**Level 1 (Regular Data Science Questions):**

1. Calculate the average Audience Rating across all movies.

87.55

1. Find the median Length in Time for movies released after the year 2000.

122.5 minutes

1. What is the range (difference between maximum and minimum) of the Rotten Tomato Rating across all movies?

82

**Level 2 (Multiple Step Hard Data Science Questions):**

1. For each 'Genre', calculate the average Rotten Tomato Rating. Which Genre has the highest average rating?

‘Family’ Genre

1. Group movies by 'Part of Series' and calculate the median Audience Rating for each group. Which group has a higher median rating?

Movies that are part of a series

1. Identify the movie with the longest Length in Time. What is the name and Length in Time of this movie?

‘Gone with the Wind’ with 238 minutes

1. Calculate the average Rotten Tomato Rating for movies with an 'Age Rating' of 'R'. How does it compare to the average rating of movies with 'PG-13'?

Movies with an ‘R’ rating have a higher average

1. For each decade, calculate the total number of movies released. Which decade saw the highest number of releases?

The 2010s

1. Divide movies into two groups based on the median 'Audience Rating' (above or below). Compare the average 'Length in Time' between these two groups. Which group has a longer average length?

Movies above the median Audience Rating

**Level 3 (Multistep Data Analysis Aspects of data science/ML):**

1. Using a clustering algorithm (e.g., K-Means), cluster the movies based on Rotten Tomato Rating and Audience Rating. How many clusters provide the most meaningful grouping?

7 clusters

1. Predict the 'Genre' of a movie using a decision tree classifier based on Rotten Tomato Rating, Audience Rating, and Length in Time. What is the accuracy of this model?

11.11%

1. Identify outliers in the dataset with respect to Length in Time and Rotten Tomato Rating using an Isolation Forest algorithm. How many outliers are detected?

29 Outliers in the dataset

1. Implement Principal Component Analysis (PCA) on the numerical columns (Release Year, Rotten Tomato Rating, Audience Rating, Length in Time) and determine how many components explain 90% of the variance.

3 components

1. Using the features Rotten Tomato Rating, Audience Rating, Length in Time, train a Support Vector Machine (SVM) classifier to predict whether a movie is part of a series. What is the precision of this model?

0%

1. Apply a time series forecasting model (e.g., ARIMA) to predict the next 5 values for 'Rotten Tomato Rating' if it were a time series. What are these 5 predicted values?

Next 5 values: [52.28, 58.78, 57.89, 58.01, 57.99]