Insights on IMDB data

The Link to github repository : - https://github.com/Devandra21/Insights_on_IMDB_Revidly

Submitted By:-

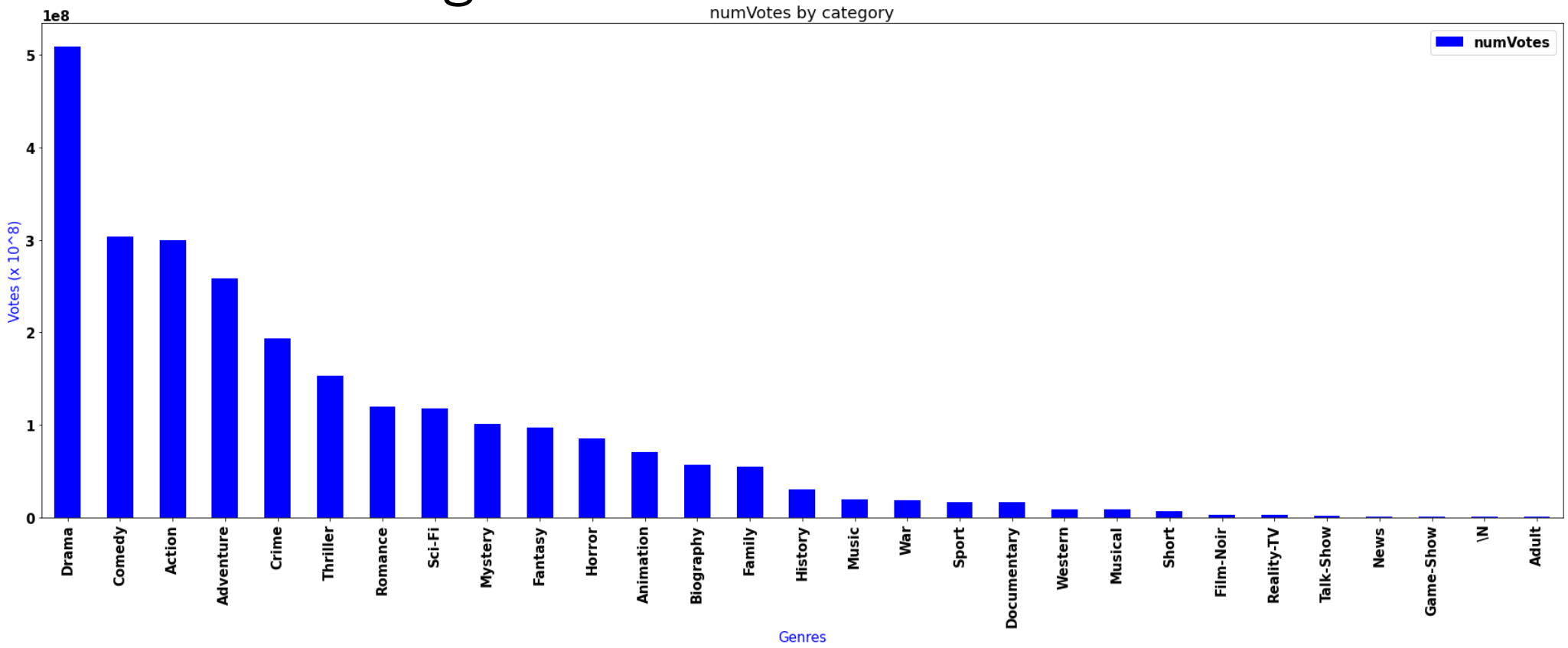
Devandra Jain

3rd Year, B.Tech

Electronics and Communication Engineering

NIT, Trichy

Which Genre gets the maximum votes?

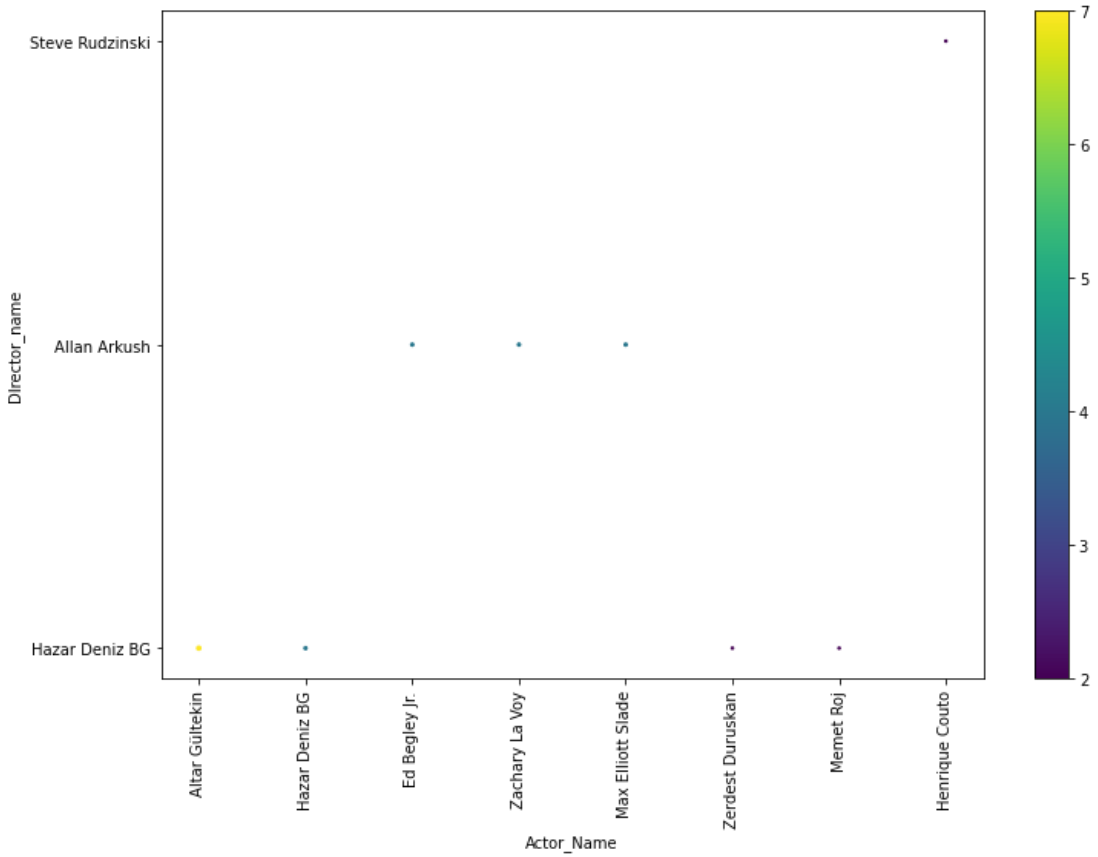


From title.basics just genre and title constant were extracted and kept in dataframe then from title.ratings, title and numVotes were kept in dataframe , both of the dataframe were merged and then using groupby and sum total number of votes for each genre were found, finally for each genre bar graph was plotted and found that **Drama** Genre got maximum votes.

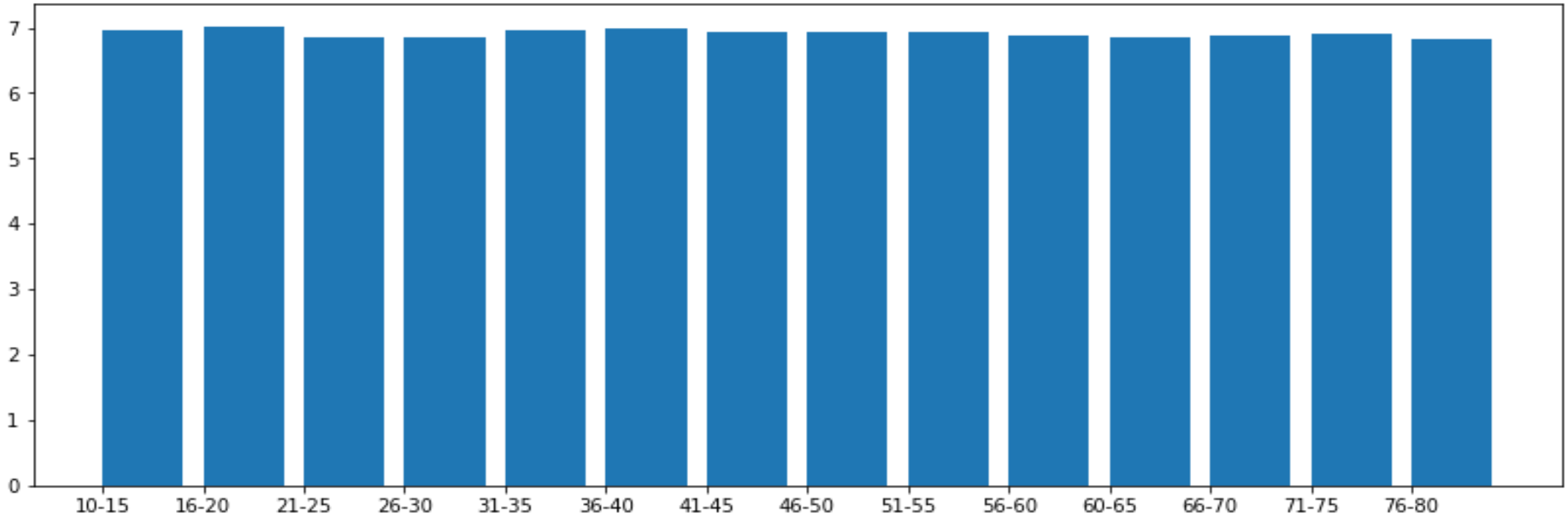
Which actor - director pair is most successful (in terms of IMDB ratings)?

The most successful actor-director pair in terms of imdb rating is **Altar Gültekin(actor)-Hazar Deniz BG(director)**. They got 7 times average rating as 10 when number if votes were also too high.

From the ratings having maxVotes less than 50 were removed as for analysing at least 50 votes should be there, after that all the actor linked to particular title constant were found using name.basics and title.principals , using title.crew and title.principals all the directors linked to a particular title was found. Finally both actors and directors were merged getting actor director pairs. Merging this with the average ratings. Then keeping all pairs that have avegage rating as 10 . Finally for each pair count of number of ratings as 10 were found and we got the pair with maximum 10count to be the most successful in terms of imdb rating.



Male leads are most successful in which age bracket?



As we can see in the graph above , male leads is most successful in the age bracket **16-20**.
In the name.basics primary profession and birthyear of each individual is given, from that using tolist and stack all the different professions were splitted and only actors details were kept in dataframe. Then in all types of gene whichever actor has worked is merged using inner joint along with the start year if that genre. Then substracting birthyear and start year gives age. Then using genres title id ratings were merged using inner joint and keeping only those ratings that has votes greater than 50. Ages between 10-80 are only kept in the dataframe. For different ages rating is listed from which different age groups are formed and mean average rating is found in each age group from which the above chart is made.

For TV shows, what is the avg episode duration of the best TV shows?

43.024 minutes

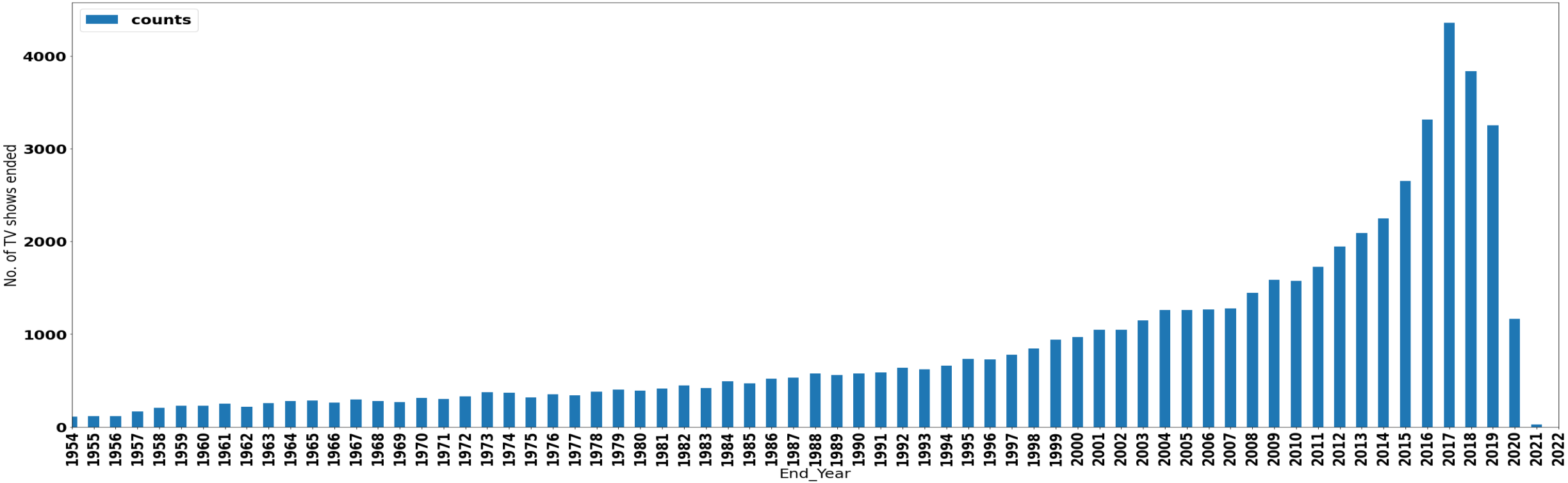
From title.basics only tv related titles are kept from titleType, then only tconst and runtime is kept in dataframe, then all the null values are removed. Finally average of all the runtime is been found.

Best TV shows last for how many seasons?

Maximum of best tvshows lasted for 1 year only but a few good lasted for 18,4,5,19,11,6,10 years.

From the title.basics only the titleType related to tv is kept remaining along with startyear and endyear, number of seasons lasted is calculated using years = (endyear – start year). Then using title.ratings only with number of votes greater than 50 is merged with the years one. The tvshows only greater than 1 year is kept as below one year it can be a tvshort. Finally everything is sorted according to ratings.

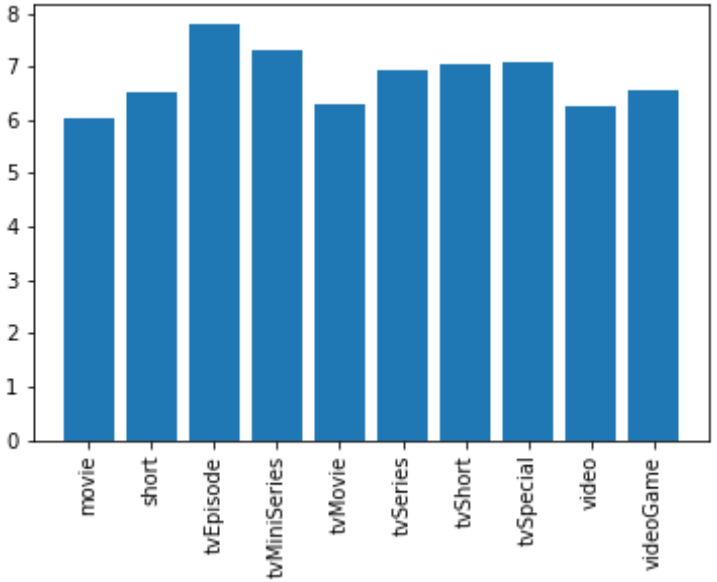
In which year maximum TV shows ended?



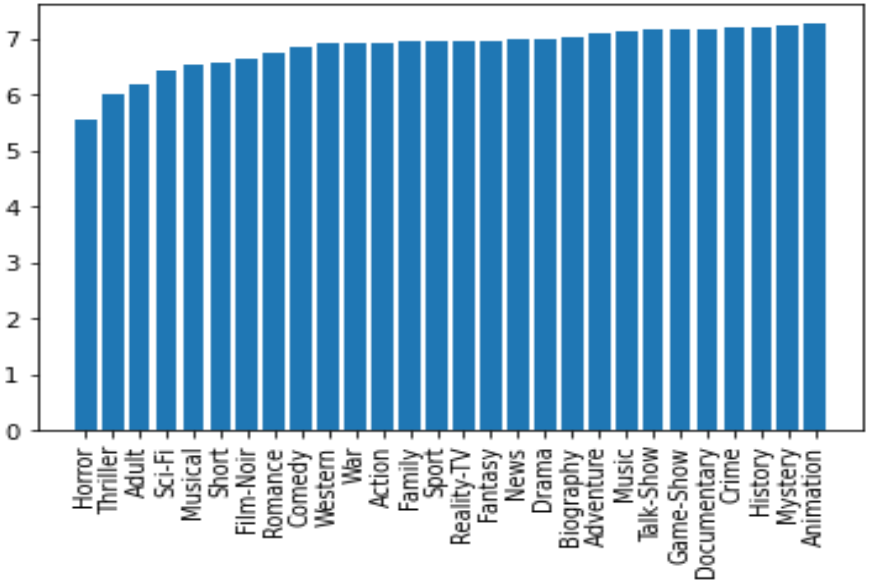
From the above graph it is clear that maximum tvshows ended in **2017**.

From title.basics only tvshows titleType is taken. Then in tconst and endyear is taken in which using groupby it is sorted according to end of year and the the above graph is drawn.

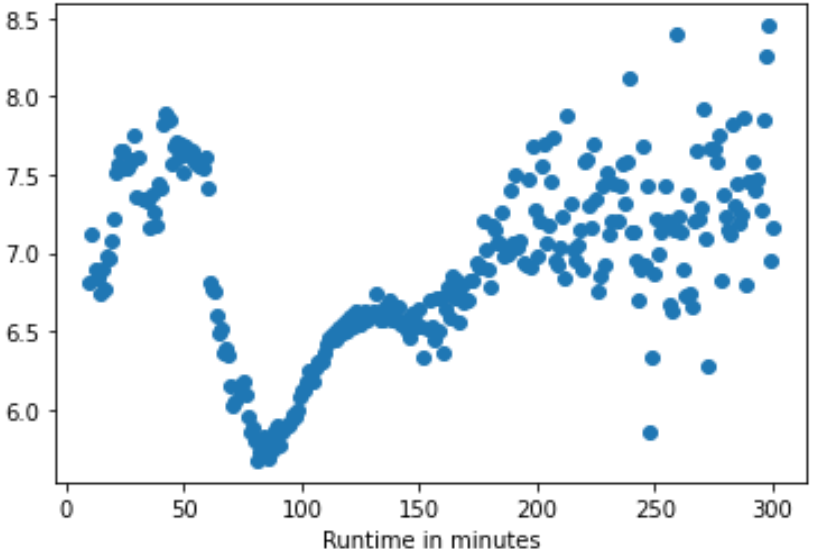
What factors lead to higher rating?



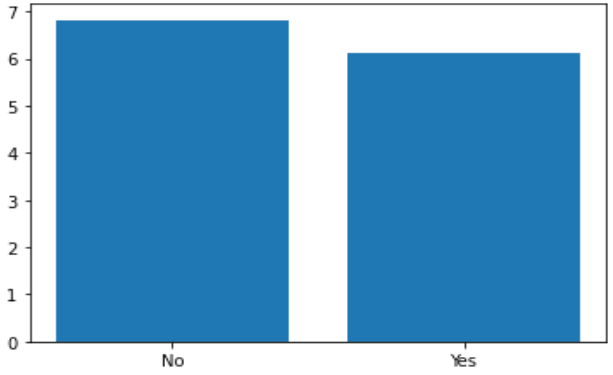
Effect of **cinematography**



Effect of **genre**



Effect of **runtime**



Effect of **isadult**

For all these factors title.basics is merged with title.ratings then entries with more than 50 votes are kept then based on the factor columns are kept and sorted accordingly and the graphs are generated.