1. Introduction:

In this analysis, we examine a dataset containing information about movies. The columns in the dataset include showdowns, number of release, first release, gross and total gross. The dataset was scrapped from <https://www.boxofficemojo.com/showdown/?ref_=bo_nb_cso_tab>, and our goal is to gain insights into how different variables like number of release, gross of the first release affect the total gross. The dataset consists of 582 observations with five columns; namely, “showdowns”, “Number of Release”, “Total Gross”, “First Release” and “Gross”. There are no missing values in the dataset. The “Number of Release” is a categorical variable with 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 categories of release.

2. Summary Statistics:

The mean “Total Gross” for all movies is $690 million.

The median “Total Gross” is $465 million.

The minimum “Total Gross” is $6 million while the maximum is $4211 million.

The category with 8 “Number of Release” has the highest mean of $3371

The category with 6 “Number of Release” has the highest Total Gross of $127271 million

3. Data Distributions:

The distribution of box office revenue is right-skewed, with a few high-grossing movies driving the average higher.

4. Relationships and Patterns:

There is a positive correlation between the Gross of the first release and the Total Gross, indicating that higher acceptance and revenue for the first release influenced the acceptance and revenue of the subsequent release on the series.

5. Outliers and Anomalies:

Several movies with exceptionally high box office revenue values are identified as outliers.

The outliers are within the categories of 2, 3, 4, 5, and 6 “Number of Release”

There is no time dimension in this dataset as it does not include release dates.

7. Key Findings and Insights:

The presence of outliers suggests that a few movies have an outsized impact on the average box office revenue.

7. Limitations and Assumptions:

This analysis assumes that the dataset is representative of the broader movie industry and that there are no significant data quality issues. The dataset didn’t capture genre and this placed a limitation as we were unable to explore the different sub-classes. The absence of a time dimension limits our ability to explore trends over time. Furthermore, it is important to note the absence of budget variable and the impact it would have added to the analysis.

8. Next Steps:

Further analysis could involve examining the impact of specific actors, directors, or studios on box office success. A deeper dive into marketing strategies and release timing could provide more insights into revenue generation.

9. Conclusion:

In conclusion, our exploratory data analysis of the movie dataset suggests a positive relationship between first Gross and the Total Gross. “Number of Release” did not impact the overall revenue. It's important to note that this analysis represents an initial exploration, and further analysis is warranted to uncover more nuanced insights into the movie industry.