Investigate_a_Dataset

February 17, 2021

1 TMDB 5000 Movie Dataset- Genres- The higher the ranking the more revenue

Does the higher the ranking of a movie have a direct correlation to the revenue made? Which Genre has made the most money in the 2000s?

Introduction: The hypothesis is: when the reviews are high for a movie, does the movie make more money? In the 2000s which genre made the most money?

```
In [1]: import pandas as pd
        import numpy as np
        import datetime as dt
        % matplotlib inline
        df_movies = pd.read_csv('tmdb-movies.csv')
```

Data Wrangling

> **Tip**: In this section of the report, you will load in the data, check for cleanliness, and then trim and clean your dataset for analysis. Make sure that you document your steps carefully and justify your cleaning decisions.

1.0.1 General Properties

```
In [2]: df_movies.head()
```

```
Out[2]:       id    imdb_id  popularity     budget     revenue  \
        0  135397  tt0369610   32.985763  150000000  1513528810
        1   76341  tt1392190   28.419936  150000000   378436354
        2  262500  tt2908446   13.112507  110000000   295238201
        3  140607  tt2488496   11.173104  200000000  2068178225
        4  168259  tt2820852    9.335014  190000000  1506249360

                         original_title  \
        0                Jurassic World
        1             Mad Max: Fury Road
        2                      Insurgent
        3  Star Wars: The Force Awakens
        4                      Furious 7
```

```
                                                cast  \
0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...
2  Shailene Woodley|Theo James|Kate Winslet|Ansel...
3  Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...
4  Vin Diesel|Paul Walker|Jason Statham|Michelle ...

                                        homepage          director  \
0                    http://www.jurassicworld.com/   Colin Trevorrow
1                      http://www.madmaxmovie.com/     George Miller
2    http://www.thedivergentseries.movie/#insurgent  Robert Schwentke
3  http://www.starwars.com/films/star-wars-episod...       J.J. Abrams
4                          http://www.furious7.com/         James Wan

                         tagline      ...       \
0             The park is open.      ...
1            What a Lovely Day.      ...
2    One Choice Can Destroy You      ...
3  Every generation has a story.     ...
4           Vengeance Hits Home      ...

                                        overview runtime  \
0  Twenty-two years after the events of Jurassic ...     124
1  An apocalyptic story set in the furthest reach...     120
2  Beatrice Prior must confront her inner demons ...     119
3  Thirty years after defeating the Galactic Empi...     136
4  Deckard Shaw seeks revenge against Dominic Tor...     137

                                     genres  \
0  Action|Adventure|Science Fiction|Thriller
1  Action|Adventure|Science Fiction|Thriller
2         Adventure|Science Fiction|Thriller
3   Action|Adventure|Science Fiction|Fantasy
4                     Action|Crime|Thriller

                                production_companies release_date vote_count  \
0  Universal Studios|Amblin Entertainment|Legenda...       6/9/15       5562
1  Village Roadshow Pictures|Kennedy Miller Produ...      5/13/15       6185
2  Summit Entertainment|Mandeville Films|Red Wago...      3/18/15       2480
3          Lucasfilm|Truenorth Productions|Bad Robot     12/15/15       5292
4  Universal Pictures|Original Film|Media Rights ...       4/1/15       2947

   vote_average  release_year    budget_adj   revenue_adj
0           6.5          2015  1.379999e+08  1.392446e+09
1           7.1          2015  1.379999e+08  3.481613e+08
2           6.3          2015  1.012000e+08  2.716190e+08
3           7.5          2015  1.839999e+08  1.902723e+09
```

```
4           7.3          2015  1.747999e+08  1.385749e+09

[5 rows x 21 columns]
```

Tip: You should *not* perform too many operations in each cell. Create cells freely to explore your data. One option that you can take with this project is to do a lot of explorations in an initial notebook. These don't have to be organized, but make sure you use enough comments to understand the purpose of each code cell. Then, after you're done with your analysis, create a duplicate notebook where you will trim the excess and organize your steps so that you have a flowing, cohesive report.

Tip: Make sure that you keep your reader informed on the steps that you are taking in your investigation. Follow every code cell, or every set of related code cells, with a markdown cell to describe to the reader what was found in the preceding cell(s). Try to make it so that the reader can then understand what they will be seeing in the following cell(s).

1.0.2 Data Cleaning (Replace this with more specific notes!)

In [3]:
```
# Firstly we need to remove ""|"" from the list of cast and replace it with ","
df_movies['cast'] = df_movies['cast'].str.replace('|',',')
df_movies['genres'] = df_movies['genres'].str.replace('|',',')
df_movies =df_movies.dropna()
df_movies.head()
```

Out[3]:
```
       id    imdb_id  popularity     budget     revenue  \
0  135397  tt0369610   32.985763  150000000  1513528810
1   76341  tt1392190   28.419936  150000000   378436354
2  262500  tt2908446   13.112507  110000000   295238201
3  140607  tt2488496   11.173104  200000000  2068178225
4  168259  tt2820852    9.335014  190000000  1506249360

                original_title  \
0                Jurassic World
1            Mad Max: Fury Road
2                     Insurgent
3  Star Wars: The Force Awakens
4                     Furious 7

                                                cast  \
0  Chris Pratt,Bryce Dallas Howard,Irrfan Khan,Vi...
1  Tom Hardy,Charlize Theron,Hugh Keays-Byrne,Nic...
2  Shailene Woodley,Theo James,Kate Winslet,Ansel...
3  Harrison Ford,Mark Hamill,Carrie Fisher,Adam D...
4  Vin Diesel,Paul Walker,Jason Statham,Michelle ...

                                      homepage          director  \
0                   http://www.jurassicworld.com/   Colin Trevorrow
1                     http://www.madmaxmovie.com/     George Miller
```

```
2    http://www.thedivergentseries.movie/#insurgent  Robert Schwentke
3  http://www.starwars.com/films/star-wars-episod...       J.J. Abrams
4                          http://www.furious7.com/         James Wan

                        tagline      ...       \
0             The park is open.      ...
1            What a Lovely Day.      ...
2    One Choice Can Destroy You      ...
3  Every generation has a story.     ...
4           Vengeance Hits Home      ...

                                          overview runtime  \
0  Twenty-two years after the events of Jurassic ...     124
1  An apocalyptic story set in the furthest reach...     120
2  Beatrice Prior must confront her inner demons ...     119
3  Thirty years after defeating the Galactic Empi...     136
4  Deckard Shaw seeks revenge against Dominic Tor...     137

                                     genres  \
0  Action,Adventure,Science Fiction,Thriller
1  Action,Adventure,Science Fiction,Thriller
2         Adventure,Science Fiction,Thriller
3   Action,Adventure,Science Fiction,Fantasy
4                      Action,Crime,Thriller

                               production_companies release_date vote_count  \
0  Universal Studios|Amblin Entertainment|Legenda...       6/9/15       5562
1  Village Roadshow Pictures|Kennedy Miller Produ...      5/13/15       6185
2  Summit Entertainment|Mandeville Films|Red Wago...      3/18/15       2480
3          Lucasfilm|Truenorth Productions|Bad Robot     12/15/15       5292
4  Universal Pictures|Original Film|Media Rights ...       4/1/15       2947

   vote_average  release_year    budget_adj   revenue_adj
0           6.5          2015  1.379999e+08  1.392446e+09
1           7.1          2015  1.379999e+08  3.481613e+08
2           6.3          2015  1.012000e+08  2.716190e+08
3           7.5          2015  1.839999e+08  1.902723e+09
4           7.3          2015  1.747999e+08  1.385749e+09

[5 rows x 21 columns]
```

Exploratory Data Analysis

Tip: Now that you've trimmed and cleaned your data, you're ready to move on to exploration. Compute statistics and create visualizations with the goal of addressing the research questions that you posed in the Introduction section. It is recommended that you be systematic with your approach. Look at one variable at a time, and then follow it up by looking at relationships between variables.

1.0.3 Research Question 1 Is the rating of a movie directly correlated to how much budget is put towards the movie? (Hypothesis: the greater the budget the better the rating)

```
In [4]: # Use this, and more code cells, to explore your data. Don't forget to add
        #   Markdown cells to document your observations and findings.
        df_movies1 = df_movies[['budget', 'original_title', 'vote_average']]
        df_movies1.head()
```

Out[4]:

	budget	original_title	vote_average
0	150000000	Jurassic World	6.5
1	150000000	Mad Max: Fury Road	7.1
2	110000000	Insurgent	6.3
3	200000000	Star Wars: The Force Awakens	7.5
4	190000000	Furious 7	7.3

when exploring the data and sorting by vote average I noticed there was more clean up to be made because the question is related to the budget therefore zeros and nulls should be removed.

```
In [5]: df_movies1= df_movies1.sort_values('vote_average',
                                           ascending=False)

In [6]: df_movies1 = df_movies1[df_movies1.budget != 0]
```

we want to get a better view so lets take all movies ranked over 8 with their budget to see if the budget and voting are correlated

```
In [7]: df_movies1 = df_movies1[df_movies1.vote_average >= 8.0]
        df_movies1.head()
```

Out[7]:

	budget	original_title	vote_average
7269	6000000	The Godfather	8.3
650	3300000	Whiplash	8.2
3826	30000000	Kill Bill: The Whole Bloody Affair	8.1
2875	185000000	The Dark Knight	8.1
2409	63000000	Fight Club	8.1

```
In [8]: df_movies1.head(15)
```

Out[8]:

	budget	original_title	vote_average
7269	6000000	The Godfather	8.3
650	3300000	Whiplash	8.2
3826	30000000	Kill Bill: The Whole Bloody Affair	8.1
2875	185000000	The Dark Knight	8.1
2409	63000000	Fight Club	8.1
10222	22000000	Schindler's List	8.1
7309	18000000	The Empire Strikes Back	8.0
8987	20000000	American History X	8.0
5914	10000000	One Direction: This Is Us	8.0
35	6000000	Room	8.0
636	14000000	The Imitation Game	8.0

```
8069     6000000                  The Usual Suspects           8.0
629    165000000                        Interstellar           8.0
2414    60000000                    The Green Mile           8.0
9      175000000                        Inside Out           8.0
```

```
In [9]: df_movies1['budget'].max()
```

```
Out[9]: 185000000
```

As we can see above in our table, it is clear that budget is not a direct reflection on the vote average. As we can see the max budget paid for a movie at $185000000 for the Dark Knight had a lower voting average than Whiplash and had a budget that was 56 times less at $3300000

1.0.4 Research Question 2 Which top genres of movies that were released had the most revenue?

```
In [10]: df_movies2 = df_movies[['genres', 'original_title', 'revenue', 'release_date' ]]
         df_movies2.head()
```

```
Out[10]:                                       genres                original_title  \
         0  Action,Adventure,Science Fiction,Thriller                Jurassic World
         1  Action,Adventure,Science Fiction,Thriller           Mad Max: Fury Road
         2         Adventure,Science Fiction,Thriller                     Insurgent
         3   Action,Adventure,Science Fiction,Fantasy  Star Wars: The Force Awakens
         4                      Action,Crime,Thriller                     Furious 7

               revenue release_date
         0  1513528810       6/9/15
         1   378436354      5/13/15
         2   295238201      3/18/15
         3  2068178225     12/15/15
         4  1506249360       4/1/15
```

Lets have revenue sorted from largest to smallest

```
In [11]: df_movies2= df_movies2.sort_values('revenue', ascending=False)
         df_movies2.head()
```

```
Out[11]:                                         genres                original_title  \
         1386   Action,Adventure,Fantasy,Science Fiction                        Avatar
         3      Action,Adventure,Science Fiction,Fantasy  Star Wars: The Force Awakens
         5231                     Drama,Romance,Thriller                       Titanic
         4361           Science Fiction,Action,Adventure                 The Avengers
         0     Action,Adventure,Science Fiction,Thriller                Jurassic World

                  revenue release_date
         1386  2781505847     12/10/09
         3     2068178225     12/15/15
         5231  1845034188     11/18/97
         4361  1519557910      4/25/12
         0     1513528810       6/9/15
```

Now that we have the columns we need as well as the sorted revenue largest to smallest, it is time to filter the year to 2015 to see which genre had earned the most money.

```
In [15]: df_movies3=df_movies2[['genres', 'revenue']]
         df_movies3.head(15)
```

Out[15]:

	genres	revenue
1386	Action,Adventure,Fantasy,Science Fiction	2781505847
3	Action,Adventure,Science Fiction,Fantasy	2068178225
5231	Drama,Romance,Thriller	1845034188
4361	Science Fiction,Action,Adventure	1519557910
0	Action,Adventure,Science Fiction,Thriller	1513528810
4	Action,Crime,Thriller	1506249360
14	Action,Adventure,Science Fiction	1405035767
3374	Adventure,Family,Fantasy	1327817822
5422	Animation,Adventure,Family	1274219009
5425	Action,Adventure,Science Fiction	1215439994
8	Family,Animation,Adventure,Comedy	1156730962
3522	Action,Science Fiction,Adventure	1123746996
4949	Adventure,Fantasy,Action	1118888979
4365	Action,Adventure,Thriller	1108561013
4363	Action,Crime,Drama,Thriller	1081041287

```
In [16]: df_movies3.tail(15)
```

Out[16]:

	genres	revenue
3670	Music	0
3667	Horror,Thriller	0
1890	Drama	0
5309	Comedy	0
1873	Music,Comedy	0
3431	Drama,Animation,Family,Comedy	0
1870	Documentary	0
1869	Comedy	0
6697	Comedy,Horror	0
1865	Documentary	0
3657	Horror,Thriller,Mystery	0
3655	Drama,Romance	0
5404	Family	0
1859	Drama,Horror,Thriller	0
2214	War,Drama,Action,Adventure,History	0

The top genres that produced the most revenue are: Action, Adventure, Fantasy, Science Fiction pulling in roughly 4 billion dollars whereas the least popular genres that received zero revenue are Drama, and Comedy.

Conclusions

> **Tip**: Finally, summarize your findings and the results that have been performed. Make sure that you are clear with regards to the limitations of your exploration. If you

haven't done any statistical tests, do not imply any statistical conclusions. And make sure you avoid implying causation from correlation!

Tip: Once you are satisfied with your work here, check over your report to make sure that it is satisfies all the areas of the rubric (found on the project submission page at the end of the lesson). You should also probably remove all of the "Tips" like this one so that the presentation is as polished as possible.

1.1 Submitting your Project

Before you submit your project, you need to create a .html or .pdf version of this notebook in the workspace here. To do that, run the code cell below. If it worked correctly, you should get a return code of 0, and you should see the generated .html file in the workspace directory (click on the orange Jupyter icon in the upper left).

Alternatively, you can download this report as .html via the **File** > **Download as** submenu, and then manually upload it into the workspace directory by clicking on the orange Jupyter icon in the upper left, then using the Upload button.

Once you've done this, you can submit your project by clicking on the "Submit Project" button in the lower right here. This will create and submit a zip file with this .ipynb doc and the .html or .pdf version you created. Congratulations!

```
In [ ]: from subprocess import call
        call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

```
Out[ ]: 0
```

```
In [ ]:
```