**Data Cleaning:**

* No duplicates were found in this dataset.
* Removed the first column with no header which was the index/serial.
* Converted all the numerical columns to number data type and date column to date data type for consistency and data integrity.
* The over view column contains 2 blank cells. Used =COUNTBLANK(). Used Find & Replace to get rid of the blanks with N/A.
* Fixed title and overview columns with TRIM() and PROPER() to have consistent string values.

**Visualizations:**

* Created a histogram for data popularity.
* Created a timeseries line graph to analyze popularity score over time.
* Created a box plot aka whiskers plot for average\_ratings.
* Created a scatter plot which shows the relationship between Vote Counts and Movie Ratings.
* Created a Pivot table to analyze genre\_id and popularity and added a slicer with a clustered pie chart.

**Findings:**

*Popularity Histogram*

* The majority of movies in the dataset have popularity scores ranging from 7 to 20.
* However, there is a notable concentration of approximately 150 movies with a popularity score around 200. This spike in popularity is an interesting observation in the dataset.
* The distribution of movie popularity is right-skewed, indicating that the mean popularity score is greater than the median popularity score.

*Popularity TimeSeries*

* Recent Surge in Popularity: The chart reveals a substantial and prolonged surge in movie popularity in the most recent time frame covered by the dataset.
* This surge suggests that movies released in the dataset's latest period have, on average, achieved significantly higher popularity scores compared to their predecessors.
* Intermittent Spikes: In addition to the overall surge, the chart displays intermittent spikes at various points in time. These spikes likely correspond to the release dates
* of specific movies that experienced temporary surges in popularity.
* Recent Movies Dominate: The dominance of recent movies in the chart's rightmost portion suggests that movies released more recently are more likely to attain higher popularity scores.
* This aligns with the industry trend where new releases often generate immediate interest and higher levels of popularity.

*Vote\_Average Box Plot*

* Central Tendency: The central tendency of movie ratings, as measured by the mean (6.8) and median (6.7), indicates that
* the typical movie in the dataset receives a rating close to 6.7 to 6.8. This suggests that the majority of movies have
* ratings clustered around this range.
* Spread of Ratings: The range of ratings in the dataset varies from a minimum of 5.7 to a maximum of 8.7.
* This demonstrates that there is a diversity of movie ratings, with some movies receiving higher scores (8.7)
* and others lower scores (5.7).
* Variability: The interquartile range (IQR), which is 0.9, represents the middle 50% of ratings in the dataset.
* This indicates that most ratings fall within a relatively narrow range, emphasizing the consistency in movie ratings.

*Ratings & Votes Scatterplot*

* Corelation: The correlation coefficient measures the strength and direction of a linear relationship between two variables. In this case,
* a positive value (0.2592) indicates a positive correlation, meaning that as vote counts increase, movie ratings tend to increase, albeit with a relatively weak positive relationship.
* The magnitude of the correlation coefficient: (0.2592) suggests that the correlation between vote counts and movie ratings is relatively weak.
* A correlation closer to 1 would indicate a strong positive linear relationship, while a correlation closer to 0 suggests a weak or no linear relationship.
* Concentration of Low Vote Counts: The dense cluster of data points near the lower end of the vote\_counts range (around 0 to 10,000) suggests
* that a significant number of movies in your dataset have relatively low vote counts. This is typical, as most movies may not receive a large number of votes.
* Wider Distribution with Higher Vote Counts: As vote\_counts increase beyond 10,000, the data points begin to disperse, indicating greater variability in movie ratings.
* This suggests that movies with higher vote counts exhibit a wider range of ratings, and there is less uniformity in the ratings of popular movies.