REPORT: ANALYSIS OF THE MOVIE DATABASE (TMDB)

The dataset I work on in this project is: **THE MOVIE DATABASE (TMDB)**

Questions:

- What's the most popular genres over years?
- what is the relationship between budget and popularity/votes

Description of steps to investigate the dataset:

- loaded dataset and the necessary packages

- checked the data frame shape, columns data types, null values, duplicates, zero values, unique values...

-cleaned the data frame by dropping unnecessary columns ,duplicates, null and zero values /rounded some columns values and change dates from strings to data format

-explored the variable vote average to see how movies are distributed en function of vote average by using **a histogram**

First question

Steps*:*

To answer the question: What's **the most popular genre over years?**

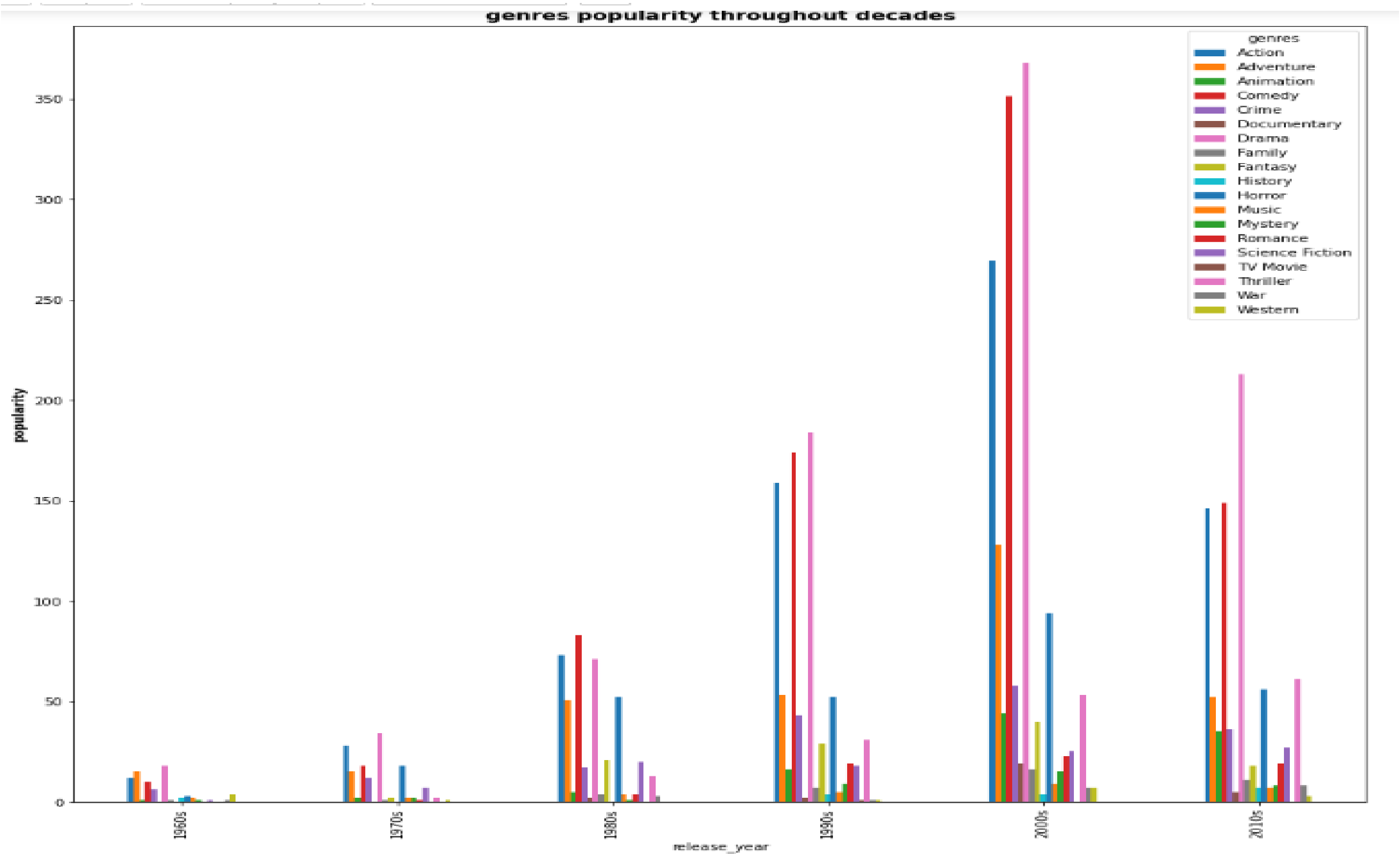
-divided the column release year to frames, every frame represent a decade

-split values of genres since there is multiple values separated by a '|' and get every value in a separate column (genre_1, genre_2,)

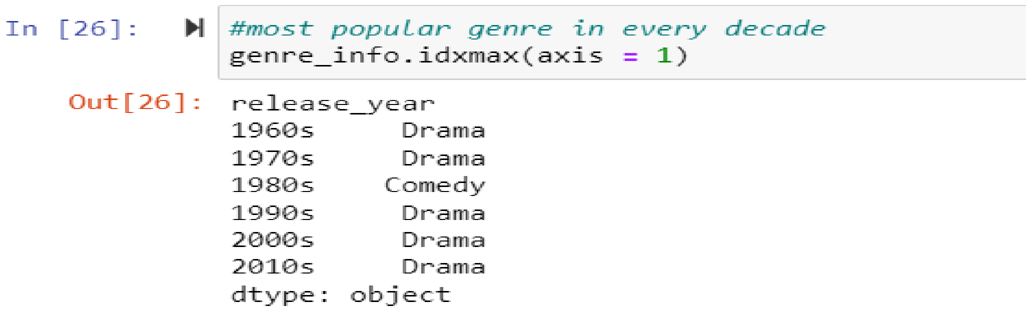
-changed old genres column with genre_1 in the last output

-and now that we have data ready, I used the function group by to group data according to genres and years

-finally I plotted the bar plot to see the most popular movies:



To check statistically result, we use the function idxmax () to find maximum values indexes (genres) for every decade:



```
In [26]:    #most popular genre in every decade
            genre_info.idxmax(axis = 1)

Out[26]:    release_year
            1960s      Drama
            1970s      Drama
            1980s     Comedy
            1990s      Drama
            2000s      Drama
            2010s      Drama
            dtype: object
```

Conclusion:

Movie genres popularity varies over time periods but drama occurs more to be the most popular

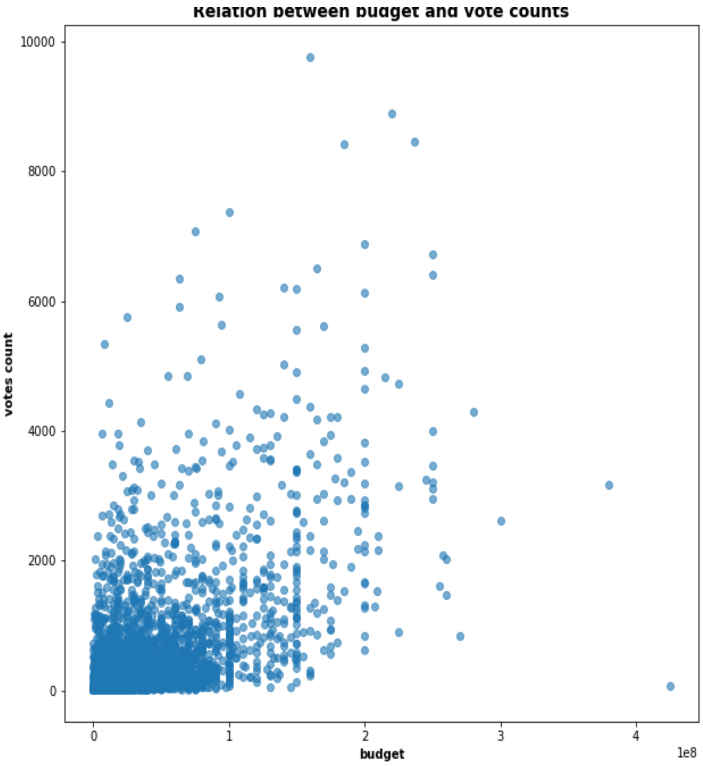
Second question:

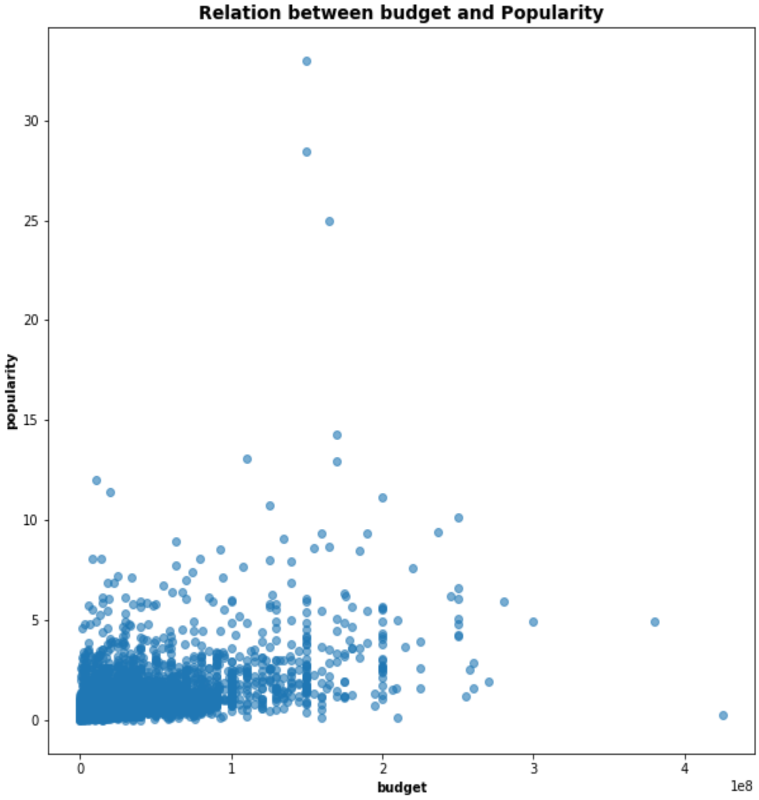
Steps:

To answer the question: what is the relationship between budget and popularity/votes?

I used function scatter_RP that plots a scatter using 2 columns and then implement it to plot for columns budget and popularity and then budget and votes count

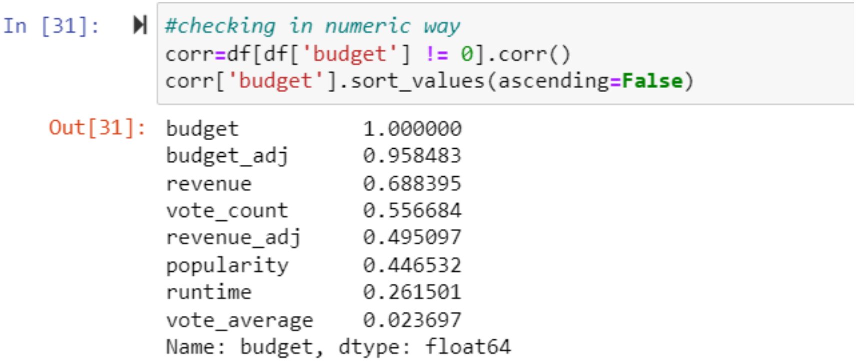
I had the following plots:





Through the plot **Relation between budget and popularity** observation , we notice that correlation between budget and popularity is not significant , while in the second plot **Relation between budget and vote counts** the 2 variable tend to have a positive significant correlation.

To check statistically the result, we generate the correlation matrix between df columns, and we see correlation values between budget and popularity / budget and votes count using function corr ()



```
In [31]:  #checking in numeric way
          corr=df[df['budget'] != 0].corr()
          corr['budget'].sort_values(ascending=False)

Out[31]:  budget          1.000000
          budget_adj      0.958483
          revenue         0.688395
          vote_count      0.556684
          revenue_adj     0.495097
          popularity      0.446532
          runtime         0.261501
          vote_average    0.023697
          Name: budget, dtype: float64
```

Conclusion:

0.45<0.5 is not a significant value, so there is no strong correlation between budget and popularity, while 0.55>0.5 is significant, hence there is strong correlation between budget and popularity