**Final Project: Movie Performance**

Alexa L. Wittlieff

College of Science and Technology, Bellevue University

DSC530-T302: Data Exploration and Analysis (2223-1)

Dr. Neerja Sahu

March 5, 2022

**Statistical/Hypothetical Question**

Does movie run time impact movie performance?

Great films bring people together for a shared emotional experience. My passion for the movies and interest in the forty-billion-dollar film industry led to the selection of my statistical question. The IMDB Movies Dataset on Kaggle featuring key performance metrics for top performing films was used in this data analysis.

**Outcome of your EDA**

Through the initial exploratory data analysis, a correlation of statistical significance was discovered between movie run time and IMDB rating. The p-value was zero. This means that the length of the movie could potentially impact the movie’s performance, keeping in mind that correlation does not necessarily equate to causation.

**What do you feel was missed during the analysis?**

A new variable that incorporates multiple measurements of movie performance to represent overall film performance would have been useful in this data analysis. Then the overall movie performance variable could have been used in the regression analysis as the dependent variable with the other key movie characteristics as independent variables. This would lessen the need to compare multiple variables to gain insight to movie performance, as I utilized IMDB rating, Meta score, and Gross as separate variables.

**Were there any variables you felt could have helped in the analysis?**

Expanding the hypothesis to identify other variables that impact movie performance may yield interesting results. For this analysis, I chose to focus on run time. However, data on the movie budget and movie theater attendance numbers for opening weekend would have been interesting to investigate, as well.

**Were there any assumptions made you felt were incorrect?**

I did not remove outliers from the data that was used for analysis. Being able to see how the data changed when outliers were removed would have been insightful. Through the visualization, the outliers became more readily apparent.

**What challenges did you face, what did you not fully understand?**

When I worked through creating the histograms, I discovered that my data was not necessarily the types that I anticipated. Therefore, I needed to go back to transform my data from objects to integers. Once the transformation was complete, my histogram results looked much more presentable. This allowed me to better understand the distribution of the data.