investigate-a-dataset-template

February 10, 2019

1 Project: Investigate a Movie Dataset

1.1 Table of Contents

Introduction

The data set analyzed in this notebook is the TMDB Movies Database. This data set originiated from Kaggle and provided by Udacity. There is information on more than 5000 movies. The information used below is popularity, revenue, bugdet, and runtime. The information chosen from the data set is to dive into what metrics are good for figuring out how to measure a movies' success.

```
In [1]: import pandas as pd
        import numpy as np
        import scipy.stats as st
        import matplotlib.pyplot as plt
        import os
        from scipy.stats import pearsonr
        import seaborn as sns
        import statsmodels.api as sm
        from statsmodels.formula.api import ols
        %matplotlib inline
```

```
/home/aurora/anaconda3/lib/python3.6/site-packages/statsmodels/compat/pandas.py:56: FutureWarn
  from pandas.core import datetools
```

```
In [2]: pd.set_option('display.float_format', lambda x: '%.3f' % x) # surppresses scietific no
```

Data Wrangling

1.1.1 General Properties

```
In [3]: # Load your data and print out a few lines. Perform operations to inspect data
        #   types and look for instances of missing or possibly errant data.
        data = pd.read_csv('../tmdb-movies.csv')
```

```
In [4]: #This allows us to look inside of the dataset
        data.head()
```

```
Out[4]:        id    imdb_id  popularity     budget     revenue  \
        0  135397  tt0369610      32.986  150000000  1513528810
        1   76341  tt1392190      28.420  150000000   378436354
        2  262500  tt2908446      13.113  110000000   295238201
        3  140607  tt2488496      11.173  200000000  2068178225
        4  168259  tt2820852       9.335  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         6.500          2015 137999939.280 1392445892.524
1         7.100          2015 137999939.280  348161292.489
2         6.300          2015 101199955.472  271619025.408
3         7.500          2015 183999919.040 1902723129.802
4         7.300          2015 174799923.088 1385748801.471

[5 rows x 21 columns]
```

```python
In [5]: data.shape #returns how many rows and column of the dataset
```

```
Out[5]: (10866, 21)
```

```python
In [6]: data.describe() #returns statistics about the numerical columns
```

```
Out[6]:              id  popularity        budget        revenue   runtime  \
count  10866.000   10866.000     10866.000      10866.000 10866.000
mean   66064.177       0.646  14625701.094    39823319.793   102.071
std    92130.137       1.000  30913213.831   117003486.582    31.381
min        5.000       0.000         0.000          0.000     0.000
25%    10596.250       0.208         0.000          0.000    90.000
50%    20669.000       0.384         0.000          0.000    99.000
75%    75610.000       0.714  15000000.000    24000000.000   111.000
max   417859.000      32.986 425000000.000  2781505847.000   900.000

       vote_count  vote_average  release_year    budget_adj    revenue_adj
count   10866.000     10866.000     10866.000     10866.000      10866.000
mean      217.390         5.975      2001.323   17551039.823   51364363.253
std       575.619         0.935        12.813   34306155.723  144632485.040
min        10.000         1.500      1960.000         0.000          0.000
25%        17.000         5.400      1995.000         0.000          0.000
50%        38.000         6.000      2006.000         0.000          0.000
75%       145.750         6.600      2011.000   20853251.084   33697095.717
max      9767.000         9.200      2015.000  425000000.000 2827123750.412
```

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).

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

In [7]:
```
# Putting all of the columns that will be used in a data frame
movie_data = data[['popularity', 'budget', 'revenue','runtime','genres']]
```

In [8]:
```
#checks the new data set
movie_data.shape
```

Out[8]:
```
(10866, 5)
```

In [9]:
```
#Drops the rows that have atleast 1 NAN
clean_mdata = movie_data.dropna()
```

In [10]:
```
#Now we take a peak at the clean dataset
clean_mdata.describe()
```

Out[10]:

	popularity	budget	revenue	runtime
count	10843.000	10843.000	10843.000	10843.000
mean	0.647	14656724.439	39907792.389	102.138
std	1.001	30938637.671	117113132.251	31.293
min	0.000	0.000	0.000	0.000
25%	0.208	0.000	0.000	90.000
50%	0.385	0.000	0.000	99.000
75%	0.715	15000000.000	24136754.000	111.000
max	32.986	425000000.000	2781505847.000	900.000

In [11]:
```
# drop if there are duplicates, just in case
clean_mdata = clean_mdata.drop_duplicates()
```

In [12]:
```
clean_mdata = clean_mdata[(clean_mdata != 0).all(1)]
```

In [13]:
```
#take another peak
clean_mdata.head()
```

Out[13]:

	popularity	budget	revenue	runtime	\
0	32.986	150000000	1513528810	124	
1	28.420	150000000	378436354	120	
2	13.113	110000000	295238201	119	
3	11.173	200000000	2068178225	136	
4	9.335	190000000	1506249360	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

```
In [14]: #function that separates the genres for each movie
         def add_genre(column):
             #will take a column, and separate the string from the '|'
             data = clean_mdata[column].str.cat(sep = '|')

             #giving pandas series and storing the values separately
             data = pd.Series(data.split('|'))

             #arranging in descending order
             count = data.value_counts(ascending = False)

             return count
```

```
In [15]: #variable to store the retured value
         genre_count = add_genre('genres')
```

```
In [16]: # Formula for gross profit: Revenue - Budget
         clean_mdata['gross_profit'] = clean_mdata['revenue'] - clean_mdata['budget']
```

```
In [17]: clean_mdata.head()
```

```
Out[17]:    popularity     budget     revenue  runtime  \
         0      32.986  150000000  1513528810      124
         1      28.420  150000000   378436354      120
         2      13.113  110000000   295238201      119
         3      11.173  200000000  2068178225      136
         4       9.335  190000000  1506249360      137

                                            genres  gross_profit
         0  Action|Adventure|Science Fiction|Thriller    1363528810
         1  Action|Adventure|Science Fiction|Thriller     228436354
         2         Adventure|Science Fiction|Thriller     185238201
         3   Action|Adventure|Science Fiction|Fantasy    1868178225
         4                     Action|Crime|Thriller    1316249360
```

```
In [18]: #Defining these variables here to see a statisitcal correlation in
         #bivariate analysis. Separating my independet and dependent variables
         pop = clean_mdata['popularity']
         ygp = clean_mdata['gross_profit']

         xgen = clean_mdata['genres']
         xrev = clean_mdata['revenue']
         xbudget = clean_mdata['budget']
```

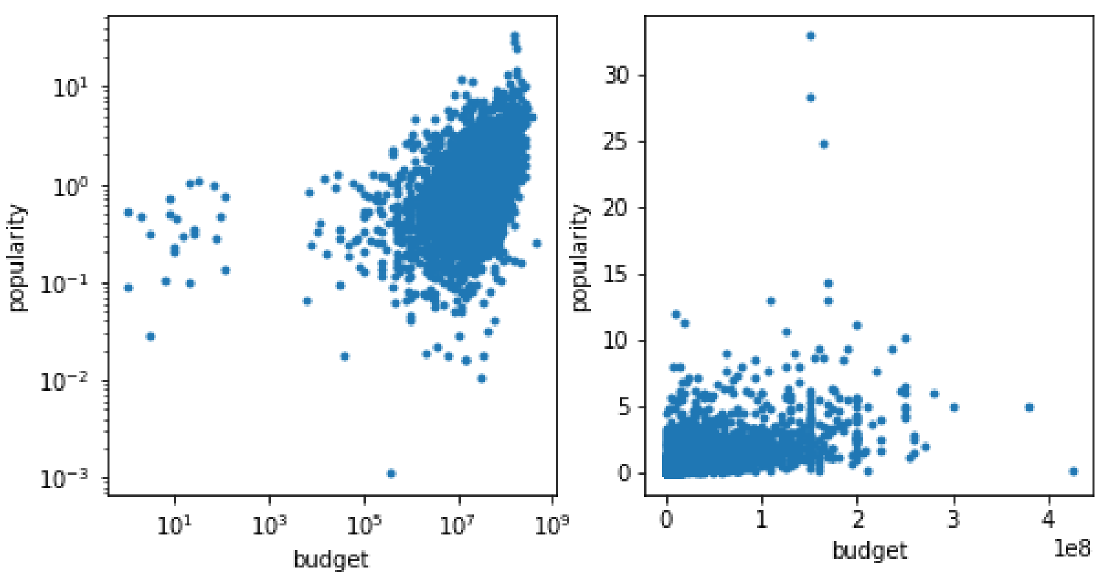
1.2 Exploratory Data Analysis

1.2.1 Research Question 1 (What is the relationship between popularity of a movie and it's budget?)

Below is a scatter plot that decribes the relationship between budget and popularity. Budget is the independent variable and popularity is the dependent variable. The graph to the left shows the relationship in log, while the plot to the right shows the relationship on the varaibles regular scale. The log transformation has spread out the data a bit, in order to see how much budget influences popularity. In most cases, the higher the budget the more popular a movie was. In a few cases, we can see that a low budget had a middle popularity. The same goes for a few high budget movies that had a very low popularity. So, overall, if the movie had a high budget it looks to correlate with moderate to high popularity. Popularity and budget are also run through a Pearson's R correlation and are statistically significant.

In [19]:
```
fig, (ax1,ax2) = plt.subplots(1, 2, figsize=(8,4))
ax1.plot(clean_mdata['budget'], clean_mdata['popularity'], '.')
ax1.set_yscale('log')
ax1.set_xscale('log')
ax1.set_xlabel('budget')
ax1.set_ylabel('popularity')
ax2.plot(clean_mdata['budget'], clean_mdata['popularity'], '.')
ax2.set_xlabel('budget')
ax2.set_ylabel('popularity')
```

Out[19]:
```
Text(0,0.5,'popularity')
```



In [20]:
```
#Budget and Population are statistically significant with an alpha of 0.05
pearsonr(xbudget, pop)
```

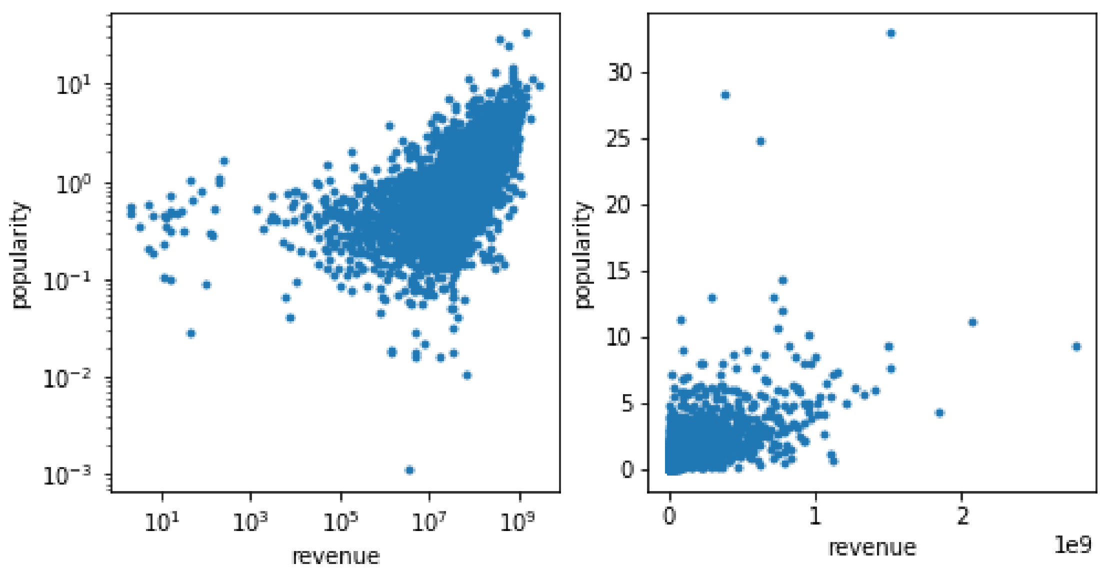
Out[20]:
```
(0.4469866930400051, 1.0513638511750907e-188)
```

1.2.2 Research Question 2 (Is there an association between a movies' revenue and how popular it is?)

Below is a scatter plot between the variables popularity and revenue. THe plot on the left is in log scale and the plot of the right is the default values. The plot that is in log shows that the higher the revenue the more popular the movie was. There were cases that show movies with high revenue were not very popular, but the majority shows that if a movie had a medium revenue then it was popular. These varaibles are statisitcally significant, which we can tell by running them with a Pearson's R correlation test.

In [21]:
```
#clean_mdata.plot(x='revenue',y='popularity',kind='scatter')
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(8,4))
ax1.plot(clean_mdata['revenue'], clean_mdata['popularity'], '.')
ax1.set_yscale('log')
ax1.set_xscale('log')
ax1.set_xlabel('revenue')
ax1.set_ylabel('popularity')
ax2.plot(clean_mdata['revenue'], clean_mdata['popularity'], '.')
ax2.set_xlabel('revenue')
ax2.set_ylabel('popularity')
```

Out[21]:
```
Text(0,0.5,'popularity')
```



In [22]:
```
pearsonr(xrev, pop)
```

Out[22]:
```
(0.6155346545546607, 0.0)
```

1.2.3 Research Question 3 (Is there a relationship between genre and gross profit?)

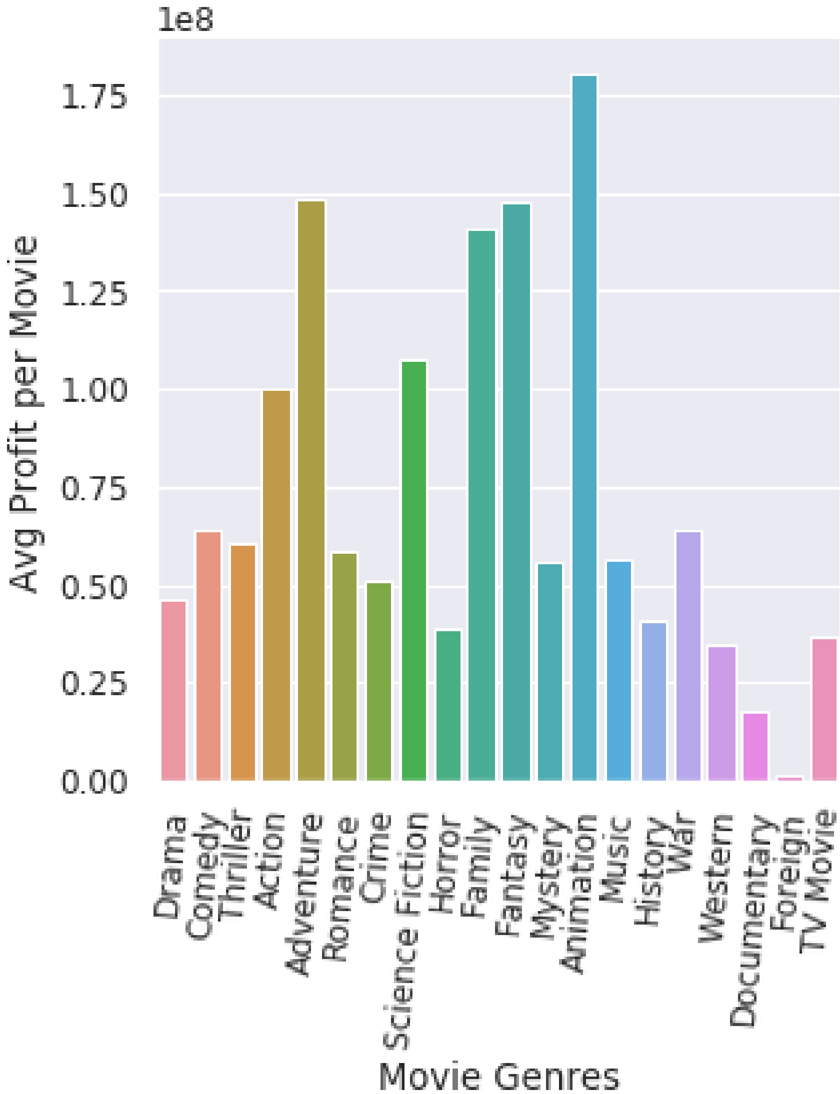
Below is a histogram of genre and gross profit. Which genre's did had the highest gross profit? Action, Adventure, Science Fiction,Family, Fantasy, and Animation. The gren's that did the worst are: Foreign, Documentary, Horror, and TV Movie.

In [23]:
```
genre_profit = genre_count.copy()
for genre in genre_count.keys():
    has_genre = clean_mdata['genres'].str.contains(genre)
    genre_profit[genre] = clean_mdata['gross_profit'][has_genre].mean()

#ax = genre_profit.plot(kind='bar', title='Avg Gross Profit By Genre')
#ax.set_ylabel('Avg Gross Profit')
#ax.set_xlabel('Genres')
```

In [24]:
```
sns.set(font_scale=1.1)
sns.set_style('darkgrid')
gp = genre_profit.to_frame('profit')
gp.reset_index(level=0, inplace=True)
gp.columns = ['genre', 'profit']
plot = sns.catplot(x='genre', y='profit', kind='bar', data=gp)
plot.set_xticklabels(rotation=85)
plot.fig.get_axes()[0].set_ylabel('Avg Profit per Movie')
plot.fig.get_axes()[0].set_xlabel('Movie Genres')
```

Out[24]:
```
Text(0.5,10.256,'Movie Genres')
```



Conclusions

Overall, important information can be gathered when looking at the budet and revenue for the popularity of movies. The above plots, on average, show that with a decent budget a movie can be popular. There are a few outliers but for most part, looking at the groupings of each movie's budget showed that a movie needs a moderate budget to do well. A movie's revenue is also important to look at. Surprisingly, a small cohert of movies that did not have a high revenue return was still considered to be popular. For the most part though, revenue is a good indicator of how popular a movie is.