Movies Dataset Analysis Summary

Questions about the Dataset

- ### *1.the Data Structure:*
    - *Does Null Values Affect The analysis?*
    - *Are there Important Missing Values?*
    - *Are there Columns with distorted data?*
    - *Are the Columns Data Type Correct?*

- ### *2.Data Cleaning:*
    - *How To Replace The Missing Values?*
    - *Is there's a column that's not important*
    - *Replacing distorted data*

- ### *3.Data Analysis:*
    - *How Can a Movie Become Successful?*
    - *Does the Director Affect The Success of The Movie?*
    - *Does The Runtime of the Movie Affect the Success of the Movie?*
    - *Which Year Had The Largest Amount Of Movies ?*
    - *Does The Budget of the Movie Affect The Revenue?*

Conclusion :

- ***1.the Data Structure:***
  - *Does Null Values Affect The analysis?*
    - *we didn't remove null values as they didn't affect the analysis negatively*
  - *Are there Important Missing Values?*
    - *Many Important Columns Had Missing Values That Was written as 0*

column	count
budget	5696
revenue	6016
runtime	31
budget_adj	5696
revenue_adj	6016

  - *Are there Columns with distorted data?*
    - *the 'genre' column had text that we couldn't work with so we extracted the important genre from the text*
  - *Are the Columns Data Type Correct?*
    - *the 'release_date' Column is not formatted correctly so we changed the data type to datetime*

- ***2.Data Cleaning:***

    - *How To Replace The Missing Values?*

        - *we tried using web scraping but the code took a long time so i tried alternatives,*
            - *dropping rows with multiple missing values*
            - *predicting the remaining values using machine learning*

    - *Is there's a column that's not important*

        - *id,tagline,homepage columns didn't have any importance in the analysis so they were removed*

    - *Replacing distorted data*
        - *the important values were taken from the data as in the 'genre' column*

- ***3.Data Analysis:***

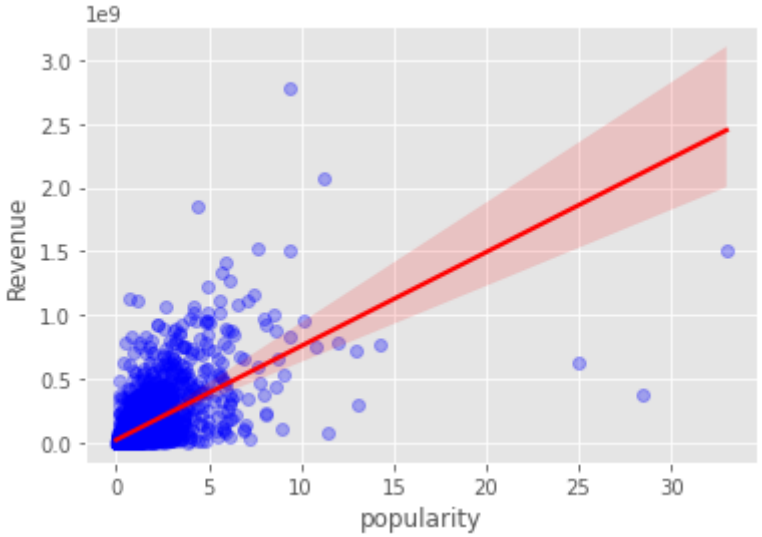
    - *How Can a Movie Become Successful?*
        - *as shown in the relation between columns and revenue some columns showed a high correlation*

            - *budget*

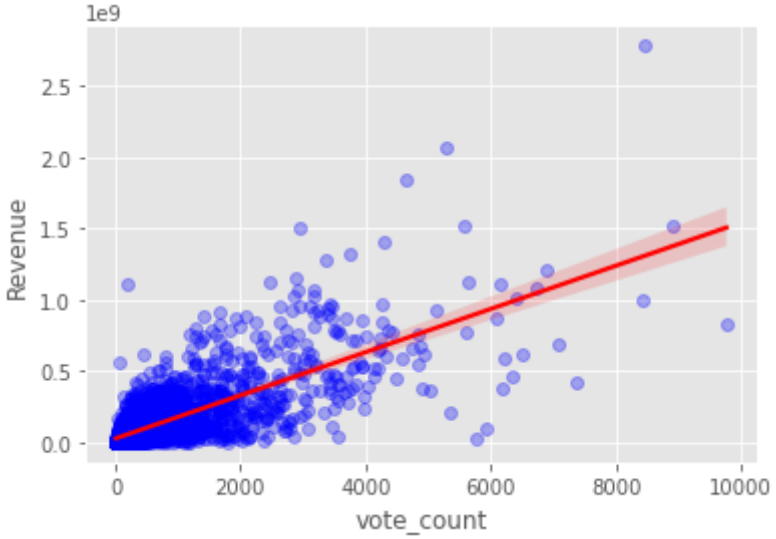
              a bigger budget is shown to result in a more successful movie (corr = 0.73490068)

- *Popularity*

  popularity correlates with the revenue so stuff like marketing campaigns for movies can affect the movie revenue success (corr=0.66335837)



- *.vote_count*

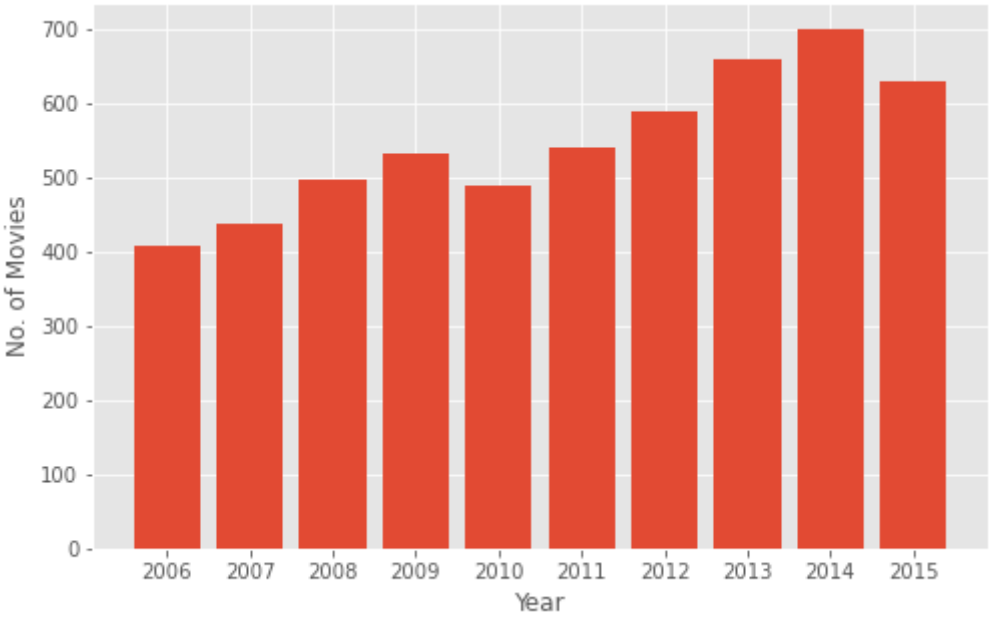


- *Does the Director Affect The Success of The Movie?*
  - *some directors had a high percentage of successful movies which makes them more likely to have a successful movie than other*

director	movies	successful movies	percentage %
Woody Allen	45	14	31
Clint Eastwood	34	21	61
Martin Scorsese	29	12	41
Steven Spielberg	29	26	89
Ridley Scott	23	15	65
Steven Soderbergh	22	10	45
Ron Howard	22	15	68
Joel Schumacher	21	13	61
Brian De Palma	20	5	25
Barry Levinson	19	10	52
Wes Craven	19	10	52
Tim Burton	19	14	73
Mike Nichols	18	9	50

- *Does The Runtime of the Movie Affect the Success of the Movie?*
  - *runtime showed no correlation with the revenue (corr =0.16283789)*

- *Which Year Had The Largest Amount Of Movies ?*
  - *as shown in the analysis year 2014 had the most movies with 700 movies*


700
600
500
400
300
200
100
0
No. of Movies
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Year