final

December 13, 2022

1 Movie Data Analysis

2 Microsoft's Entry Into Original Movie Production Exploration

2.1 Overview

This project will analyze data from the film industry to determine what first film a new film studio founded and funded my Microsoft should look like in order to produce the highest ROI(return on investment). Descriptive analysis shows that certain months of the year, directors, and genres have a higher likelyhood of resulting in a high ROI.

2.2 Business Problem

Microsoft is interested in producing their own original video content by starting a movie studio. I will analyze data scraped from various sites in regards to thousands films. This analysis will help me to identify the most optimal hiring of directors, the best month to release a film, and the genres that offer the most promising ROI for Microsoft. ## Data Understanding I'll be using data from the following sites: - imdb - The Numbers

This collection of data will generally provide me with the title of a film alongside the the date that it was released, the genre, the gross profits, and budgetary costs for producing a film. I will combine these datasets to determine the most profitable options for Microsoft to consider before investing in a founding of a movie studio.

[1]:
```
# Import necessary libraries
import pandas as pd
import sqlite3
import matplotlib.pyplot as plt
%matplotlib inline
```

[2]:
```
# Add necessary datasets
# imdb database
conn = sqlite3.connect("im.db")
q = """
SELECT
    mb.primary_title,
    mb.genres,
    p.primary_name
FROM movie_basics AS mb
```

```
JOIN directors AS d
    ON mb.movie_id = d.movie_id
JOIN persons AS p
    ON d.person_id = p.person_id
GROUP BY mb.primary_title
HAVING primary_profession LIKE '%director%'
"""
imdb = pd.read_sql(q, conn)

# The Number scraped data
tn_mb = pd.read_csv('zippedData/tn.movie_budgets.csv.gz')
```

[3]:
```
imdb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 121179 entries, 0 to 121178
Data columns (total 3 columns):
 #   Column         Non-Null Count   Dtype 
---  ------         --------------   ----- 
 0   primary_title  121179 non-null  object
 1   genres         118367 non-null  object
 2   primary_name   121179 non-null  object
dtypes: object(3)
memory usage: 2.8+ MB
```

[4]:
```
tn_mb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5782 entries, 0 to 5781
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   id                 5782 non-null   int64 
 1   release_date       5782 non-null   object
 2   movie              5782 non-null   object
 3   production_budget  5782 non-null   object
 4   domestic_gross     5782 non-null   object
 5   worldwide_gross    5782 non-null   object
dtypes: int64(1), object(5)
memory usage: 271.2+ KB
```

2.2.1 IMDB Data

By far the largest dataset I will be using throughout this project. I have already selected the records the tables movie_basics and persons provide. This edited dataset includes the title of of over 120 thousand films (under the column **primary_title**), a variety of genres (under the column **genre**), and the name of the director (under **primary_name**).

[5]:
```
imdb.head(2)
```

[5]:
```
            primary_title       genres          primary_name
0  !Women Art Revolution  Documentary  Lynn Hershman-Leeson
1       #1 Serial Killer       Horror          Stanley Yung
```

[6]:
```
imdb['genres'].value_counts()
```

[6]:
```
Documentary                  28141
Drama                        17947
Comedy                        7812
Horror                        3455
Comedy,Drama                  2949
                             ...
Biography,Fantasy,History        1
Action,Animation,Mystery         1
Reality-TV,Talk-Show             1
Adventure,Crime,Mystery          1
Drama,News,Sci-Fi                1
Name: genres, Length: 1035, dtype: int64
```

[7]:
```
imdb['primary_name'].value_counts()[:20]
```

[7]:
```
Omer Pasha                   62
Stephan Düfel                48
Rajiv Chilaka                47
Larry Rosen                  45
Graeme Duane                 44
Gérard Courant               44
Claudio Costa                42
Nayato Fio Nuala             40
Eckhart Schmidt              36
Tetsuya Takehora             33
Charlie Minn                 29
Yoshikazu Katô               27
Paul T.T. Easter             27
Narinderpal Singh Chandok    26
David DeCoteau               26
Philip Gardiner              26
Kazuyoshi Sekine             25
Manny Velazquez              25
Mototsugu Watanabe           25
Ram Gopal Varma              25
Name: primary_name, dtype: int64
```

2.2.2 The Number Data

The Number includes data for a little under 6000 films. It includes their their release date, production budget, domestic gross profit, and worldwide gross profit in dollars.

[8]:
```
tn_mb.head()
```

[8]:
```
   id  release_date                                      movie  \
0   1  Dec 18, 2009                                     Avatar
1   2  May 20, 2011  Pirates of the Caribbean: On Stranger Tides
2   3   Jun 7, 2019                                Dark Phoenix
3   4   May 1, 2015                    Avengers: Age of Ultron
4   5  Dec 15, 2017         Star Wars Ep. VIII: The Last Jedi

  production_budget domestic_gross worldwide_gross
0      $425,000,000   $760,507,625  $2,776,345,279
1      $410,600,000   $241,063,875  $1,045,663,875
2      $350,000,000    $42,762,350    $149,762,350
3      $330,600,000   $459,005,868  $1,403,013,963
4      $317,000,000   $620,181,382  $1,316,721,747
```

2.3 Data Preparation

2.3.1 Data Cleaning

IMDB Data Cleaning For IMDB database I will make the data easier to work with by removing the records that don't include any genre information since it makes up such a small percentage of the data. I will also rename the columns to make the data easier to read.

[9]:
```
# Remove the records that have NaN under the genre column
imdb = imdb.dropna()
```

[10]:
```
# Change the column names to make them more legible
imdb.rename(columns = {'primary_title': 'movie_title', 'primary_name':␣
↪'director_name'}, inplace = True)
```

The Number Data Cleaning For the The Number file I will make the data easier to work with by rename the **movie**, create a new column called **release_month** from the existing column **release_date**, removing unecessary columns, and convert the **production_budget**, **domestic_gross**, **worldwide_gross** to floats to make it easier to work with. Finally, I will reformat the **worldwide_gross** so that it becomes a number that is easier to read. I will also remove records where there is no domestic or worldwide gross profit.

[11]:
```
# Rename the movie column
tn_mb.rename(columns = {'movie': 'movie_title'}, inplace = True)

tn_mb['release_month'] = tn_mb['release_date'].str[:3]

# Remove unecessary columns
```

```
tn_mb = tn_mb.drop(columns=['id', 'release_date', 'domestic_gross'])
```

[12]:
```
# Convert production_budget, domestic_gross, and worldwide_gross to floats
tn_mb['production_budget'] = tn_mb['production_budget'].replace('[\$,]', '',␣
↪regex=True).astype(float)
tn_mb['worldwide_gross'] = tn_mb['worldwide_gross'].replace('[\$,]', '',␣
↪regex=True).astype(float)
```

[13]:
```
# Convert the worldwide_gross column to an easier to read number
pd.set_option('display.float_format','{:.2f}'.format)
```

[14]:
```
# Remove all records that have both domestic_gross and worldwide_gross == 0
tn_mb = tn_mb[ tn_mb['worldwide_gross'] != 0]
```

2.4 Mergin Datasets

Combining the data from The Number and IMDB allows me to work on a single dataset for feature engineering and analysis. I will exclude any unmatched records between the **imdb** and **tn_mb** data to ensure that there are no missing galues for the data features.

[15]:
```
print(tn_mb.info())
tn_mb.head(2)
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5415 entries, 0 to 5781
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   movie_title        5415 non-null   object
 1   production_budget  5415 non-null   float64
 2   worldwide_gross    5415 non-null   float64
 3   release_month      5415 non-null   object
dtypes: float64(2), object(2)
memory usage: 211.5+ KB
None
```

[15]:

	movie_title	production_budget	worldwide_gross	release_month
0	Avatar	425000000.00	2776345279.00	Dec
1	Pirates of the Caribbean: On Stranger Tides	410600000.00	1045663875.00	May

[16]:
```
# Merge tn_mb and bom_movie_gross on movie_title
movie_data = pd.merge(tn_mb,
                      imdb,
```

```
                      on=['movie_title'],
                      how='inner')
```

[17]:
```
movie_data.head(3)
```

[17]:
```
                                 movie_title  production_budget  \
0                                     Avatar       425000000.00
1  Pirates of the Caribbean: On Stranger Tides       410600000.00
2                                Dark Phoenix       350000000.00

   worldwide_gross release_month                    genres  director_name
0    2776345279.00           Dec                    Horror   Atsushi Wada
1    1045663875.00           May  Action,Adventure,Fantasy   Rob Marshall
2     149762350.00           Jun   Action,Adventure,Sci-Fi  Simon Kinberg
```

2.5 Feature Engineering

I create a **total_roi** to see how each film profited based on the **production_budget** and **worldwide_gross** since it includes both domestic and foreign gross profits.

At this point I will also drop unecessary column and re-order them so it becomes easier to read moving forward.

[18]:
```
# Create the roi column
movie_data['roi'] = movie_data['worldwide_gross'] -␣
 ↪movie_data['production_budget']
movie_data
```

[18]:
```
                                    movie_title  production_budget  \
0                                        Avatar       425000000.00
1     Pirates of the Caribbean: On Stranger Tides       410600000.00
2                                   Dark Phoenix       350000000.00
3                           Avengers: Age of Ultron       330600000.00
4                            Avengers: Infinity War       300000000.00
…                                              …                  …
1922                                       Krisha           30000.00
1923                                        Emily           27000.00
1924                                       Exeter           25000.00
1925                                        Clean           10000.00
1926                                         Cure           10000.00

      worldwide_gross release_month                    genres  \
0       2776345279.00           Dec                    Horror
1       1045663875.00           May  Action,Adventure,Fantasy
2        149762350.00           Jun   Action,Adventure,Sci-Fi
3       1403013963.00           May   Action,Adventure,Sci-Fi
4       2048134200.00           Apr   Action,Adventure,Sci-Fi
…                 …             …                         …
```

```
1922        144822.00           Mar                       Drama
1923          3547.00           Jan                       Drama
1924        489792.00           Sep   Horror,Mystery,Thriller
1925        138711.00           Apr       Comedy,Drama,Horror
1926         94596.00           Jul                       Drama

           director_name          roi
0           Atsushi Wada 2351345279.00
1           Rob Marshall  635063875.00
2          Simon Kinberg -200237650.00
3            Joss Whedon 1072413963.00
4          Anthony Russo 1748134200.00
…                    …           …
1922  Trey Edward Shults     114822.00
1923      Timothy McNeil     -23453.00
1924       Marcus Nispel     464792.00
1925       Graham Wright     128711.00
1926            Bill Yip      84596.00

[1927 rows x 7 columns]
```

[19]:
```
# Re-orders and drops some unecessary columns
movie_data = movie_data[['movie_title',
                         'release_month',
                         'genres',
                         'director_name',
                         'roi']]
```

[20]:
```
movie_data.head()
```

[20]:
```
                                  movie_title release_month  \
0                                      Avatar           Dec
1  Pirates of the Caribbean: On Stranger Tides           May
2                                 Dark Phoenix           Jun
3                     Avengers: Age of Ultron           May
4                      Avengers: Infinity War           Apr

                    genres  director_name          roi
0                   Horror   Atsushi Wada 2351345279.00
1  Action,Adventure,Fantasy   Rob Marshall  635063875.00
2   Action,Adventure,Sci-Fi  Simon Kinberg -200237650.00
3   Action,Adventure,Sci-Fi    Joss Whedon 1072413963.00
4   Action,Adventure,Sci-Fi  Anthony Russo 1748134200.00
```

2.6 Analysis

2.6.1 Most Profitable Month of Release

Most films in this dataset are released on the last four months of the year with October being the most likely month for a film to be released on.

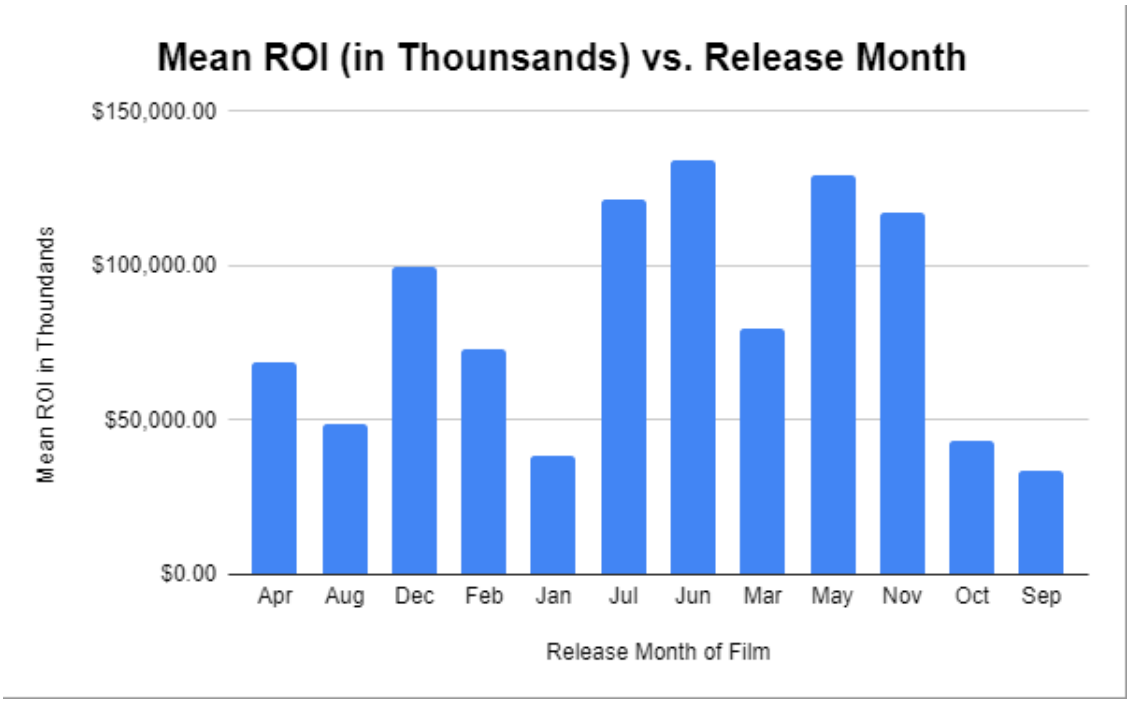
However, we see that on taking the mean average ROI for films released by month we can see that the data doesn't skew that way. **May**, **June**, and **July** offer the **highest mean average ROI for films released**. While **November** follows closely behind. This can be helpful to keep in mind in the case that the first film produced by Microsoft gets delayed we can plan to release the film in November.

The median average ROI of films released by specific months helps to clarify that, although there are many films that don't provide a high ROI (sometimes under 10,000,000 dollars in return), a film released in **July** or **November** still have a **higher median average return**.

[21]:
```
# Group data by month and show count, median, mean, max, min of roi
profit_months = movie_data[['release_month','roi']].groupby(['release_month']).
↪agg(['count','mean', 'median'])
profit_months
```

[21]:

	roi		
	count	mean	median
release_month			
Apr	168	68468781.45	8259822.50
Aug	164	48637008.09	16021216.00
Dec	195	99765770.37	17342956.00
Feb	133	72837442.39	22060480.00
Jan	121	38607968.99	18752858.00
Jul	153	121334211.44	42898100.00
Jun	153	134005810.93	29400000.00
Mar	163	79526273.07	14758389.00
May	137	129068713.41	24042224.00
Nov	169	117397373.26	35196684.00
Oct	197	43499637.74	4769209.00
Sep	174	33370936.45	7840750.50



2.6.2 Director Most Likely to Create a Film with a ROI

If we look at the top 5 most commonly seen directors in this dataset we see: - Steven Spielberg - David Gordon Green - Steven Soderbergh - Ridley Scott - Clint Eastwood

Selecting the top 5 most popular (or commonly used directors) we can determine what would be the most likely name to get audiences excited for a film since they can expect a certain direction with the production of a film.

Based on this data, I would recommend the following directors in order of most likely to provide a high ROI as the following: - Steven Spielberg - Ridley Scott - Clint Eastwood

These directors are likely to offer double what the directors in the 75% quartile could in regards to ROI. However, Clint Eastwood seems to have the lowest median average ROI of the group. This would suggest that he has some films he's directed where the ROI is significantly lower.

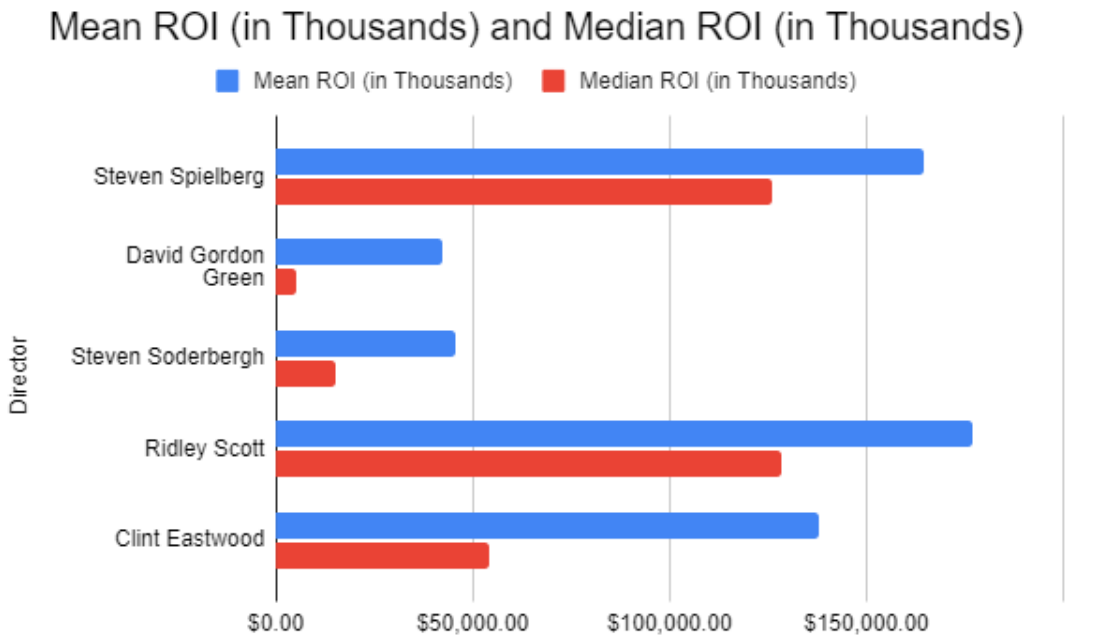
[22]:
```
# Creates the dataset grouped by directors and displays their average ROI for␣
 ↪the total films they've produced
profit_directors_avg = movie_data[['director_name','roi']].
 ↪groupby(['director_name']).agg(['count','mean', 'median'])
```

[23]:
```
# Most popular director in the dataset
profit_directors_avg.sort_values(by=('roi','count'), ascending=False).head(5)
```

[23]:

	roi		
	count	mean	median
director_name			
Steven Spielberg	8	164754974.38	126123609.00
David Gordon Green	8	42364834.50	5273421.50

```
Steven Soderbergh      7  45581726.86  15264271.00
Ridley Scott           7 176967321.29 128314513.00
Clint Eastwood         6 137916026.67  54133150.00
```



[24]:
```
# Most popular director in the dataset
profit_directors_avg.describe()
```

[24]:

	roi		
	count	mean	median
count	1416.00	1416.00	1416.00
mean	1.36	61111504.97	59463083.43
std	0.83	141651999.07	141344644.73
min	1.00	-200237650.00	-200237650.00
25%	1.00	-1897597.75	-1956212.25
50%	1.00	12716453.50	12393289.25
75%	1.00	61806325.38	61049344.25
max	8.00	2351345279.00	2351345279.00

2.6.3 ROI Based on Genre

Movies can have a combination of genres. However, as we can see by displaying the top 10 most common genres (we will discount documentaries since we can see that they do not offer a high ROI based on the top 10 most profitable film genres) some films can include a **combination of genres** that make it something unique all on their own.

Although the data would suggest that a film with a combination of genres like **Adventure**, **Drama**, and **Sport** would deliver the highest ROI I would consider this more of an outlier. Instead I would

consider films that offer the following combination of genres: - Comedy, and Mystery - Action, Adventure, and Sci-Fi - Adventure, and Fantasy

These genres are the most likely type of films to produce a high ROI. However, films with the combination genres of **Action**, **Adventure**, and **Sci-Fi** falls under one of the top 10 most common genre combinations

[25]:
```
# Creates the dataset grouped by genres and displays their average ROI for the␣
 ↪total films they've produced
profit_genre_avg = movie_data[['genres','roi']].groupby(['genres']).
 ↪agg(['count','mean', 'median'])
```

Films with a **Drama** are by far the most common to be produced based on the data gathered.

[26]:
```
# Most popular genres
profit_genre_avg.sort_values(by=('roi','count'), ascending=False).head(11)
```

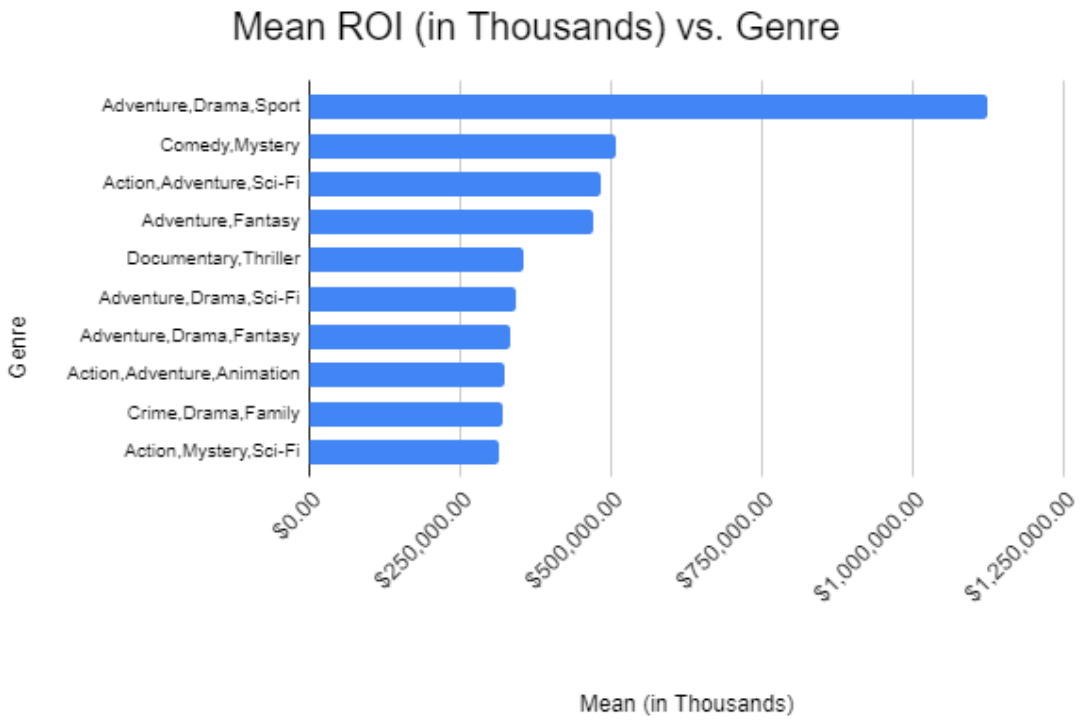
[26]:

	roi		
	count	mean	median
genres			
Drama	161	26100081.86	4152584.00
Documentary	80	38019683.12	3657215.00
Comedy	62	36365263.87	14067620.00
Drama,Romance	56	35333689.96	12624836.50
Comedy,Drama,Romance	53	25724635.98	4416951.00
Adventure,Animation,Comedy	53	262811628.42	151091610.00
Comedy,Drama	51	30998083.96	12141617.00
Action,Adventure,Sci-Fi	51	483139106.27	369076069.00
Action,Crime,Drama	40	33897239.35	26192531.00
Comedy,Romance	37	59860290.03	31623819.00
Action,Adventure,Fantasy	36	229761302.75	136332777.50

[27]:
```
# Top 10 genres with the highest mean ROI
profit_genre_avg.sort_values(by=('roi','mean'), ascending=False).head(10)
```

[27]:

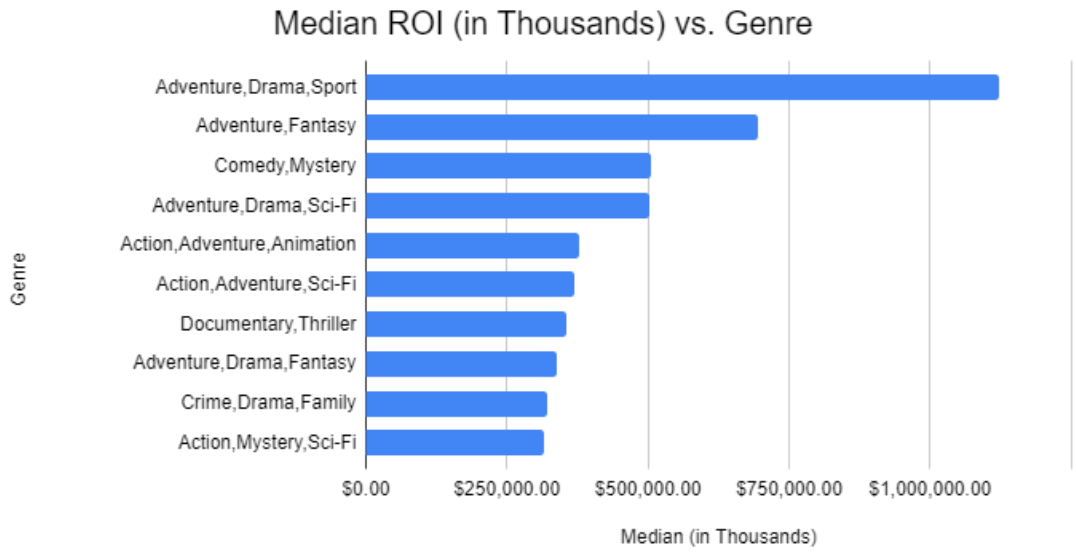
	roi		
	count	mean	median
genres			
Adventure,Drama,Sport	1	1122469910.00	1122469910.00
Comedy,Mystery	1	506464305.00	506464305.00
Action,Adventure,Sci-Fi	51	483139106.27	369076069.00
Adventure,Fantasy	3	469544026.33	695577621.00
Documentary,Thriller	1	354683805.00	354683805.00
Adventure,Drama,Sci-Fi	3	343699429.67	501379375.00
Adventure,Drama,Fantasy	4	334192689.25	338601398.50
Action,Adventure,Animation	17	322257606.47	377599142.00
Crime,Drama,Family	1	321116343.00	321116343.00
Action,Mystery,Sci-Fi	1	314319861.00	314319861.00



[28]:
```
# Top 10 genres with the highest median ROI
profit_genre_avg.sort_values(by=('roi','median'), ascending=False).head(10)
```

[28]:

	roi		
	count	mean	median
genres			
Adventure,Drama,Sport	1	1122469910.00	1122469910.00
Adventure,Fantasy	3	469544026.33	695577621.00
Comedy,Mystery	1	506464305.00	506464305.00
Adventure,Drama,Sci-Fi	3	343699429.67	501379375.00
Action,Adventure,Animation	17	322257606.47	377599142.00
Action,Adventure,Sci-Fi	51	483139106.27	369076069.00
Documentary,Thriller	1	354683805.00	354683805.00
Adventure,Drama,Fantasy	4	334192689.25	338601398.50
Crime,Drama,Family	1	321116343.00	321116343.00
Action,Mystery,Sci-Fi	1	314319861.00	314319861.00



[29]:
```
conn.close()
```

2.7 Conclusions

This analysis leads to three recommendations for increasing the likelyhood that a film produced by a new film studio founded and funded by Microsoft will result in a high ROI: * **Films should be released in the months of either May, June, and July.** With a small exception of November. However, this should only be considred if the production of a film runs behind since data shows that November still does not offer as high an ROI as the the May, June, and July do. * **The directors that should be most sought after to direct the first few films should be Steven Spielberg, Ridley Scott, or Clint Eastwood.** These directors are the most frequently used directors in the films in this dataset. This gives the high ROI they've been able to deliver with their films more validity and reliability of being capable of reproducing or improving on new projects like the ones a film studio run by Microsoft would likely fund. * **A film with the combination of genres of Action, Adventure, and Sci-Fi should be one of the first to be produced by a potentially new film studio founded by Microsoft.** The data suggest that genres of **Comedy and Mystery**, as well as **Adventure and Fantasy** are also likely candidates. However, in this dataset they did not fall under the top 10 most common genre combinations regardless of their high ROI.

2.8 Next Steps

Further analyses could offer additional insight and improve the likelyhood of a new film studio founded by Micorsoft to be a success would be: * **An analysis films based on individual genres.** This means looking into films that have a combination of genres, but including them into grouped data that have similar genres (i.e. a comendy and thriller movie would be grouped into data for both comedy and thriller films). To see how these relationships affect the ROI of certain aspects of a film. This type of investigation could use already available data. * **Checking how critic ratings are related to ROI** This would require additional data from different sources. * **Checking how movie ratings (i.e. films rated G to R ratings) affect the success of a film.**

This would help to determine the type of audiences a new film produced by Microsofts potentially new film studio should target. This would require additional data from different sources.

[ ]: