NAME: Joyce Kibii

STUDY: DSC Part Time

GIT REPOSITORY: https://github.com/purpleturtle11/dsc-phase-1-project

MOVIE ANALYSIS FOR MICROSOFT STUDIO

BUSINESS BACKGROUND

Microsoft Corporation is a multinational technology corporation.

The company is 47 years old.

Major products made by Microsoft include:

- Computer software
- Consumer electronics
- Personal computers

BUSINESS PROBLEM

Microsoft has decided to have a new movie studio to create movies. This is however not their field of expertise and they need pointers to making the right type of movie.

- I have explored the nature of the moves that are currently showing best performance at the box office.
- I also displayed how length of film, genre and customer feedback translates to good performance of a movie.

SOLUTION PROVIDED

Using exploratory data analysis method:

I have explored the nature of the movies that are currently showing best performance at the box office.

I also displayed how length of film, genre and customer feedback translates to good performance of a movie.

TECHNOLOGIES USED

The following Python libraries were used in the analysis:

1. Pandas
2. Numpy

To create visualization:

1. Matplotlib
2. Seaborn

DATA UNDERSTANDING

Data used was from from two files from IMDb.

1. The first dataset contains information on title basics: 'title.basics.csv'

   It has 6 columns namely: **tconst, primary_title, original_title, start_year, runtime_minutes and genres**

2. The second dataset contains information on title ratings: 'title.ratings.csv'

   It has three columns namely: **tconst, averagerating and numvotes**

In order to further understand the data, the following aspects were assessed

1. The number of columns in each dataset.
2. The general information contained in each dataset. E.g:
    - The data type of each column
    - The none null count
3. Descriptive statistics for numerical columns in each dataset. E.g:
    - Mean
    - Standard deviation
    - Median
    - Quartiles

We also checked for factors that would affect our analysis.

1. **Outliers:** These are extreme values that would be outside the normal range. We handled them by either replacing it with the median or capping them at a specific value.

2. **Duplicates:** Multiple records affect results therefore we ensured our data did not have any duplicated records.

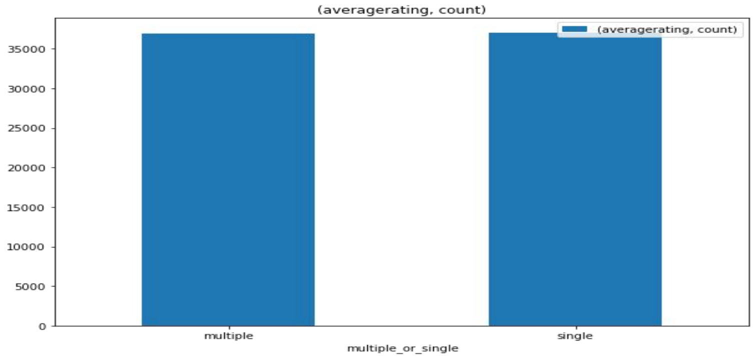
3. **Missing values:** These were indicated by Nan value in the dataset. We handled them by first identifying the percentage of missing values in a column. This helped in making an informed decision on whether to drop the row or fill it with a given value.

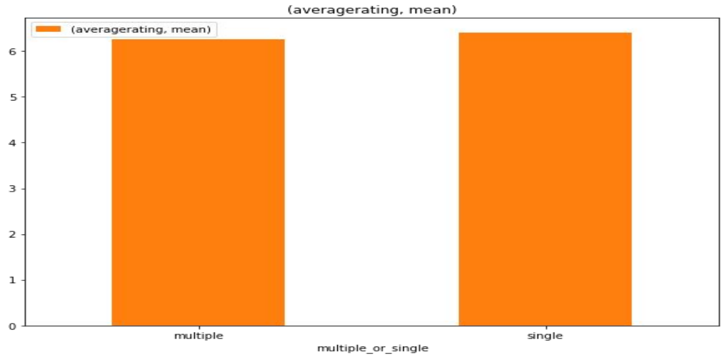
ANALYSIS AND EVALUATION

1) **What is the preferred movie genre**

a) Some genres are produced more than others. Eg Documentary and Drama each contributed to 21% of all movies produced.

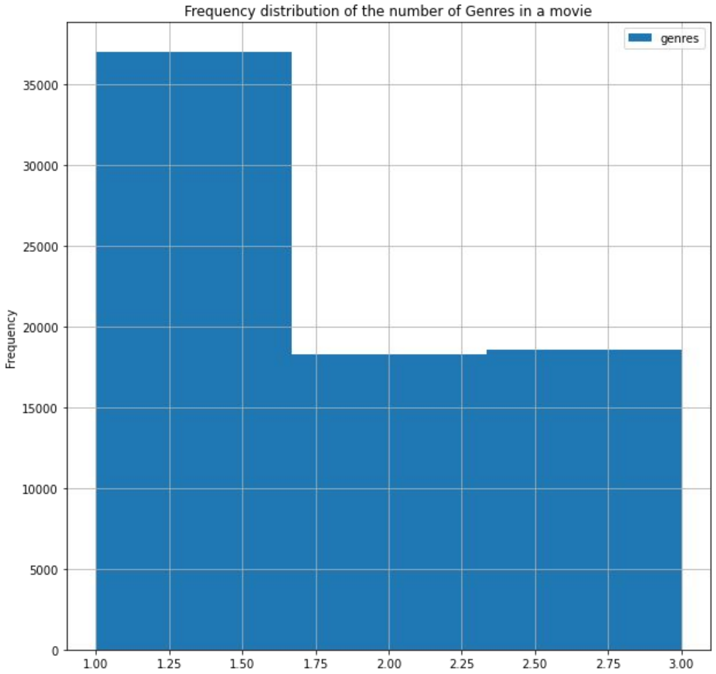
b) Movies with a single genre on average had a slightly higher rating compared to the ones with multiple genres, as seen in the graphs shown.





More Visualizations

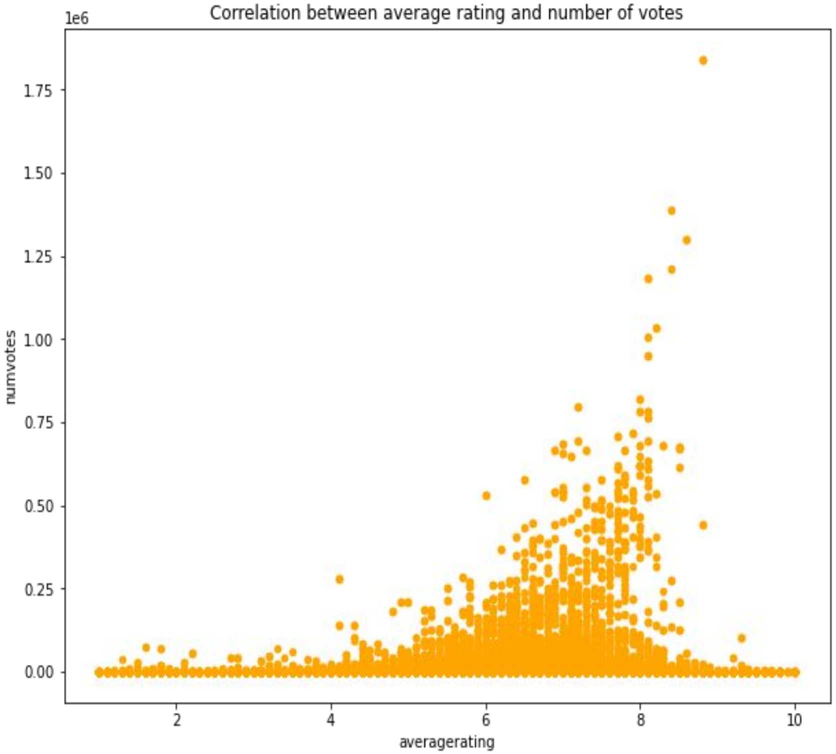
The graph shown highlights the distribution of the number of genres in a movie. The highest frequency being movies that have a single frequency.



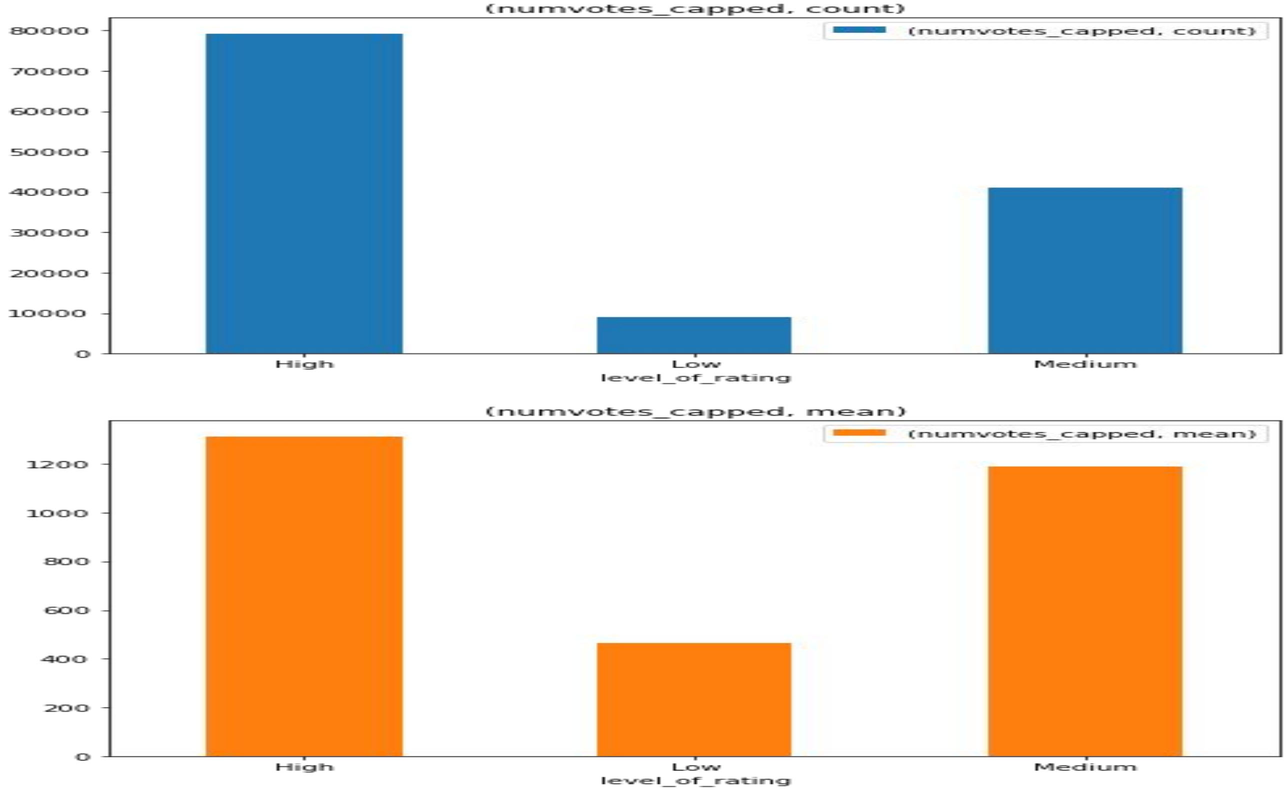
2. **Assessing the Impact of number of votes for a movie on the ratings.**

High ratings from our analysis increase as the number of votes for them increase. consequently, movies with lower ratings have fewer votes.

As seen in the diagram, there’s a positive relationship between the average rating and the number of votes.



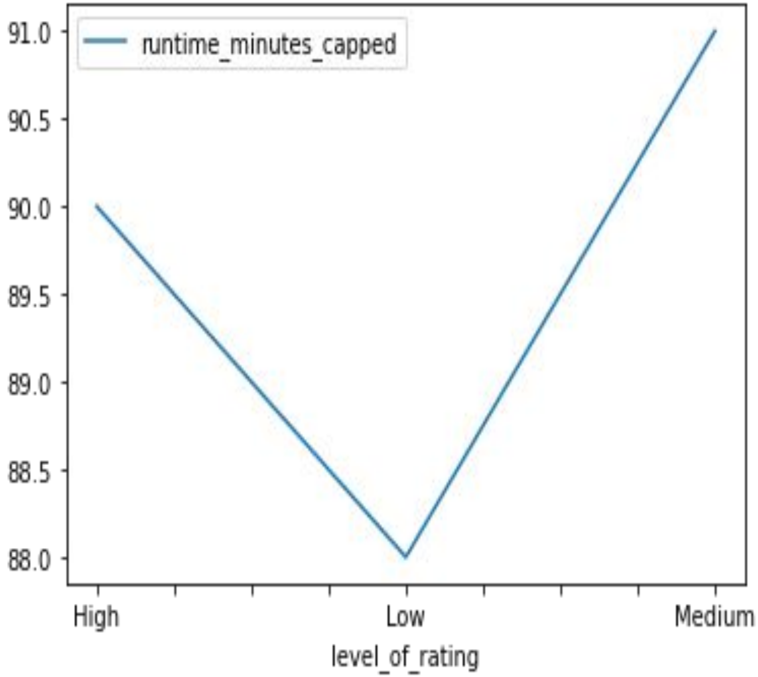
More Visualizations



3. How the length of a movie affects the rating.

The median time for movies that rated highly was 90 minutes as illustrated by the line graph.

Optimal time is therefore a factor to consider when thinking of the success of a movie.



CONCLUSION

This analysis leads to a few recommendations for the head of Microsoft's new movie studio.

1. Produce a genre that the market demands most (as indicated by what producers are producing more of). Documentaries ranking highest, may be prefered by consumers due to the fact that they are based on real life. A single genre will also give the company an edge in the market.

2. Invest in consumer feedback. Movies that had high ratings also had many people accessing them and rating them. Microsoft, being a tech company can leverage that and ensure it has algorithms for fetching feedback from consumers.

3. Have an optimal time for the length of the movie. Many movies that rate highly are 90 minutes long. Therefore this seems to be the optimal length for successful viewership by consumers.

NEXT STEPS

Further analysis would give more insights to the Microsoft studio by:

1. **Modelling the impact of change in title on the ratings.** This data is available in the given dataset.