Movie_Industry_Analysis

April 24, 2022

1 Movie Industry Analysis

1.1 Overview

The purpose of this project is to study the movie industry and generate advice on how a new entrant to the industry might create and follow a strategy that is aligned with corporate interests and will assure success.

1.2 Business understanding

A large, multinational, technology company has engaged us to explore industry movie trends and identify what type of films are successful at the Box Office and to help them formulate a film strategy that is relevant and optimal for their business concerns.

1.3 Data understanding

This project pursues an analysis of objective data that ties film characteristics and attributes to a film's financial performance. The profitability of a movie was analyzed with respect to:
- The director
- Production budget
- Length of a movie

Two databases were chosen for analysis:
The Internet Movie Database (IMDb)
Provides information on movie title, year of release, running time (length), director, actors, etc.
The Numbers (TN)
Provides information on a movie's financial performance including release date , production budget and domestic and worldwide gross.

1.4 Data assumptions and preparation

We distinguished movies with the same title by the year in which it was released.
The exact duplicate entries were removed.
In calculation of average rating, many movies had too few. We only looked at movies that had more than 500 votes.

The TN database contained financial information and had the movie name in common with the IMDb. It had also a full release date as opposed to only the release year in

IMDb. We calculated the release year in the TN database as a separate column and then used the movie name and release year to join these two databases. Note: The TN database addition of the Release Year column was carried out in a separate Jupyter Notebook called "Exploration_tn_movie_budgets.ipynb" inthe folder called "Study". The enhanced dataset was exported as .csv file to teh data folder called "zippedData" and is called "NzModified_tn.movie_budgets.csv". This .csv file is imported in this notebook.

The resultant data set contains information on a little over 2,000 movies which is sufficient to identify trends.
Profit was calculated as:
- Profit = (Domestic_Gross + Worldwide_Gross) − Production_Budget
- Profit Percentage = Profit / Production_Cost * 100.
This Profit Percentage is a key measure which we use to judge success.

1.5 Initial exploration of databases

All exporation and initial analyses were carried out in Jupyter Notebooks in the folder called "Study".
**This notebook contains only code pertinent to data included in the final analysis and presentation.

[1]:
```
import pandas as pd
import sqlite3
from matplotlib import pyplot as plt
import seaborn as sns
```

1.6 Connect to im.db database.

Join movie_basics, directors, persons and movie_ratings tables and with Number of votes greater than 500.

[2]:
```
conn = sqlite3.connect('./zippedData/im.db')
```

[3]:
```
query = """
SELECT DISTINCT mb.primary_title AS "Movie Name", mb.start_year AS "Release␣
↪Year", mb.runtime_minutes AS "Length", mb.genres AS "Genres", p.primary_name␣
↪AS "Director", mr.averagerating AS "Avg Rating", mr.numvotes AS "Number of␣
↪Votes"

FROM movie_basics mb

    LEFT JOIN directors d
        ON mb.movie_id = d.movie_id
    LEFT JOIN persons p
        ON d.person_id = p.person_id
    LEFT JOIN movie_ratings mr
        ON mb.movie_id = mr.movie_id
```

```
WHERE mr.numvotes > 500

ORDER BY mr.averagerating DESC

"""
imdb_df = pd.read_sql_query(query, conn)
len(imdb_df)
```

[3]: 15595

[4]:
```
imdb_df.head()
```

[4]:
```
                          Movie Name  Release Year  Length  \
0  Once Upon a Time … in Hollywood          2019   159.0
1                          Eghantham          2018   125.0
2           Yeh Suhaagraat Impossible          2019    92.0
3                 Ananthu V/S Nusrath          2018   149.0
4                 Ekvtime: Man of God          2018   132.0

                     Genres              Director  Avg Rating  Number of Votes
0              Comedy,Drama     Quentin Tarantino         9.7             5600
1                     Drama       Arsel Arumugam         9.7              639
2                    Comedy       Abhinav Thakur         9.6              624
3      Comedy,Drama,Family    Sudheer Shanbhogue         9.6              808
4  Biography,Drama,History  Nikoloz Khomasuridze         9.6             2604
```

1.7 Read in csv file derived from tn_movie database

As mentioned above, the following csv file was created in a separate Jupyter Notebook called "Exploration_tn_movie_budgets". It has an added column containing just the release year as opposed to the full date.

[5]:
```
tn_df = pd.read_csv('./zippedData/NzModified_tn.movie_budgets.csv')
len(tn_df)
```

[5]: 5782

[6]:
```
tn_movie_list = tn_df['movie']
tn_df['movie'][0]
```

[6]: 'Avatar'

[7]:
```
imdb_df.keys()
```

[7]:
```
Index(['Movie Name', 'Release Year', 'Length', 'Genres', 'Director',
       'Avg Rating', 'Number of Votes'],
      dtype='object')
```

```
[8]: imdb_df['Movie Name'].head()
```

```
[8]: 0    Once Upon a Time … in Hollywood
     1                         Eghantham
     2             Yeh Suhaagraat Impossible
     3                  Ananthu V/S Nusrath
     4                  Ekvtime: Man of God
     Name: Movie Name, dtype: object
```

1.8 Merge imdb and tn dataframes with inner join on movie names

```
[9]: merged_df = imdb_df.merge(tn_df, left_on="Movie Name", right_on="movie",␣
      ↪how='inner')
```

```
[10]: merged_df.head()
```

```
[10]:     Movie Name  Release Year  Length                      Genres  \
     0  Frankenstein          2011   130.0                       Drama
     1  Frankenstein          2015    89.0   Horror,Sci-Fi,Thriller
     2     Inception          2010   148.0  Action,Adventure,Sci-Fi
     3    Coriolanus          2014   192.0         Drama,History,War
     4    Coriolanus          2011   123.0        Drama,Thriller,War

                Director  Avg Rating  Number of Votes  Unnamed: 0  id  \
     0        Danny Boyle         9.0             1832        1302   3
     1       Bernard Rose         5.1             2089        1302   3
     2  Christopher Nolan         8.8          1841066         137  38
     3    Tim Van Someren         8.7             1347        3698  99
     4      Ralph Fiennes         6.1            29654        3698  99

       release_date         movie production_budget domestic_gross  \
     0   Nov 4, 1994  Frankenstein       $45,000,000    $22,006,296
     1   Nov 4, 1994  Frankenstein       $45,000,000    $22,006,296
     2  Jul 16, 2010     Inception      $160,000,000   $292,576,195
     3  Jan 20, 2012    Coriolanus       $10,000,000       $749,641
     4  Jan 20, 2012    Coriolanus       $10,000,000       $749,641

       worldwide_gross  year
     0    $112,006,296  1994
     1    $112,006,296  1994
     2    $835,524,642  2010
     3      $2,179,623  2012
     4      $2,179,623  2012
```

```
[11]: merged_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2112 entries, 0 to 2111
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Movie Name         2112 non-null   object
 1   Release Year       2112 non-null   int64
 2   Length             2111 non-null   float64
 3   Genres             2112 non-null   object
 4   Director           2112 non-null   object
 5   Avg Rating         2112 non-null   float64
 6   Number of Votes    2112 non-null   int64
 7   Unnamed: 0         2112 non-null   int64
 8   id                 2112 non-null   int64
 9   release_date       2112 non-null   object
 10  movie              2112 non-null   object
 11  production_budget  2112 non-null   object
 12  domestic_gross     2112 non-null   object
 13  worldwide_gross    2112 non-null   object
 14  year               2112 non-null   int64
dtypes: float64(2), int64(5), object(8)
memory usage: 264.0+ KB
```

This is a significantly reduced data set compared to the imdb database but still has sufficient number of records to derive trends.

Sort merged data by "domestic_gross' in descending order

[12]:
```python
merged_df.sort_values(by=['domestic_gross'], ascending=False)
```

[12]:
```
                Movie Name  Release Year  Length                      Genres  \
385      Christopher Robin          2018   104.0  Adventure,Animation,Comedy
1401              Hercules          2014    98.0    Action,Adventure,Fantasy
1022     Olympus Has Fallen         2013   119.0             Action,Thriller
1521      The Green Hornet          2011   119.0          Action,Comedy,Crime
1179            Date Night          2010    88.0         Comedy,Crime,Romance
…                        …             …       …                           …
1932              The Veil          2016    93.0                      Horror
1933              The Veil          2017    93.0     Action,Adventure,Sci-Fi
1691              Survivor          2015    96.0       Action,Crime,Thriller
1935           Dawn Patrol          2014    88.0              Drama,Thriller
1602   Queen of the Desert          2015   128.0   Adventure,Biography,Drama

               Director  Avg Rating  Number of Votes  Unnamed: 0  id  \
385        Marc Forster         7.3            52737         642  43
1401       Brett Ratner         6.0           137287         707   8
1022      Antoine Fuqua         6.5           235443         708   9
1521      Michel Gondry         5.8           148622         328  29
```

```
1179         Shawn Levy         6.3          144683          998   99
…                  …          …              …         …   ..
1932          Phil Joanou         4.8            6895         4563   64
1933    Brent Ryan Green         3.5            1236         4563   64
1691       James McTeigue         5.6           28614         2710   11
1935    Daniel Petrie Jr.         4.8             615         4631   32
1602        Werner Herzog         5.7            8529         1621   22

      release_date                movie production_budget domestic_gross  \
385    Aug 3, 2018    Christopher Robin       $75,000,000    $99,215,042
1401  Jun 13, 1997             Hercules       $70,000,000    $99,112,101
1022  Mar 22, 2013    Olympus Has Fallen       $70,000,000    $98,927,592
1521  Jan 14, 2011      The Green Hornet      $110,000,000    $98,780,042
1179   Apr 9, 2010           Date Night       $55,000,000    $98,711,404
…              …                    …                 …              …
1932  Dec 31, 2015             The Veil        $4,000,000            $0
1933  Dec 31, 2015             The Veil        $4,000,000            $0
1691  May 29, 2015             Survivor       $20,000,000            $0
1935   Jun 5, 2015          Dawn Patrol        $3,500,000            $0
1602  Apr 14, 2017  Queen of the Desert       $36,000,000            $0

     worldwide_gross  year
385     $197,504,758  2018
1401    $250,700,000  1997
1022    $172,878,928  2013
1521    $229,155,503  2011
1179    $152,269,033  2010
…                …     …
1932              $0  2015
1933              $0  2015
1691      $1,703,281  2015
1935              $0  2015
1602      $1,578,543  2017

[2112 rows x 15 columns]
```

Gross and budget columns contain string values. Convert them to floats.

[13]:
```python
merged_df['float_production_budget'] = merged_df['production_budget'].
↪replace('[\$,]', '', regex=True).astype(float)
merged_df['float_domestic_gross'] = merged_df['domestic_gross'].
↪replace('[\$,]', '', regex=True).astype(float)
merged_df['float_worldwide_gross'] = merged_df['worldwide_gross'].
↪replace('[\$,]', '', regex=True).astype(float)
```

Calculate profit percentage and create 'profit percent' column
Profit = (Domestic_Gross + Worldwide_Gross) – Production_Budget
Profit Percentage = Profit / Production_Cost * 100

```
[14]: merged_df['profit percent'] = (merged_df['float_domestic_gross'] +␣
      ↪merged_df['float_worldwide_gross'] - merged_df['float_production_budget']) /␣
      ↪merged_df['float_production_budget'] * 100
```

```
[15]: merged_df.head()
```

```
[15]:      Movie Name  Release Year  Length                    Genres  \
      0  Frankenstein          2011   130.0                     Drama
      1  Frankenstein          2015    89.0   Horror,Sci-Fi,Thriller
      2     Inception          2010   148.0  Action,Adventure,Sci-Fi
      3    Coriolanus          2014   192.0        Drama,History,War
      4    Coriolanus          2011   123.0       Drama,Thriller,War

                  Director  Avg Rating  Number of Votes  Unnamed: 0  id  \
      0       Danny Boyle         9.0             1832        1302   3
      1      Bernard Rose         5.1             2089        1302   3
      2  Christopher Nolan        8.8          1841066         137  38
      3    Tim Van Someren        8.7             1347        3698  99
      4     Ralph Fiennes         6.1            29654        3698  99

        release_date         movie production_budget domestic_gross  \
      0   Nov 4, 1994  Frankenstein       $45,000,000    $22,006,296
      1   Nov 4, 1994  Frankenstein       $45,000,000    $22,006,296
      2  Jul 16, 2010     Inception      $160,000,000   $292,576,195
      3  Jan 20, 2012    Coriolanus       $10,000,000       $749,641
      4  Jan 20, 2012    Coriolanus       $10,000,000       $749,641

        worldwide_gross  year  float_production_budget  float_domestic_gross  \
      0    $112,006,296  1994               45000000.0            22006296.0
      1    $112,006,296  1994               45000000.0            22006296.0
      2    $835,524,642  2010              160000000.0           292576195.0
      3      $2,179,623  2012               10000000.0              749641.0
      4      $2,179,623  2012               10000000.0              749641.0

         float_worldwide_gross  profit percent
      0            112006296.0      197.805760
      1            112006296.0      197.805760
      2            835524642.0      605.063023
      3              2179623.0      -70.707360
      4              2179623.0      -70.707360
```

1.9 Create new df of merged_df grouped by "Director"

Create a new series (column) of the count of movies directed by each director. Calculate the means of each element in the newly grouped df called director_means. Add the count series to the director_means df.

[16]:
```
director_count = merged_df.groupby(by='Director')['Director'].count()
director_means = merged_df.groupby(by='Director')[['Avg Rating', 'Number of␣
↪Votes', 'float_production_budget', 'float_domestic_gross',␣
↪'float_worldwide_gross', 'profit percent']].mean()

director_means['count'] = director_count
director_means
```

[16]:
```
                   Avg Rating  Number of Votes  float_production_budget  \
Director
Aaron Hann                6.0          30645.0                2000000.0
Aaron Seltzer             3.4          43984.0               20000000.0
Aaron T. Wells            3.5           2230.0                 500000.0
Abby Kohn                 5.4          39936.0               32000000.0
Abdolreza Kahani          7.0            903.0                4000000.0
…                           …                …                        …
Zackary Adler             5.0           1723.0                2500000.0
Zak Forsman               5.2            846.0                  50000.0
Zal Batmanglij            6.7          33095.5                3317500.0
Zhigang Yang              7.1            581.0               70000000.0
Zsófia Szilágyi           7.2            501.0               15000000.0

                   float_domestic_gross  float_worldwide_gross  profit percent  \
Director
Aaron Hann                      10024.0                10024.0      -98.997600
Aaron Seltzer                36661504.0             81424988.0      490.432460
Aaron T. Wells                      0.0                    0.0     -100.000000
Abby Kohn                    48795601.0             91553797.0      338.591869
Abdolreza Kahani                    0.0                63180.0      -98.420500
…                                     …                      …               …
Zackary Adler                       0.0                    0.0     -100.000000
Zak Forsman                         0.0                    0.0     -100.000000
Zal Batmanglij                1341332.0              1728702.0      250.960751
Zhigang Yang                 55011732.0             94973540.0      114.264674
Zsófia Szilágyi              13843771.0             59168692.0      386.749753

                   count
Director
Aaron Hann             1
Aaron Seltzer          1
Aaron T. Wells         1
Abby Kohn              1
Abdolreza Kahani       1
…                      …
Zackary Adler          1
Zak Forsman            1
Zal Batmanglij         2
```

```
Zhigang Yang         1
Zsófia Szilágyi      1

[1426 rows x 7 columns]
```

1.10 Sort of Directors by Count of Movie Releases

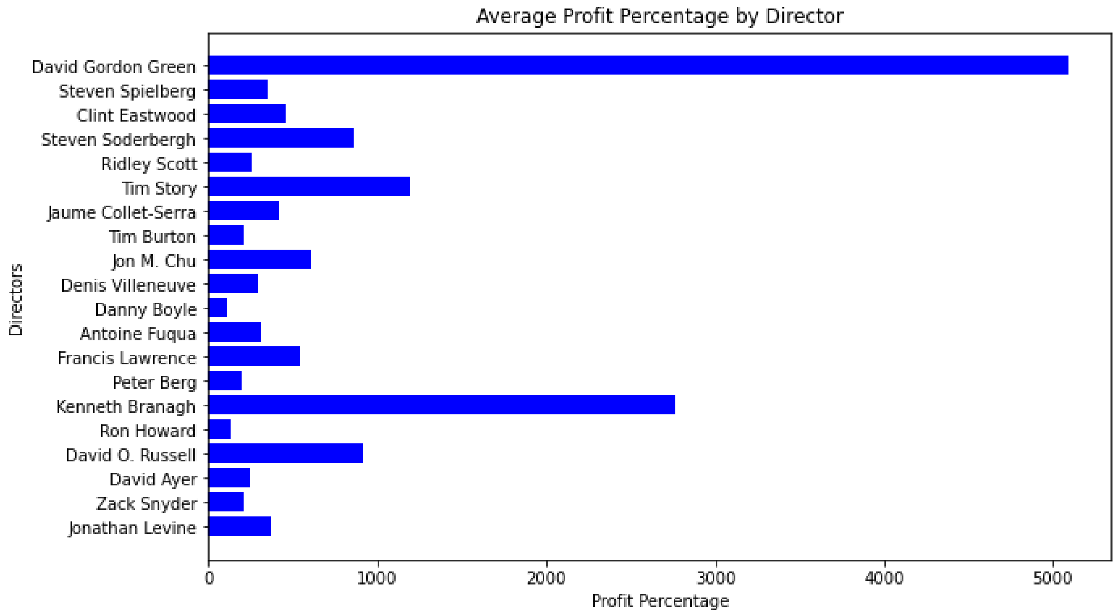
Sort director_means df by count to list in order of most prolific directors. Only top 20 directors in sorted list will be used in the profit percentage and busget charts below.

[17]:
```
director_sort_by_count_df = director_means.sort_values(by=['count', 'Avg␣
↪Rating','profit percent','float_production_budget'], ascending=False)
```

[18]:
```
director_sort_by_count_df.index
```

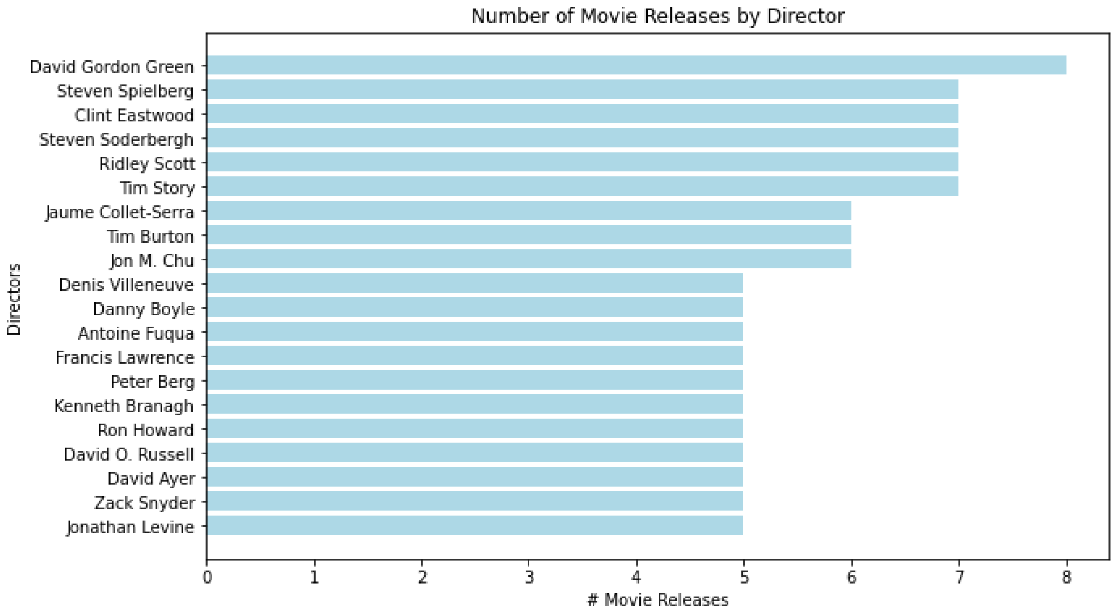
[18]:
```
Index(['David Gordon Green', 'Steven Spielberg', 'Clint Eastwood',
       'Steven Soderbergh', 'Ridley Scott', 'Tim Story', 'Jaume Collet-Serra',
       'Tim Burton', 'Jon M. Chu', 'Denis Villeneuve',
       …
       'Matthew R. Anderson', 'Jamie Buckner', 'Timothy Woodward Jr.',
       'Glenn Ciano', 'David Winning', 'David DeCoteau', 'Kaizad Gustad',
       'Frédéric Auburtin', 'Justin Price', 'Lawrence Kasanoff'],
      dtype='object', name='Director', length=1426)
```

[19]:
```
fig, ax = plt.subplots(figsize=(10,6))

ax.barh(director_sort_by_count_df.index[0:20],␣
↪director_sort_by_count_df['profit percent'].head(20), color='blue')
ax.invert_yaxis()
ax.set_title("Average Profit Percentage by Director")
ax.set_xlabel('Profit Percentage')
ax.set_ylabel('Directors')
plt.savefig('./Images//Profit_Percentage_by_Director_for_High_Movie_Count.
↪png',bbox_inches='tight')
```



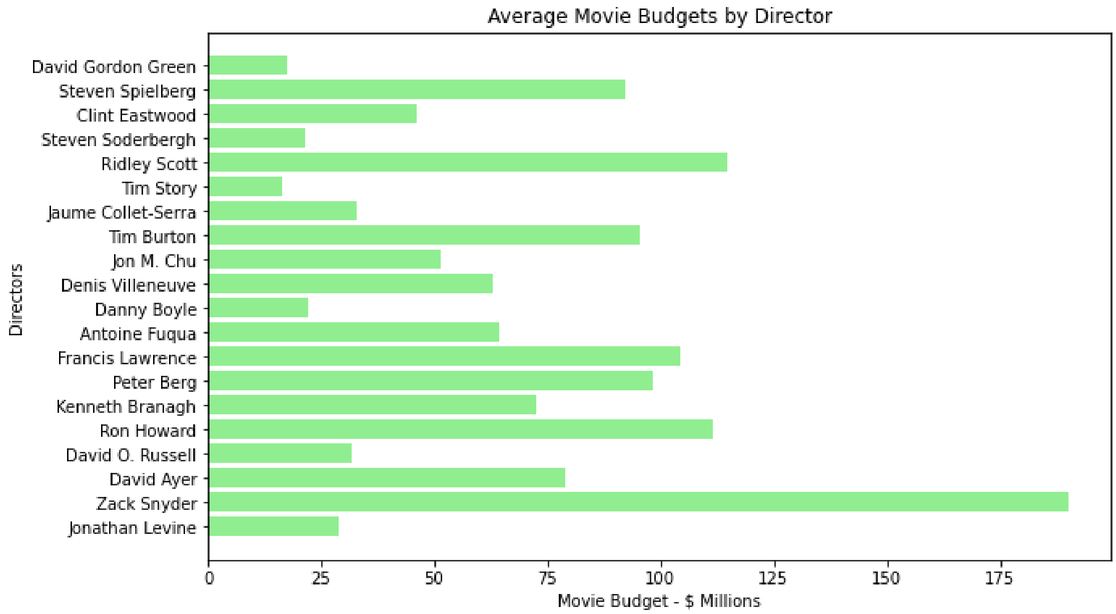
This chart shows that of the top 20 directors by movie count, all of them were very profitable exceeding 100% and couple in the range of thousands of percent. Clearly, there is a correlation between directors with a track record and their profitability.

[20]:
```
fig, ax = plt.subplots(figsize=(10,6))

ax.barh(director_sort_by_count_df.index[0:20],␣
 ↪director_sort_by_count_df['count'].head(20), color='lightblue')
ax.invert_yaxis()
ax.set_title("Number of Movie Releases by Director")
ax.set_xlabel('# Movie Releases')
ax.set_ylabel('Directors')
plt.savefig('./Images/High_Count_of_Movie_Releases_by_Director.
 ↪png',bbox_inches='tight')
```



This chart was generated to show the track record of the top directors. As is apparent, they all have made multiple movies.

[21]:
```
fig, ax = plt.subplots(figsize=(10,6))

ax.barh(director_sort_by_count_df.index[0:20],
 ↪director_sort_by_count_df['float_production_budget'].head(20)/1000000,
 ↪color='lightgreen')
ax.invert_yaxis()
ax.set_title("Average Movie Budgets by Director")
ax.set_xlabel('Movie Budget - $ Millions')
ax.set_ylabel('Directors')
plt.savefig('./Images/
 ↪Movie_Budgets_of_Directors_with_High_Count_of_Movie_Releases.
 ↪png',bbox_inches='tight')
```



The previous charts raised the question as to why some directors can be phenominally profitable and others who are known for making well received movies have not earned as high a profit. This chart along with some general knowledge about the directors provides a clue. David Gordon Green is known for making horror movies. Horror movies do not require a high budget because they do not generally require elaborate special effects and since much happens at night or atleast in the dark, the movie sets do not need to be as detailed or intricate. The likes of Steven Spielberg and Ridley Scott are known for Sci Fi movies which definitely require elaborate special effects and the creation of movie sets reflecting the otherworldly environments that are central to the movies. ALmost by definition a Sci Fi has to require a higher budget.
This project did not focus on genre but this data is implying that profitability by genre should be separately investigated.

1.11 Sort of Directors by Profitability

Take the previously sorted director_means df and sort it by 'profit percent'. The chart will show most profitable directors regardless of number of movies made.

[22]:
```
director_sort_by_profit_df = director_means.sort_values(by=['profit percent'],␣
↪ascending=False)
director_sort_by_profit_df.head(20)
```

[22]:
```
                  Avg Rating  Number of Votes  float_production_budget  \
Director
Travis Cluff        4.200000        17763.000                 100000.0
Chris Lofing        4.200000        17763.000                 100000.0
Brandon Camp        6.400000         2779.000                 500000.0
```

Levan Gabriadze	5.600000	62043.000	1000000.0
Nate Parker	6.400000	18442.000	5055000.0
Tod Williams	5.700000	93122.000	3000000.0
Jamie Buckner	2.900000	557.000	5000000.0
William Brent Bell	5.100000	51239.500	5500000.0
Bradley Parker	5.000000	60304.000	1000000.0
Franck Khalfoun	6.100000	32534.000	350000.0
Jordan Peele	7.400000	251492.500	12500000.0
David Gordon Green	6.312500	63319.125	17290625.0
Fabrice Gobert	6.400000	971.000	5000000.0
Alex Kendrick	6.750000	14651.000	2500000.0
Lynn Shelton	6.700000	24780.000	120000.0
Jessica Cameron	4.800000	554.000	3500000.0
James Wan	7.175000	312458.000	92875000.0
Dan Trachtenberg	7.200000	260383.000	5000000.0
Josh Boone	7.700000	315135.000	12000000.0
Henry Joost	5.633333	82293.000	10000000.0

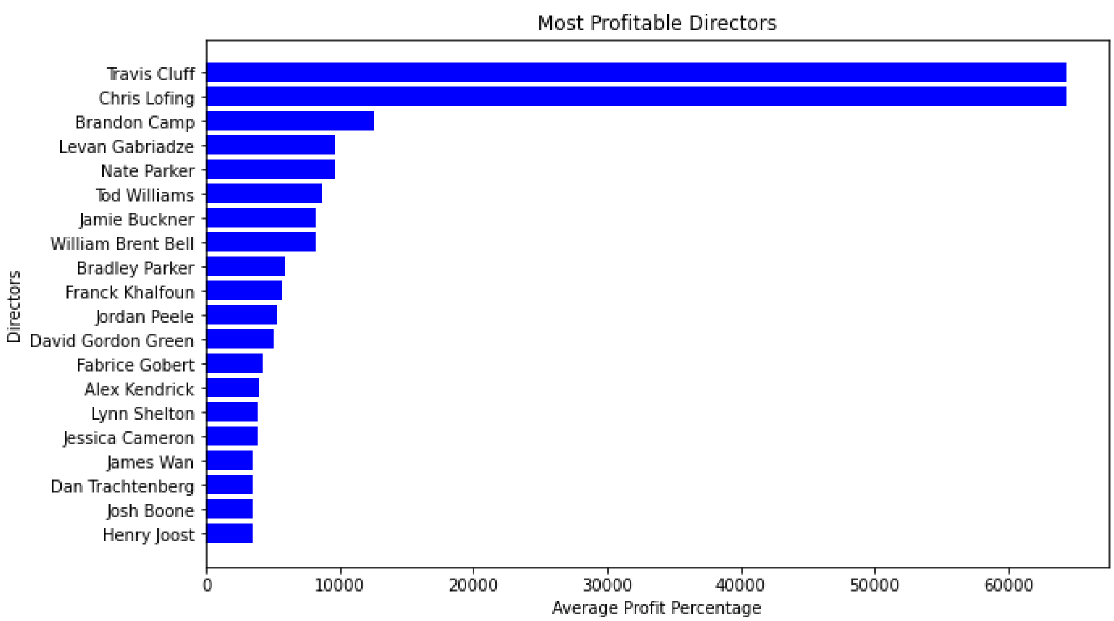
	float_domestic_gross	float_worldwide_gross \
Director		
Travis Cluff	2.276441e+07	4.165647e+07
Chris Lofing	2.276441e+07	4.165647e+07
Brandon Camp	3.155956e+07	3.155956e+07
Levan Gabriadze	3.278964e+07	6.436420e+07
Nate Parker	1.293078e+07	1.394551e+07
Tod Williams	8.475291e+07	1.775120e+08
Jamie Buckner	1.381416e+08	2.789648e+08
William Brent Bell	4.454125e+07	8.499022e+07
Bradley Parker	1.811964e+07	4.241172e+07
Franck Khalfoun	1.000000e+07	1.000000e+07
Jordan Peele	1.755238e+08	2.547891e+08
David Gordon Green	4.018337e+07	5.965546e+07
Fabrice Gobert	6.726884e+07	1.488065e+08
Alex Kendrick	5.115617e+07	5.458056e+07
Lynn Shelton	1.597486e+06	3.090593e+06
Jessica Cameron	4.141102e+07	9.512734e+07
James Wan	2.198695e+08	7.708721e+08
Dan Trachtenberg	7.208300e+07	1.082864e+08
Josh Boone	1.248724e+08	3.071668e+08
Henry Joost	6.550426e+07	1.401700e+08

	profit percent	count
Director		
Travis Cluff	64320.884000	1
Chris Lofing	64320.884000	1
Brandon Camp	12523.824000	1
Levan Gabriadze	9615.384300	1

```
Nate Parker              9609.217430       2
Tod Williams             8642.164633       1
Jamie Buckner            8242.127820       1
William Brent Bell       8171.324290       2
Bradley Parker           5953.136100       1
Franck Khalfoun          5614.285714       1
Jordan Peele             5287.129260       2
David Gordon Green       5088.186541       8
Fabrice Gobert           4221.506900       1
Alex Kendrick            4005.458558       2
Lynn Shelton             3806.732500       1
Jessica Cameron          3801.095971       1
James Wan                3511.753553       4
Dan Trachtenberg         3507.388420       1
Josh Boone               3500.326533       1
Henry Joost              3467.306370       3
```

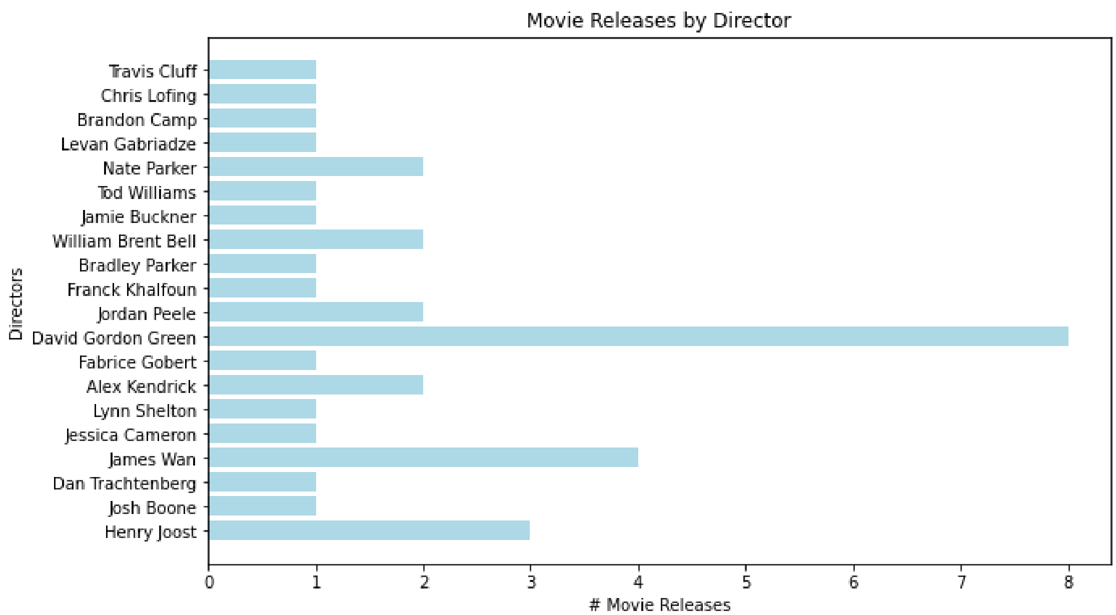
[23]:

```
fig, ax = plt.subplots(figsize=(10,6))

ax.barh(director_sort_by_profit_df.index[0:20],␣
 ↪director_sort_by_profit_df['profit percent'].head(20), color='blue')
ax.invert_yaxis()
ax.set_title("Most Profitable Directors")
ax.set_xlabel('Average Profit Percentage')
ax.set_ylabel('Directors')
plt.savefig('./Images/Most_Profitable_Directors.png')
```



Just looking at profitability without sorting by movie counts, there are many directors that are showing profitability in the thousands of percent, some well over 10,000 percent. However, the number of movies that they have made is limited. Let's also generate the movie count for these most profitable directors. See the next chart below.

[24]:
```
fig, ax = plt.subplots(figsize=(10,6))

ax.barh(director_sort_by_profit_df.index[0:20],␣
↪director_sort_by_profit_df['count'].head(20), color='lightblue')
ax.invert_yaxis()
ax.set_title("Movie Releases by Director")
ax.set_xlabel('# Movie Releases')
ax.set_ylabel('Directors')
plt.savefig('./Images/Count_of_Movie_Releases_by_Most_Profitable_Directors.
↪png',bbox_inches='tight')
```



So, many of these highly profitable directors have only made one movie. Perhaps, there was some political, societal or environmental anomaly that brought focus on their single movie and it generated a huge percentage of profitability. A single point of data does not indicate a trend. They may continue to make profitable movies or they may be relegated in history as One-Hit Wonders.

1.12 Movie Length vs Profitability

Sort of original merged_df by Movie Length

This analysis was initiated with the question as to whether other factors like the length of the movie may correlate with profit or lack thereof. This following sorted data set

will be used to look at the effect of movie length on profitability.

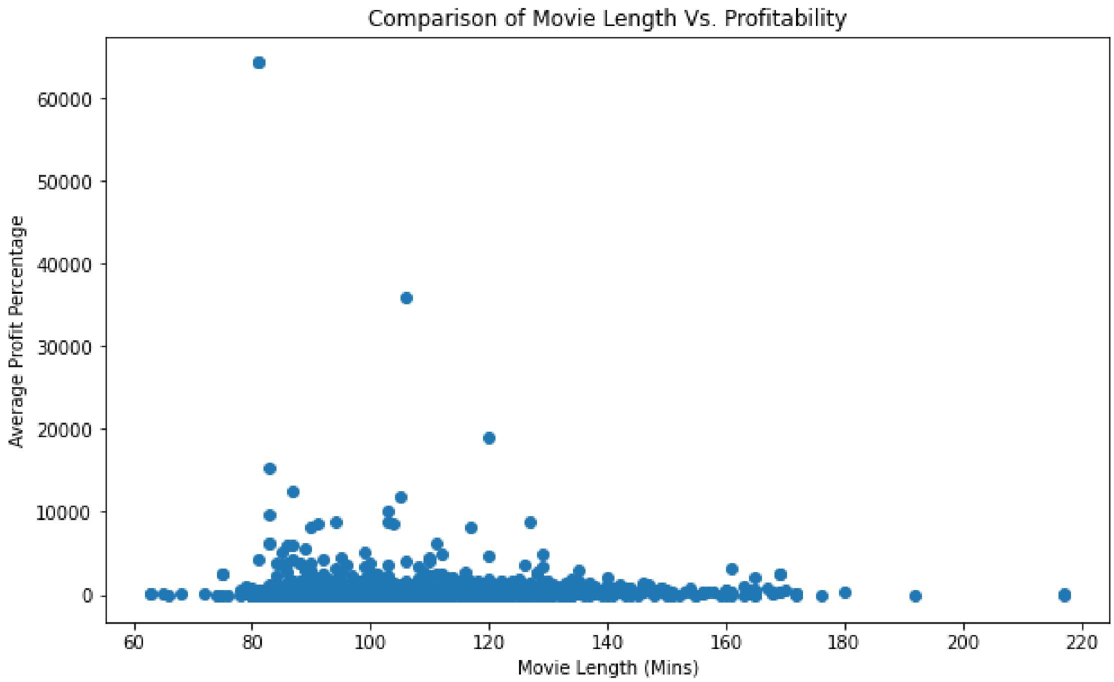
[25]:
```
merged_sort_by_Length = merged_df.sort_values(by=['Length'], ascending=False)
merged_sort_by_Length
```

[25]:
```
                    Movie Name  Release Year  Length  \
5                       Hamlet          2015   217.0
6                       Hamlet          2015   217.0
3                   Coriolanus          2014   192.0
30     The Wolf of Wall Street          2013   180.0
762            Jab Tak Hai Jaan         2012   176.0
…                          …             …       …
1802                   Aroused          2013    66.0
714                Unstoppable          2013    65.0
414            Winnie the Pooh          2011    63.0
415            Winnie the Pooh          2011    63.0
1889               The Bachelor         2016     NaN

                           Genres             Director  Avg Rating  \
5                           Drama          Robin Lough         8.6
6                           Drama          Robin Lough         8.6
3               Drama,History,War     Tim Van Someren         8.7
30          Biography,Crime,Drama      Martin Scorsese         8.2
762                 Drama,Romance          Yash Chopra         6.8
…                             …                    …          …
1802                  Documentary     Deborah Anderson         5.3
714                   Documentary         Darren Doane         4.3
414   Adventure,Animation,Comedy  Stephen J. Anderson         7.2
415   Adventure,Animation,Comedy             Don Hall         7.2
1889               Comedy,Romance  Antonis Sotiropoulos         5.1

      Number of Votes  Unnamed: 0  id  release_date                    movie  \
5                1587        2831  32  Dec 25, 1996                   Hamlet
6                1587        4933  34  May 12, 2000                   Hamlet
3                1347        3698  99  Jan 20, 2012               Coriolanus
30            1035358         375  76  Dec 25, 2013  The Wolf of Wall Street
762             48364        3763  64  Nov 13, 2012         Jab Tak Hai Jaan
…                  …           …  ..           …                       …
1802              596        5663  64   May 3, 2013                  Aroused
714               551         418  19  Nov 12, 2010              Unstoppable
414             19605        1938  39  Jul 15, 2011          Winnie the Pooh
415             19605        1938  39  Jul 15, 2011          Winnie the Pooh
1889              895        2473  74   Nov 5, 1999              The Bachelor

     production_budget domestic_gross worldwide_gross  year  \
5          $18,000,000     $4,501,094      $7,129,670  1996
6           $2,000,000     $1,577,287      $2,419,669  2000
```

```
3          $10,000,000       $749,641     $2,179,623  2012
30        $100,000,000   $116,900,694   $389,870,414  2013
762         $9,200,000     $3,047,539     $5,806,666  2012
…                  …              …              …     …
1802          $150,000             $0             $0  2013
714        $95,000,000    $81,562,942   $165,720,921  2010
414        $30,000,000    $26,692,846    $50,145,607  2011
415        $30,000,000    $26,692,846    $50,145,607  2011
1889       $21,000,000    $21,731,001    $36,882,378  1999

      float_production_budget  float_domestic_gross  float_worldwide_gross  \
5                  18000000.0             4501094.0              7129670.0
6                   2000000.0             1577287.0              2419669.0
3                  10000000.0              749641.0              2179623.0
30                100000000.0           116900694.0            389870414.0
762                 9200000.0             3047539.0              5806666.0
…                         …                     …                      …
1802                 150000.0                   0.0                    0.0
714                95000000.0            81562942.0            165720921.0
414                30000000.0            26692846.0             50145607.0
415                30000000.0            26692846.0             50145607.0
1889               21000000.0            21731001.0             36882378.0

      profit percent
5         -35.384644
6          99.847800
3         -70.707360
30        406.771108
762        -3.758641
…                …
1802     -100.000000
714       160.298803
414       156.128177
415       156.128177
1889      179.111329

[2112 rows x 19 columns]
```

[26]:

```
fig, ax = plt.subplots(figsize=(10,6))

ax.scatter(x=merged_sort_by_Length['Length'],y=merged_sort_by_Length['profit␣
↪percent'])
ax.set_title("Comparison of Movie Length Vs. Profitability")
ax.set_xlabel('Movie Length (Mins)')
ax.set_ylabel('Average Profit Percentage')
plt.savefig('./Images/Movie_Length_Vs_Profitability.png',bbox_inches='tight')
```



The top few high profit movies (outliers) are skewing the results. Let's only look at movies under 10K% profit by not including the first 10 movies in the sorted list.

[27]:
```
merged_sort_by_Profit_Percent_under_10k = merged_df.sort_values(by=['profit␣
↪percent'], ascending=False)[10:-1]
merged_sort_by_Profit_Percent_under_10k.head(20)
```

[27]:
```
                Movie Name  Release Year  Length                      Genres  \
252                   Home          2016   103.0                       Drama
255                   Home          2015    94.0  Adventure,Animation,Comedy
1593  Paranormal Activity 2         2010    91.0                      Horror
152                Get Out          2017   104.0     Horror,Mystery,Thriller
381                  Split          2016    90.0       Comedy,Romance,Sport
380                  Split          2016   117.0             Horror,Thriller
1525  Paranormal Activity 3         2011    83.0     Horror,Mystery,Thriller
1526  Paranormal Activity 3         2011    83.0     Horror,Mystery,Thriller
268              Moonlight          2016   111.0                       Drama
1662       The Last Exorcism         2010    87.0      Drama,Horror,Thriller
1892       Chernobyl Diaries         2012    86.0     Horror,Mystery,Thriller
1354                 Maniac          2012    89.0             Horror,Thriller
1765              Annabelle          2014    99.0     Horror,Mystery,Thriller
1586              The Purge          2013    85.0             Horror,Thriller
436     Beauty and the Beast         2014   112.0      Drama,Fantasy,Romance
434     Beauty and the Beast         2017   129.0      Family,Fantasy,Musical
974               War Room          2015   120.0                       Drama
```

904	You're Next	2011	95.0	Action,Comedy,Horror
750	Sinister	2012	110.0	Horror,Mystery,Thriller
729	A Ghost Story	2017	92.0	Drama,Fantasy,Romance

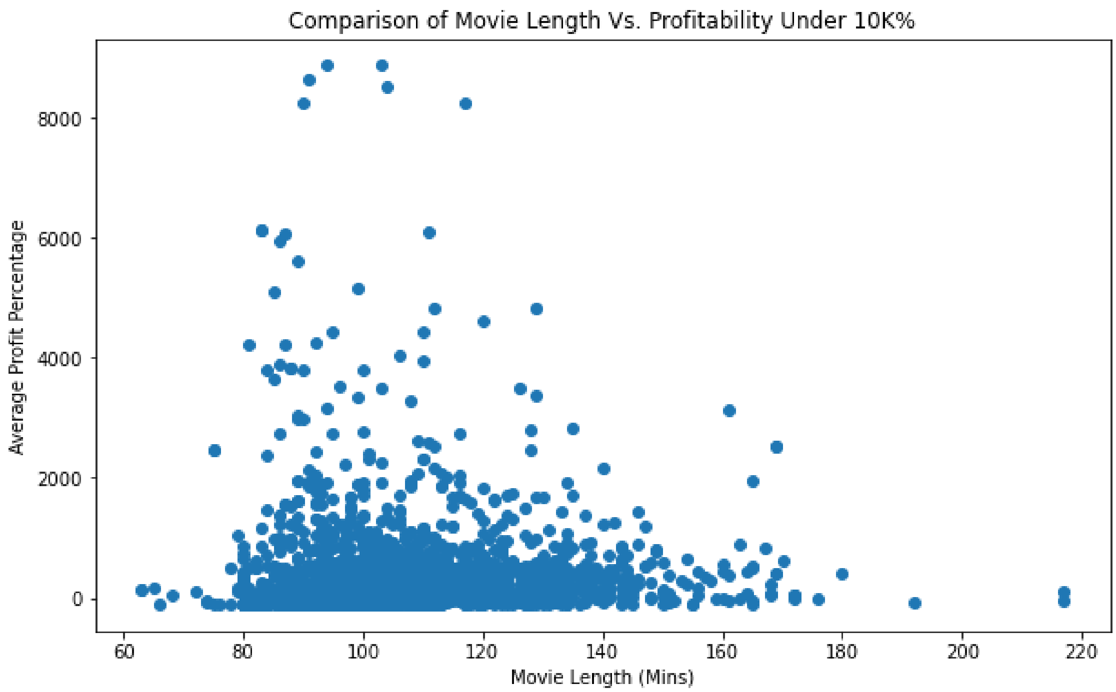
	Director	Avg Rating	Number of Votes	Unnamed: 0	id \
252	Fien Troch	7.2	811	5459	60
255	Tim Johnson	6.6	85831	5459	60
1593	Tod Williams	5.7	93122	4664	65
152	Jordan Peele	7.7	400474	4248	49
381	Jamie Buckner	2.9	557	4249	50
380	M. Night Shyamalan	7.3	358543	4249	50
1525	Henry Joost	5.8	85689	4250	51
1526	Ariel Schulman	5.8	85689	4250	51
268	Barry Jenkins	7.4	227964	5063	64
1662	Daniel Stamm	5.6	45815	5014	15
1892	Bradley Parker	5.0	60304	5217	18
1354	Franck Khalfoun	6.1	32534	5527	28
1765	John R. Leonetti	5.4	122039	4083	84
1586	James DeMonaco	5.7	183549	4666	67
436	Christophe Gans	6.4	18100	2485	86
434	Bill Condon	7.2	238325	2485	86
974	Alex Kendrick	6.5	11716	4665	66
904	Adam Wingard	6.6	79451	5216	17
750	Scott Derrickson	6.8	198345	4668	69
729	David Lowery	6.8	46280	5685	86

	release_date	movie	production_budget	domestic_gross \
252	Apr 23, 2009	Home	$500,000	$15,433
255	Apr 23, 2009	Home	$500,000	$15,433
1593	Oct 20, 2010	Paranormal Activity 2	$3,000,000	$84,752,907
152	Feb 24, 2017	Get Out	$5,000,000	$176,040,665
381	Jan 20, 2017	Split	$5,000,000	$138,141,585
380	Jan 20, 2017	Split	$5,000,000	$138,141,585
1525	Oct 21, 2011	Paranormal Activity 3	$5,000,000	$104,028,807
1526	Oct 21, 2011	Paranormal Activity 3	$5,000,000	$104,028,807
268	Oct 21, 2016	Moonlight	$1,500,000	$27,854,931
1662	Aug 27, 2010	The Last Exorcism	$1,800,000	$41,034,350
1892	May 25, 2012	Chernobyl Diaries	$1,000,000	$18,119,640
1354	Jan 1, 1980	Maniac	$350,000	$10,000,000
1765	Oct 3, 2014	Annabelle	$6,500,000	$84,273,813
1586	Jun 7, 2013	The Purge	$3,000,000	$64,473,115
436	Nov 13, 1991	Beauty and the Beast	$20,000,000	$376,057,266
434	Nov 13, 1991	Beauty and the Beast	$20,000,000	$376,057,266
974	Aug 28, 2015	War Room	$3,000,000	$67,790,117
904	Aug 23, 2013	You're Next	$1,000,000	$18,494,006
750	Oct 12, 2012	Sinister	$3,000,000	$48,086,903
729	Jul 7, 2017	A Ghost Story	$100,000	$1,594,798

```
     worldwide_gross  year  float_production_budget  float_domestic_gross  \
252      $44,793,168  2009                 500000.0               15433.0
255      $44,793,168  2009                 500000.0               15433.0
1593    $177,512,032  2010                3000000.0            84752907.0
152     $255,367,951  2017                5000000.0           176040665.0
381     $278,964,806  2017                5000000.0           138141585.0
380     $278,964,806  2017                5000000.0           138141585.0
1525    $207,039,844  2011                5000000.0           104028807.0
1526    $207,039,844  2011                5000000.0           104028807.0
268      $65,245,512  2016                1500000.0            27854931.0
1662     $70,165,900  2010                1800000.0            41034350.0
1892     $42,411,721  2012                1000000.0            18119640.0
1354     $10,000,000  1980                 350000.0            10000000.0
1765    $256,862,920  2014                6500000.0            84273813.0
1586     $91,266,581  2013                3000000.0            64473115.0
436     $608,431,132  1991               20000000.0           376057266.0
434     $608,431,132  1991               20000000.0           376057266.0
974      $73,975,239  2015                3000000.0            67790117.0
904      $26,887,177  2013                1000000.0            18494006.0
750      $87,727,807  2012                3000000.0            48086903.0
729       $2,769,782  2017                 100000.0             1594798.0

      float_worldwide_gross  profit percent
252              44793168.0     8861.720200
255              44793168.0     8861.720200
1593            177512032.0     8642.164633
152             255367951.0     8528.172320
381             278964806.0     8242.127820
380             278964806.0     8242.127820
1525            207039844.0     6121.373020
1526            207039844.0     6121.373020
268              65245512.0     6106.696200
1662             70165900.0     6077.791667
1892             42411721.0     5953.136100
1354             10000000.0     5614.285714
1765            256862920.0     5148.257431
1586             91266581.0     5091.323200
436             608431132.0     4822.441990
434             608431132.0     4822.441990
974              73975239.0     4625.511867
904              26887177.0     4438.118300
750              87727807.0     4427.157000
729               2769782.0     4264.580000
```

[28]:
```
fig, ax = plt.subplots(figsize=(10,6))
```

```
ax.
↪scatter(x=merged_sort_by_Profit_Percent_under_10k['Length'],y=merged_sort_by_Profit_Percent_
↪percent'])
ax.set_title("Comparison of Movie Length Vs. Profitability Under 10K%")
ax.set_xlabel('Movie Length (Mins)')
ax.set_ylabel('Average Profit Percentage')
plt.savefig('./Images/Movie_Length_Vs_Profitability_Under_10K.
↪png',bbox_inches='tight')
```



The multithousand percent profit movies are still skewing the results. Let's only look at movies under 2K% profit by not including the first 81 movies in the sorted list. This will get a better view of the remaining 2000+ movies.

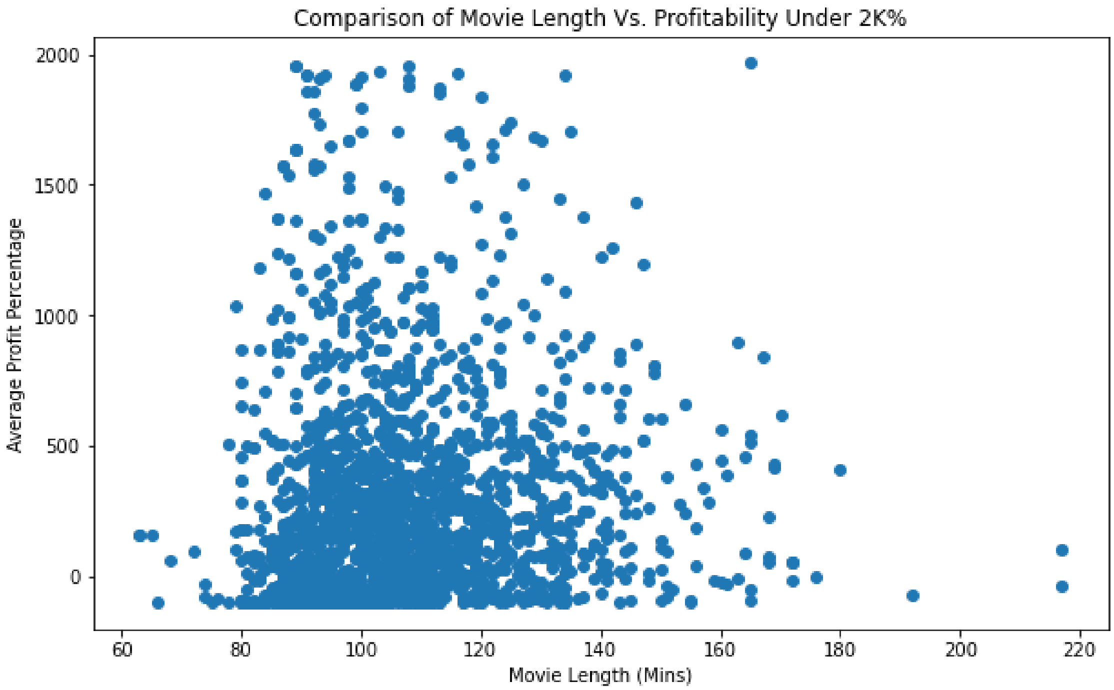
[29]:
```
merged_sort_by_Profit_Percent_under_2k = merged_df.sort_values(by=['profit␣
↪percent'], ascending=False)[82:-1]
merged_sort_by_Profit_Percent_under_2k.head()
```

[29]:
```
                          Movie Name  Release Year  Length  \
95                           Boyhood          2014   165.0
263   Kevin Hart: Laugh at My Pain          2011    89.0
262   Kevin Hart: Laugh at My Pain          2011    89.0
539                         The Gift          2015   108.0
1021             The Purge: Anarchy          2014   103.0

                       Genres         Director  Avg Rating  Number of Votes  \
```

```
95                    Drama  Richard Linklater        7.9           315584
263       Comedy,Documentary          Tim Story        7.4             5081
262       Comedy,Documentary       Leslie Small        7.4             5081
539   Drama,Mystery,Thriller      Joel Edgerton        7.1           123834
1021    Action,Horror,Sci-Fi     James DeMonaco        6.5           126203

      Unnamed: 0  id  release_date                          movie  \
95          4484  85  Jul 11, 2014                        Boyhood
263         5374  75   Sep 9, 2011  Kevin Hart: Laugh at My Pain
262         5374  75   Sep 9, 2011  Kevin Hart: Laugh at My Pain
539         4262  63   Aug 7, 2015                       The Gift
1021        3770  71  Jul 18, 2014            The Purge: Anarchy

     production_budget domestic_gross worldwide_gross  year  \
95          $4,000,000    $25,379,975     $57,273,049  2014
263           $750,000     $7,706,436      $7,712,436  2011
262           $750,000     $7,706,436      $7,712,436  2011
539         $5,000,000    $43,787,265     $58,978,477  2015
1021        $9,000,000    $71,562,550    $111,534,881  2014

      float_production_budget  float_domestic_gross  float_worldwide_gross  \
95                  4000000.0            25379975.0             57273049.0
263                  750000.0             7706436.0              7712436.0
262                  750000.0             7706436.0              7712436.0
539                 5000000.0            43787265.0             58978477.0
1021                9000000.0            71562550.0            111534881.0

      profit percent
95        1966.32560
263       1955.84960
262       1955.84960
539       1955.31484
1021      1934.41590
```

[30]:

```
fig, ax = plt.subplots(figsize=(10,6))

ax.
↪scatter(x=merged_sort_by_Profit_Percent_under_2k['Length'],y=merged_sort_by_Profit_Percent_
↪percent'])
ax.set_title("Comparison of Movie Length Vs. Profitability Under 2K%")
ax.set_xlabel('Movie Length (Mins)')
ax.set_ylabel('Average Profit Percentage')
plt.savefig('./Images/Movie_Length_Vs_Profitability_Under_2K.
↪png',bbox_inches='tight')
```



Looking at the vast majority of the movies in our original data set, we do not see a strong correlation. There is some loose indication that most movies are under 2 to 2.5 hours. High profits can be generated in a movie of only 1.5 hours. So if the story does not require a long movie there is no need to stretch the movie into that longer timeframe.

1.13 Recommendations

We offer the following 3 recommendations:
- Work with Directors with track records of consistent profit.**
- Give NEW directors a smaller budget to prove themselves.
- Need to investigate further the effect of Genre on budgets and therefore profitability.

1.14 Other Observations

Microsoft has:
- Diverse product lines.
- Sells internationally.
- Has a large international employee base.

Movie decisions may be driven by:
- Product placement considerations.
- Overseas markets.

US Market prefers shorter movies. Need to study other markets.

1.15 Next Steps

Investigate non-US markets.
Identify opportunities for product placement.
Formulate marketing strategies to drive product placement.
Engage movie industry drivers (producers, directors, writers, etc.) to consider how to move Microsoft interests further.

[ ]:

```

```