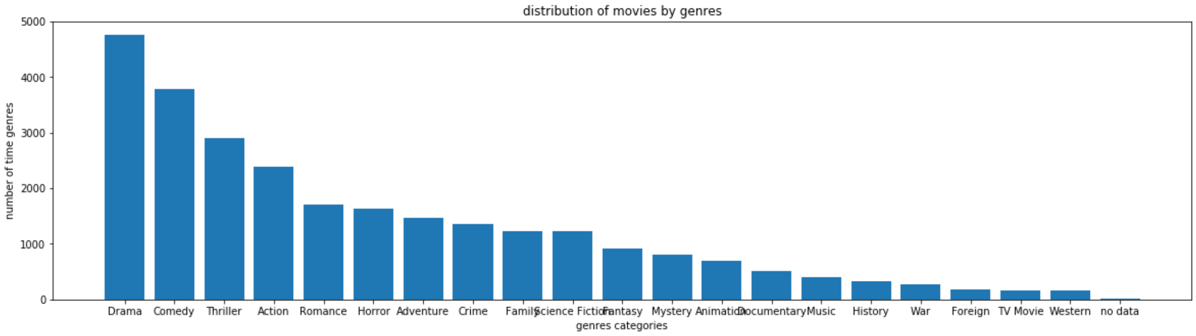
Question that I want to investigate

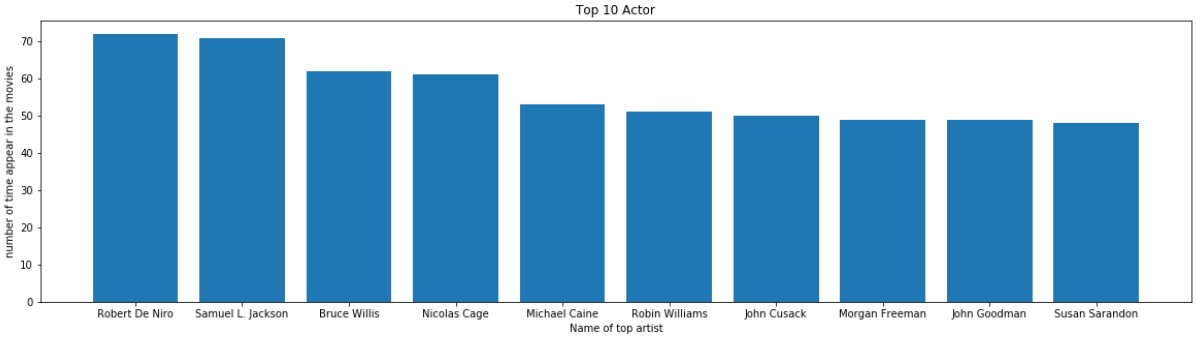
1. movie distribution by genre
2. Actor and actress with highest movies role
3. What movie categories generate most average profit?
4. Genre with highest rating and popularity, do it have relationship with the 5. profit that category make?
5. is the movie with that popular actress/actor have high impact to the popularity of the movie compare to average??
6. Top grossing film
7. Is increase in budget leaded to higher rating? (correlation)
8. Do higher rating increase the number of runtime? (correlation)

Distribution of movie by genre



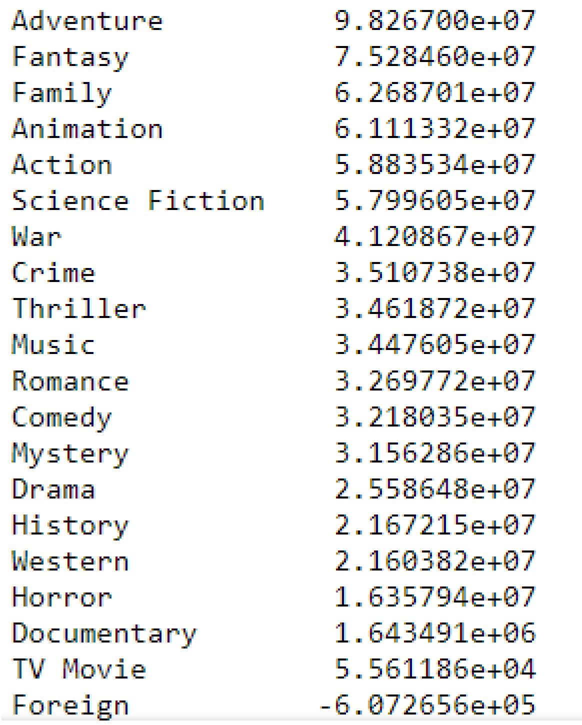
Look like most movies were mainly on drama and comedy genres, while documentary, music, history , war and foreign got lower populace

Finding top 10 actor



The most appearing in film movies were Robert who have appear in 70 movies!! Comparing to most of the actor in 5000 movies in database that got only 1 or 2 movies in their career

Genre with highest profit

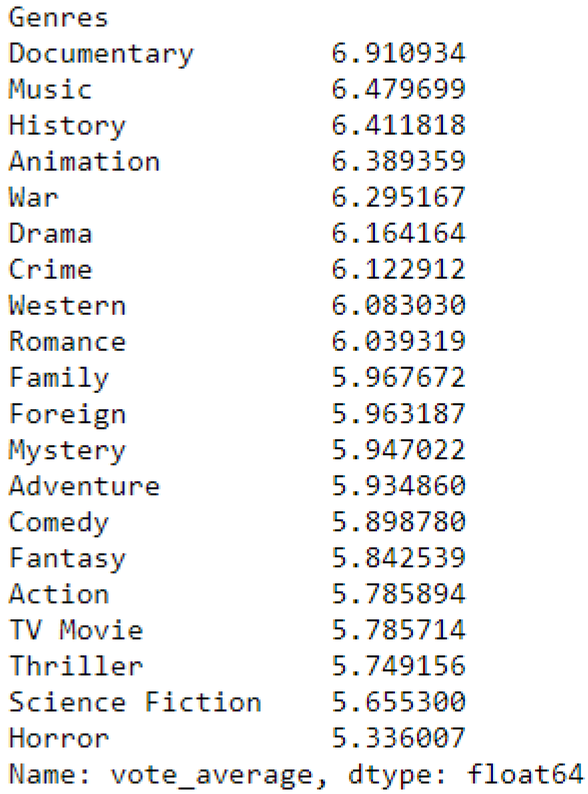


```
Adventure          9.826700e+07
Fantasy            7.528460e+07
Family             6.268701e+07
Animation          6.111332e+07
Action             5.883534e+07
Science Fiction    5.799605e+07
War                4.120867e+07
Crime              3.510738e+07
Thriller           3.461872e+07
Music              3.447605e+07
Romance            3.269772e+07
Comedy             3.218035e+07
Mystery            3.156286e+07
Drama              2.558648e+07
History            2.167215e+07
Western            2.160382e+07
Horror             1.635794e+07
Documentary        1.643491e+06
TV Movie           5.561186e+04
Foreign           -6.072656e+05
```

Why number of movies that based on adventure were in median number the average profit made by these film were considering the highest! But why? Is it because it gain higher popularity? Or rating compare to other type?

- ➔ As we can see from here drama and comedy movies where most film distribution are at doesn’t get that much of a profit…

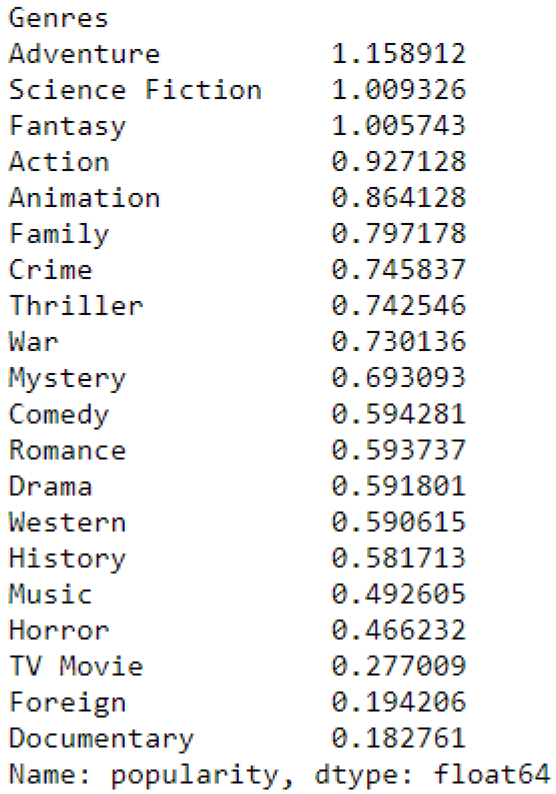
Genre with highest rating



```
Genres
Documentary        6.910934
Music              6.479699
History            6.411818
Animation          6.389359
War                6.295167
Drama              6.164164
Crime              6.122912
Western            6.083030
Romance            6.039319
Family             5.967672
Foreign            5.963187
Mystery            5.947022
Adventure          5.934860
Comedy             5.898780
Fantasy            5.842539
Action             5.785894
TV Movie           5.785714
Thriller           5.749156
Science Fiction    5.655300
Horror             5.336007
Name: vote_average, dtype: float64
```

➔ Surprisingly the genres with highest vote score were Documentary! Which is one of the lowest number of movies distribution!!

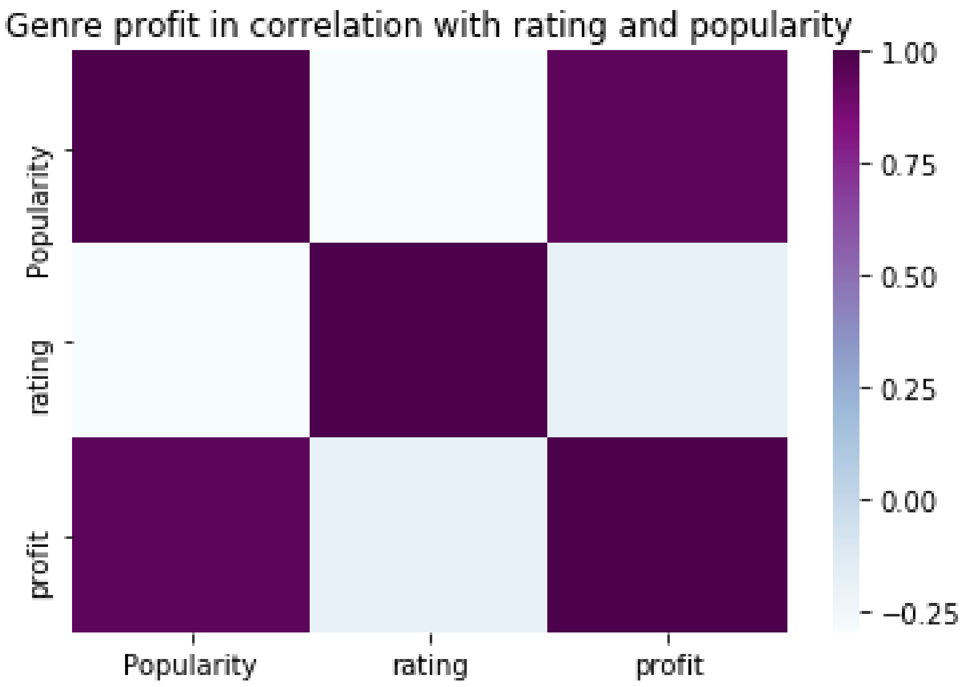
Genre with highest popularity



```
Genres
Adventure          1.158912
Science Fiction    1.009326
Fantasy            1.005743
Action             0.927128
Animation          0.864128
Family             0.797178
Crime              0.745837
Thriller           0.742546
War                0.730136
Mystery            0.693093
Comedy             0.594281
Romance            0.593737
Drama              0.591801
Western            0.590615
History            0.581713
Music              0.492605
Horror             0.466232
TV Movie           0.277009
Foreign            0.194206
Documentary        0.182761
Name: popularity, dtype: float64
```

- ➔ But the popularity show difference result??? The adventure movie (in which was #1 in profit were consider most popular!)

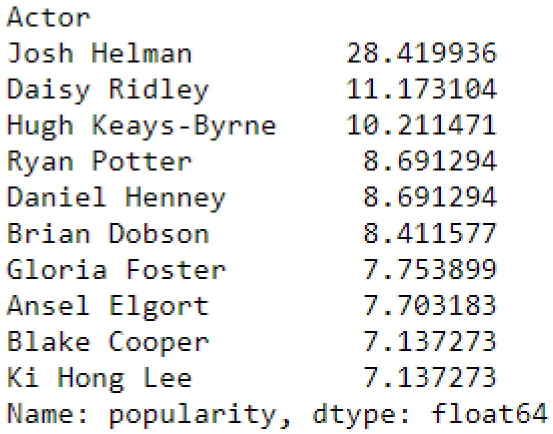
Finding correlation between these 3 metric in describing genre… as we see that higher average rating of the genre didn't lead to higher profit..(as genre with higher rating were one of the low profit generation genre)



Look like there's a high correlation between profit and movie popularity! Not a rating that the movie receive.. and rating and popularity has negative correlation! (But is this true… we would explore more in the overall correlation table, as this's based on genre average)

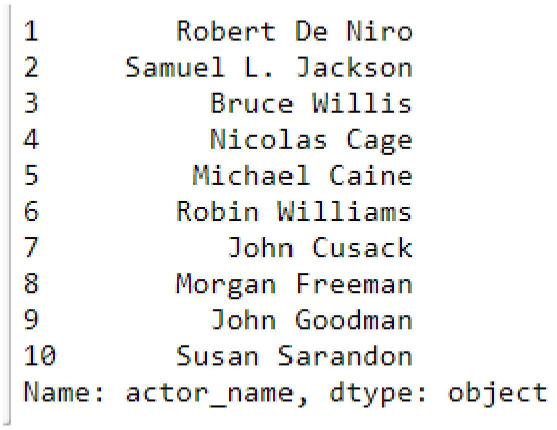
is the movie with that popular actress/actor have higher impact of popularity compare to average??

Looking at top grossing actor…

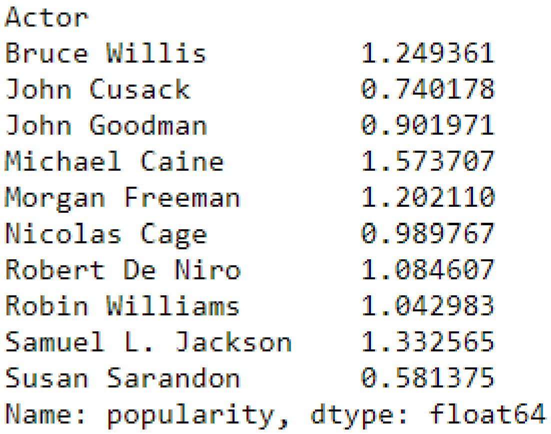


```
Actor
Josh Helman          28.419936
Daisy Ridley         11.173104
Hugh Keays-Byrne     10.211471
Ryan Potter           8.691294
Daniel Henney         8.691294
Brian Dobson          8.411577
Gloria Foster         7.753899
Ansel Elgort          7.703183
Blake Cooper          7.137273
Ki Hong Lee           7.137273
Name: popularity, dtype: float64
```

Comparing to those top 10 actor name?



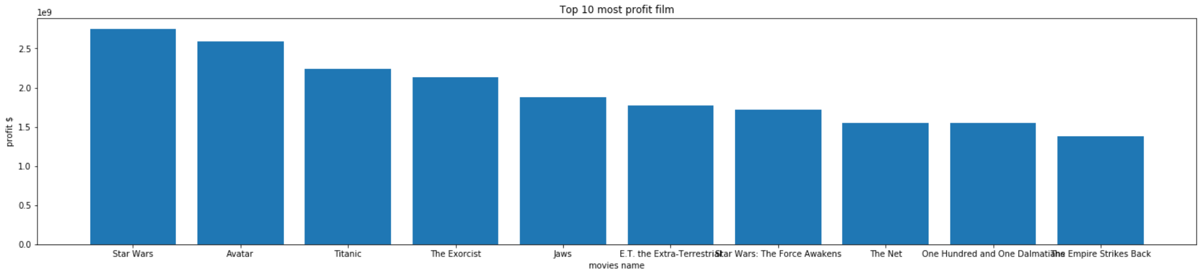
```
1         Robert De Niro
2      Samuel L. Jackson
3           Bruce Willis
4           Nicolas Cage
5          Michael Caine
6         Robin Williams
7            John Cusack
8         Morgan Freeman
9           John Goodman
10        Susan Sarandon
Name: actor_name, dtype: object
```



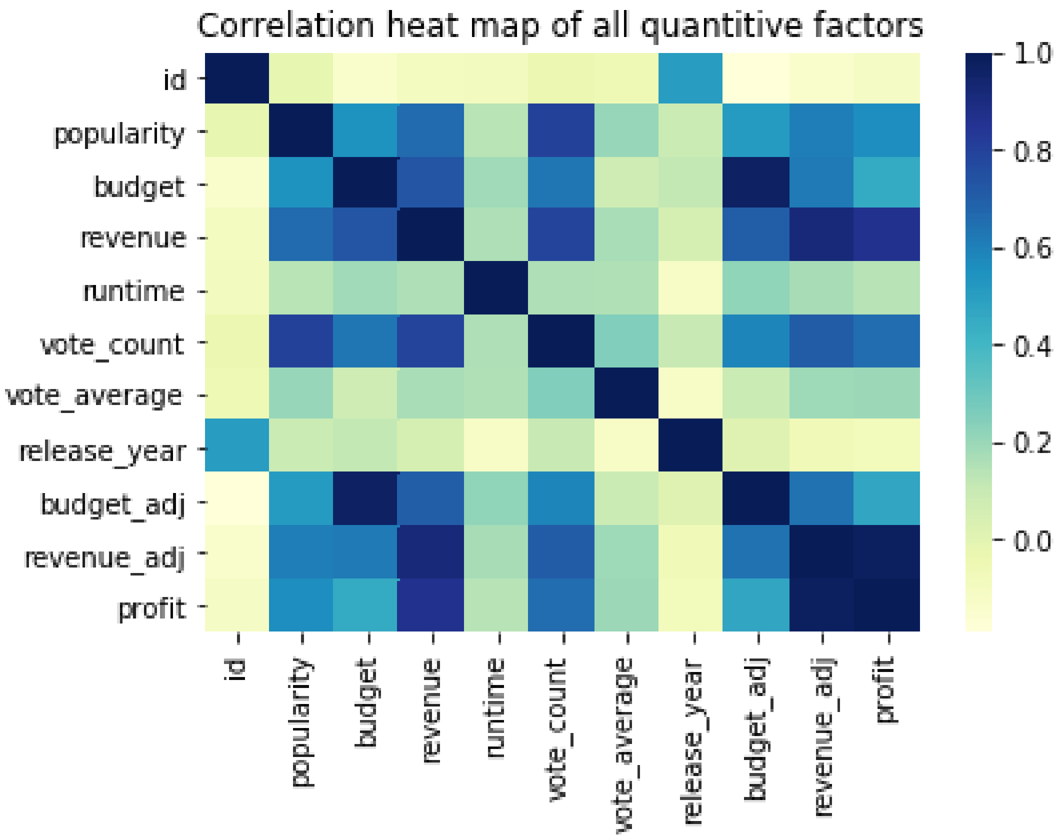
```
Actor
Bruce Willis         1.249361
John Cusack          0.740178
John Goodman         0.901971
Michael Caine        1.573707
Morgan Freeman       1.202110
Nicolas Cage         0.989767
Robert De Niro       1.084607
Robin Williams       1.042983
Samuel L. Jackson    1.332565
Susan Sarandon       0.581375
Name: popularity, dtype: float64
```

Seem like no.. ended up that the popularity of the movie doesn't depend on the actor popularity alone (assuming higher popularity actor = most frequent appear actor)

Top 10 most profitable film



# do higher budget leaded to higher average rating??? (correlation)



No higher budget doesn’t cauase the vote_average to rise.. but it do have positive correlation with the popularity with correlation at 0.6.. which mean that as the budget rise there’s a tendency that the popularity of the movei would increase too

#Do higher rating increase the number of runtime? (correlation)

Looking from the table of correlation, no factor did contribute a significant impact to the number of run time of particular movies

Conclusion

1. movie distribution by genre: highest number of movie create in Drama and comedy category
2. Actor and actress with highest movies role: Robert De niro ranked as number 1 most appearing actor
3. What movie categories generate most average profit?: Adventure category got most average profit (with highest popularity)
4. Genre with highest rating and popularity, do it have relationship with the profit that category make? -- as we see that higher average rating of the genre didn’t lead to higher profit..(as genre with higher rating were one of the low profit generation genre)

➔ Limitation in this analysis is that we’re judging it based on the average that the category have, ignoring the fact that there might be a blockbuster movie that outlier its category ex. Starwar

5. is the movie with that popular actress/actor have high impact to the popularity of the movie compare to average??

   Seem like no.. ended up that the popularity of the movie doesn't depend on the actor popularity alone (assuming higher popularity actor = most frequent appear actor)

6. Top grossing film: Most profitable film were starwar
7. Is increase in budget leaded to higher rating? (correlation)

   No higher budget doesn’t cause the vote_average to rise.. but it do have positive correlation with the popularity with correlation at 0.6.. which mean that as the budget rise there’s a tendency that the popularity of the movei would increase too

   (remark: there might be other factor that contribute to the changes here as well which has not been test in this finding)

8. Do higher rating increase the number of runtime? (correlation)

   Looking from the table of correlation, no factor did contribute a significant impact to the number of run time of particular movies

   (remark: there might be other factor that contribute to the changes here as well which has not been test in this finding)

Limitation Report

Limitation in the report is that we didn’t consider any other things that could impact the #4, #5 question in to the analysis, therefore the conclusion here might not be fully complete. There might be other factor that could enhance the research result such as the director talent that could lead to higher profit, average rating, or popularity by the users.

The Data have come with unequal values, some are with some missing values in a column, so we have to clean the data before using it.

More than this the data have come up with a list of value in the column such as Actor,director,genre with | in between each and every variable, therefore we have to separate the data using split and stack it onto the same movies. While doing all this I’ve notice that the list # of each row weren’t the same as well so we have to eliminate and clean the data again.

When matching the data I’ve come to notice that the unequal length in data would cost the data discrepancy, therefore we have to drop the variable that weren’t use in the result, and classify those as “NO data” in order to separate it from the rest of the variable