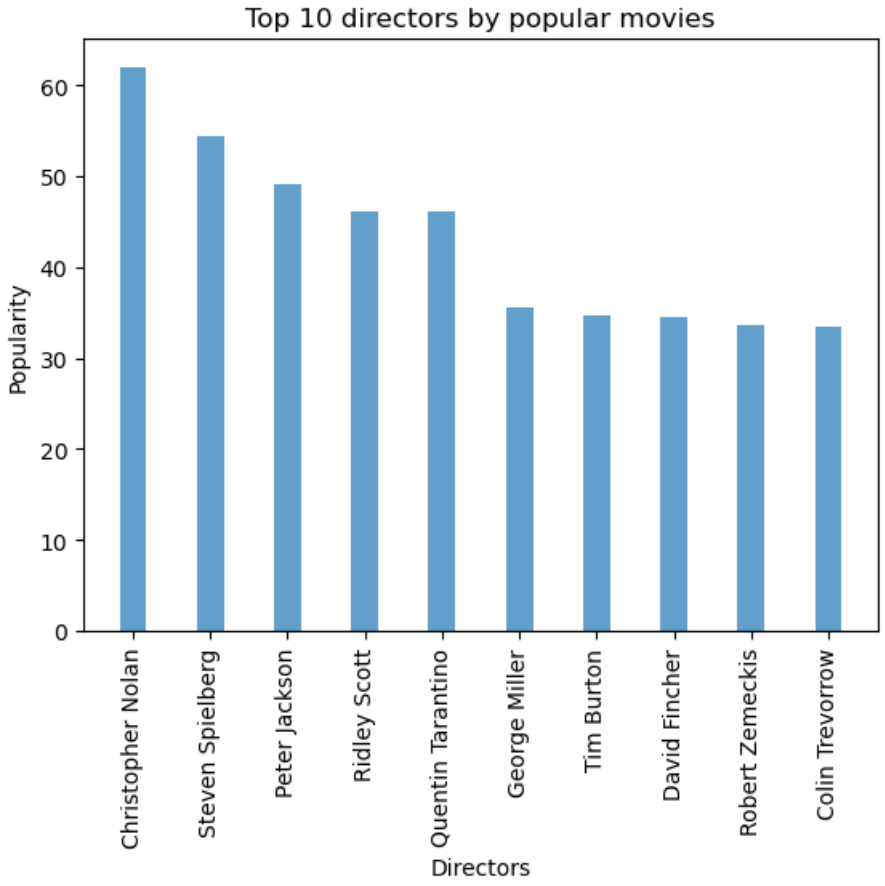
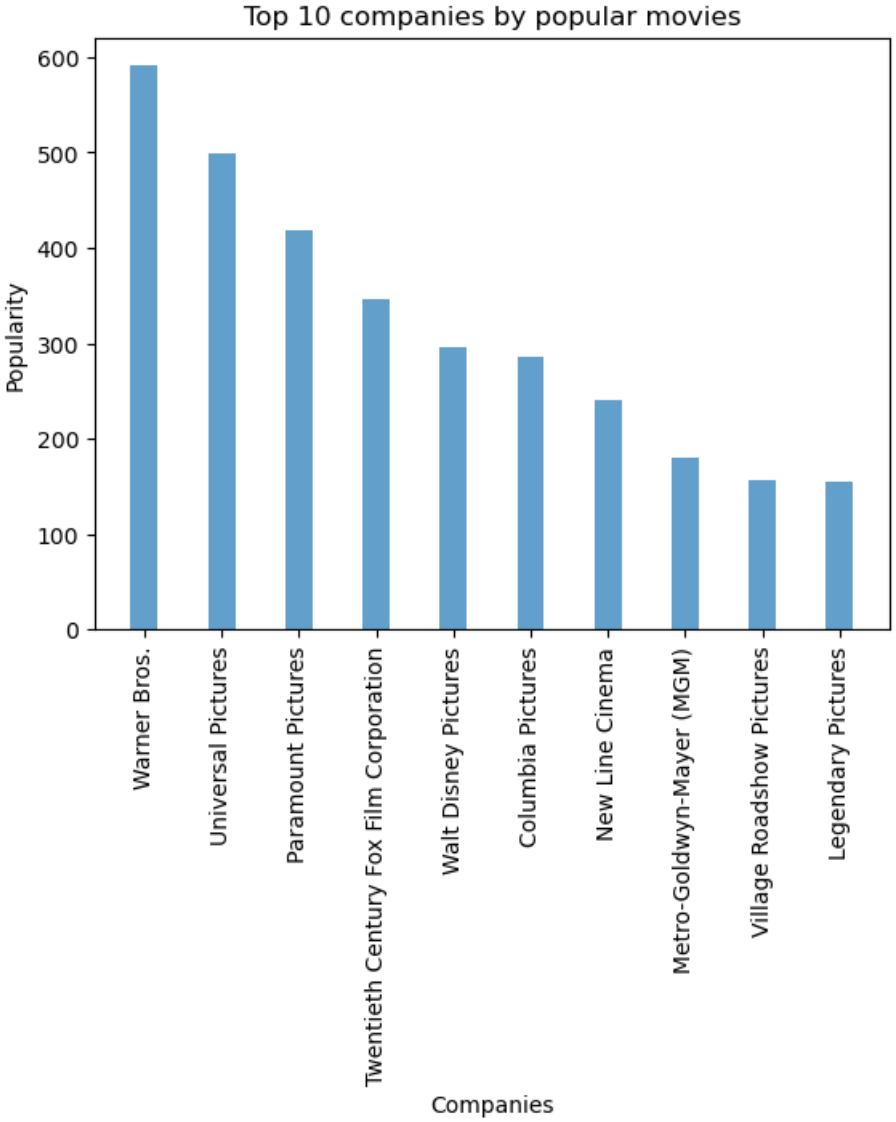
Project: Investigate a Dataset (TMDB movie data)

- .In this report I will use TMDB dataset to answer some questions:
    1. Which director made the popular movies?
    2. Which production company made the most popular movies?
    3. Which genre is the most popular among audience?
    4. In which year released the highest number of movies?
    5. Which production company achieved the highest profits over years?
    6. Which genre achieved the highest number of profits?
    7. Relations between variables?
- To answer the questions above I manipulate the original data to make a simple dataframe which I will use to answer intended question:
    1. For question one, I made dataframe contains the following information (movie title, directors, release year, popularity), then I check the data, Then I sum the popularity score of movies for each director then arrange Directors descendingly and plot a bar chart with top 10 directors
       And I found Christopher Nolan is the top director with most popular Movies as the sum of popularity score of his movies equals 61.955.

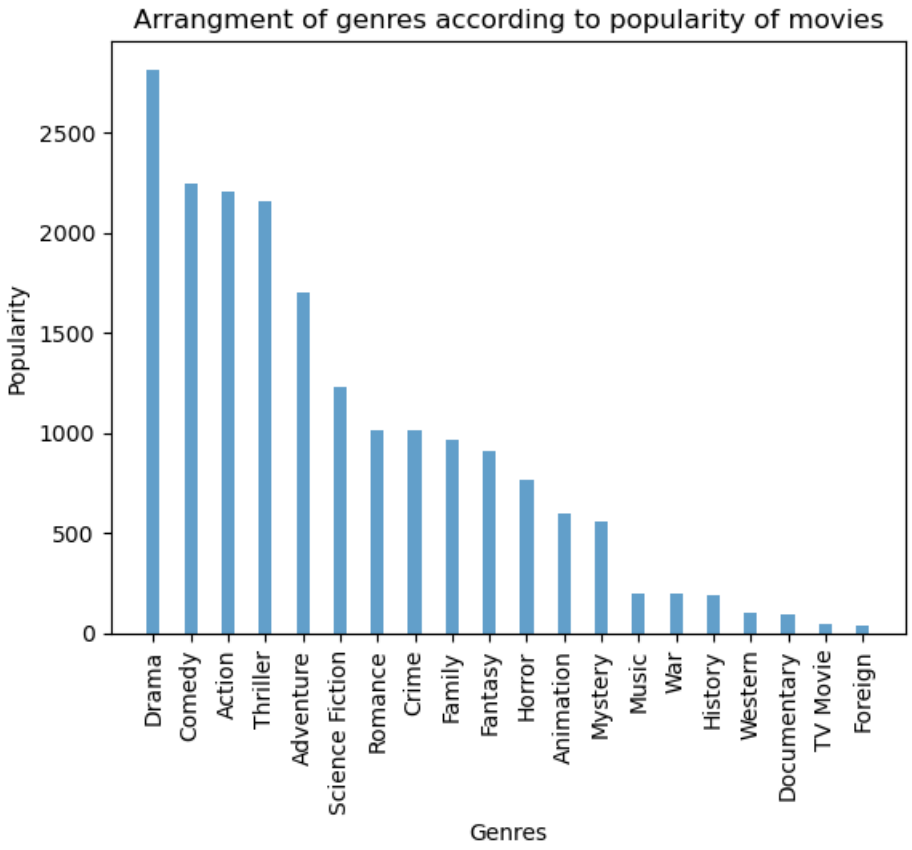


    2. For question two, first I make a function to split strings with "|" delimeter In a new dataframe and merge this dataframe with the original one and drop original string, then use the function to make a dataframe with split companies produced each movie, then I arrange the companies

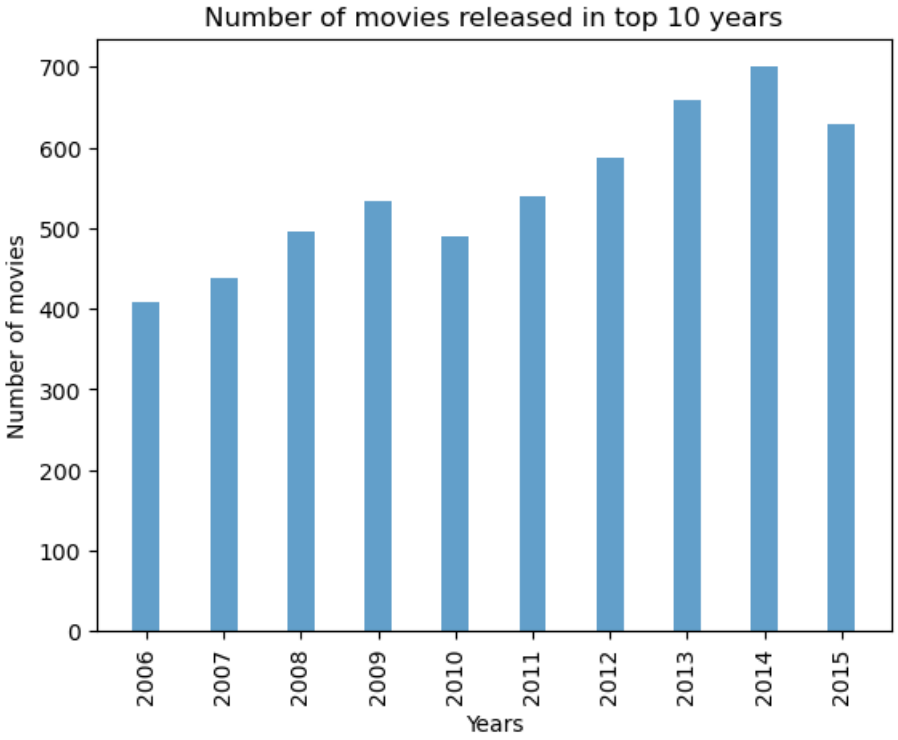
descendingly according to the sum of the popularity score, then I plot the top 10 companies and found Warner Bros is the top company as the sum of the popularity score of movies it produced equals 590.824.



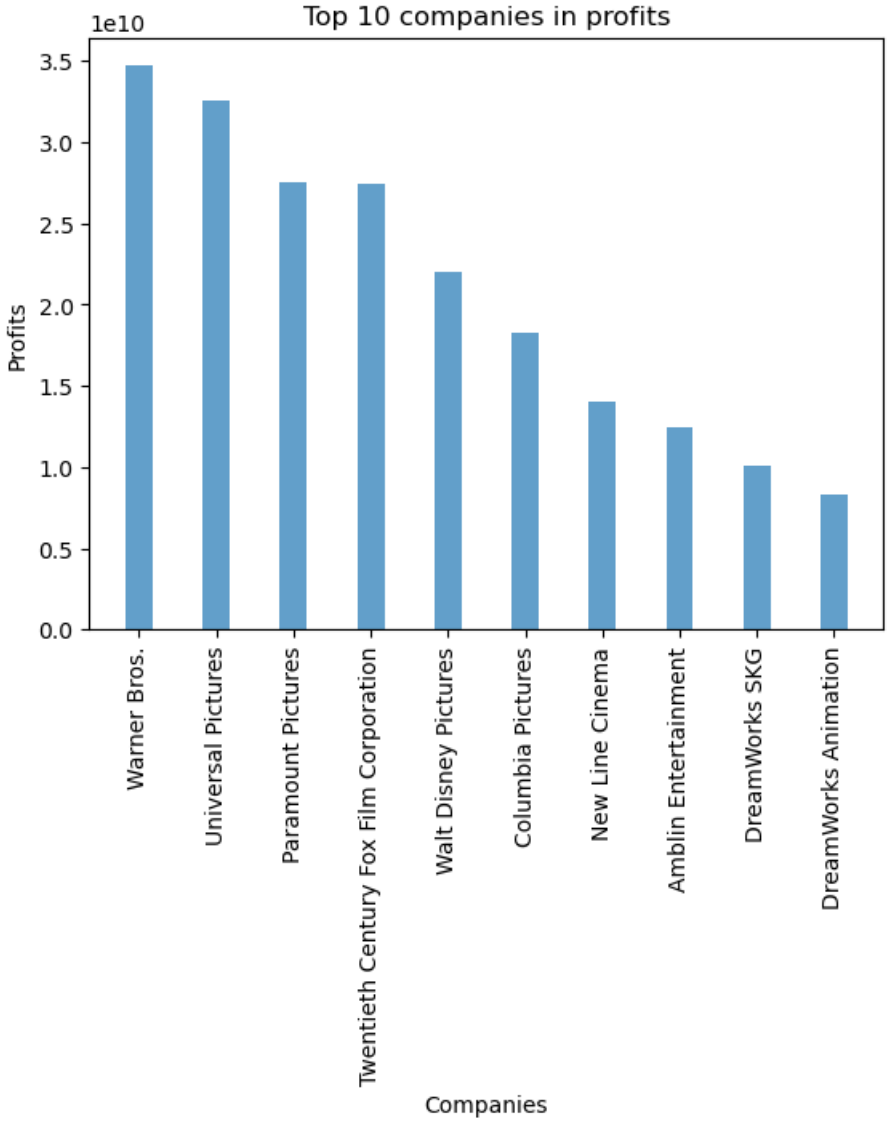
3. For question three, I use split function to split genres of each movie, the arrange genres descendingly according to the sum of the popularity score then I plot genres and found that drama is the most popular genre among audience with sum popularity score of it's movies equals 2815.517.



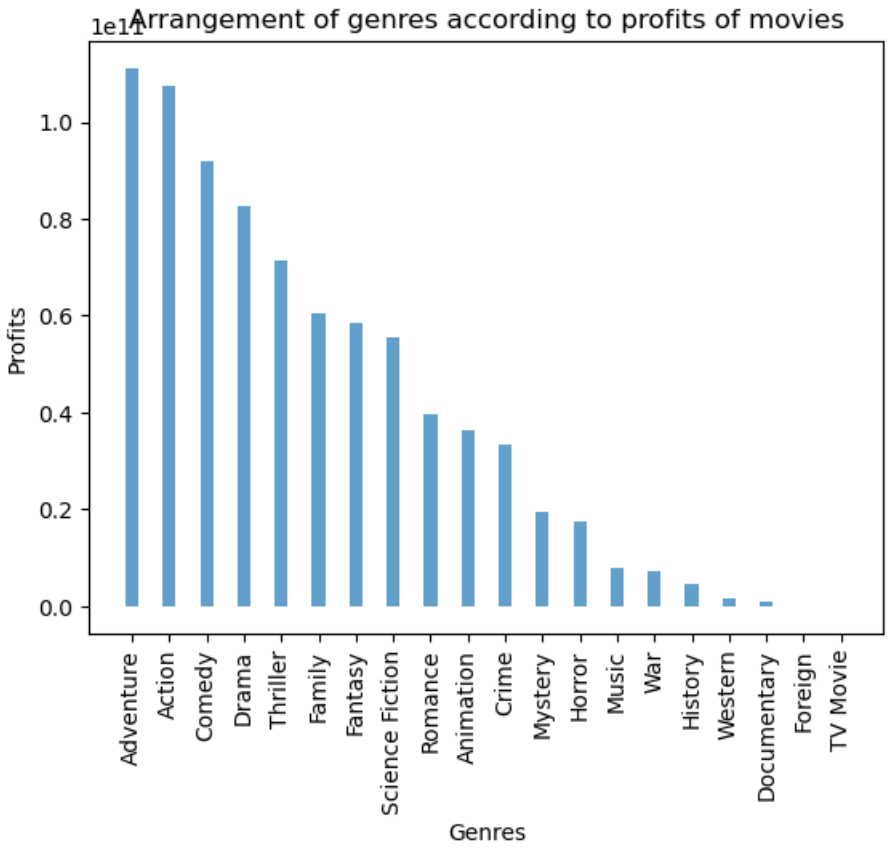
4. For question four, I make a dataframe contains release year of each movie then arrange years descendingly according to the sum of the number of movies released in that year and plot a bar chart with top 10 years and found the top year is 2014 with 700 movie released in that year.



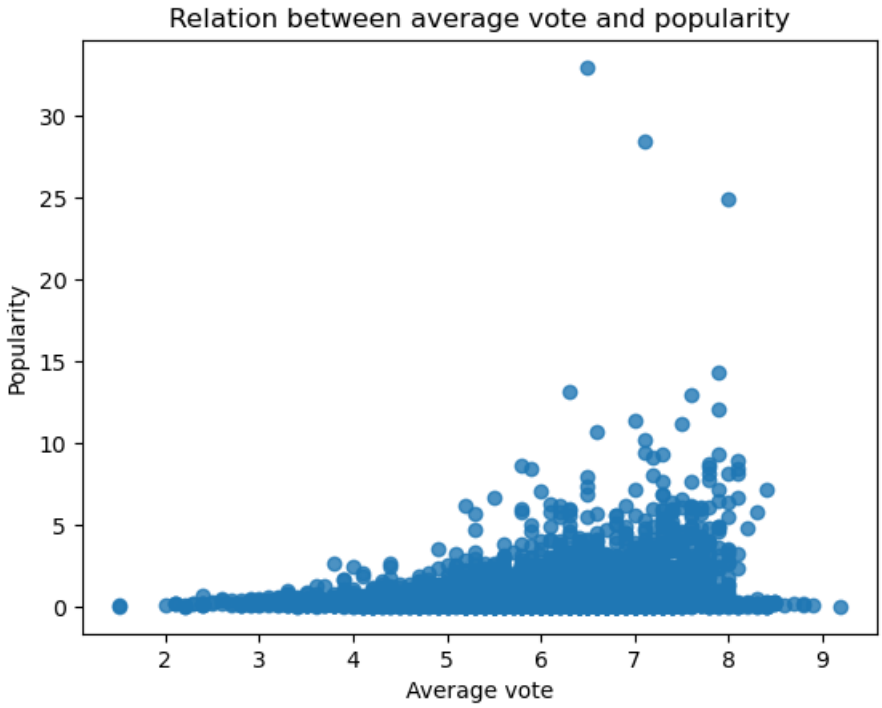
5. For question five, I use the dataframe of split companies produced each movie, then make a new column called profit (budget – revenue) then arrange companies descendingly according to profits if achieved from it's produced movies, then plot a bar chart with top 10 companies and found the top company is Warner Bros. as the sum of it's profits over years equals 34703823331 dollars.



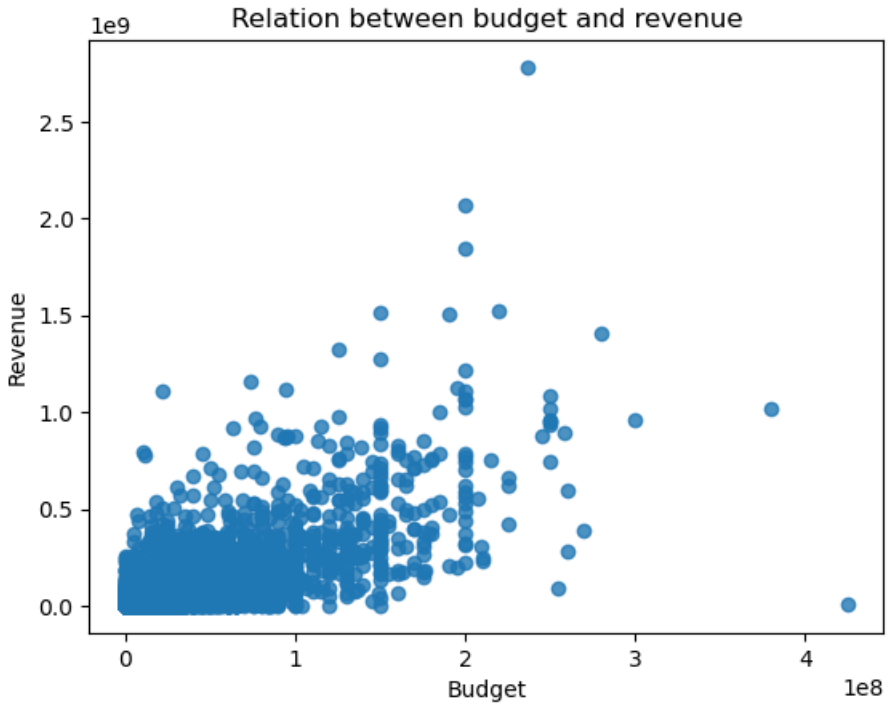
6. For question six, I use the dataframe of split genres of each movie then make a new column called profit (budget – revenue) then arrange genres descendingly according to profits it achieved from it's movies, then plot a bar chart with genres and found the top genre is Adventures. as the sum of it's profits over years equals 111199018978 dollars.



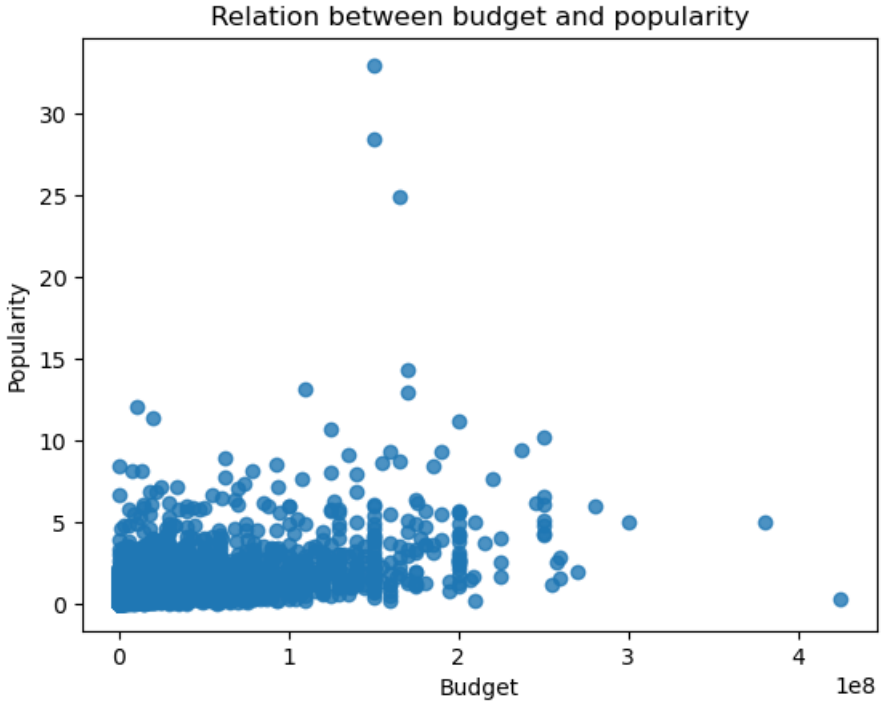
7. For question seven, I make a function to plot scatter plots between variables, I found the relation between avg. voting and popularity is positive correlation



And the relation between budget and revenue is +ve correlation



And the relation between budget and popularity is +ve correlation



- All plots (bar – scatter) are made by functions to avoid repititve code.
- Limitation: Around half of the data the budget value equals zero, so if I drop zero budget I will lose around half of my data , and the median equals zero so it is useless to replace zero budget with zero.

- Data wrangling:
  1. I load the dataset csv file using pandas and make id as index column to make it easy to merge dataframes.
  2. I check the dimension of the dataframe, datatypes, number of unique values of each variable in the dataframe.
  3. Check null values and the sum of them for each variable and the sum of dublicated rows.
  4. Overview statistics of data briefly
  5. Overview the distribution of data briefly by histogram and I found around of half of data has zero budget, and check this.
  6. I clean duplicated rows
  7. I clean useless columns(imdb id, homepage, keywords, overview, budget_adj, revenue_adj) and check cleaning process.
  8. I split strings with "|" delimeter with split function to use this variables(genres-production companies) in calculation and put split string in new dataframe.
  9. Add profit column in a new dataframe