Introduction:
This report seeks to answer the questions: Is there evidence to suggest that the volume of released movies has increased over the years due to the pursuit of profit, potentially leading to a sacrifice in the overall quality of movies? Using a dataset containing information about movies, including their release year, runtime, ratings, and gross earnings. The dataset was cleaned and processed using Python libraries such as Pandas, Matplotlib, and SciPy.

Data Overview:
The dataset consists of 10,000 entries, with columns representing various attributes of each movie, such as name, year of release, runtime, rating, and gross earnings. Initially, there were missing values in the 'Gross' column, which were handled by removing rows with missing data.

Methodology:

Data Import and Overview:

Pandas' read_csv() function is used to import the movie dataset, and the head() function is used to display the first few rows.

Data Cleaning and Renaming:

The rename() function is used to rename columns, and the dropna() function is used to remove rows with missing values.

Exploratory Data Analysis (EDA):

Scatter Plot:

Matplotlib's scatter() function is used to create scatter plots of rating over time and gross earnings over time.

Bar Plot:

Matplotlib's bar() function is used to create bar plots of the number of movies per year and average gross earnings per year.

Pie Chart Visualization:

Matplotlib's pie() function is used to create pie charts categorizing movies into "Poor Quality" and "Good Quality" based on ratings.

Data Grouping and Visualization:

Grouping:

Pandas' groupby() function is used to group movies by year.

Visualization:

Matplotlib's plot() function is used to visualize the grouped data.

Categorization and Visualization by Time Periods:

Categorization:

Pandas' cut() function is used to categorize movies into "Early Period" and "Recent" based on the median year.

Visualization:

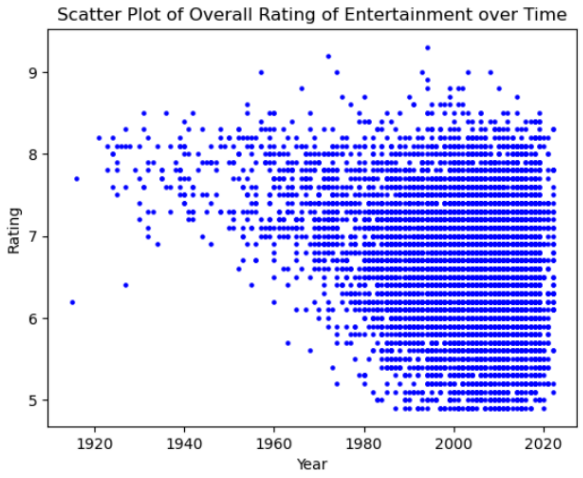
Matplotlib's pie() function is used to create pie charts visualizing the quality of movies in each time period.

Linear Regression Analysis:

Scipy's linregress() function is used to perform linear regression analysis, and Matplotlib's annotate() function is used to annotate the plot with regression equation and statistics.

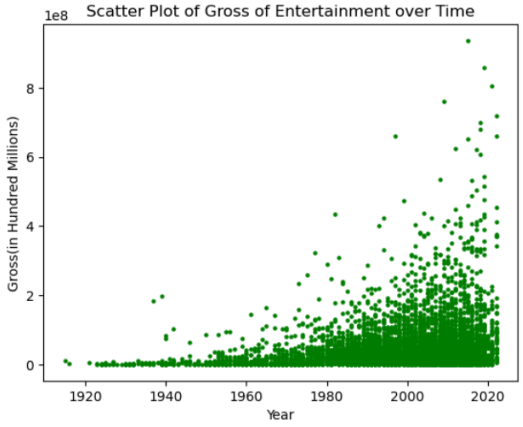
Exploratory Data Analysis:

Scatter Plot of Rating Over Time:



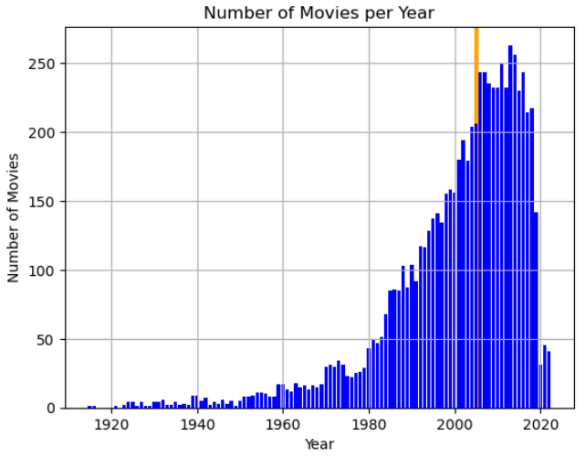
- The scatter plot illustrates the trend of movie ratings over the years. There is a noticeable variation in ratings across different years.
- The ratings appear to be concentrated between 6 and 10, indicating a generally positive reception of movies.
- The volume of plotted points over time is increasing, as well as the variability in rating.

Scatter Plot of Gross Earnings Over Time:



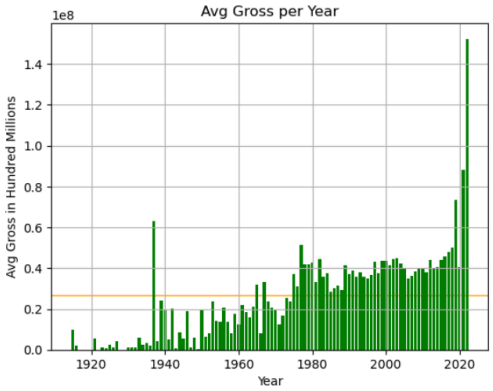
- This scatter plot displays the gross earnings of movies over the years.
- As time increases so does the gross revenue
- Data is skewed to the right exhibiting a positive relationship between year progression and Gross Revenue

Bar Plot of Number of Movies Per Year:



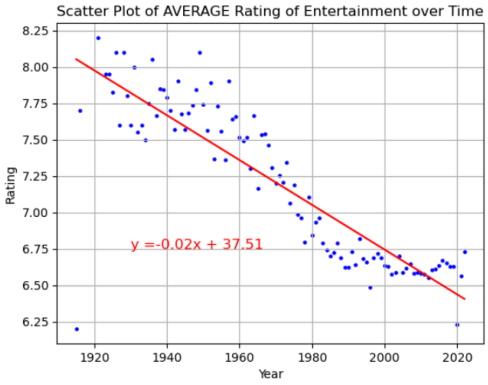
- This bar plot shows the distribution of the number of movies released each year.
- The number of movies released per year has increased over time, with peaks observed in recent years.
- Data is skewed to the right exhibiting a positive relationship between year progression and number of movies made
- This can further be backed by the median indicated at 2005, despite the data set spanning from 1915-2023

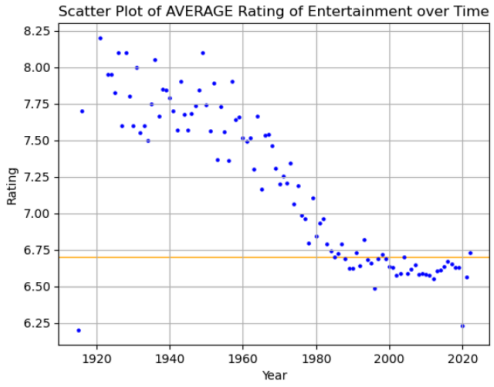
Bar Plot of Average Gross Earnings Per Year:



- The bar plot illustrates the average gross earnings of movies per year.
- The average gross (in Hundred Millions) is 0.27, and as it can be seen in the bar graph that all the years after 1974 are above the mean, whilst most years prior to 1974 are below except for a few outliers
- This represents the increase in revenue generated on average in the recent movies produced

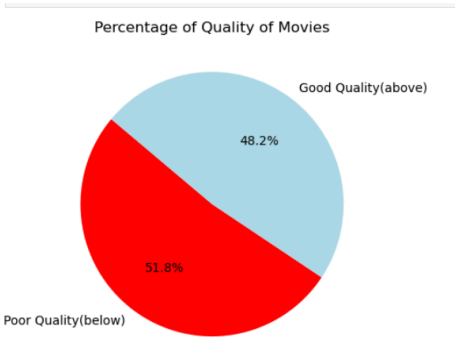
Scatter Plot of Average Rating Over Time with Linear Regression:





- This scatter plot depicts the average rating of movies over time, along with a linear regression line.
- The linear regression analysis reveals a significant negative correlation between the average rating and the year of release.
    - Indicating a decrease in quality of movies over time within our dataset
- An r value of -0.867 suggests that as the year of release increases , the average rating of movies tends to decrease. This indicates a negative trend where movies released in later years tend to have lower ratings compared to those released earlier.
- A p value of 2.72e-32 suggests that there is an extremely low probability of observing the observed relationship if there were truly no relationship between these variables. Providing strong evidence to reject the null hypothesis and supporting that there is a significant negative correlation between the average rating of movies and the year of release.

Pie Chart of Quality of Movies:

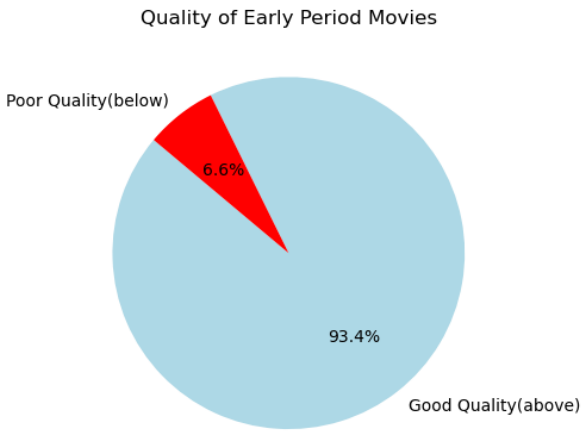


- The pie chart categorizes movies into two groups based on their ratings: "Poor Quality" and "Good Quality."

- The majority of movies fall into the "Good Quality" category, indicating a generally positive reception among viewers.

Pie Charts for Early Period Movies:

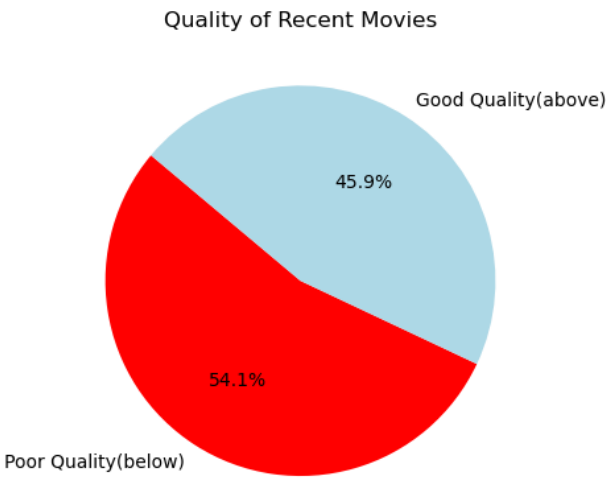
- First the Data was divided among the average year over the time coverage of our data. This middle value was 1969.
- The Pie chart was then made to determine percentage of “Poor Quality” and “Good Quality” movies during the “Early Period”(1915-1969)



- In the Early Period, we see that a majority of movies (93.4%) were of “Good Quality”(above 6.70 on IMDB Rating)

Pie Charts for Recent Period Movies:

- Conversely, a separate Pie Chart had to be made to determine percentage of “Poor Quality” and “Good Quality” movies during the “Recent Period”(1970-2023)



- Comparatively, we see a majority of movies (54.1%) were of "Poor Quality"(Below 6.70 on IMDB Rating)

Conclusion:

- Our overall question was, is there evidence to suggest that the volume of released movies increases over time while the overall quality decreases over time in the pursuit of profit?
- We had achieved a p-value of 2.719835163770518e-32 showing there is statistical significance to know there is a negative correlation between quality(rating) of movie and Year of release.
- Over the years, the movie industry has had a significant rise in the number of movies produced annually. With the rising number of movies, the average ratings of these films have a downward trend, suggesting while more movies are being made, many are failing to meet the quality standards set by previous films. The linear regression analysis indicates a strong negative correlation between the year of release and the average rating of movies, implying as time progresses, the quality decreases, indicating a potential trade-off between quantity and quality in the movie industry. The analysis of gross earnings per year reinforces the suggestion that financial success is prioritized over quality.

Limitations:

The observed trend of the movie industry producing more films per year over time and experiencing greater financial success in recent years warrants further investigation. While the data suggests a correlation between increasing film production and revenue, it's essential to acknowledge the possibility of various underlying factors influencing these trends. Factors such as inflation, changes in audience behavior, advancements in technology, and accessibility to movies could significantly impact these outcomes. Additionally, it's important to note that the dataset under scrutiny likely represents a sample rather than an exhaustive catalog of all films released during the specified period (1915-2023), which could introduce biases in the analysis. Further study is necessary to explore these complexities comprehensively and ascertain the true drivers behind the observed patterns in the movie industry.