Part 2 Content

- Import Libraries
- Load myData CSV File
- Q1: Is there a correlation between the production budget and the profit?
- Q2: Is there a correlation between the ratings and the profit?
- Q3: Which directors, writers, actors and actresses make the most profit (High Budget)?
- Q4: Which genres make the most profit (High Budget)?

Import Libraries

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

from functions import *
```

Load myData CSV File

In [2]:
```
#myData.csv file is created in Part 1 python file. It combined all required
df = pd.read_csv('zippedData/myData.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53305 entries, 0 to 53304
Data columns (total 18 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Unnamed: 0         53305 non-null  int64  
 1   release_date       53305 non-null  object 
 2   movie              53305 non-null  object 
 3   production_budget  53305 non-null  float64
 4   domestic_gross     53305 non-null  float64
 5   worldwide_gross    53305 non-null  float64
 6   tconst             53305 non-null  object 
 7   genres             53265 non-null  object 
 8   averagerating      53305 non-null  float64
 9   numvotes           53305 non-null  int64  
 10  nconst             53305 non-null  object 
 11  category           53305 non-null  object 
 12  characters         21488 non-null  object 
 13  primary_name       53305 non-null  object 
 14  death_year         1243 non-null   float64
 15  release_year       53305 non-null  int64  
 16  profit_gross       53305 non-null  float64
 17  profit_rate        53305 non-null  float64
dtypes: float64(7), int64(3), object(8)
memory usage: 7.3+ MB
```

In [3]:
```
#df[df.primary_name =='Emma Thompson']
```

Q1: Is there a correlation between the production budget and the profit?

In [4]:
```
selected_columns1 = ['movie', 'production_budget', 'averagerating', 'profit
df1 = df[selected_columns1].copy()

#Remove duplicate movies
df1.drop_duplicates(subset='movie', keep='first', inplace=True)

#Remove outliers with very low budget and very high profit_rate
df1.drop(df1[(df1.production_budget <1000000) | (df1.profit_rate >30)].inde

# Remove movies with release year before 1990
df1.drop(df1[df1.release_year < 1990].index, inplace=True)

print(df1.shape)
df1.head(10)
```

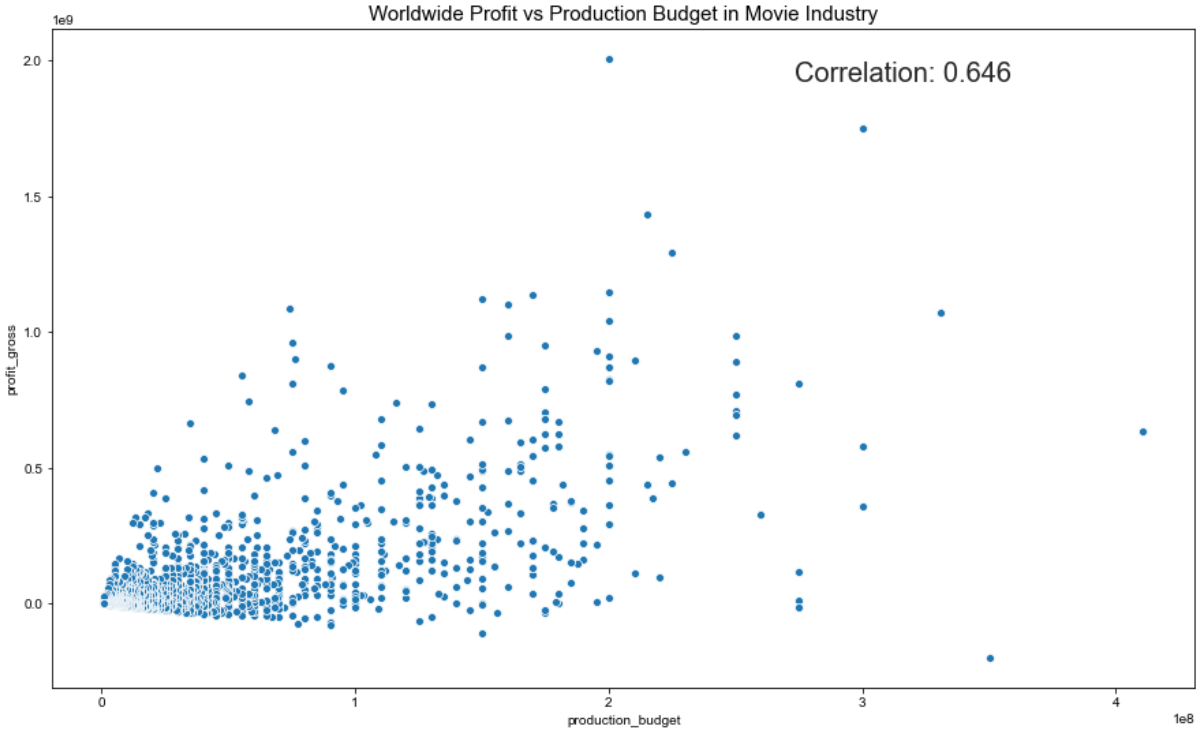
```
(1733, 6)
```

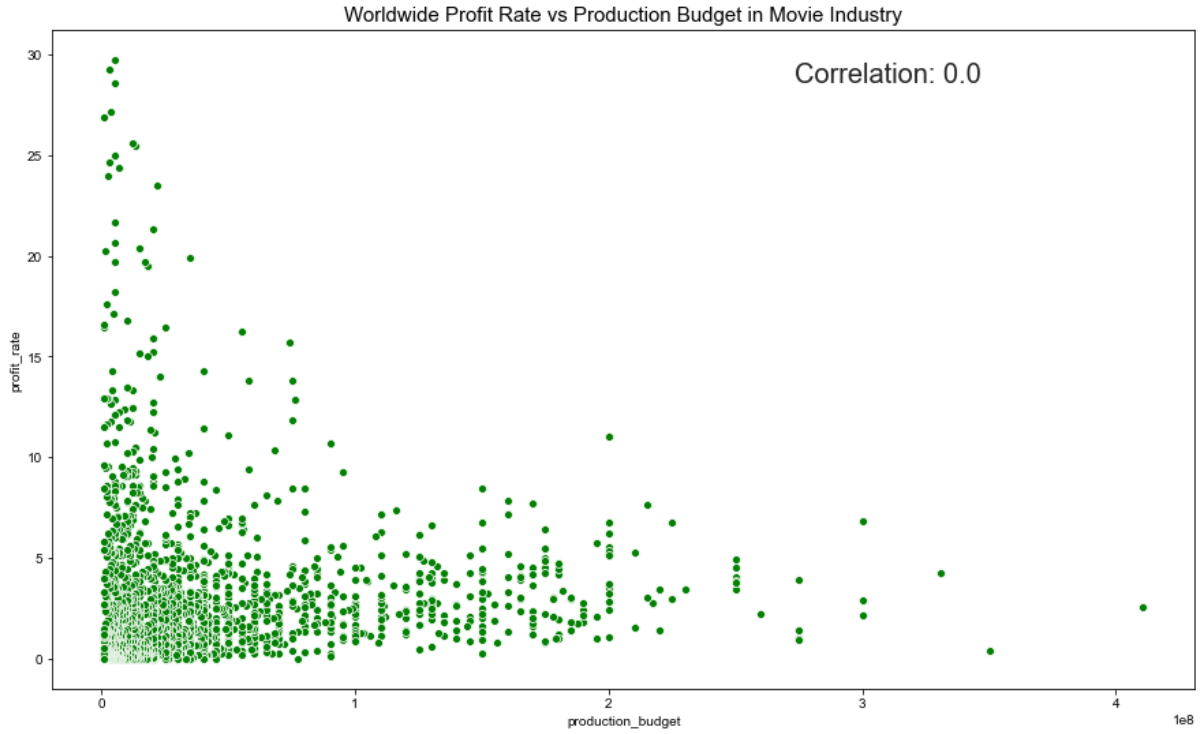
Out[4]:

	movie	production_budget	averagerating	profit_gross	profit_rate	release_year
0	Pirates of the Caribbean: On Stranger Tides	410600000.0	6.6	6.350639e+08	2.546673	2011
30	Dark Phoenix	350000000.0	6.0	-2.002376e+08	0.427892	2019
60	Avengers: Age of Ultron	330600000.0	7.3	1.072414e+09	4.243841	2015
90	Avengers: Infinity War	300000000.0	8.5	1.748134e+09	6.827114	2018
91	Justice League	300000000.0	6.5	3.559452e+08	2.186484	2017
119	Spectre	300000000.0	6.8	5.796209e+08	2.932070	2015
180	The Dark Knight Rises	275000000.0	8.4	8.094391e+08	3.943415	2012
181	Solo: A Star Wars Story	275000000.0	7.0	1.181513e+08	1.429641	2018
192	John Carter	275000000.0	6.6	7.778100e+06	1.028284	2012
208	The Lone Ranger	275000000.0	6.4	-1.499788e+07	0.945462	2013

In [5]:

```
budget = df1['production_budget']
profit= df1['profit_gross']
profitRate = df1['profit_rate']

corr0 = round(profit.corr(budget), 3)
corr1 = round(profitRate.corr(budget), 3)

fig, axes = plt.subplots(2, 1, figsize=(15, 20))
plt.subplots_adjust(hspace=0.25)
sns.set(font_scale=1.5)

sns.scatterplot(ax=axes[0], x=budget, y=profit)
axes[0].set_title('Worldwide Profit vs Production Budget in Movie Industry'
axes[0].text(0.65, 0.95,"Correlation: " + str(corr0), transform=axes[0].tra

sns.scatterplot(ax=axes[1], x=budget, y=profitRate, color='green')
axes[1].set_title('Worldwide Profit Rate vs Production Budget in Movie Indu
axes[1].text(0.65, 0.95,"Correlation: " + str(corr1), transform=axes[1].tra

plt.savefig('figures/budget-profit1.png')
```


Worldwide Profit vs Production Budget in Movie Industry
1e9
Correlation: 0.646
profit_gross
production_budget
1e8

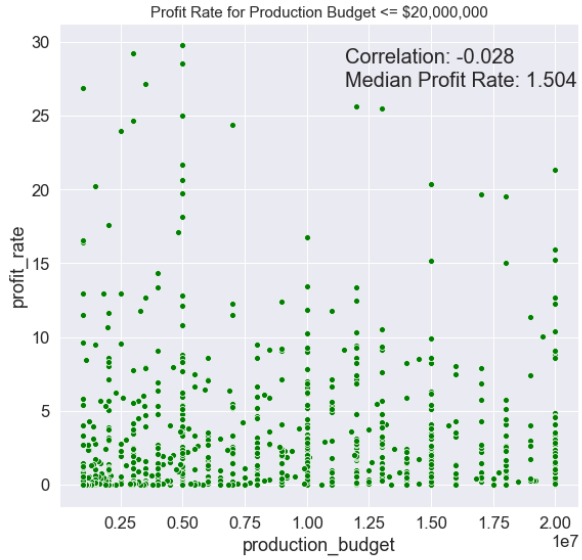

Worldwide Profit Rate vs Production Budget in Movie Industry
Correlation: 0.0
profit_rate
production_budget
1e8

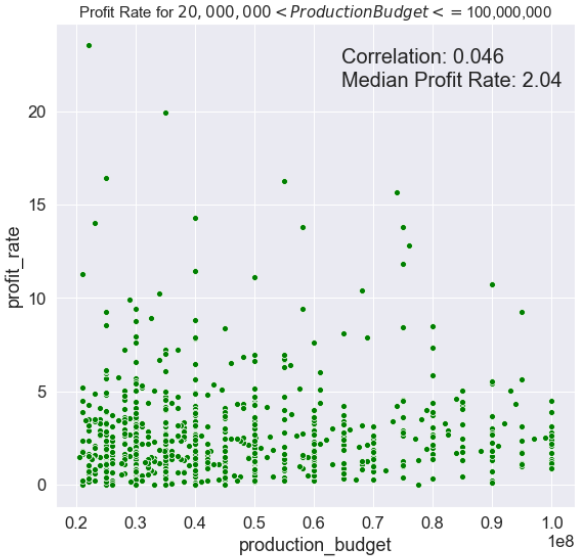
In [9]:

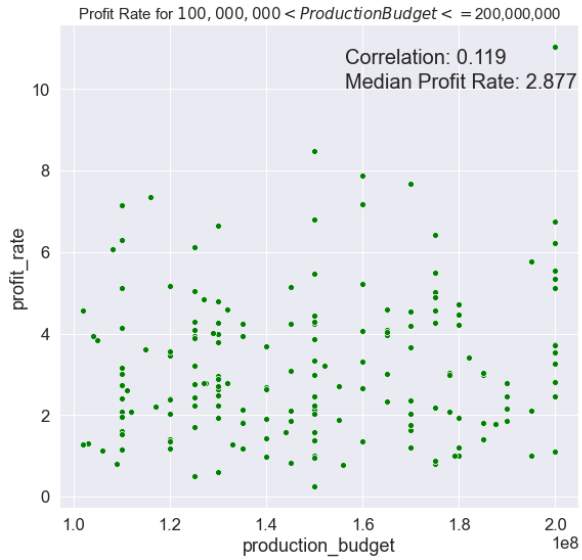
```
budget = df1['production_budget']
profitRate_1= df1[df1.production_budget<=20000000]['profit_rate']
profitRate_2= df1[(df1.production_budget>20000000) & (df1.production_budget
profitRate_3= df1[(df1.production_budget>100000000) & (df1.production_budge
profitRate_4= df1[(df1.production_budget>200000000)]['profit_rate']

profitRateList = [profitRate_1, profitRate_2, profitRate_3, profitRate_4]
corr = []
profit_median = []
profitRateList[0].corr(budget)

for plist in profitRateList:
    profit_median.append(round(plist.median(), 3))
    corr.append(round(plist.corr(budget), 3))

#print(corr)
#print(profit_mean)
```

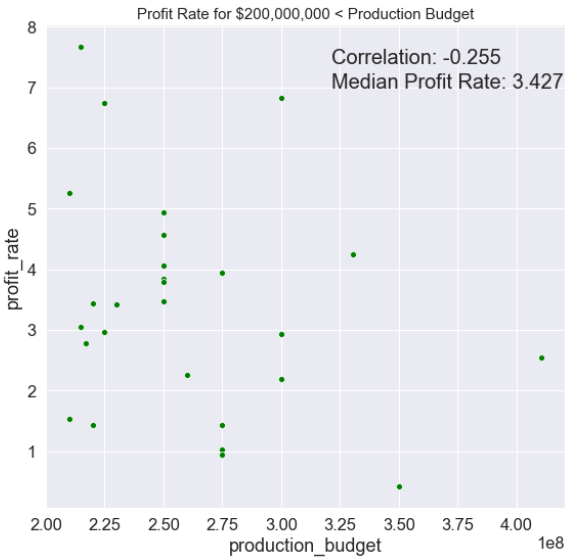
In [10]:

```
fig, axes = plt.subplots(2, 2, figsize=(20, 20))
plt.subplots_adjust(hspace=0.3)
sns.set(font_scale=1.5)

sns.scatterplot(ax=axes[0, 0], x=budget, y=profitRateList[0], color='green'
axes[0, 0].set_title('Profit Rate for Production Budget <= $20,000,000', fo
axes[0, 0].text(0.55, 0.95,"Correlation: " + str(corr[0]), transform=axes[0
                verticalalignment='top')
axes[0, 0].text(0.55, 0.90,"Median Profit Rate: " + str(profit_median[0]),
                fontsize=20, verticalalignment='top')

sns.scatterplot(ax=axes[0, 1], x=budget, y=profitRateList[1], color='green'
axes[0, 1].set_title('Profit Rate for $20,000,000 < Production Budget <= $1
axes[0, 1].text(0.55, 0.95,"Correlation: " + str(corr[1]), transform=axes[0
                verticalalignment='top')
axes[0, 1].text(0.55, 0.90,"Median Profit Rate: " + str(profit_median[1]),
                fontsize=20, verticalalignment='top')

sns.scatterplot(ax=axes[1, 0], x=budget, y=profitRateList[2], color='green'
axes[1, 0].set_title('Profit Rate for $100,000,000 < Production Budget <= $
axes[1, 0].text(0.55, 0.95,"Correlation: " + str(corr[2]), transform=axes[1
axes[1, 0].text(0.55, 0.90,"Median Profit Rate: " + str(profit_median[2]),

sns.scatterplot(ax=axes[1, 1], x=budget, y=profitRateList[3], color='green'
axes[1, 1].set_title('Profit Rate for $200,000,000 < Production Budget', fo
axes[1, 1].text(0.55, 0.95,"Correlation: " + str(corr[3]), transform=axes[1
axes[1, 1].text(0.55, 0.90,"Median Profit Rate: " + str(profit_median[3]),

plt.savefig('figures/budget-profit2.png')
```





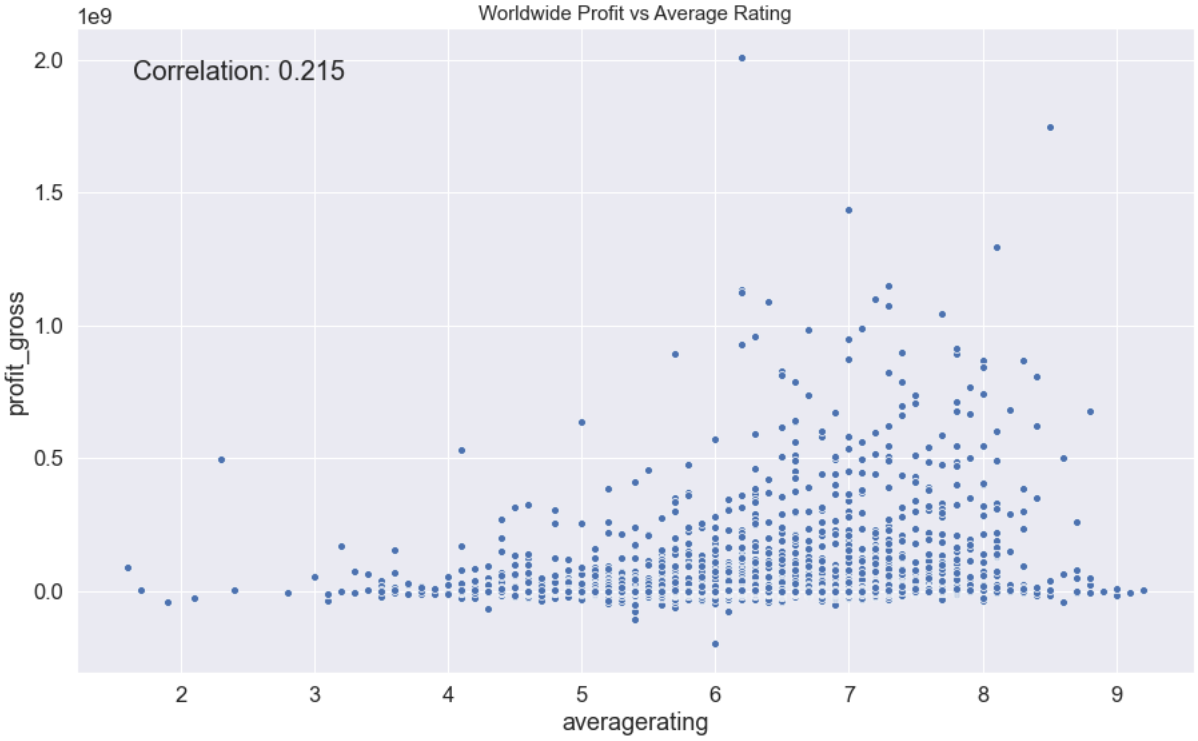


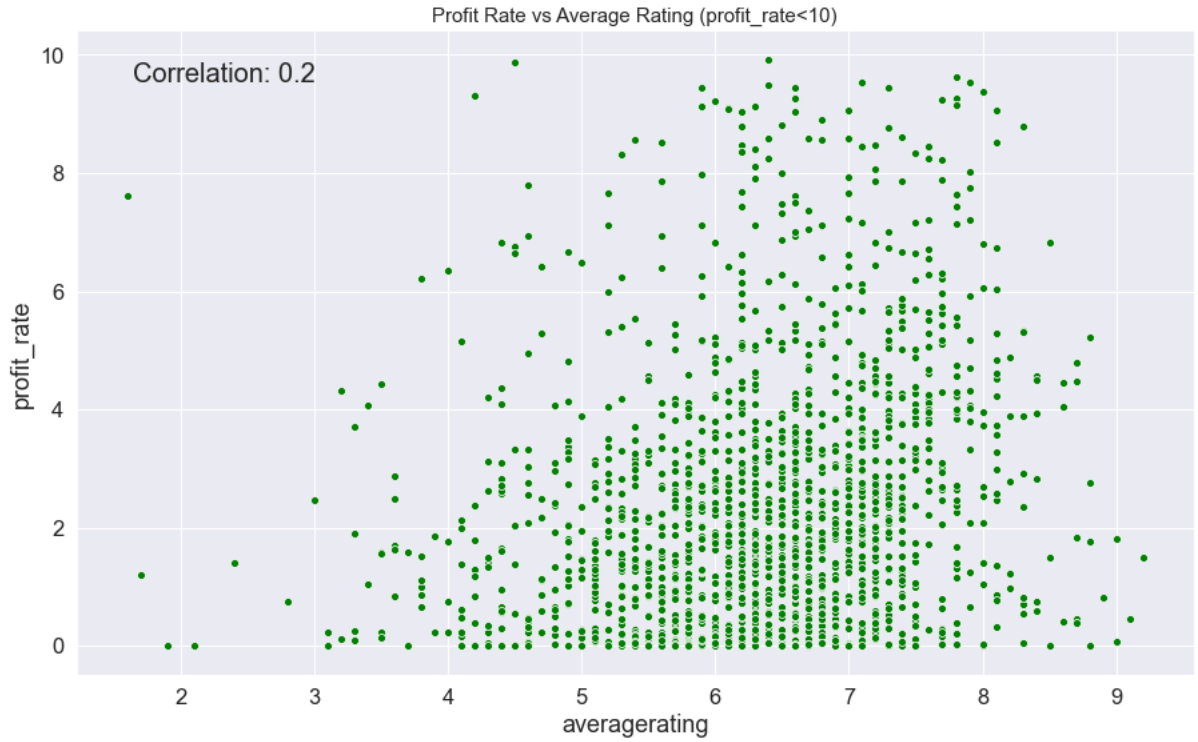


Q2: Is there a correlation between the ratings and the profit?

In [11]:

```
rating = df1['averagerating']
profit= df1['profit_gross']
profitRate = df1[df1.profit_rate < 10]['profit_rate'] # remove outliers

corr0 = round(profit.corr(rating), 3)
corr1 = round(profitRate.corr(rating), 3)

fig, axes = plt.subplots(2, 1, figsize=(15, 20))
plt.subplots_adjust(hspace=0.25)
sns.set(font_scale=1.5)

sns.scatterplot(ax=axes[0], x=rating, y=profit)
axes[0].set_title('Worldwide Profit vs Average Rating', fontsize=15)
axes[0].text(0.05, 0.95,"Correlation: " + str(corr0), transform=axes[0].tra

sns.scatterplot(ax=axes[1], x=rating, y=profitRate, color='green')
axes[1].set_title('Profit Rate vs Average Rating (profit_rate<10)', fontsiz
axes[1].text(0.05, 0.95,"Correlation: " + str(corr1), transform=axes[1].tra

plt.savefig('figures/rating-profit.png')
```





Q3: Which directors, writers, actors and actresses make the most profit (high budget)?

In [12]:
```
# Copy dataframe with selected columns
selected_columns2 = ['category', 'primary_name', 'death_year', 'movie', 'pr
df2 = df[selected_columns2].copy() ## Select columns

# Clean duplicates
df2.drop_duplicates(subset=None, keep='first', inplace=True)

#Select the movies with production_budget > $100,000,000
df2 = df2[(df2.production_budget > 100000000)]

# Remove the dead people
df2 = df2[df2.death_year.isnull()]

# Drop the movies with release year before 1990
df2.drop(df2[df2.release_year < 1990].index, inplace=True)

print(df2.shape)
df2.head(20)
```

```
(2158, 8)
```

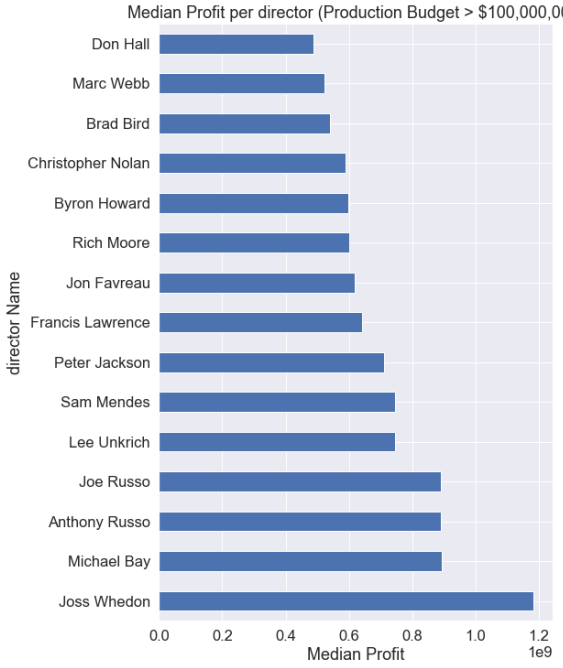
Out[12]:

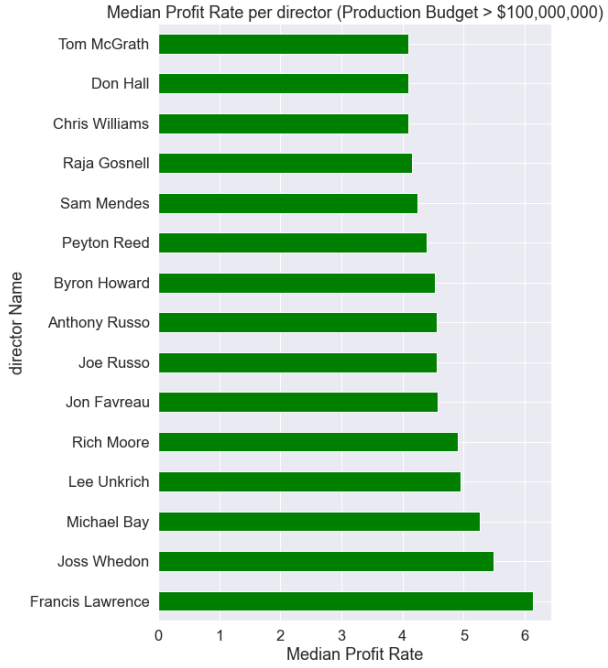
	category	primary_name	death_year	movie	production_budget	profit_gross	profit_rate
0	writer	Tim Powers	NaN	Pirates of the Caribbean: On Stranger Tides	410600000.0	6.350639e+08	2.546673
1	director	Rob Marshall	NaN	Pirates of the Caribbean: On Stranger Tides	410600000.0	6.350639e+08	2.546673
2	writer	Ted Elliott	NaN	Pirates of the Caribbean: On Stranger Tides	410600000.0	6.350639e+08	2.546673
3	actor	Johnny Depp	NaN	Pirates of the Caribbean: On Stranger Tides	410600000.0	6.350639e+08	2.546673
7	actor	Ian McShane	NaN	Pirates of the Caribbean: On Stranger Tides	410600000.0	6.350639e+08	2.546673

	category	primary_name	death_year	movie	production_budget	profit_gross	profit_rate
8	writer	Jay Wolpert	NaN	Pirates of the Caribbean: On Stranger Tides	410600000.0	6.350639e+08	2.546673
10	actor	Geoffrey Rush	NaN	Pirates of the Caribbean: On Stranger Tides	410600000.0	6.350639e+08	2.546673
12	actress	Penélope Cruz	NaN	Pirates of the Caribbean: On Stranger Tides	410600000.0	6.350639e+08	2.546673
17	writer	Terry Rossio	NaN	Pirates of the Caribbean: On Stranger Tides	410600000.0	6.350639e+08	2.546673
20	writer	Stuart Beattie	NaN	Pirates of the Caribbean: On Stranger Tides	410600000.0	6.350639e+08	2.546673
31	writer	John Byrne	NaN	Dark Phoenix	350000000.0	-2.002376e+08	0.427892
33	writer	Chris Claremont	NaN	Dark Phoenix	350000000.0	-2.002376e+08	0.427892
39	director	Simon Kinberg	NaN	Dark Phoenix	350000000.0	-2.002376e+08	0.427892
43	actor	James McAvoy	NaN	Dark Phoenix	350000000.0	-2.002376e+08	0.427892
45	actress	Jennifer Lawrence	NaN	Dark Phoenix	350000000.0	-2.002376e+08	0.427892
47	actor	Michael Fassbender	NaN	Dark Phoenix	350000000.0	-2.002376e+08	0.427892
49	actor	Nicholas Hoult	NaN	Dark Phoenix	350000000.0	-2.002376e+08	0.427892
60	actor	Mark Ruffalo	NaN	Avengers: Age of Ultron	330600000.0	1.072414e+09	4.243841
61	actor	Robert Downey Jr.	NaN	Avengers: Age of Ultron	330600000.0	1.072414e+09	4.243841
63	actor	Chris Evans	NaN	Avengers: Age of Ultron	330600000.0	1.072414e+09	4.243841

In [14]:
```
# Director, call function
categoryStudy(df2, 'director', 15, 12, 'high')
```

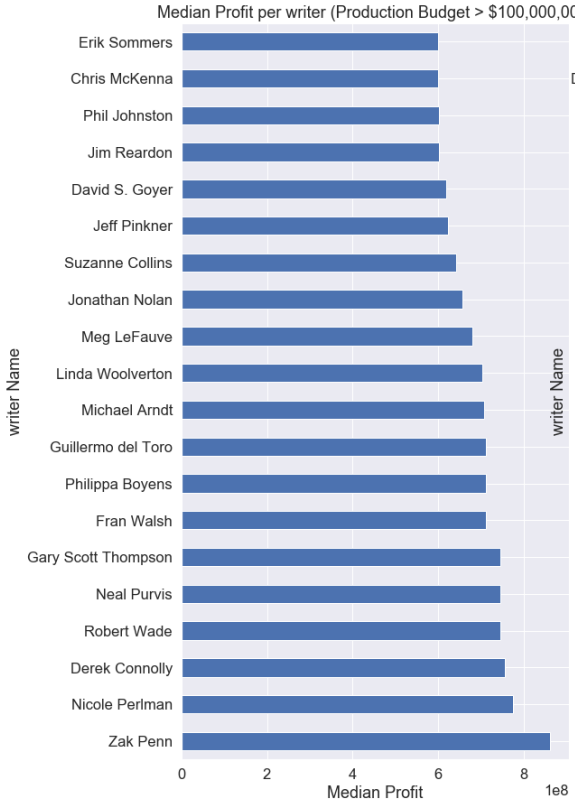
```
Total number: 51
Best director List:  ['Byron Howard', 'Joe Russo', 'Francis Lawrence', 'J
oss Whedon', 'Anthony Russo', 'Rich Moore', 'Michael Bay', 'Don Hall', 'S
am Mendes', 'Jon Favreau', 'Lee Unkrich']
```

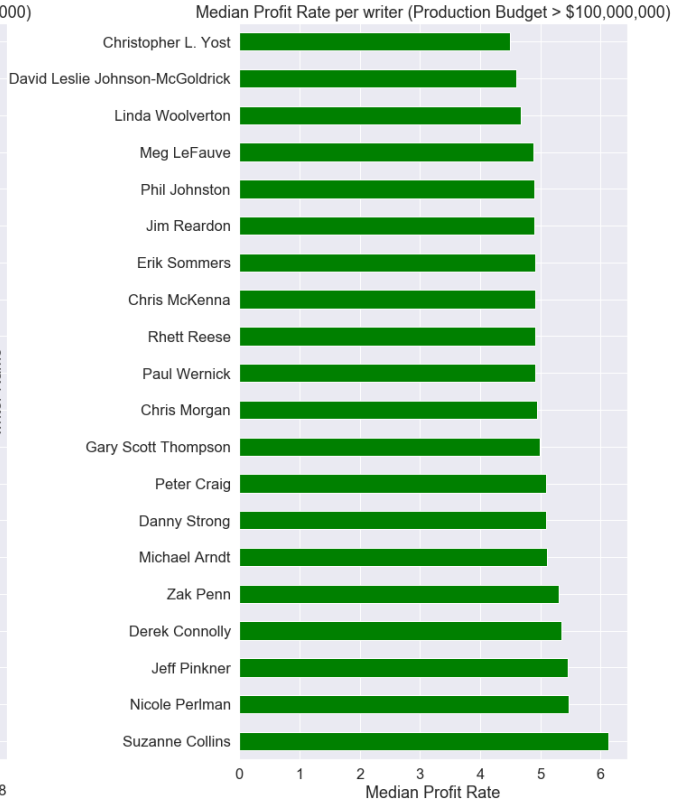




In [14]:
```
# Writer, , call function
categoryStudy(df2, 'writer', 20, 15, 'high')
```

```
Total number: 110
Best writer List:  ['Nicole Perlman', 'Derek Connolly', 'Erik Sommers',
'Meg LeFauve', 'Linda Woolverton', 'Phil Johnston', 'Zak Penn', 'Gary Sco
tt Thompson', 'Suzanne Collins', 'Chris McKenna', 'Michael Arndt', 'Jim R
eardon', 'Jeff Pinkner']
```





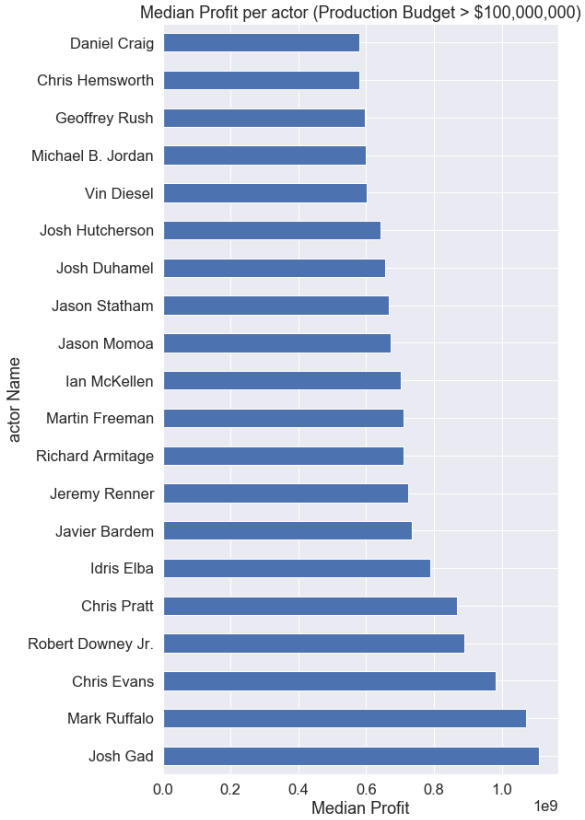
In [38]:
```
df2[df2.primary_name == 'Cate Blanchett']
```

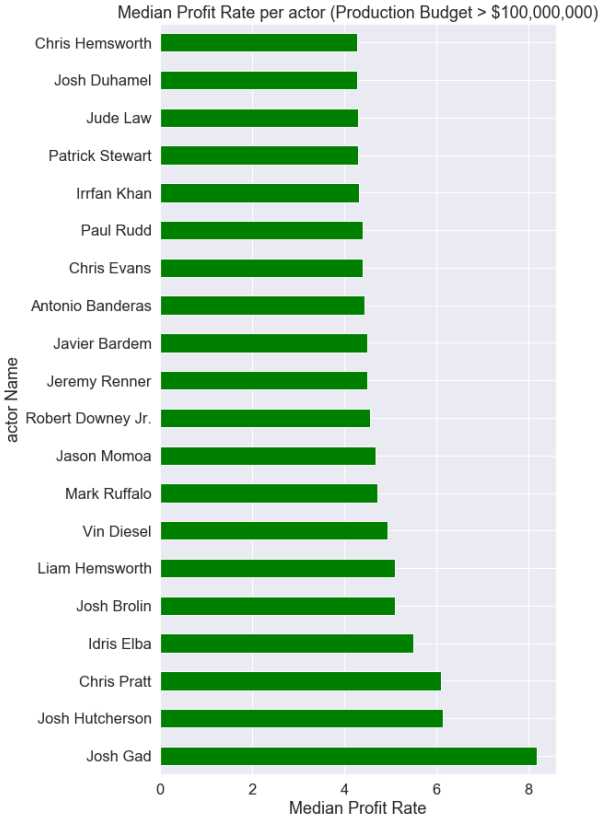
Out[38]:

	category	primary_name	death_year	movie	production_budget	profit_gross	profit_rate	r
450	actress	Cate Blanchett	NaN	The Hobbit: The Battle of the Five Armies	250000000.0	695577621.0	3.782310	
769	actress	Cate Blanchett	NaN	Robin Hood	210000000.0	112459006.0	1.535519	
1657	actress	Cate Blanchett	NaN	Thor: Ragnarok	180000000.0	666980024.0	4.705445	
3883	actress	Cate Blanchett	NaN	How to Train Your Dragon 2	145000000.0	469586270.0	4.238526	
4875	actress	Cate Blanchett	NaN	How to Train Your Dragon: The Hidden World	129000000.0	390258283.0	4.025258	

In [15]:
```
# Actor
categoryStudy(df2, 'actor', 20, 15, 'high')
```

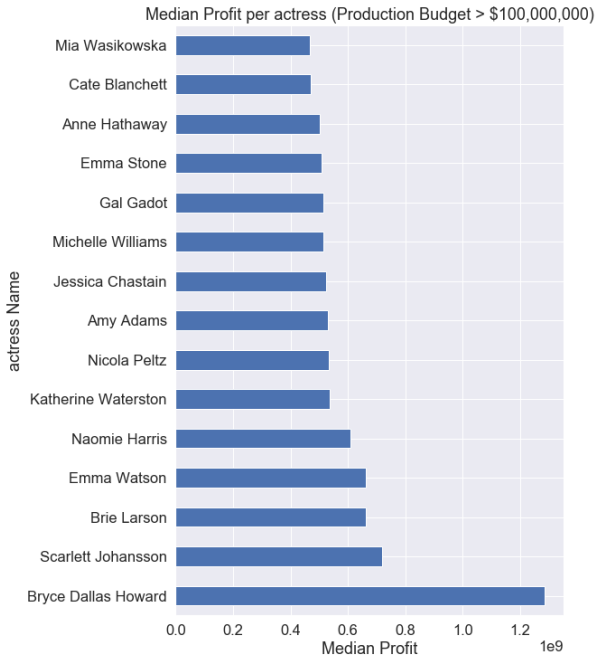
```
Total number: 113
Best actor List:  ['Javier Bardem', 'Josh Duhamel', 'Vin Diesel', 'Idris
Elba', 'Josh Hutcherson', 'Chris Hemsworth', 'Chris Pratt', 'Josh Gad',
'Jason Momoa', 'Mark Ruffalo', 'Robert Downey Jr.', 'Jeremy Renner', 'Chr
is Evans']
```

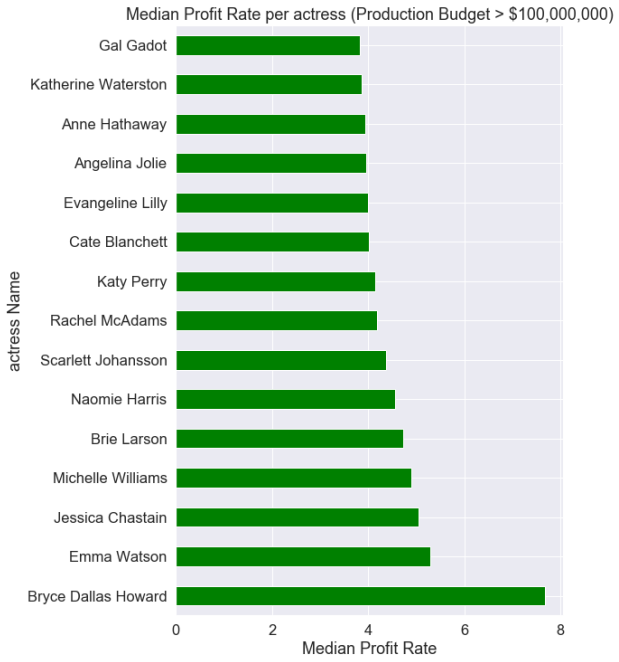




In [16]:
```
# Actress, call function
categoryStudy(df2, 'actress', 15, 12, 'high')
```

```
Total number: 45
Best actress List:  ['Jessica Chastain', 'Cate Blanchett', 'Katherine Wat
erston', 'Michelle Williams', 'Emma Watson', 'Scarlett Johansson', 'Brie
Larson', 'Gal Gadot', 'Bryce Dallas Howard', 'Naomie Harris', 'Anne Hatha
way']
```





Q4: Which genre brings highest profit (high budget)?

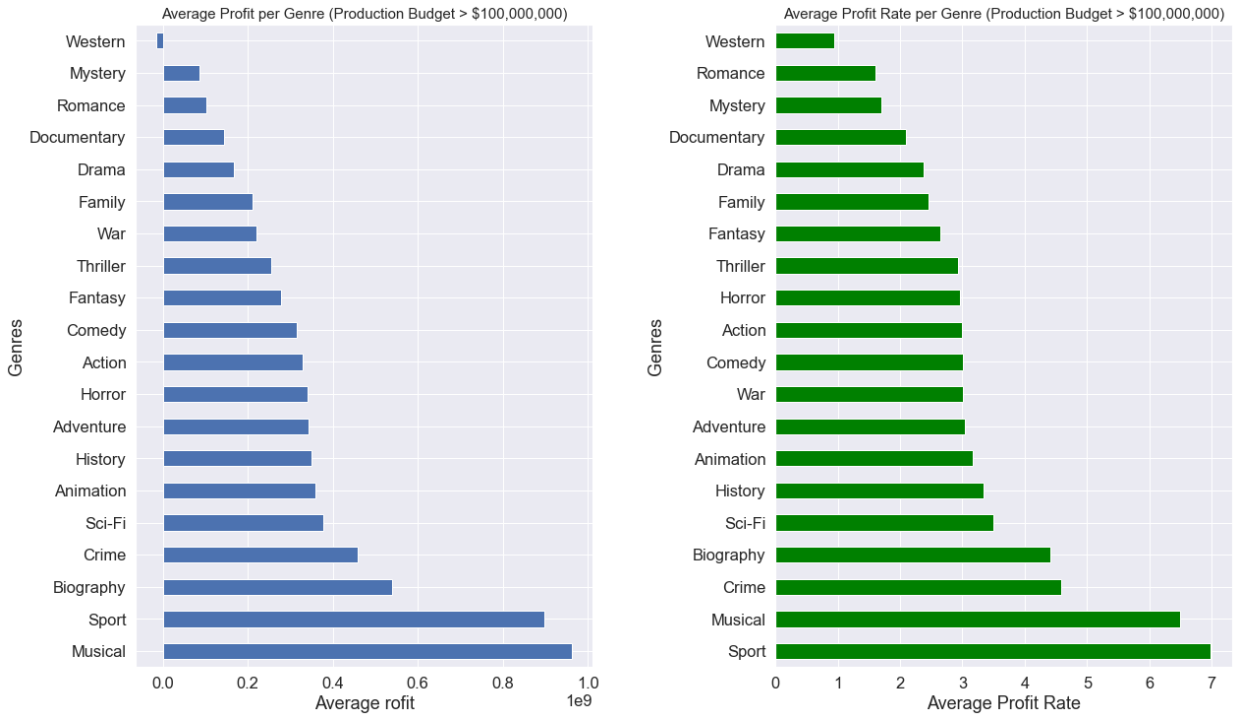
In [15]:
```
# Copy dataframe with selected columns
selected_columns3 = ['movie', 'production_budget', 'genres', 'profit_gross'
df3 = df[selected_columns3].copy() ## Select columns

# Clean duplicates
df3.drop_duplicates(subset=None, keep='first', inplace=True)

#Select the movies with production_budget > $100,000,000
df3 = df3[(df3.production_budget > 100000000)]

#Drop the movies with release year before 2000
df3.drop(df3[df3.release_year < 1990].index, inplace=True)

print(df3.shape)
df3.head(20)
```

```
(632, 6)
```

Out[15]:

	movie	production_budget	genres	profit_gross	profit_rate	release_year
0	Pirates of the Caribbean: On Stranger Tides	410600000.0	Action	6.350639e+08	2.546673	2011
3	Pirates of the Caribbean: On Stranger Tides	410600000.0	Fantasy	6.350639e+08	2.546673	2011
4	Pirates of the Caribbean: On Stranger Tides	410600000.0	Adventure	6.350639e+08	2.546673	2011
30	Dark Phoenix	350000000.0	Adventure	-2.002376e+08	0.427892	2019
32	Dark Phoenix	350000000.0	Sci-Fi	-2.002376e+08	0.427892	2019
33	Dark Phoenix	350000000.0	Action	-2.002376e+08	0.427892	2019
60	Avengers: Age of Ultron	330600000.0	Sci-Fi	1.072414e+09	4.243841	2015
61	Avengers: Age of Ultron	330600000.0	Adventure	1.072414e+09	4.243841	2015
62	Avengers: Age of Ultron	330600000.0	Action	1.072414e+09	4.243841	2015
90	Avengers: Infinity War	300000000.0	Action	1.748134e+09	6.827114	2018
91	Justice League	300000000.0	Fantasy	3.559452e+08	2.186484	2017
92	Justice League	300000000.0	Action	3.559452e+08	2.186484	2017
93	Justice League	300000000.0	Adventure	3.559452e+08	2.186484	2017
94	Avengers: Infinity War	300000000.0	Adventure	1.748134e+09	6.827114	2018
100	Avengers: Infinity War	300000000.0	Sci-Fi	1.748134e+09	6.827114	2018
119	Spectre	300000000.0	Adventure	5.796209e+08	2.932070	2015
140	Spectre	300000000.0	Action	5.796209e+08	2.932070	2015
141	Spectre	300000000.0	Thriller	5.796209e+08	2.932070	2015
180	The Dark Knight Rises	275000000.0	Thriller	8.094391e+08	3.943415	2012
181	Solo: A Star Wars Story	275000000.0	Action	1.181513e+08	1.429641	2018

In [16]:
```
# Plot genres
genre_series1 = df3.groupby('genres')['profit_gross'].median().sort_values(
genre_series2 = df3.groupby('genres')['profit_rate'].median().sort_values(a

fig, axes = plt.subplots(1, 2, figsize=(20, 12))
plt.subplots_adjust(wspace=0.4)
sns.set(font_scale=1.5)

genre_series1.plot.barh(ax=axes[0])
axes[0].set_title('Average Profit per Genre (Production Budget > $100,000,0
axes[0].set_xlabel('Average rofit')
axes[0].set_ylabel('Genres')

genre_series2.plot.barh(ax=axes[1], color='green')
axes[1].set_title('Average Profit Rate per Genre (Production Budget > $100,
axes[1].set_xlabel('Average Profit Rate')
axes[1].set_ylabel('Genres')

plt.savefig('figures/genres-profit1_highBudget.png')
```



In [17]:
```
# Full list of genres in descending order
df3.groupby('genres').count().sort_values(by='movie', ascending=False)
```

Out[17]:

genres	movie	production_budget	profit_gross	profit_rate	release_year
Adventure	172	172	172	172	172
Action	137	137	137	137	137
Sci-Fi	60	60	60	60	60
Comedy	52	52	52	52	52
Animation	47	47	47	47	47
Fantasy	42	42	42	42	42
Drama	38	38	38	38	38
Family	27	27	27	27	27
Thriller	17	17	17	17	17
Horror	7	7	7	7	7
Crime	6	6	6	6	6
Documentary	6	6	6	6	6
Mystery	5	5	5	5	5
Romance	4	4	4	4	4
History	3	3	3	3	3
Musical	2	2	2	2	2
Sport	2	2	2	2	2
Biography	2	2	2	2	2
War	1	1	1	1	1
Western	1	1	1	1	1

In [18]:
```
df3[df3.genres == 'Sport']
```

Out[18]:

	movie	production_budget	genres	profit_gross	profit_rate	release_year
3148	Frozen	150000000.0	Sport	1.122470e+09	8.483133	2013
3301	Wonder Woman	150000000.0	Sport	6.711334e+08	5.474223	2017

In [19]:
```
# Select the popular genres names with higher number of movies
popular_genres = list(df3.groupby('genres').count().sort_values(by='movie',
popular_genres
```

Out[19]:
```
['Adventure',
 'Action',
 'Sci-Fi',
 'Comedy',
 'Animation',
 'Fantasy',
 'Drama',
 'Family',
 'Thriller']
```

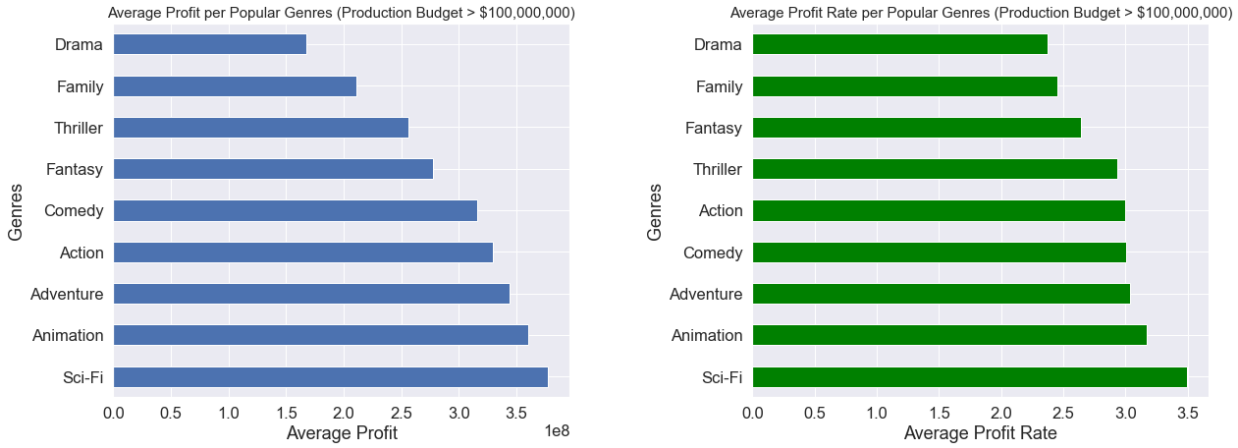
In [24]:
```
# Filter the data for the popular genres
df3_pop = df3.loc[df3['genres'].isin(popular_genres)]
df3
#df3_pop[df3_pop.genres == 'Comedy']
```

Out[24]:

	movie	production_budget	genres	profit_gross	profit_rate	release_year
0	Pirates of the Caribbean: On Stranger Tides	410600000.0	Action	635063875.0	2.546673	2011
3	Pirates of the Caribbean: On Stranger Tides	410600000.0	Fantasy	635063875.0	2.546673	2011
4	Pirates of the Caribbean: On Stranger Tides	410600000.0	Adventure	635063875.0	2.546673	2011
30	Dark Phoenix	350000000.0	Adventure	-200237650.0	0.427892	2019
32	Dark Phoenix	350000000.0	Sci-Fi	-200237650.0	0.427892	2019
...	...	...	...	...	...	...
6338	Bumblebee	102000000.0	Action	363195589.0	4.560741	2018
6339	Bumblebee	102000000.0	Adventure	363195589.0	4.560741	2018
6340	Bumblebee	102000000.0	Sci-Fi	363195589.0	4.560741	2018
6342	Cloud Atlas	102000000.0	Action	28673154.0	1.281109	2012
6344	Cloud Atlas	102000000.0	Mystery	28673154.0	1.281109	2012

632 rows × 6 columns

In [21]:
```
# plots for popular genres
genre_series1 = df3_pop.groupby('genres')['profit_gross'].median().sort_val
genre_series2 = df3_pop.groupby('genres')['profit_rate'].median().sort_valu

fig, axes = plt.subplots(1, 2, figsize=(20, 7))
plt.subplots_adjust(wspace=0.4)
sns.set(font_scale=1.5)

genre_series1.plot.barh(ax=axes[0])
axes[0].set_title('Average Profit per Popular Genres (Production Budget > $
axes[0].set_xlabel('Average Profit')
axes[0].set_ylabel('Genres')

genre_series2.plot.barh(ax=axes[1], color='green')
axes[1].set_title('Average Profit Rate per Popular Genres (Production Budge
axes[1].set_xlabel('Average Profit Rate')
axes[1].set_ylabel('Genres')

plt.savefig('figures/genres-profit2_highBudget.png')
```



In [ ]:

In [ ]: