# Project 2: Investigating movie dataset

The dataset I analyzed is tmdb-movie dataset

## Cleaning data:

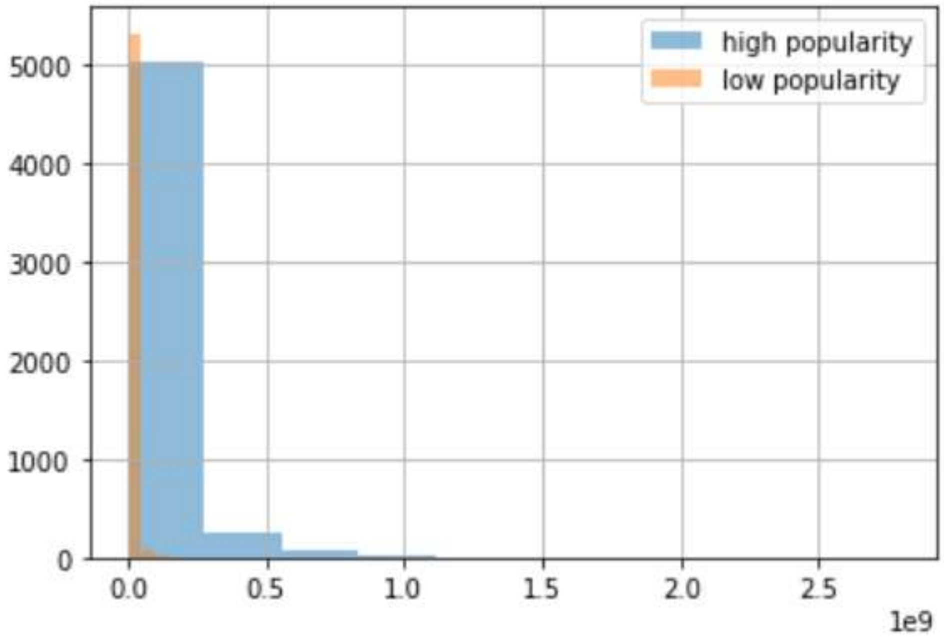
I loaded in my predownloaded tmdb-movies dataset and inspected it. I took look the first 5 rows of data to understand its structure, checked how many rows and columns in total and the name of each column. Later, I took look at the summary statistics of the entire dataframe. I pose some questions.

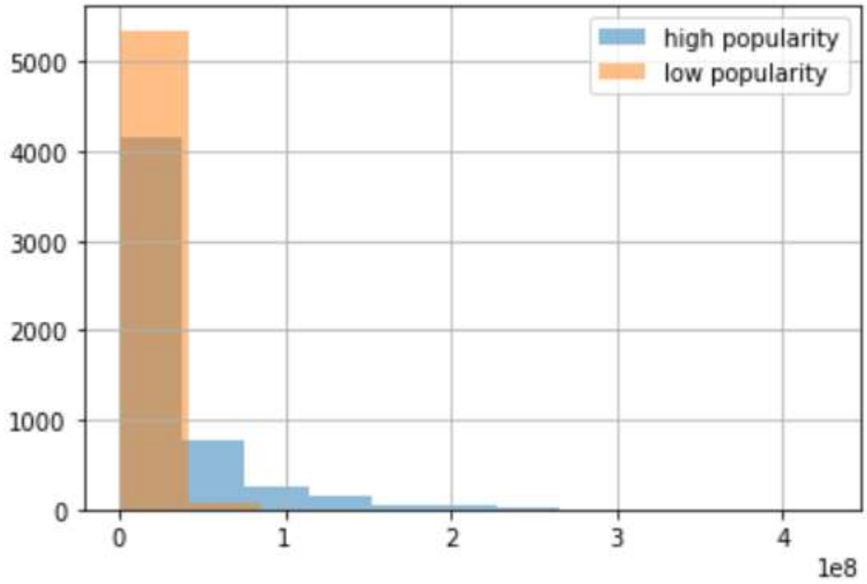
To justify my questons, variable 'popularity' would be the dependent variable and variables such as'budget', 'revenue', 'runtime', 'genres', 'vote average', and 'revenue-adj' would be used as my independent variables. After having the dependent and independent variables for my analysis, I started checking the missing values. I found variables 'cast', 'homepage', 'director', 'tagline', 'keywords', 'overview', 'genres', 'production_companies' having some missing values but majority of them are not in my analysis. I only need to consider the missing values of 'genres' as all other columns not related to my analysis would be dropped.I started inspecting the rows having null values in column ''genres', I made the histogram to compare to the histgram of original dataframe, I found the distrubution of the histgram for null values in genres has no big difference with the histgram of original dataframe plus some columns are marked as 0 which is not useful for my research. I decide to drop all those rows.

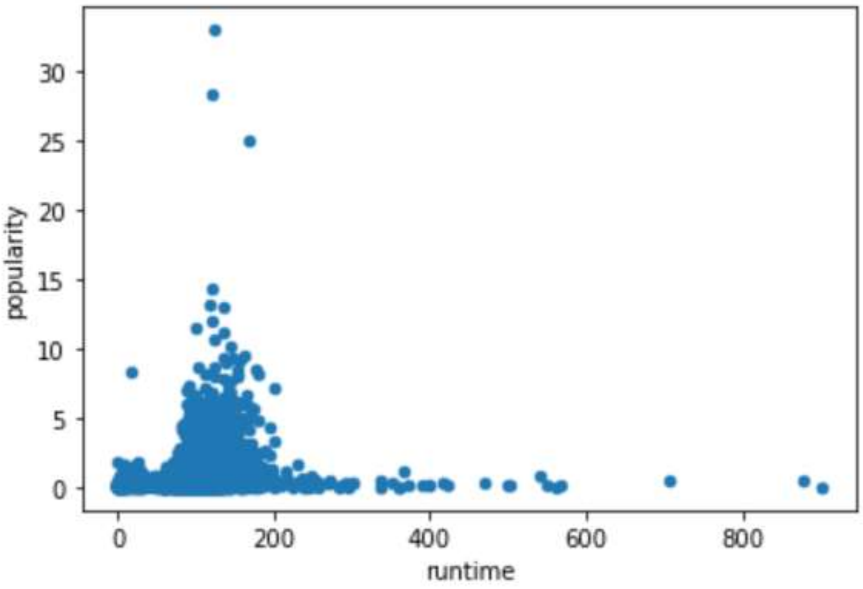
## Two questions:

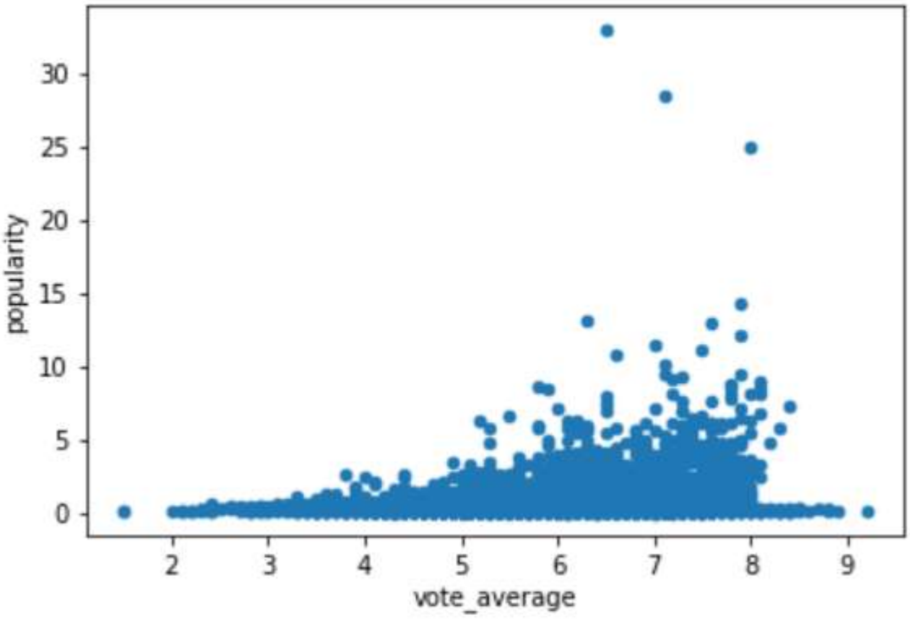
1. What’s the trend among the movies with high popularity and how they from the movie with low popularity
2. Which genre is more likely to have high popularity, action or comedy?

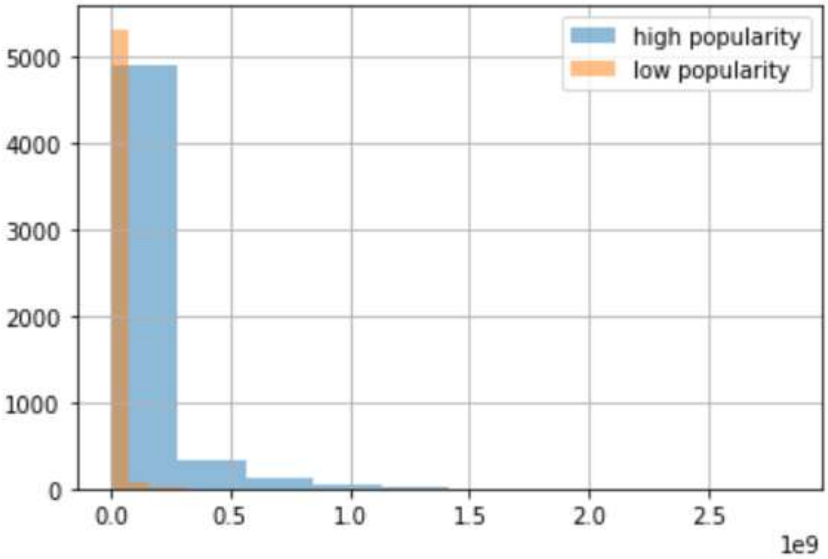
## Visulaization:

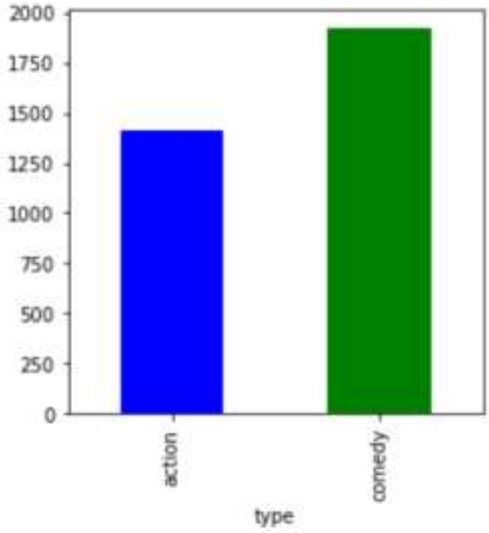

high popularity
low popularity
5000
4000
3000
2000
1000
0
0.0
0.5
1.0
1.5
2.0
2.5
1e9


high popularity
low popularity
5000
4000
3000
2000
1000
0
0
1
2
3
4
1e8


popularity
30
25
20
15
10
5
0
0
200
400
600
800
runtime


popularity
30
25
20
15
10
5
0
2
3
4
5
6
7
8
9
vote_average





## Conclusion:

Based on my findings, the varaiable popularity is correlated with the variables revenue, revenue_adj and genre. There is no relationship between popularity and runtime, vote_average, budget. In other words, the popularity of movie is based on how much revenue the moving gonna making, how much revenue of its associated movie making, and its genre. However, there is still some limitation of my exploration. Once the sample getter bigger and bigger, the correlation between the variable may change.