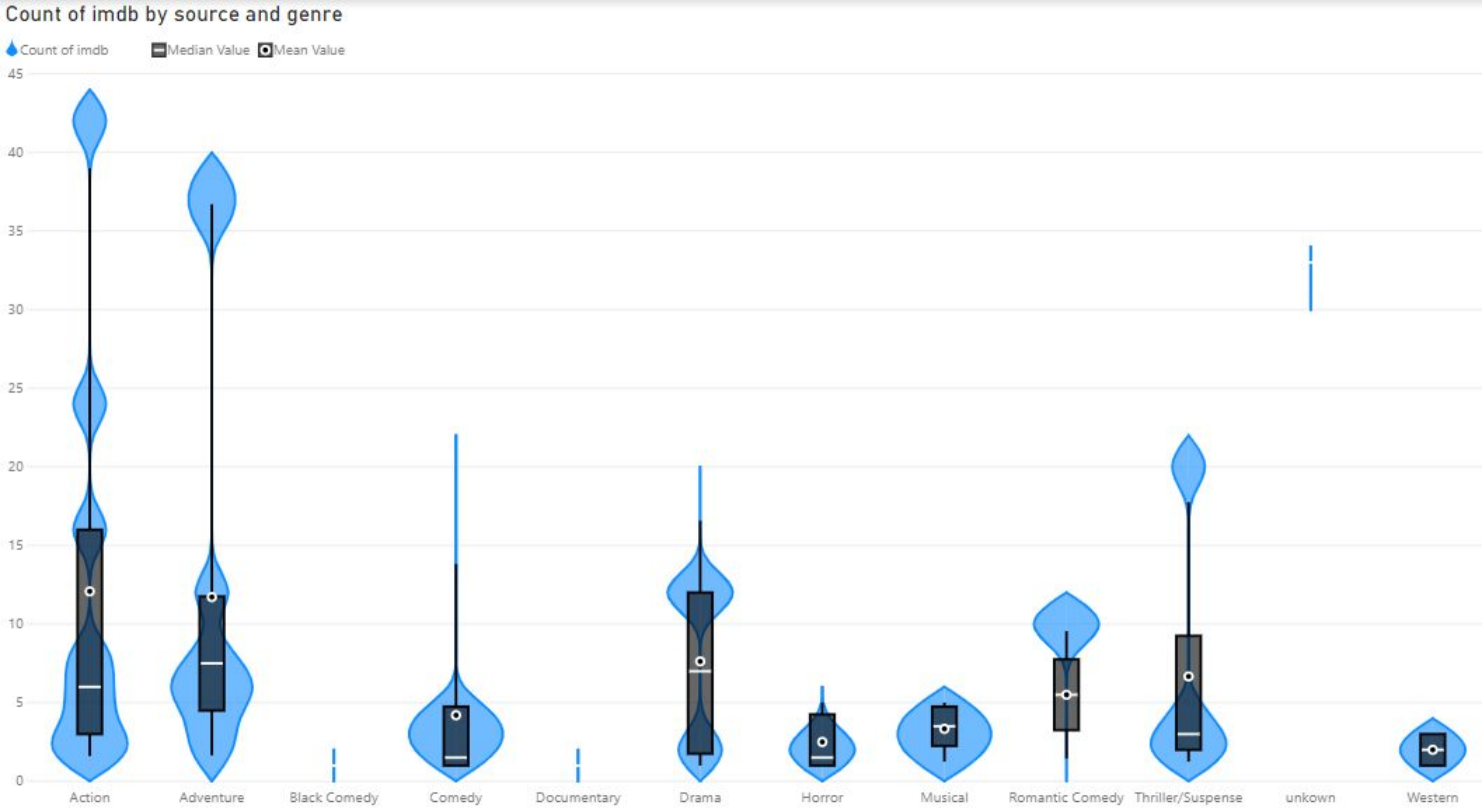
best_performed	Australia
genre	Action
First production_company max per best_performed	3 Arts Entertainment

Pie & Donut Charts: Relationship between genres and sources.





Box Plot: IMDb ratings by genre and source.



Bubble Charts: Movie distribution by genre, source, and rating.

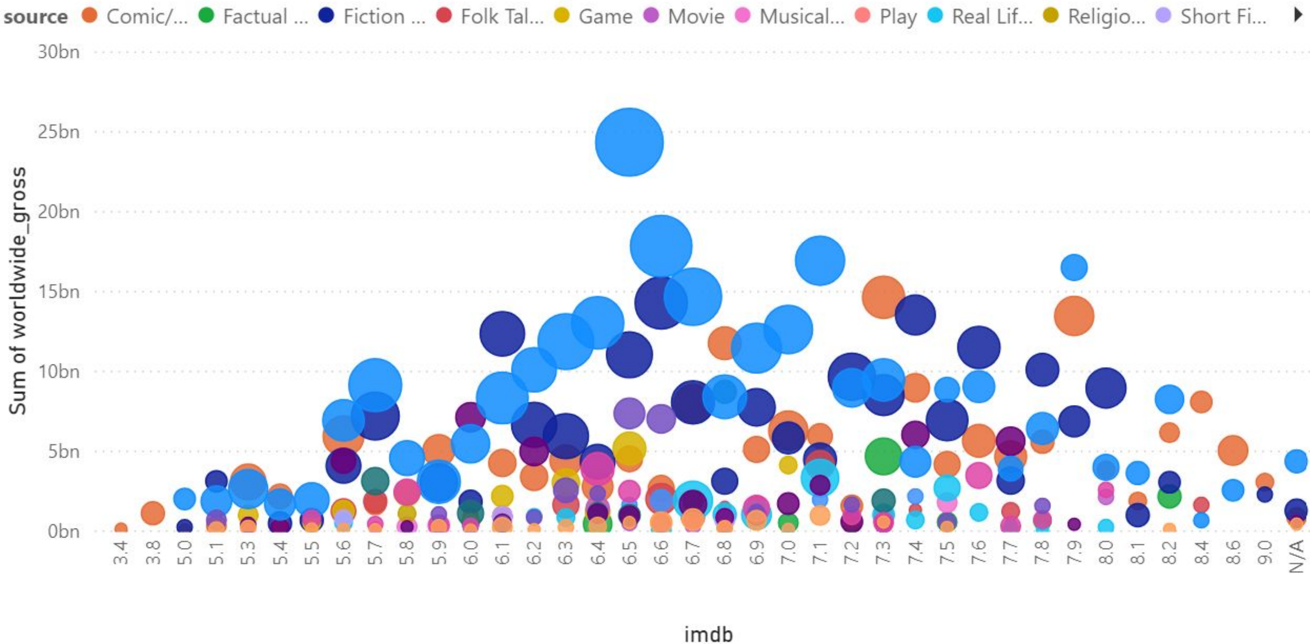
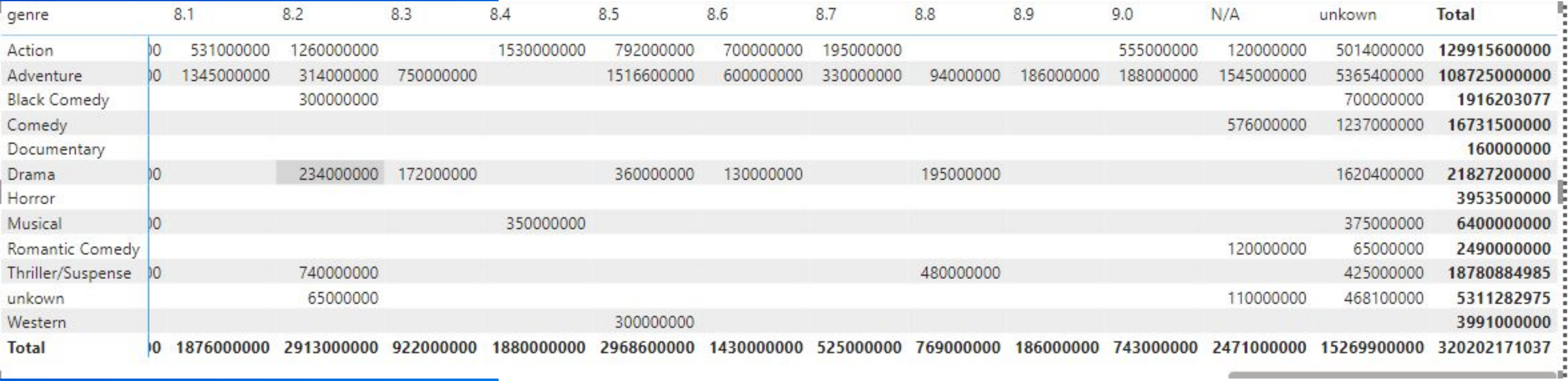
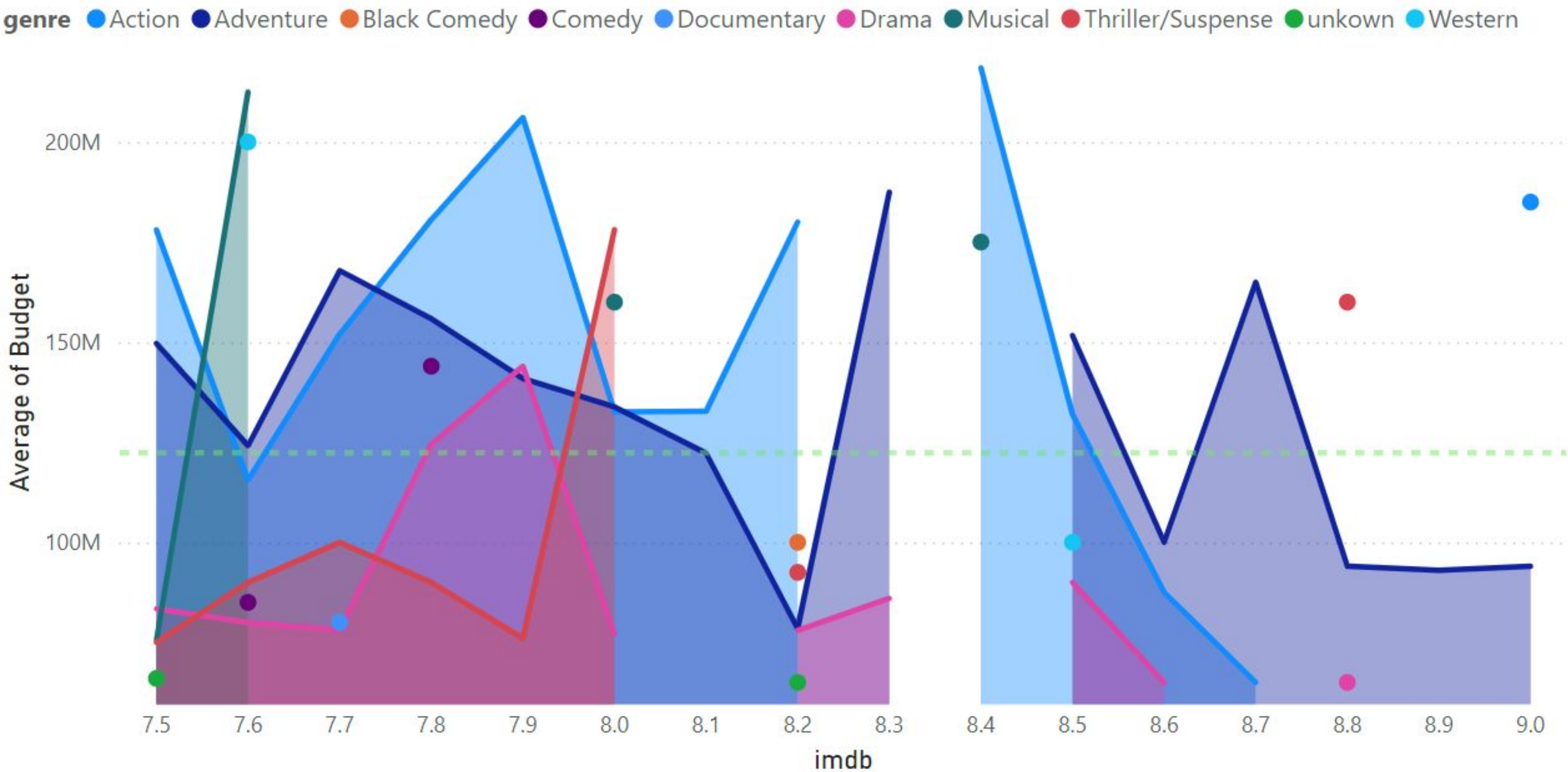


Table for Genre and Imdb ratings with values for Budget.



genre		8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0	N/A	unkown	**Total**
Action	0	531000000	1260000000		1530000000	792000000	700000000	195000000			555000000	120000000	5014000000	**129915600000**
Adventure	0	1345000000	314000000	750000000		1516600000	600000000	330000000	94000000	186000000	188000000	1545000000	5365400000	**108725000000**
Black Comedy			300000000										700000000	**1916203077**
Comedy												576000000	1237000000	**16731500000**
Documentary														**160000000**
Drama	0		234000000	172000000		360000000	130000000		195000000				1620400000	**21827200000**
Horror														**3953500000**
Musical	0				350000000								375000000	**6400000000**
Romantic Comedy												120000000	65000000	**2490000000**
Thriller/Suspense	0		740000000						480000000				425000000	**18780884985**
unkown			65000000									110000000	468100000	**5311282975**
Western						300000000								**3991000000**
Total	**0**	**1876000000**	**2913000000**	**922000000**	**1880000000**	**2968600000**	**1430000000**	**525000000**	**769000000**	**186000000**	**743000000**	**2471000000**	**15269900000**	**32020217103 7**

Graph for Genre and Imdb ratings with values for Budget.



Key Findings from the Analysis

Key findings include a strong correlation between genre and IMDb ratings, regional performance insights, and the ranking of top production companies, highlighting trends in the movie industry.

- Genre Correlation: Drama correlates highly with IMDb ratings, while adventure links to higher profits.
- IMDb Ratings: Most movies fall within the 6-7 IMDb rating range.
- Regional Performance: Action movies perform best in China; China and the UK lead in live-action and animation.
- Top Production Companies: Walt Disney, Warner Bros, Marvel Studios, Universal Pictures, Paramount, and DreamWorks rank as global leaders.
- Genre & Source Performance: Action is the top genre, followed by adventure, with original screenplays as the leading source.
- Action Movies have the highest Average Budgets and Adventure genre movies have average budgets around $125M.

Challenges Encountered

Challenges included developing a bot for dynamic pages, writing precise XPaths, using multiple tabs in Selenium, extracting country performance data, and handling financial data anomalies, which were all addressed during the project.

- Developing a bot to handle dynamic pages and navigate complex HTML.
- Writing precise XPaths for elements without clear indexing.
- Using multiple tabs in Selenium for data collection.
- Extracting country performance from charts with regex.
- Handling "$0" values in financial data with regex and predictive imputation.
- Training Random Forest Model for Missing Values.
- Plotting Correlation Matrix including categorical variable.