Microsoft Movie Industry Analysis

Author: Jonah Devoy

Overview

For my exploratory data analysis project, I have been tasked with using exploratory data analysis to generate insights for business stakeholders. My goal for this project was to investigate what genre of films is the most profitable and deliver these findings to Microsoft's new movie studio executives. The analysis results, shown by visualizations and descriptive statistics revealed the most profitable movie genres and recommended the best directors to hire for each genre. Microsoft's new movie studio can use my report to produce movies that accumulate the highest profit and tailor those movies to specific genres and what directors are the best to use for each of those selected genres.

Business Problem

It was brought to my attention that Microsoft has noticed other big enterprises creating original video content and they would like to enter the exciting world of movie-making with their new movie studio. One problem is that Microsoft is uncertain of where to start. I have been recruited with exploring what types of films Microsofts new movie studio should produce to be successful. Using my findings I transformed those results into actionable insights so that Microsoft's new movie studio will have direction to be profitable. To aid Microsofts studio, I looked at the highest-profiting genre of movies, which genre of movies should be produced, and what directors to use for the selected genres. The three factors I based my analysis on are:

- What movie genres produce the highest profits?
- What movie genres should be produced to avoid competition?
- What director should be hired to produce their movies?

Data Understanding

I used two data sources for my analysis to determine the best recommendations for Microsofts new movie studio.

- tn.movie_budgets.csv: This compressed csv was extracted from the zipped data folder provided in the dsc-phase-1-project-v2-4.git Github repository located at [https://github.com/learn-co-curriculum/dsc-phase-1-project-v2-4.git (https://github.com/learn-co-curriculum/dsc-phase-1-project-v2-4.git)](https://github.com/learn-co-curriculum/dsc-phase-1-project-v2-4.git). I forked the repository and extracted all the files to my local machine where I could work on the tn.movie_budgets.csv on my local machine's Jupyter Notebook.
- imdb_top_1000.csv: This data set is taken from Kaggles EDA on IMDB Movies Dataset located at https://www.kaggle.com/code/harshitshankhdhar/eda-on-imdb-movies-dataset/notebook

(https://www.kaggle.com/code/harshitshankhdhar/eda-on-imdb-movies-dataset/notebook). I downloaded the data to my local machine and moved the csv file to the zipped data folder with the rest of the given csv files. I chose to utilize this data set because of the information it gave such as movie directors, top four stars per movie, and gross income of the movie.

Data Preparation

In [1]:
```
# Importing the proper packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from ast import literal_eval
from collections import Counter
```

In [2]:
```
# Import the csv
data = pd.read_csv('/Users/jdapeman/Documents/Flatiron/microsoft_eda/data/i
```

In [3]:
```
data.head()
```

Out[3]:

	Poster_Link	Series_Title	Released_Year	Certificate	Runtime	Genre
0	https://m.media-amazon.com/images/M/MV5BMDFkYT...	The Shawshank Redemption	1994	A	142 min	Drama
1	https://m.media-amazon.com/images/M/MV5BM2MyNj...	The Godfather	1972	A	175 min	Crime, Drama
2	https://m.media-amazon.com/images/M/MV5BMTMxNT...	The Dark Knight	2008	UA	152 min	Action, Crime, Drama
3	https://m.media-amazon.com/images/M/MV5BMWMwMG...	The Godfather: Part II	1974	A	202 min	Crime, Drama
4	https://m.media-amazon.com/images/M/MV5BMWU4N2...	12 Angry Men	1957	U	96 min	Crime, Drama

Renaming the column titles and droping column Poster_Link and Overview

data columns key:

- Movie_Title: Name of the movie
- Year_released: Year at which that movie released
- Certificate_Grade: Certificate earned by that movie
- Runtime: Total runtime of the movie
- Genre: Genre of the movie
- IMDB_Rating: Rating of the movie at IMDB site
- Meta_score: Score earned by the movie
- Director: Name of the Director
- Star1,2,3,4: Name of the Stars
- Vote_Count: Total number of votes
- Gross: Money earned by that movie

In [4]:
```
data.rename(columns={"Series_Title": "Movie_Title", "Released_Year": "Year_
"Certificate": "Certificate_Grade", "No_of_Votes": "Vote_Count"}, inplace=T
```

In [5]:
```
# Removing unecessary columns
data = data.drop('Poster_Link', axis=1)
data = data.drop('Overview', axis=1)
```

In [6]:
```
data.head()
```

Out[6]:

	Movie_Title	Year_released	Certificate_Grade	Runtime	Genre	IMDB_Rating	Meta_score	Dire
0	The Shawshank Redemption	1994	A	142 min	Drama	9.3	80.0	F Daral
1	The Godfather	1972	A	175 min	Crime, Drama	9.2	100.0	Fra I Cop
2	The Dark Knight	2008	UA	152 min	Action, Crime, Drama	9.0	84.0	Christo N
3	The Godfather: Part II	1974	A	202 min	Crime, Drama	9.0	90.0	Fra I Cop
4	12 Angry Men	1957	U	96 min	Crime, Drama	9.0	96.0	Sic Lu

In [7]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Movie_Title        1000 non-null   object 
 1   Year_released      1000 non-null   object 
 2   Certificate_Grade  899 non-null    object 
 3   Runtime            1000 non-null   object 
 4   Genre              1000 non-null   object 
 5   IMDB_Rating        1000 non-null   float64
 6   Meta_score         843 non-null    float64
 7   Director           1000 non-null   object 
 8   Star1              1000 non-null   object 
 9   Star2              1000 non-null   object 
 10  Star3              1000 non-null   object 
 11  Star4              1000 non-null   object 
 12  Vote_Count         1000 non-null   int64  
 13  Gross              831 non-null    object 
dtypes: float64(2), int64(1), object(11)
memory usage: 109.5+ KB
```

In [8]:
```
# Changing the 'Gross' column datatype from a object to an int64 and NAN to
data['Gross'] = data['Gross'].str.replace(',', '')
data['Gross'] = data['Gross'].replace(np.nan, 0)
data['Gross'] = data['Gross'].astype('int64')
```

In [9]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Movie_Title        1000 non-null   object 
 1   Year_released      1000 non-null   object 
 2   Certificate_Grade  899 non-null    object 
 3   Runtime            1000 non-null   object 
 4   Genre              1000 non-null   object 
 5   IMDB_Rating        1000 non-null   float64
 6   Meta_score         843 non-null    float64
 7   Director           1000 non-null   object 
 8   Star1              1000 non-null   object 
 9   Star2              1000 non-null   object 
 10  Star3              1000 non-null   object 
 11  Star4              1000 non-null   object 
 12  Vote_Count         1000 non-null   int64  
 13  Gross              1000 non-null   int64  
dtypes: float64(2), int64(2), object(10)
memory usage: 109.5+ KB
```

In [10]:
```
data['Gross'].describe()
```

Out[10]:
```
count    1.000000e+03
mean     5.653688e+07
std      1.032382e+08
min      0.000000e+00
25%      4.457098e+05
50%      1.070275e+07
75%      6.153989e+07
max      9.366622e+08
Name: Gross, dtype: float64
```

In [11]:
```
# replaced all 'Gross' values equaling zero with the median gross
data.loc[data['Gross'] == 0, 'Gross'] = 10702752
```

In [12]:
```
data.sort_values(by=['Movie_Title'], inplace=True)
data.head()
```

Out[12]:

	Movie_Title	Year_released	Certificate_Grade	Runtime	Genre	IMDB_Rating	Meta_score	
754	(500) Days of Summer	2009	UA	95 min	Comedy, Drama, Romance	7.7	76.0	
4	12 Angry Men	1957	U	96 min	Crime, Drama	9.0	96.0	
215	12 Years a Slave	2013	A	134 min	Biography, Drama, History	8.1	96.0	M
84	1917	2019	R	119 min	Drama, Thriller, War	8.3	78.0	
114	2001: A Space Odyssey	1968	U	149 min	Adventure, Sci-Fi	8.3	84.0	

In [13]:
```
# import data from tn.movie_budgets.csv specifically for the 'production_bu
df_budgets = pd.read_csv('/Users/jdapeman/Documents/Flatiron/microsoft_eda/
```

In [14]:
```
# Sorting values by movie name and renaming the column 'movie' to match 'Mo
df_budgets.sort_values(by=['movie'], inplace=True)
df_budgets.rename(columns = {'movie' : 'Movie_Title'}, inplace=True)
df_budgets.head()
```

Out[14]:

	id	release_date	Movie_Title	production_budget	domestic_gross	worldwide_gross
5115	16	Nov 20, 2015	#Horror	$1,500,000	$0	$0
3954	55	Jul 17, 2009	(500) Days of Summer	$7,500,000	$32,425,665	$34,439,060
4253	54	Mar 11, 2016	10 Cloverfield Lane	$5,000,000	$72,082,999	$108,286,422
3447	48	Nov 11, 2015	10 Days in a Madhouse	$12,000,000	$14,616	$14,616
3262	63	Mar 31, 1999	10 Things I Hate About You	$13,000,000	$38,177,966	$60,413,950

In [15]:
```
# Math operations cannot be done on the prodcution_budget, domestic_gross,
# they are strings. Here I stripped them of $ and thier commas, allowing me
# string to an int
df_budgets['production_budget'] = df_budgets['production_budget'].str.repla
df_budgets['production_budget'] = df_budgets['production_budget'].str.repla
df_budgets['production_budget'] = df_budgets['production_budget'].astype(in

df_budgets['domestic_gross'] = df_budgets['domestic_gross'].str.replace(r'$
df_budgets['domestic_gross'] = df_budgets['domestic_gross'].str.replace(r',
df_budgets['domestic_gross'] = df_budgets['domestic_gross'].astype(int)

df_budgets['worldwide_gross'] = df_budgets['worldwide_gross'].str.replace(r
df_budgets['worldwide_gross'] = df_budgets['worldwide_gross'].str.replace(r
df_budgets['worldwide_gross'] = df_budgets['worldwide_gross'].astype(int)

df_budgets.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5782 entries, 5115 to 2701
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_Title        5782 non-null   object
 3   production_budget  5782 non-null   int64 
 4   domestic_gross     5782 non-null   int64 
 5   worldwide_gross    5782 non-null   int64 
dtypes: int64(4), object(2)
memory usage: 316.2+ KB

/var/folders/yd/tybxdv4901xbv_l8x4p1g7tm0000gn/T/ipykernel_39243/92744644
3.py:4: FutureWarning: The default value of regex will change from True t
o False in a future version. In addition, single character regular expres
sions will *not* be treated as literal strings when regex=True.
  df_budgets['production_budget'] = df_budgets['production_budget'].str.r
eplace(r'$', '')
/var/folders/yd/tybxdv4901xbv_l8x4p1g7tm0000gn/T/ipykernel_39243/92744644
3.py:8: FutureWarning: The default value of regex will change from True t
o False in a future version. In addition, single character regular expres
sions will *not* be treated as literal strings when regex=True.
  df_budgets['domestic_gross'] = df_budgets['domestic_gross'].str.replace
(r'$', '')
/var/folders/yd/tybxdv4901xbv_l8x4p1g7tm0000gn/T/ipykernel_39243/92744644
3.py:12: FutureWarning: The default value of regex will change from True 
to False in a future version. In addition, single character regular expre
ssions will *not* be treated as literal strings when regex=True.
  df_budgets['worldwide_gross'] = df_budgets['worldwide_gross'].str.repla
ce(r'$', '')
```

In [16]:
```
# I combined the two dataframes, data and df_budgets. For my perameters I c
# 'Movie_title' because I wanted to keep the titles from both data sets
profit_per_cat_data = pd.merge(data, df_budgets, left_on = 'Movie_Title', r
profit_per_cat_data.head()
```

Out[16]:

	Movie_Title	Year_released	Certificate_Grade	Runtime	Genre	IMDB_Rating	Meta_score	Di
0	(500) Days of Summer	2009	UA	95 min	Comedy, Drama, Romance	7.7	76.0	
1	12 Angry Men	1957	U	96 min	Crime, Drama	9.0	96.0	S I
2	12 Years a Slave	2013	A	134 min	Biography, Drama, History	8.1	96.0	Mc(
3	2001: A Space Odyssey	1968	U	149 min	Adventure, Sci-Fi	8.3	84.0	S K
4	21 Grams	2003	UA	124 min	Crime, Drama, Thriller	7.6	70.0	Alej Ii

In [17]:
```
profit_per_cat_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 521 entries, 0 to 520
Data columns (total 19 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Movie_Title        521 non-null    object 
 1   Year_released      521 non-null    object 
 2   Certificate_Grade  511 non-null    object 
 3   Runtime            521 non-null    object 
 4   Genre              521 non-null    object 
 5   IMDB_Rating        521 non-null    float64
 6   Meta_score         503 non-null    float64
 7   Director           521 non-null    object 
 8   Star1              521 non-null    object 
 9   Star2              521 non-null    object 
 10  Star3              521 non-null    object 
 11  Star4              521 non-null    object 
 12  Vote_Count         521 non-null    int64  
 13  Gross              521 non-null    int64  
 14  id                 521 non-null    int64  
 15  release_date       521 non-null    object 
 16  production_budget  521 non-null    int64  
 17  domestic_gross     521 non-null    int64  
 18  worldwide_gross    521 non-null    int64  
dtypes: float64(2), int64(6), object(11)
memory usage: 81.4+ KB
```

In [18]:
```
# Dropping rows with zero value for 'domestic_gross'. There are five zero v
profit_per_cat_data = profit_per_cat_data[profit_per_cat_data['domestic_gro
```

In [19]:
```
# created a column titled 'Total_Domestic_Gross', consisting of the average
profit_per_cat_data['Total_Domestic_Gross'] = (profit_per_cat_data['Gross']
profit_per_cat_data.head()
```

Out[19]:

	Movie_Title	Year_released	Certificate_Grade	Runtime	Genre	IMDB_Rating	Meta_score	Di
0	(500) Days of Summer	2009	UA	95 min	Comedy, Drama, Romance	7.7	76.0	
2	12 Years a Slave	2013	A	134 min	Biography, Drama, History	8.1	96.0	Mc(
3	2001: A Space Odyssey	1968	U	149 min	Adventure, Sci-Fi	8.3	84.0	S K
4	21 Grams	2003	UA	124 min	Crime, Drama, Thriller	7.6	70.0	Alej I
5	25th Hour	2002	R	135 min	Drama	7.6	68.0	

In [20]:
```
# From here I dropped the Gross, domestic_gross, and id columns
profit_per_cat_data = profit_per_cat_data.drop('Gross', axis=1)
profit_per_cat_data = profit_per_cat_data.drop('domestic_gross', axis=1)
profit_per_cat_data = profit_per_cat_data.drop('id', axis=1)
profit_per_cat_data.columns
```

Out[20]:
```
Index(['Movie_Title', 'Year_released', 'Certificate_Grade', 'Runtime', 'G
enre',
       'IMDB_Rating', 'Meta_score', 'Director', 'Star1', 'Star2', 'Star
3',
       'Star4', 'Vote_Count', 'release_date', 'production_budget',
       'worldwide_gross', 'Total_Domestic_Gross'],
      dtype='object')
```

In [21]:
```
# Now we can make a profit column
profit_per_cat_data['Profit'] = profit_per_cat_data['worldwide_gross'] - pr
```

In [22]:
```
# Seperating the genre types in the 'Genre' column
profit_per_cat_data['Genre'] = profit_per_cat_data['Genre'].str.strip('()')
profit_per_cat_data = profit_per_cat_data.explode('Genre')
profit_per_cat_data.head()
```

Out[22]:

	Movie_Title	Year_released	Certificate_Grade	Runtime	Genre	IMDB_Rating	Meta_score	Dir
0	(500) Days of Summer	2009	UA	95 min	Comedy	7.7	76.0	
0	(500) Days of Summer	2009	UA	95 min	Drama	7.7	76.0	
0	(500) Days of Summer	2009	UA	95 min	Romance	7.7	76.0	
2	12 Years a Slave	2013	A	134 min	Biography	8.1	96.0	McC
2	12 Years a Slave	2013	A	134 min	Drama	8.1	96.0	McC

In [23]:
```
profit_per_cat_data['Genre'].value_counts()
```

Out[23]:
```
 Drama         208
Drama          138
Action         106
 Adventure      89
 Thriller       73
Comedy          60
 Sci-Fi         58
Biography       56
 Romance        54
Adventure       49
Crime           48
 Comedy         47
 Crime          42
Animation       41
 Mystery        38
 Fantasy        31
 Family         30
 History        29
 War            21
 Sport          15
 Music          12
Horror          12
 Biography      11
 Action         10
 Western        10
 Horror          9
 Musical         7
 Film-Noir       4
Mystery          3
Western          2
Family           1
Name: Genre, dtype: int64
```

In [24]:
```
# There are double counts of genres becasue of whitespace attached to thier
profit_per_cat_data['Genre'] = profit_per_cat_data['Genre'].str.strip()
```

In [25]:
```
profit_per_cat_data['Genre'].value_counts()
```

Out[25]:
```
Drama          346
Adventure      138
Action         116
Comedy         107
Crime           90
Thriller        73
Biography       67
Sci-Fi          58
Romance         54
Mystery         41
Animation       41
Fantasy         31
Family          31
History         29
Horror          21
War             21
Sport           15
Music           12
Western         12
Musical          7
Film-Noir        4
Name: Genre, dtype: int64
```

In [26]:
```
# There is a 'PG' value in the 'Year_relesed' column for the movie Apollo 1
# change the year from a string to an integer.
profit_per_cat_data.loc[profit_per_cat_data['Year_released'] == 'PG', 'Year
profit_per_cat_data[profit_per_cat_data['Year_released'] == 'PG'].value_cou
profit_per_cat_data['Year_released'] = profit_per_cat_data['Year_released']
```

In [27]:
```
# Finding what movies are relased after 1999
saturation_data = profit_per_cat_data[profit_per_cat_data['Year_released']
saturation_data_v2 = saturation_data[['Year_released', 'Genre']].copy()
saturation_data_v2
```

Out[27]:

	Year_released	Genre
0	2009	Comedy
0	2009	Drama
0	2009	Romance
2	2013	Biography
2	2013	Drama
...	...	...
518	2016	Adventure
518	2016	Comedy
520	2006	Drama
520	2006	Thriller
520	2006	War

716 rows × 2 columns

In [28]:
```
# Creating a table of the top 5 profiable movies to check for oversaturatio
saturation_data_v3 = saturation_data_v2[(saturation_data_v2['Genre'] == "Ad
                    (saturation_data_v2['Genre'] == "Family") | (saturation_d
                    (saturation_data_v2['Genre'] == "Animation")]
saturation_data_v3
```

Out[28]:

	Year_released	Genre
6	2006	Action
7	2007	Action
15	2013	Fantasy
23	2000	Adventure
32	2006	Action
...	...	...
514	2014	Adventure
517	2009	Adventure
517	2009	Fantasy
518	2016	Animation
518	2016	Adventure

211 rows × 2 columns

In [29]:
```
#Top directors for movies after the 21st century
directors = profit_per_cat_data[profit_per_cat_data['Year_released'] > 1999
directors_v2 = directors[['Year_released', 'Genre', 'Director', 'Profit']].
directors_v3 = directors_v2[(directors_v2['Genre'] == "Adventure") | (direc
                        (directors_v2['Genre'] == "Family") | (directors_v2['Genr
                        (directors_v2['Genre'] == "Animation")]
directors_v3
```

Out[29]:

	Year_released	Genre	Director	Profit
6	2006	Action	Zack Snyder	394161935
7	2007	Action	James Mangold	23171825
15	2013	Fantasy	Richard Curtis	77309178
23	2000	Adventure	Cameron Crowe	-12628809
32	2006	Action	Mel Gibson	81032272
...	...	...	...	...
514	2014	Adventure	Bryan Singer	547862775
517	2009	Adventure	Ruben Fleischer	78636596
517	2009	Fantasy	Ruben Fleischer	78636596
518	2016	Animation	Byron Howard	869429616
518	2016	Adventure	Byron Howard	869429616

211 rows × 4 columns

In [30]:
```
# selecting Action movie directors
b = directors_v3[directors_v3['Genre'] == 'Action']
b_v2 = b.sort_values(ascending=False, by=['Profit']).head()
b_v2
```

Out[30]:

	Year_released	Genre	Director	Profit
37	2009	Action	James Cameron	2351345279
38	2018	Action	Anthony Russo	1748134200
374	2012	Action	Joss Whedon	1292935897
438	2003	Action	Peter Jackson	1047403341
207	2018	Action	Brad Bird	1042520711

In [31]:
```
# selecting Adventure movie directors
c = directors_v3[directors_v3['Genre'] == 'Adventure']
c_v2 = c.sort_values(ascending=False, by=['Profit']).head()
c_v2
```

Out[31]:

	Year_released	Genre	Director	Profit
37	2009	Adventure	James Cameron	2351345279
38	2018	Adventure	Anthony Russo	1748134200
374	2012	Adventure	Joss Whedon	1292935897
438	2003	Adventure	Peter Jackson	1047403341
207	2018	Adventure	Brad Bird	1042520711

In [32]:
```
# selecting Family movie directors
d = directors_v3[directors_v3['Genre'] == 'Family']
d_v2 = d.sort_values(ascending=False, by=['Profit']).head()
d_v2
```

Out[32]:

	Year_released	Genre	Director	Profit
185	2005	Family	Mike Newell	747099794
186	2009	Family	David Yates	685213767
187	2004	Family	Alfonso Cuarón	666907323
94	2017	Family	Lee Unkrich	623008101
512	2017	Family	Stephen Chbosky	284604712

In [33]:
```
# selecting Fantasy movie directors
e = directors_v3[directors_v3['Genre'] == 'Fantasy']
e_v2 = e.sort_values(ascending=False, by=['Profit']).head()
e_v2
```

Out[33]:

	Year_released	Genre	Director	Profit
37	2009	Fantasy	James Cameron	2351345279
417	2012	Fantasy	Peter Jackson	767003568
185	2005	Fantasy	Mike Newell	747099794
418	2013	Fantasy	Peter Jackson	710366855
187	2004	Fantasy	Alfonso Cuarón	666907323

In [34]:
```
# selecting Animation movie directors
f = directors_v3[directors_v3['Genre'] == 'Animation']
f_v2 = f.sort_values(ascending=False, by=['Profit']).head()
f_v2
```

Out[34]:

	Year_released	Genre	Director	Profit
207	2018	Animation	Brad Bird	1042520711
518	2016	Animation	Byron Howard	869429616
495	2010	Animation	Lee Unkrich	868879522
147	2003	Animation	Andrew Stanton	842429370
211	2015	Animation	Pete Docter	679235992

In [35]:
```
# making a data frame of the highest profiting directors in each genre
to_concat = [b_v2, c_v2, d_v2, e_v2, f_v2]
new_director_list = pd.concat(to_concat)
new_director_list
```

Out[35]:

	Year_released	Genre	Director	Profit
37	2009	Action	James Cameron	2351345279
38	2018	Action	Anthony Russo	1748134200
374	2012	Action	Joss Whedon	1292935897
438	2003	Action	Peter Jackson	1047403341
207	2018	Action	Brad Bird	1042520711
37	2009	Adventure	James Cameron	2351345279
38	2018	Adventure	Anthony Russo	1748134200
374	2012	Adventure	Joss Whedon	1292935897
438	2003	Adventure	Peter Jackson	1047403341
207	2018	Adventure	Brad Bird	1042520711
185	2005	Family	Mike Newell	747099794
186	2009	Family	David Yates	685213767
187	2004	Family	Alfonso Cuarón	666907323
94	2017	Family	Lee Unkrich	623008101
512	2017	Family	Stephen Chbosky	284604712
37	2009	Fantasy	James Cameron	2351345279
417	2012	Fantasy	Peter Jackson	767003568
185	2005	Fantasy	Mike Newell	747099794
418	2013	Fantasy	Peter Jackson	710366855
187	2004	Fantasy	Alfonso Cuarón	666907323
207	2018	Animation	Brad Bird	1042520711
518	2016	Animation	Byron Howard	869429616
495	2010	Animation	Lee Unkrich	868879522
147	2003	Animation	Andrew Stanton	842429370
211	2015	Animation	Pete Docter	679235992

In [36]:
```
#Profitability of top 5 movies
top_5_profits = profit_per_cat_data[(profit_per_cat_data['Genre'] == "Adven
                      (profit_per_cat_data['Genre'] == "Family") | (profit_per_
                      (profit_per_cat_data['Genre'] == "Animation")]
```

In [37]:
```
top_5_profits.head()
```

Out[37]:

	Movie_Title	Year_released	Certificate_Grade	Runtime	Genre	IMDB_Rating	Meta_score	Dir
3	2001: A Space Odyssey	1968	U	149 min	Adventure	8.3	84.0	St Ku
6	300	2006	A	117 min	Action	7.6	52.0	S
7	3:10 to Yuma	2007	A	122 min	Action	7.7	76.0	J Ma
10	A Christmas Story	1983	U	93 min	Family	7.9	77.0	
15	About Time	2013	R	123 min	Fantasy	7.8	55.0	Ri (

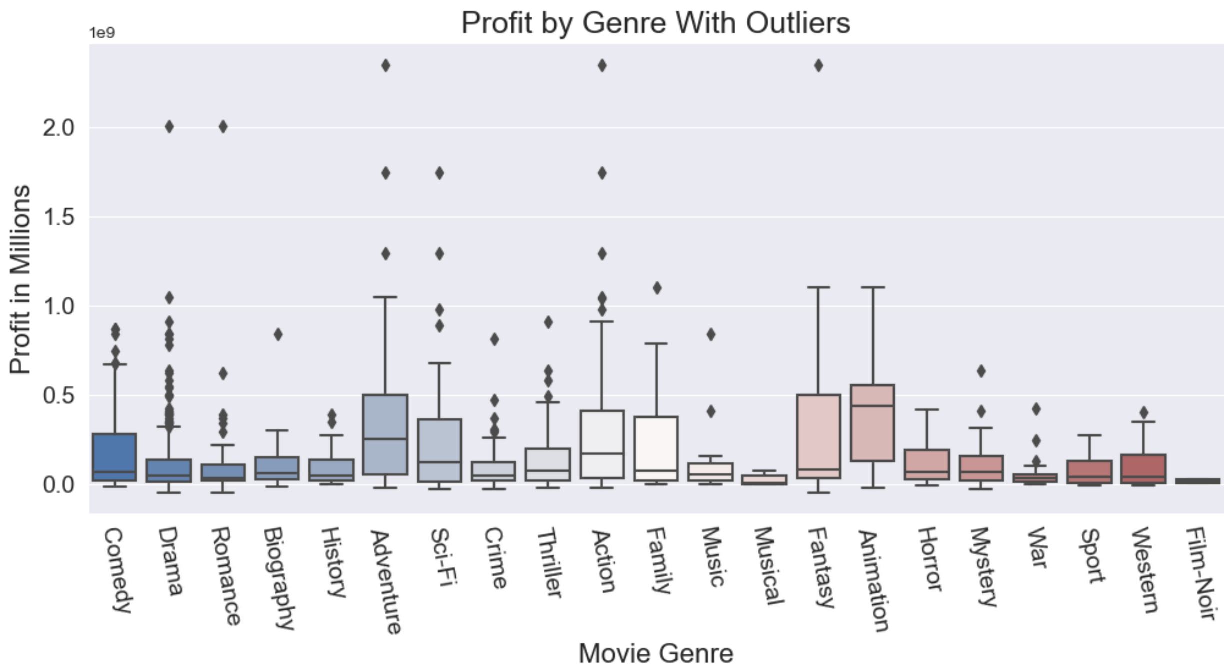
Data Analysis

After all the data has been assembled I plotted my visualizations to determine the relationships between the elements I was investigating and their relationships. Below are the visual descriptions of my three key questions:

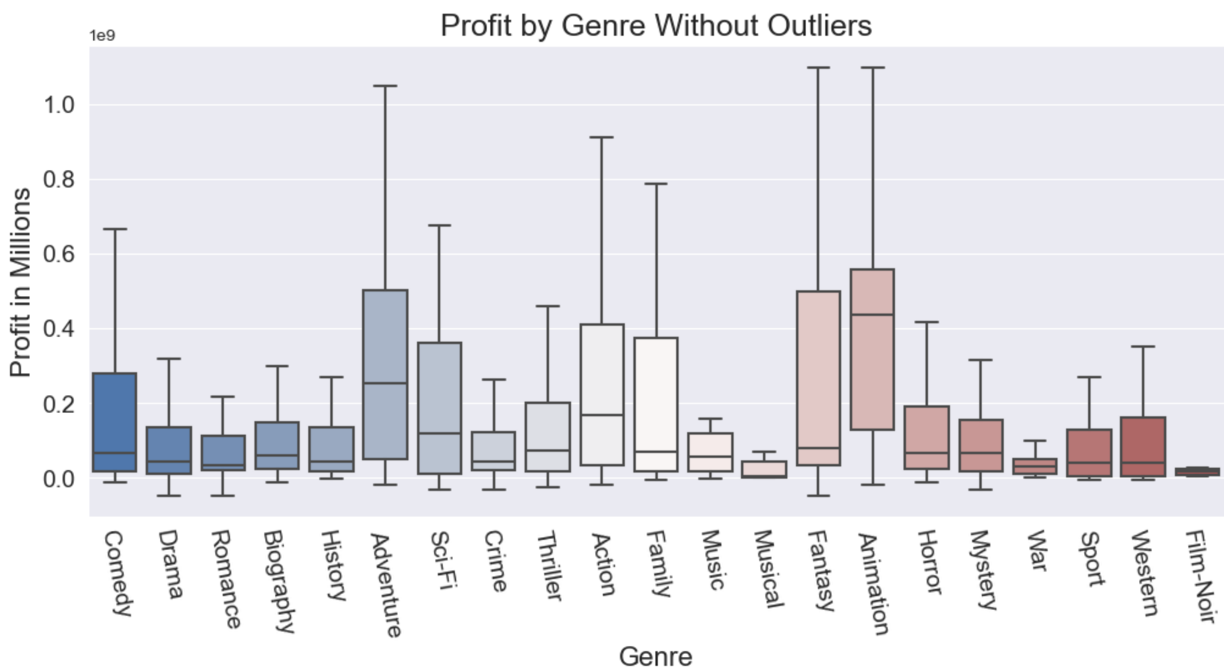
- What movie genres produce the highest profits?
- What movie genres should be produced to avoid competition?
- What director should be hired to produce their movies?

Which genre of movies genreates the highest profit?

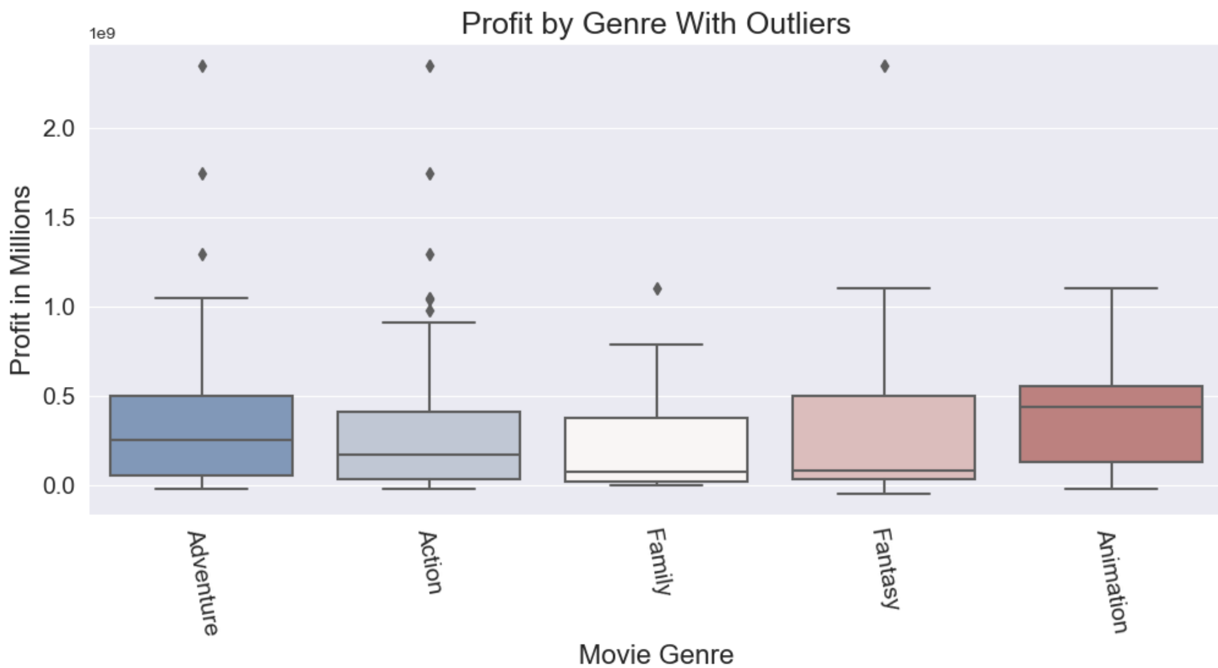
In [38]:
```
plt.figure(figsize=(12,5))
sns.set_style('darkgrid')
sns.boxplot(x='Genre', y='Profit', data=profit_per_cat_data, palette='vlag'
plt.xticks(rotation=-80)
plt.ylabel('Profit in Millions', fontsize=16)
plt.xlabel('Movie Genre', fontsize = 16)
plt.title('Profit by Genre With Outliers', fontsize = 18)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14);
```



In [39]:
```
plt.figure(figsize=(12,5))
sns.set_style('darkgrid')
sns.boxplot(x='Genre', y='Profit', data=profit_per_cat_data, showfliers=Fal
plt.xticks(rotation=-80)
plt.ylabel('Profit in Millions', fontsize=16)
plt.xlabel('Genre', fontsize = 16)
plt.title('Profit by Genre Without Outliers', fontsize = 18)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14);
```



In [40]:
```
plt.figure(figsize=(12,5))
sns.set_style('darkgrid')
sns.boxplot(x='Genre', y='Profit', data=top_5_profits, palette='vlag')
plt.xticks(rotation=-80)
plt.ylabel('Profit in Millions', fontsize=16)
plt.xlabel('Movie Genre', fontsize = 16)
plt.title('Profit by Genre With Outliers', fontsize = 18)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14);
```

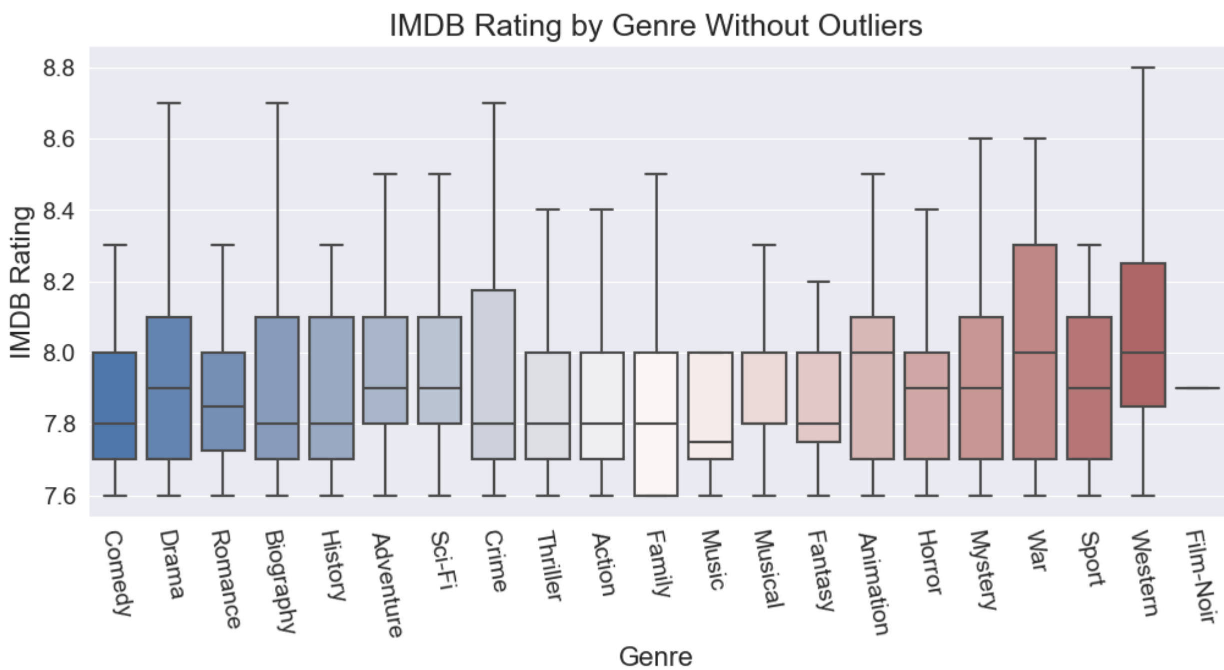


In [41]:
```
plt.figure(figsize=(12,5))
sns.set_style('darkgrid')
sns.boxplot(x='Genre', y='Profit', data=top_5_profits, showfliers=False, pa
plt.xticks(rotation=-80)
plt.ylabel('Profit in Millions', fontsize=16)
plt.xlabel('Genre', fontsize = 16)
plt.title('Profit by Genre Without Outliers', fontsize = 18)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14);
```

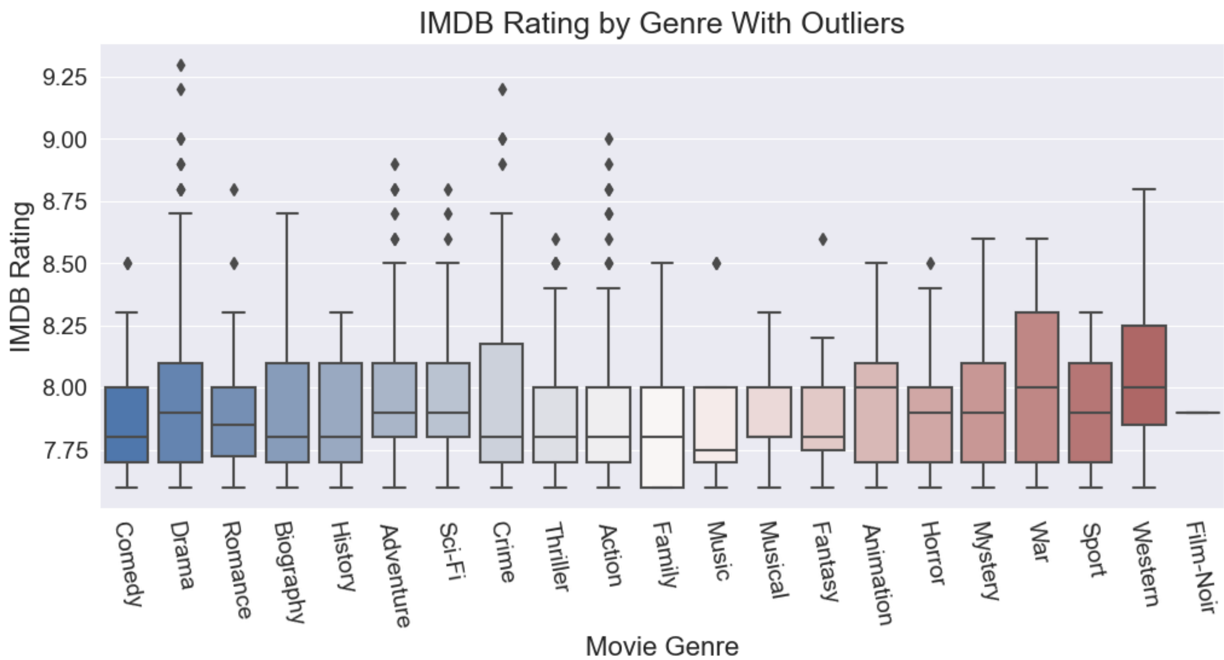
The results indicate that the highest proffiting genres are Adventure, Action, Family, Fantasy and Animation.

Lets take a quick look at IMDB rating for the movie genres.

In [42]:
```
plt.figure(figsize=(12,5))
sns.set_style('darkgrid')
sns.boxplot(x='Genre', y='IMDB_Rating', data=profit_per_cat_data, showflier
plt.xticks(rotation=-80)
plt.ylabel('IMDB Rating', fontsize=16)
plt.xlabel('Genre', fontsize = 16)
plt.title('IMDB Rating by Genre Without Outliers', fontsize = 18)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14);
```



In [43]:
```
plt.figure(figsize=(12,5))
sns.set_style('darkgrid')
sns.boxplot(x='Genre', y='IMDB_Rating', data=profit_per_cat_data, palette='
plt.xticks(rotation=-80)
plt.ylabel('IMDB Rating', fontsize=16)
plt.xlabel('Movie Genre', fontsize = 16)
plt.title('IMDB Rating by Genre With Outliers', fontsize = 18)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14);
```



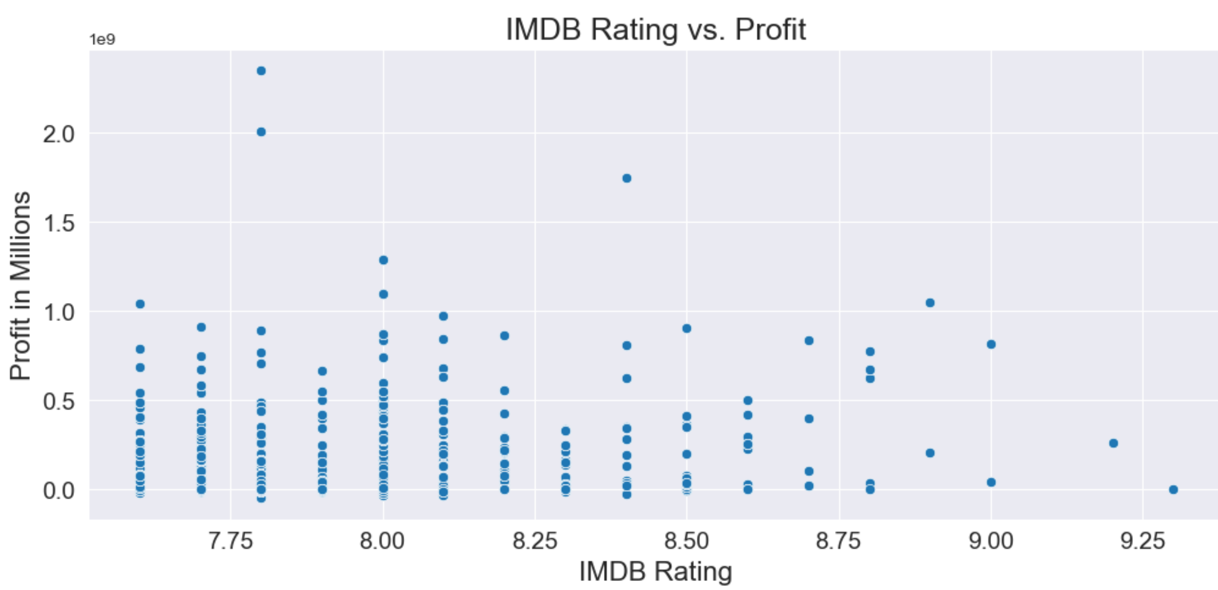
Is there any correlation between IMDB rating and Proffitability?

In [44]:
```
correlation_data = profit_per_cat_data[['Profit', 'IMDB_Rating']].copy()
```

In [45]:
```
# make a scatterplot of x = 'IMDB_Rating', y = 'Profit' from correlation_da
plt.figure(figsize=(12, 5))
sns.set_style('darkgrid')
sns.scatterplot(data=correlation_data, x='IMDB_Rating', y='Profit', palette
plt.xlabel('IMDB Rating', fontsize=16)
plt.ylabel('Profit in Millions', fontsize=16)
plt.title('IMDB Rating vs. Profit', fontsize=18)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
```

```
/var/folders/yd/tybxdv4901xbv_l8x4p1g7tm0000gn/T/ipykernel_39243/39698225
65.py:4: UserWarning: Ignoring `palette` because no `hue` variable has be
en assigned.
  sns.scatterplot(data=correlation_data, x='IMDB_Rating', y='Profit', pal
ette='vlag')
```

Out[45]:
```
(array([-5.0e+08,  0.0e+00,  5.0e+08,  1.0e+09,  1.5e+09,  2.0e+09,
         2.5e+09]),
 [Text(0, -500000000.0, '-0.5'),
  Text(0, 0.0, '0.0'),
  Text(0, 500000000.0, '0.5'),
  Text(0, 1000000000.0, '1.0'),
  Text(0, 1500000000.0, '1.5'),
  Text(0, 2000000000.0, '2.0'),
  Text(0, 2500000000.0, '2.5')])
```



In [46]:
```
correlation_data.corr()
```

Out[46]:

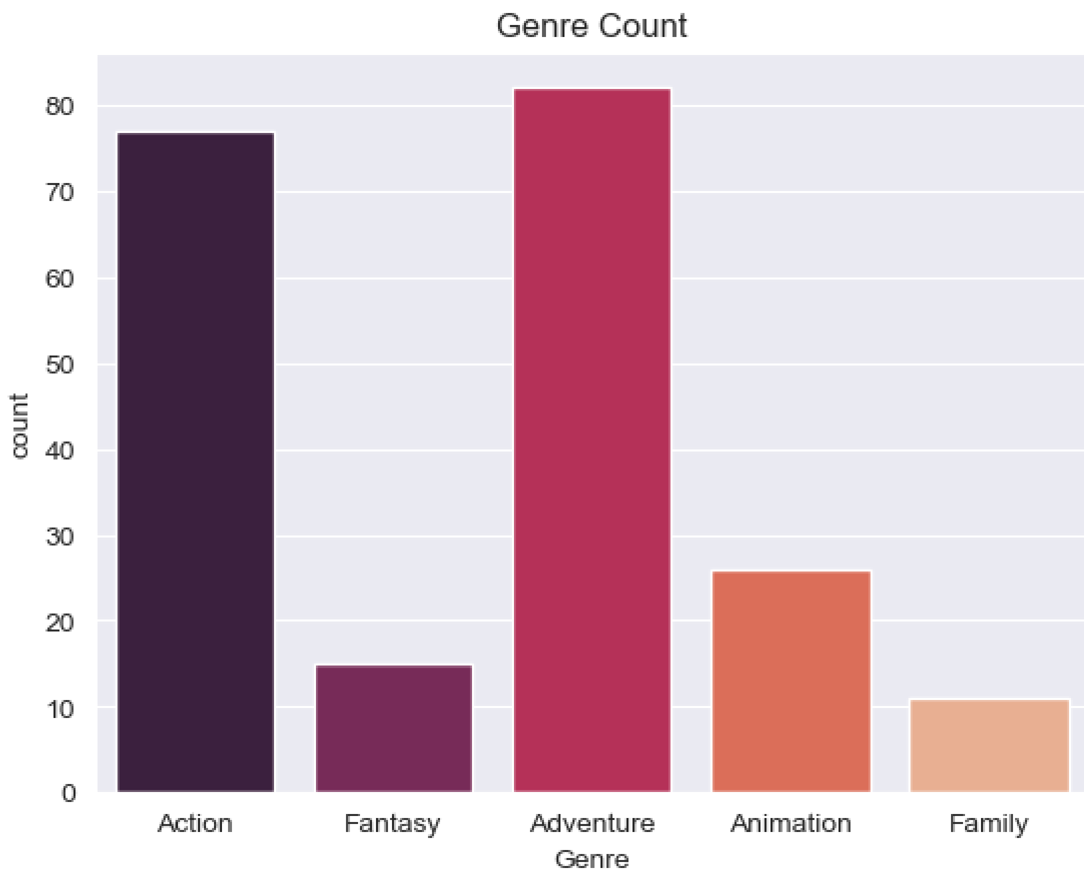
	Profit	IMDB_Rating
Profit	1.000000	0.163062
IMDB_Rating	0.163062	1.000000

From the results there seems to be little correlation between IMDB ratings and profit.

Which of the 5 most popular genre of movies would stand out against competeing movie types?

In [47]:
```
sns.countplot(x=saturation_data_v3['Genre'], palette='rocket').set_title("G
```

Out[47]:
```
Text(0.5, 1.0, 'Genre Count')
```



The results indicate that Action and Adventure are genres relased often in the 21st century. Fantasy

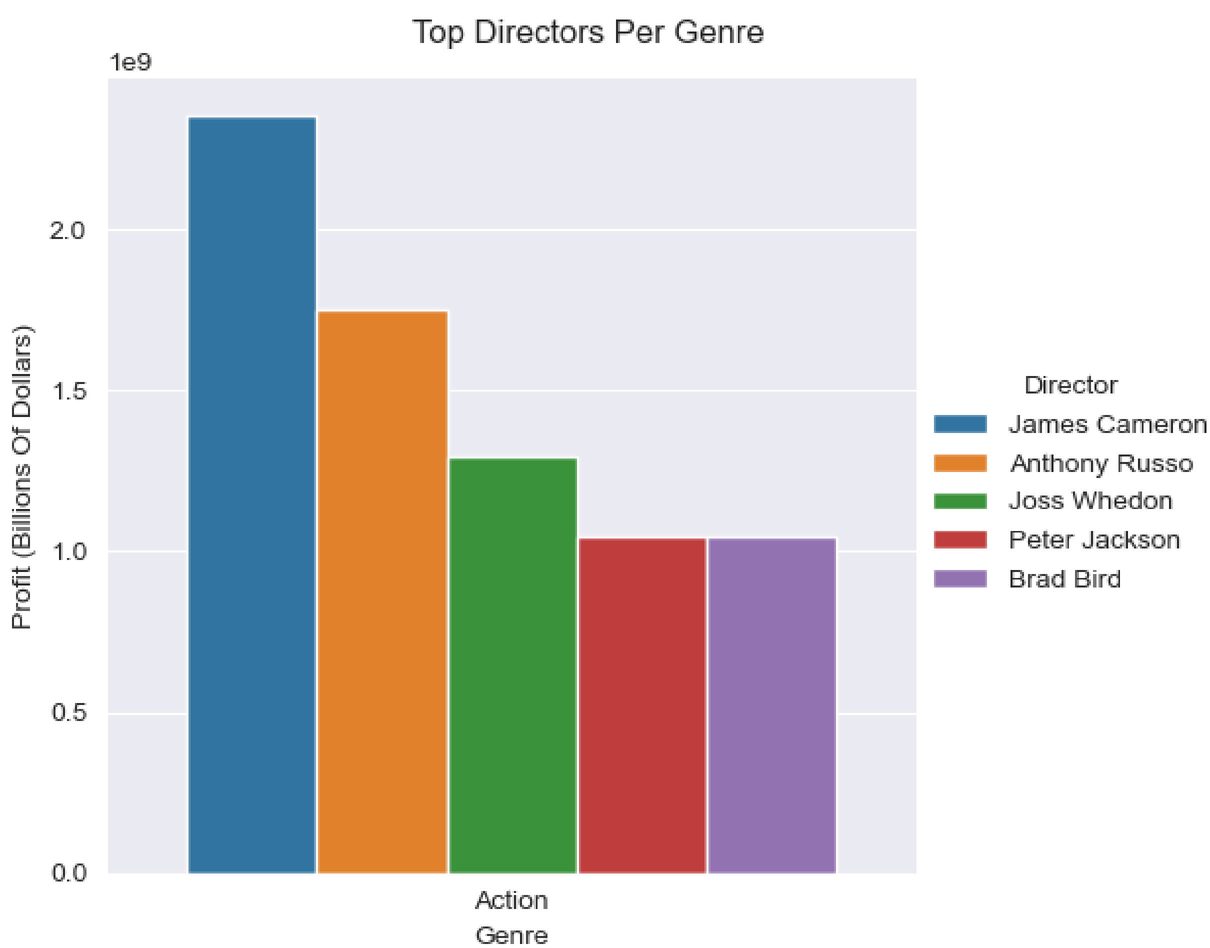
Animation and Family are genred that are not oversaturated in the market.

Who are the top five directors to hire for action, fantasy, adventure, animation, and

family movies?

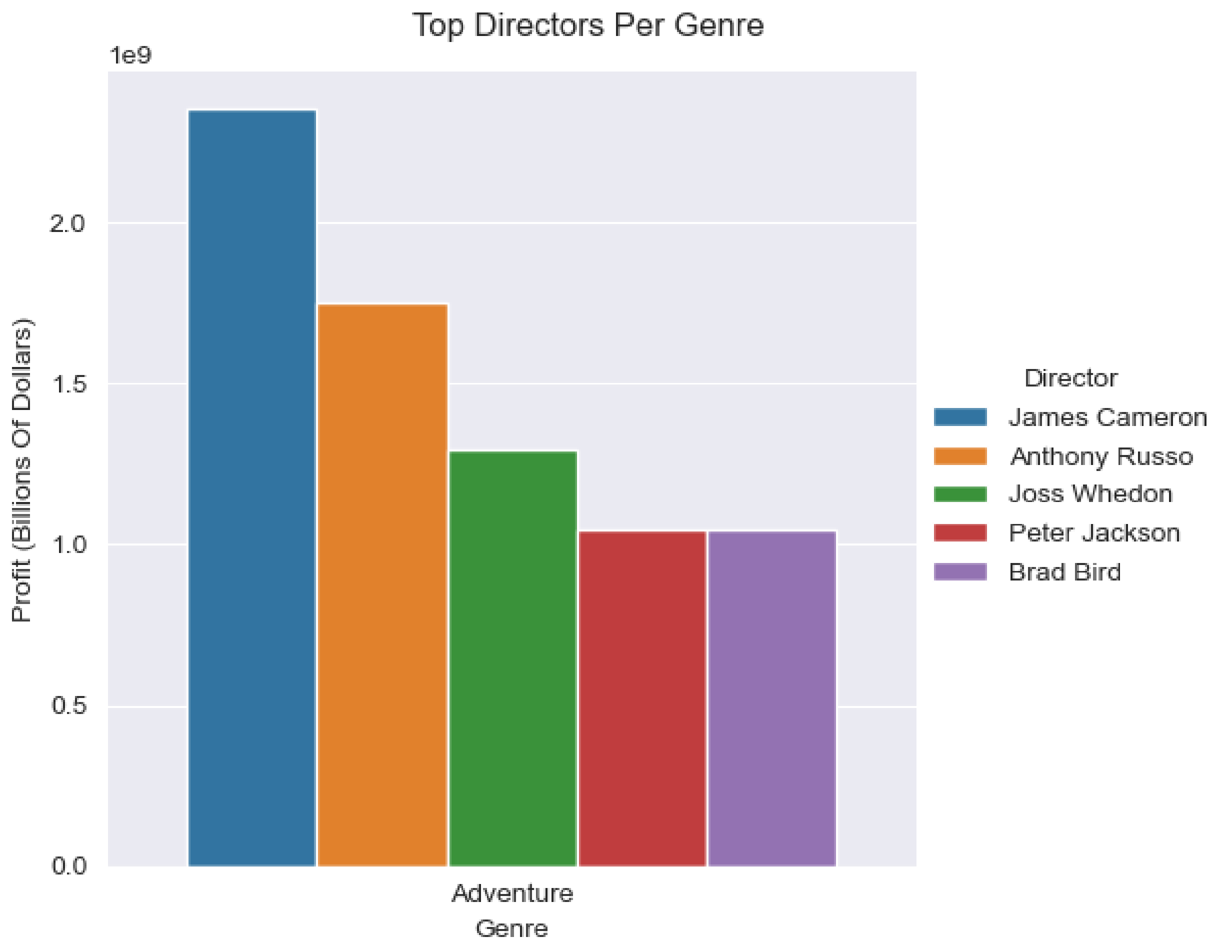
In [48]:
```
cat = sns.catplot(data=b_v2, x='Genre', y='Profit', hue='Director', kind="b
cat.fig.subplots_adjust(top=0.92)
cat.fig.suptitle('Top Directors Per Genre')
cat.set_ylabels('Profit (Billions Of Dollars)')

plt.show()
```



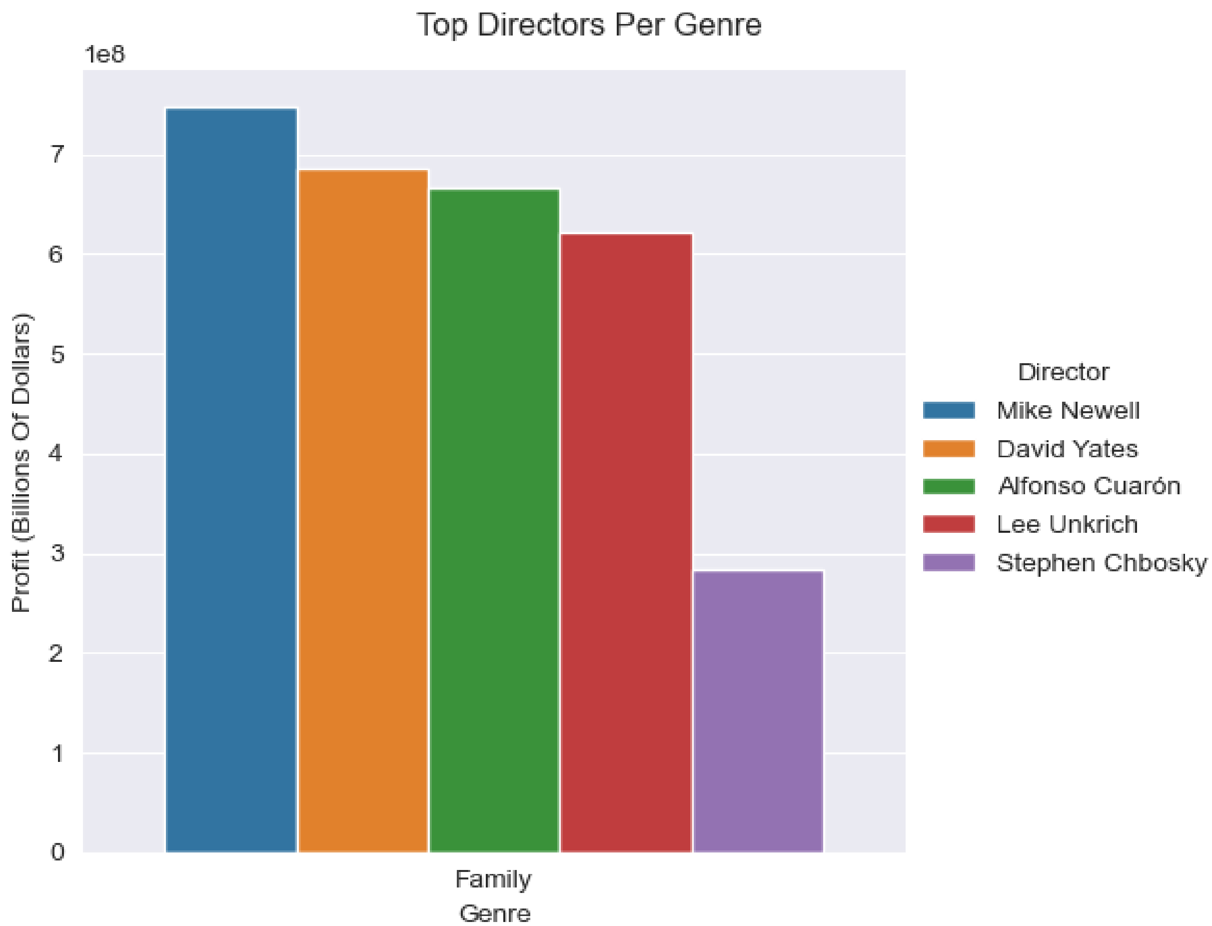
In [49]:
```
cat = sns.catplot(data=c_v2, x='Genre', y='Profit', hue='Director', kind="b
cat.fig.subplots_adjust(top=0.92)
cat.fig.suptitle('Top Directors Per Genre')
cat.set_ylabels('Profit (Billions Of Dollars)')

plt.show()
```



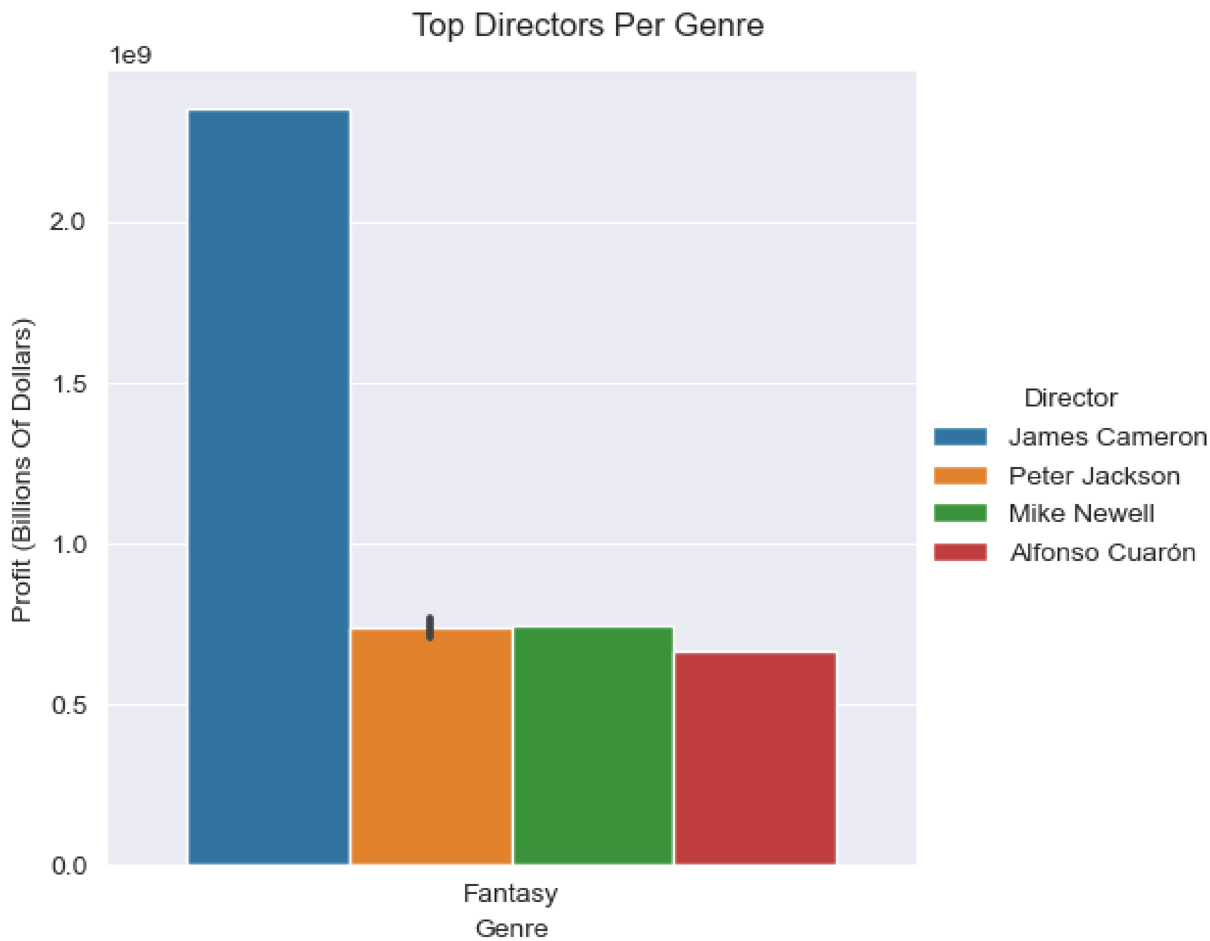
In [50]:
```
cat = sns.catplot(data=d_v2, x='Genre', y='Profit', hue='Director', kind="b
cat.fig.subplots_adjust(top=0.92)
cat.fig.suptitle('Top Directors Per Genre')
cat.set_ylabels('Profit (Billions Of Dollars)')

plt.show()
```



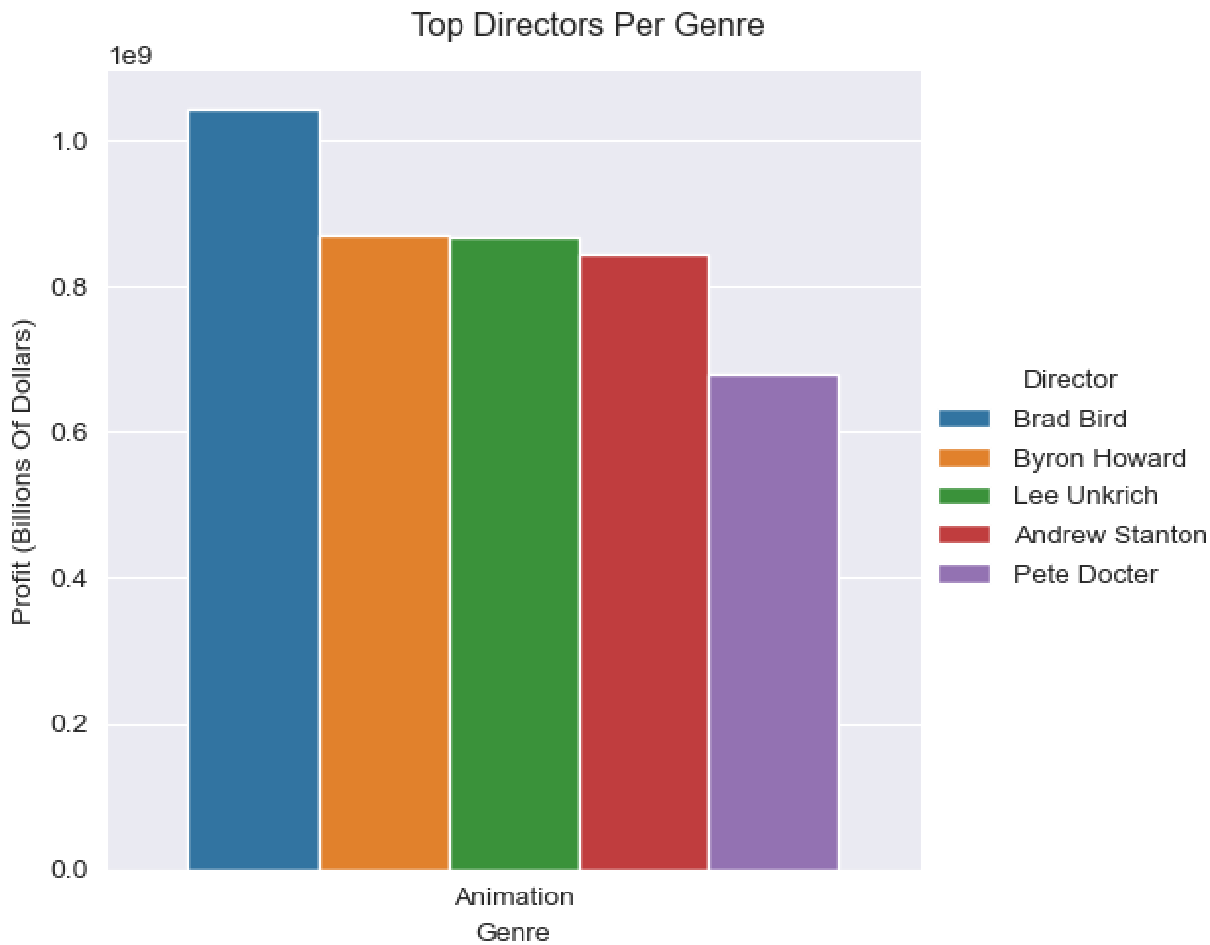
In [51]:
```
cat = sns.catplot(data=e_v2, x='Genre', y='Profit', hue='Director', kind="b
cat.fig.subplots_adjust(top=0.92)
cat.fig.suptitle('Top Directors Per Genre')
cat.set_ylabels('Profit (Billions Of Dollars)')

plt.show()
```



In [52]:
```
cat = sns.catplot(data=f_v2, x='Genre', y='Profit', hue='Director', kind="b
cat.fig.subplots_adjust(top=0.92)
cat.fig.suptitle('Top Directors Per Genre')
cat.set_ylabels('Profit (Billions Of Dollars)')

plt.show()
```



The findings indicate that for action, fantasy, and adventure movies James Cameron should be hired as director. For Animation movies, Brad Bird should be hired, and for Family movies, Microsoft should recruit Mike Newell.

Conclusion

I would suggest that Microsoft produce the following types of movies.

Movie Option #1

- A fantasy movie directed by James Cameron.
- Potential profit of $2,351,345,279

Movie Option #2

- An animation movie directed by Brad Bird.
- Potential profit of $1,042,520,711

Movie Option #3

- A family movie directed by Mike Newell.
- Potential profit of $747,099,794