Investigate Dataset

(TMDb movie data)

The data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue the data saved as csv file extracted from kaggle website

The questions

1-When I saw the Data my first question was how much growth in the revenue of movies made since 1960 till 2015 ?

So Igrouped the data with release year column to determinethe sum of revenues for each year

2- the I saw that there a different types of genres in movies so I decide to know how many movies are made with (Adventure,Action,Scince Fiction,Thriller,Fantasy,Crime) in it is genre?

So I count the movies which have any genre of those in his properties

Data wrangle

Then I start to investigate data

First I have noticed that some columns have multiple values separated with("|") like genre column because the move may represent more than one genre so I split the columns to four columns

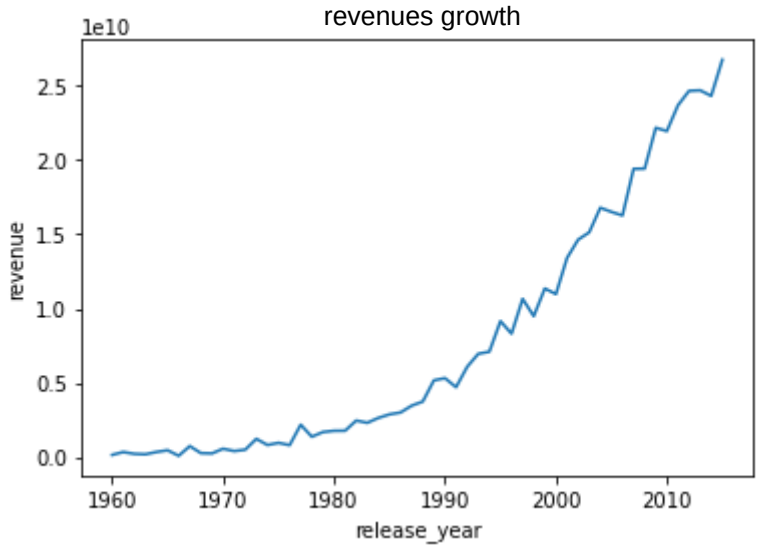
Second there is some columns has (_adj) at the end of it name so I changed it to Us dollars

Third the id column datatype is (int) and it should be a string because we don`t want to do any calculation with it while grouping data

Fourth there was a duplicated movie with id 42194 and that will mislead the count so I deleted it

Conclusion

1-in the first question the growth in movies revenues is increasing over time it starts on 1960 with less than 0.5 billion (145005000) and increased until it reach more than 26 billon (26762450518) at 2015



2-in the second one the number of movies of has Thriller in genres are made the most with more than 2500 movie (2787movie) and the movies with fantasy in it`s genres are made the least among the options with less than 1000 (875 movie)