Data Analyst Nanodegree

Investigate a Dataset Project

Dataset Chosen: TMDB Movies

Introduction:

In this project, I've investigated a TMDb movies database which has collection of details of about 10k+ movies, including their details of budget, revenue, release dates, etc. and then i communicate my findings about it. I've used the Python libraries NumPy, pandas, and Matplotlib to make my analysis easier.

Questions:

1- Which genres are most popular from year to year?

2- show the most directors who has made most films and the highest revenue they have made.

3- what is the most genre each director has made his flims of and what is the mean vote average each of them has gotten on his flims of that genre

4- show the change in the animation genre during the decades

-A description of what you did to investigate those questions

I have cleaned the dataframe , wrangle it and created other dataframes from the main dataframe contain the columns I need to investigate for each question.

-Documentation of any data wrangling you did

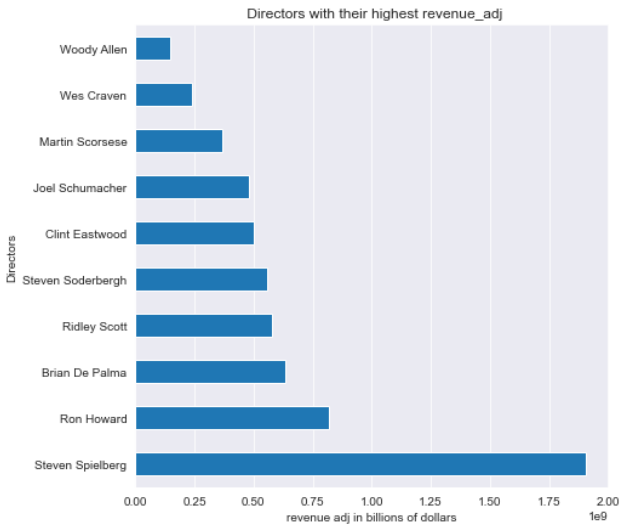
I've split the genres columns on | and put them in list, then I have turned it into a dataframe and removed and renamed columns. I merged the two dataframe I've created together then I ounted the genres with each year then I have putted it in a csv file and saved it, I removed and renamed columns then I sorted values to get the answer for the question and visualized it.

-Summary statistics and plots communicating your final results

First question's answers:

	release_year	genre	count
1050	2015	Drama	260
1030	2014	Drama	284
1010	2013	Drama	253
989	2012	Drama	232
969	2011	Drama	214
948	2010	Drama	210
928	2009	Drama	224
907	2008	Drama	233
887	2007	Drama	197
867	2006	Drama	197
846	2005	Drama	182
826	2004	Drama	141
803	2003	Comedy	111
786	2002	Drama	130
762	2001	Comedy	101
744	2000	Drama	101
724	1999	Drama	113
704	1998	Drama	108

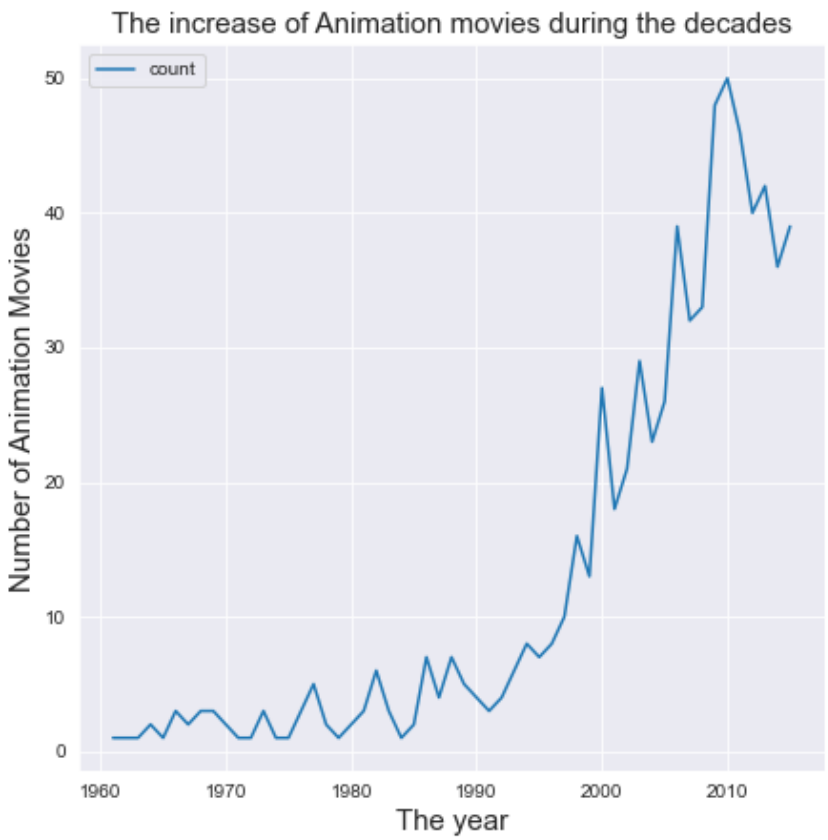
Second question visualization



Third question answer

	genre_x	director	count	vote_average
1235	Comedy	Woody Allen	38	6.484211
256	Drama	Clint Eastwood	27	6.462963
1372	Drama	Martin Scorsese	23	6.908696
1485	Horror	Wes Craven	18	5.816667
617	Drama	Ridley Scott	17	6.452941
932	Drama	Steven Soderbergh	17	6.129412
1079	Drama	Steven Spielberg	15	6.946667
180	Thriller	Brian De Palma	15	6.586667
448	Drama	Joel Schumacher	14	6.264286
786	Drama	Ron Howard	14	6.542857

Fourth question visualization



Conclusion

from the fourth question's answer it seems to me like that Martin Scorsese could be the best director for the Drama genre for getting the second best vote Average and the second most director of the count of making drama movies.

from the first question's answer we can see that the drama genre is usually the most popular.

Limitations

especially for older movies there are a lot of missing data which could affect the honesty of this analysing