PROJECT

DESCRIPTION

For this project, I chose to analyze an IMDB dataset with 1000 movies records from 2006 to 2016. I picked this topic because there has been many discussion on whether online movies aggregator databases can affect a movie's revenue or not. Also, because I love movies.

MAIN

QUESTIONS

Much has been stated about the relation between a movie genre and its gross box office so I went for 3 other relationships: rating vs revenue, rating vs votes and revenue vs votes.

My main questions were:
1) Do higher ratings mean higher revenue?
2) Do more votes mean higher rating?
3) Are higher-grossing movies watchers more prone to vote?

WORKFLOW

The workflow for this project was as listed below:
1) choose a topic, research and make questions about it;
2) choose and get data suitable to answer the questions;
3) clean and prepare the data to use as a dataset;
4) perform exploratory data analysis to have a deeper knowledge about the data;
5) perform statistical analysis in search for correlations;
6) do meaningful visualizations;
7) draw conclusions.

PARENT

DATASET

For this project, I used a dataset from Kaggle with the 1000 most popular movies on IMDB from 2006 to 2016.
The data fields included in the dataset are organized into twelve columns. The columns info is summarized in the table below.

Dtype	Column	Description
numerical	Rank	movie rank order
object	Title	the title of the film
object	Genre	a comma-separated list of genres
object	Description	brief one-sentence movie summary
object	Director	the name of the film's director
object	Actors	a comma-separated list of the main stars of the film
numerical	Year	the year the film released as na integer
numerical	Runtime (min)	the duration of the film in minutes
numerical	Rating	user rating for the movie (0 to 10)
numerical	Votes	number of votes
numerical	Revenue (Millions)	movie revenue in millions
numerical	Metascore	an aggregated average of critic scores. Values are betwwen 0 and 100. Higher scores represent positive reviews

The descriptions given in the above table are self explanatory, nevertheless it is important to clarify IMDB's popularity rank.
IMDB ranks the popularity of a particular title based on the number of times the title's page has been visited during the refering period of time.

CLEANIND AND

MANIPULATING THE DATA

The original dataset from Kaggle had 12 columns and 1000 records.

After a general overview on the dataset, I noticed that there were some missing values on the revenue and metascore fields. I kept the records with no metascore info, since I wasn't performing any analysis on that, but I dropped the records with no revenue info, for the oposite reason.

In the original dataset, the total number of votes per movie range from 178 to 1 791 916 votes and the mean is 134 654 votes. Since I don't consider that movies with few votes are relevant for my analysis, I decided to establish a minimum number of votes for a movie to be kept in the dataset.

For that I adopted IMDB's criterion for a movie to appear in the Top 250 best movies which establishes a minimum 30 000 votes. Thus, I dropped all movie records with less votes than this.

I also added some new columns, based on the existing ones, needed to achieve the goals of my analysis.

In the end of the data cleaning, I had 16 columns and 756 records to work with.

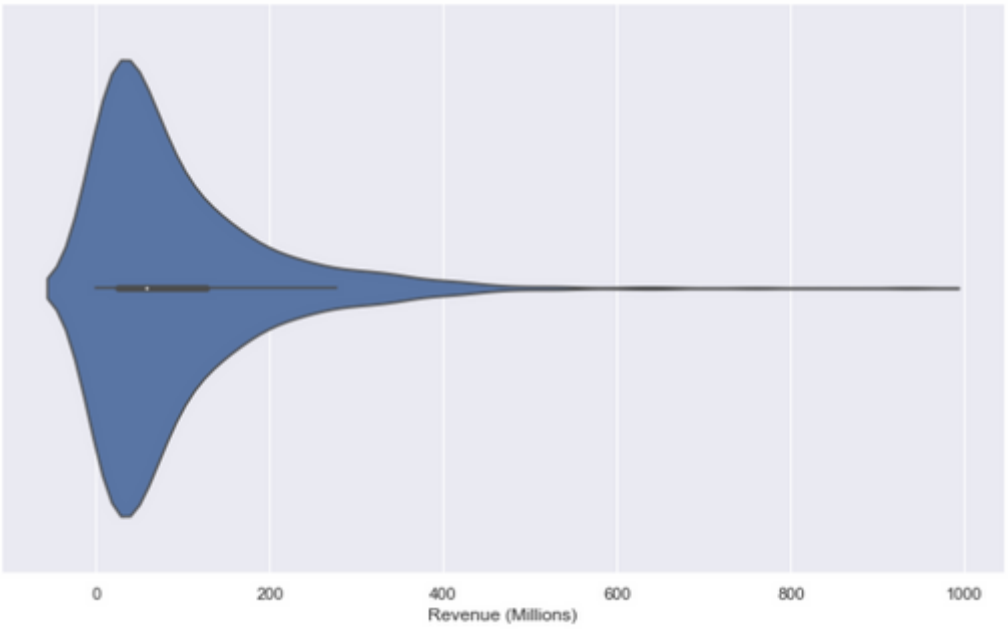
For the data manipulation, I used nothig but usual methods and functions, the one thing that stood out was the procedure used to define the director's gender. The procedure is explained in the notebook along the the code.

PERFORMING THE

ANALYSIS

I started with some exploratory data analysis to get some general insights about my dataset, specially revenue, rating and votes.

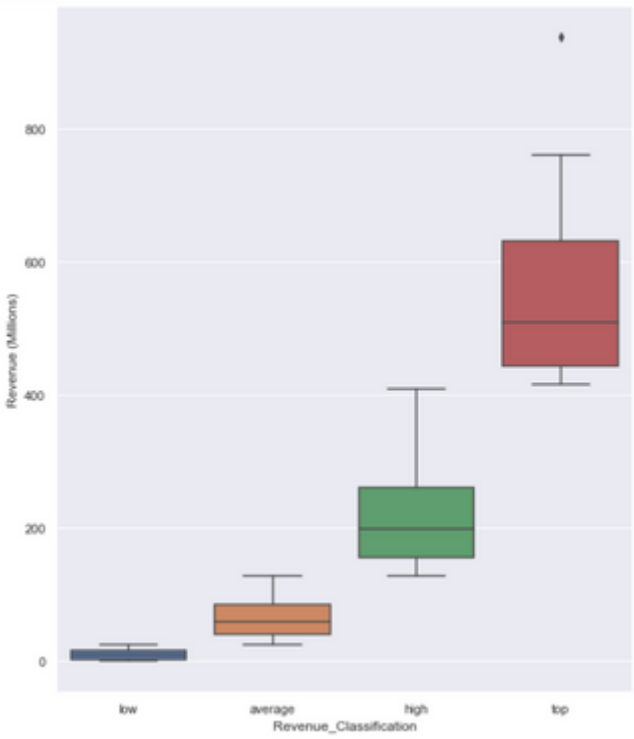
Revenue
The plot bellow shows the distribution of the movies revenue which is very spread.



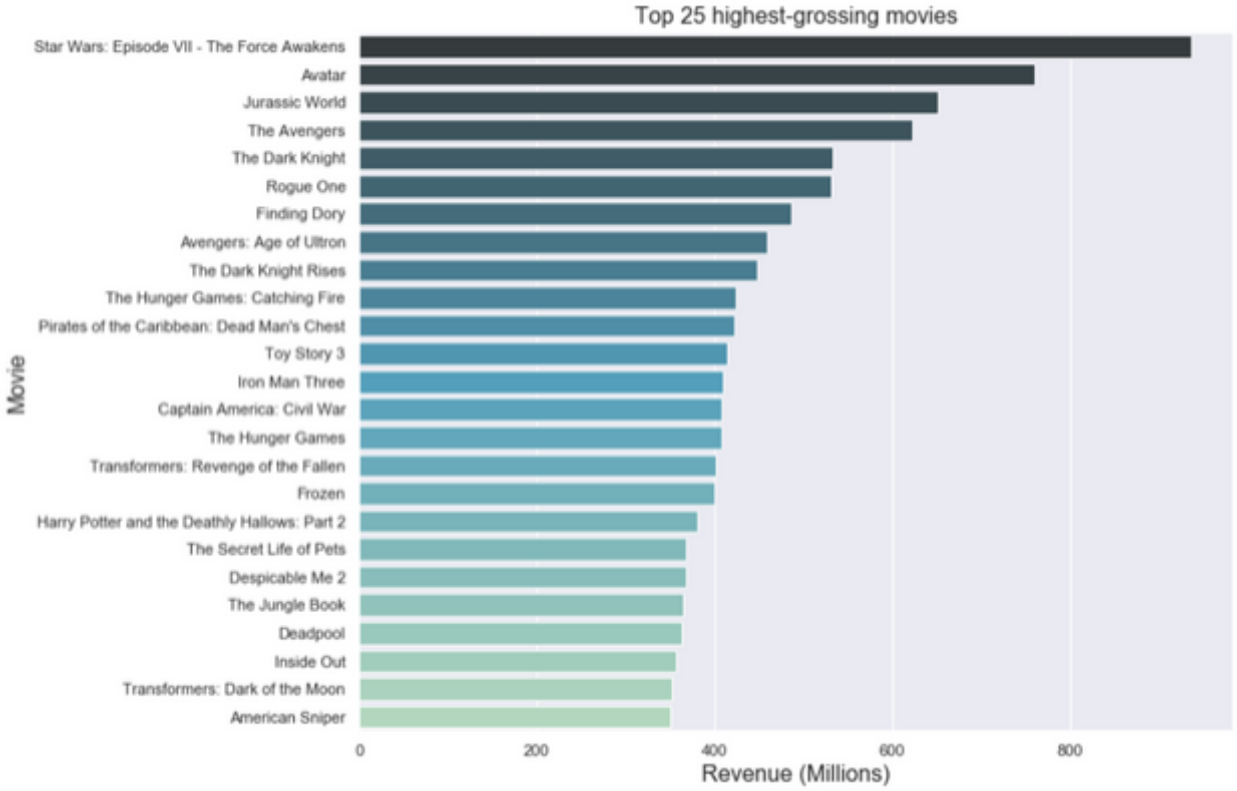
Because of the data dispersion, I binned the revenue data based on its statistics summary. The bins were defined as follows:

- up to 25% percentile = low revenue;
- 25% to 75% percentile = average revenue;
- 75% to mean + 3 standard deviations = high revenue;
- beyond mean + 3 standard deviations = top revenue.

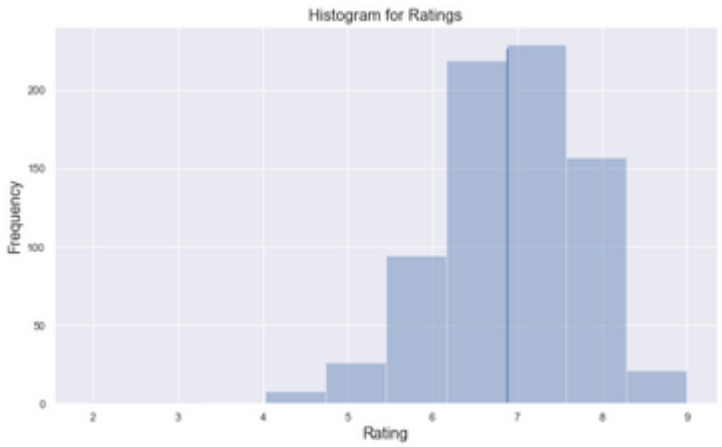
The side plot shows the binning results.

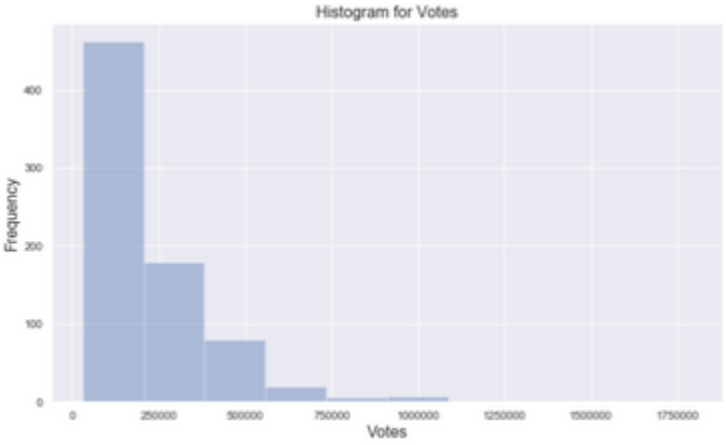


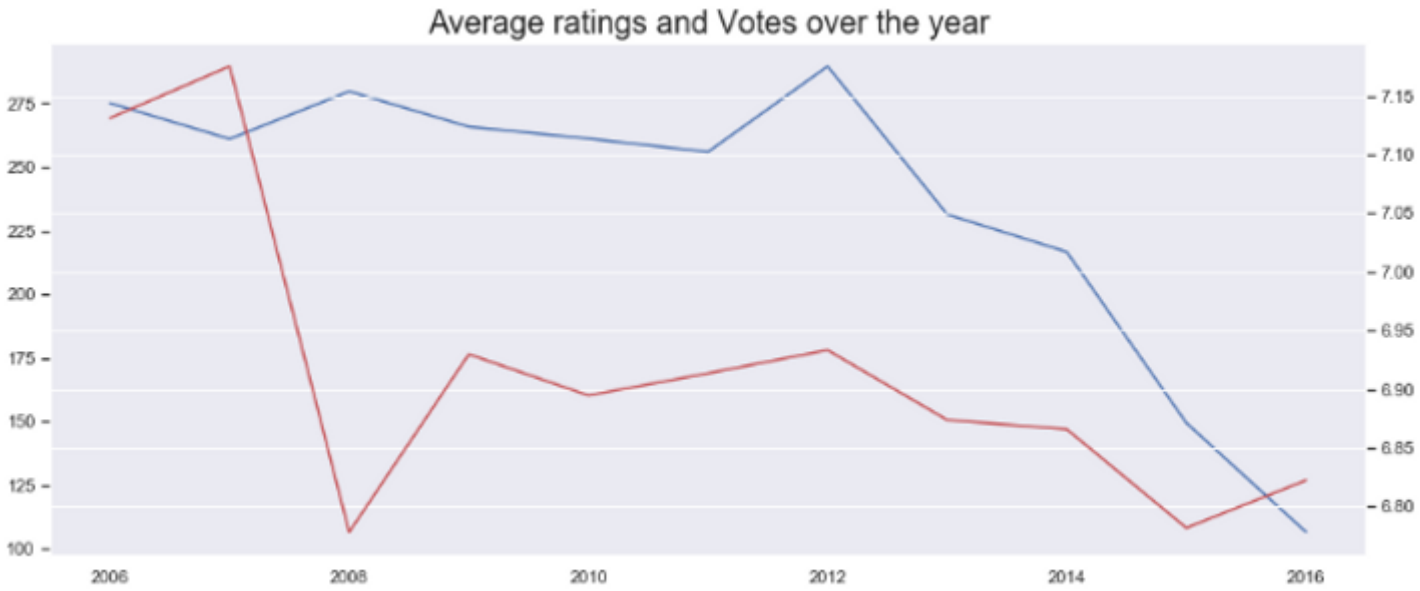
The plot bellow shows the top 25 highest grossing movies on the dataset.

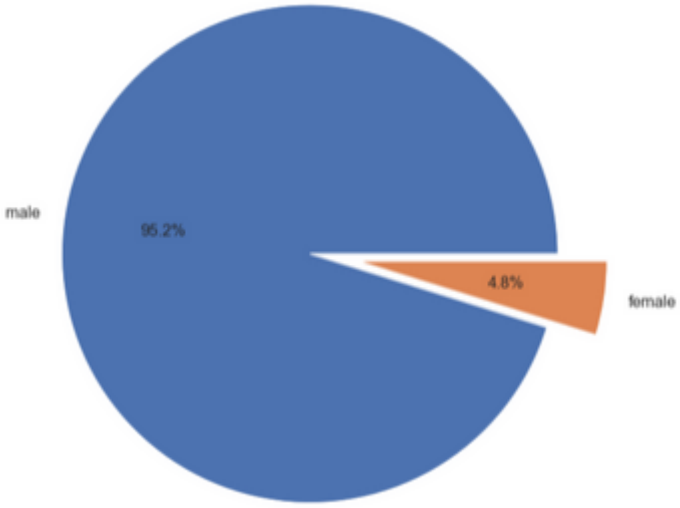


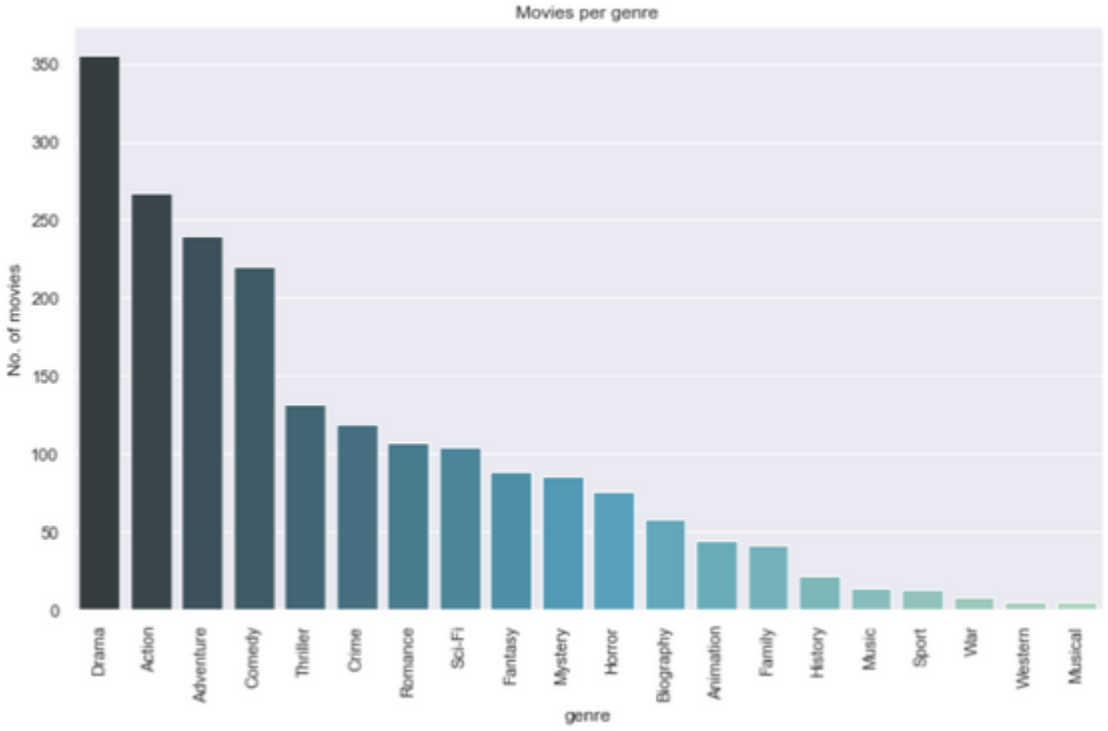
Below some general data visualizations are shown.

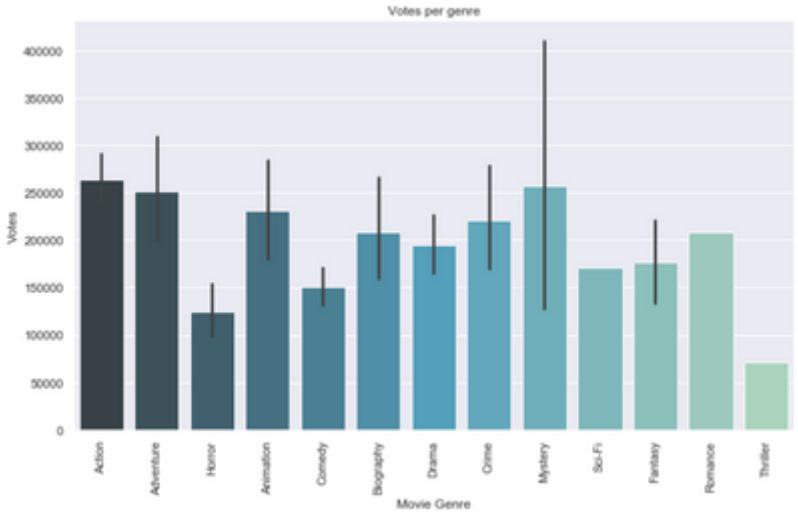


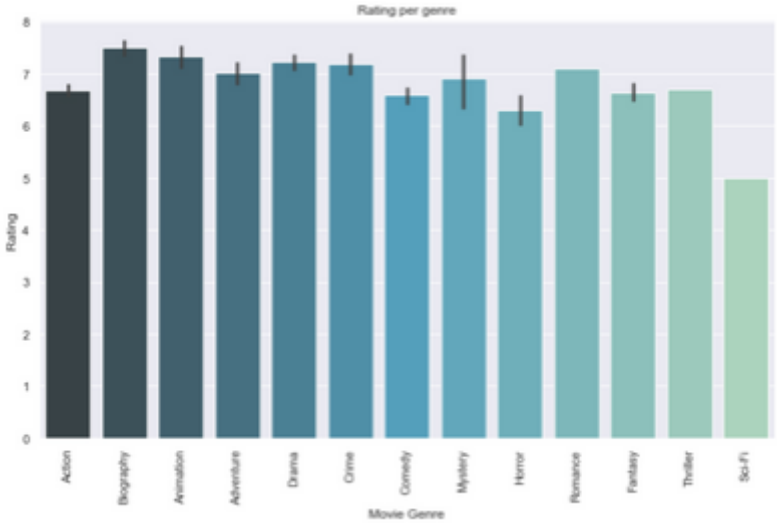







Movies per genre
No. of movies
0
50
100
150
200
250
300
350
Drama
Action
Adventure
Comedy
Thriller
Crime
Romance
Sci-Fi
Fantasy
Mystery
Horror
Biography
Animation
Family
History
Music
Sport
War
Western
Musical
genre


Votes per genre
Votes
0
50000
100000
150000
200000
250000
300000
350000
400000
Action
Adventure
Horror
Animation
Comedy
Biography
Drama
Crime
Mystery
Sci-Fi
Fantasy
Romance
Thriller
Movie Genre

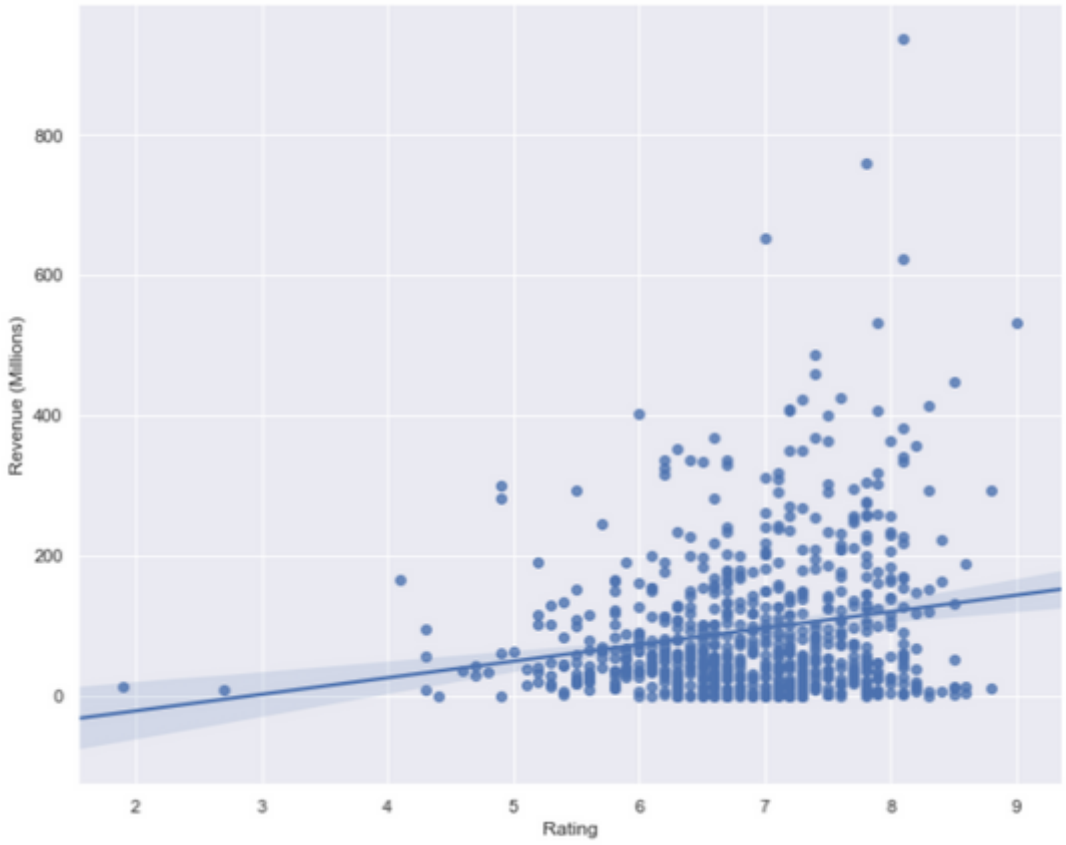

Rating per genre
Rating
0
1
2
3
4
5
6
7
8
Action
Biography
Animation
Adventure
Drama
Crime
Comedy
Mystery
Horror
Romance
Fantasy
Thriller
Sci-Fi
Movie Genre

RESULTS AND

CONCLUSIONS

Do higher ratings mean higher revenue?

The point of this question was to understand if better rated movies are also higher gross revenue movies.

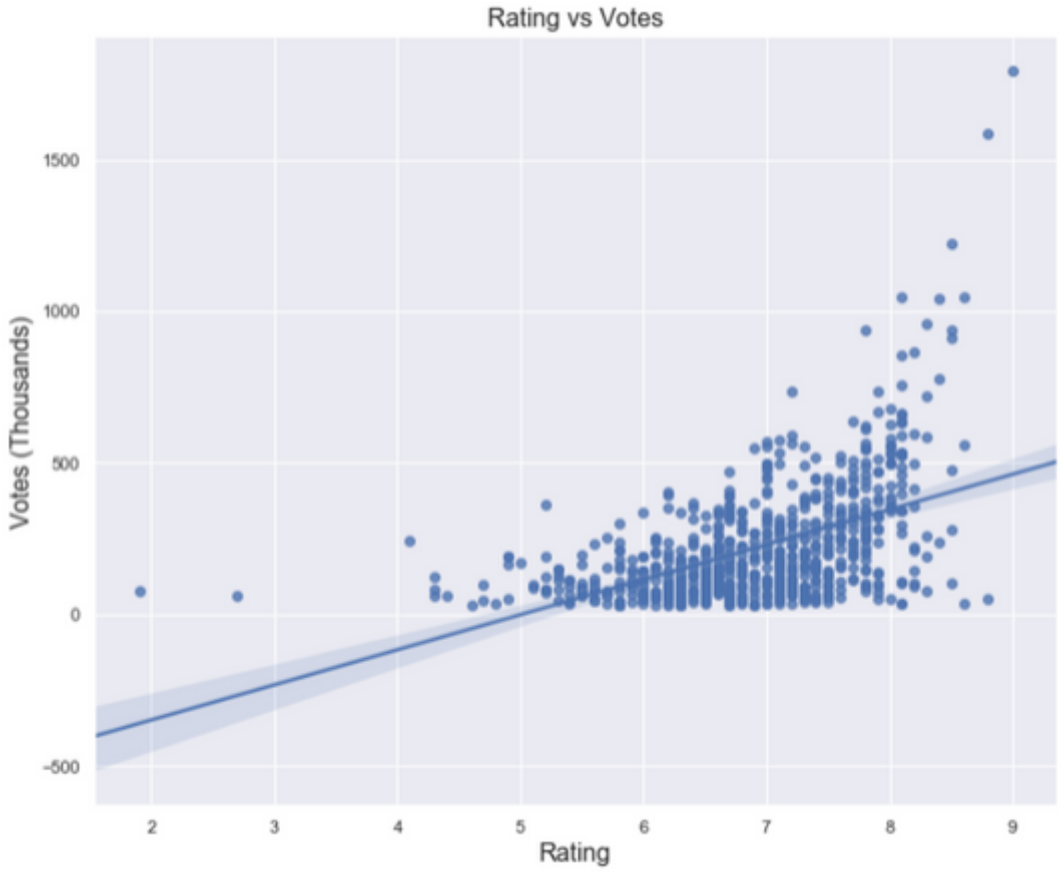


As the scatter plot shows, most of the movies fall within the rating range from 6 up to 8 on the bottom area which means average or low revenue movies. In this visualization, the angle of the plotted point cloud is pretty flat which indicates a weak correlation.

Actually, the correlation coefficient between these two variables is 0.19, less than 0.5, which confirms that there is no correlation and so the fact of being a better movie doesn't mean that it has a higher revenue.

Do more votes mean higher rating?

With this question, I wanted to check if a movie gets more votes because it is a better movie or not.



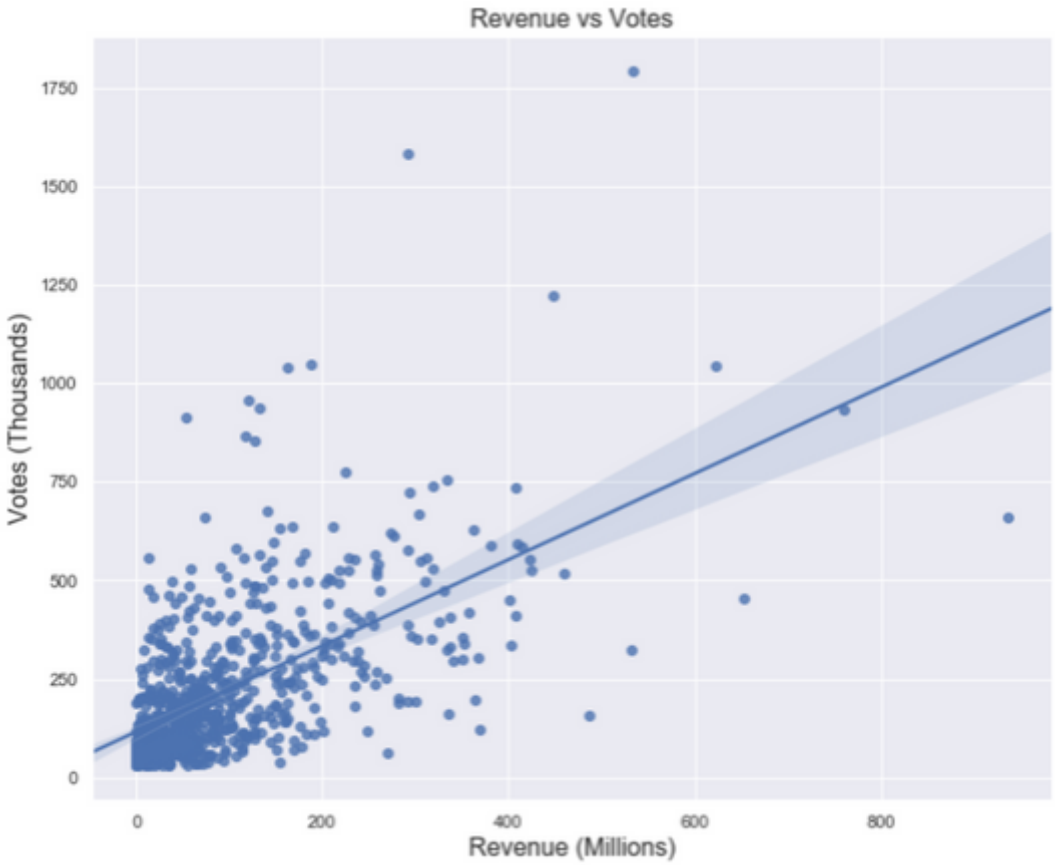
The scatter plot shows that most of the higher rated movies have less than 500 000 votes and concentrate on the bottom area. In this case, the correlation coefficient is 0.51 which means that there is a correlation although not a meaningful one.

Although there is no linear correlation, there may be a non-linear correlation, as the visualization kind of sugests an exponencial relationship.

So, as next steps, I'll try to fit an exponential curve in this distribuion.

Are higher-grossing movies watchers more prone to vote?

The point of this question was to understand if people who to go watch better rated movies are more prone to vote for the movie.



Looking to the scatter plot, this time the points cloud is at an angle and the points seem to fall closer and along the regression line which means that there is a correlation, although probably not a strong one.

This was actually the best correlation coefficient that I got: 0.60.

So, in this case, there is a relationship between the revenue and the number of votes and it can be said that the more votes a movie has, the higher its revenue. This result makes sense because more votes mean more people that watched the movie and that means more gross box office revenue.

Correlation Matrix

To summarize what has been said, find below the correlation matrix that illustrates the level of correlation between the variables analyzed in this project.