# Case Study Rubric – Replication of Project

**DS 4002**
**Individual Assignment**

**General Description:** This rubric outlines what the student will be doing, the goals, and the exact steps to successfully complete this assignment and replicate a case study from a previous DS 4002 student.

**Why am I doing this?** This is an opportunity for the student to gain experience replicating a project from a topic they may or may not be familiar with, and for the student to use their skills learned in data science courses to synthesize lessons learned from classes. Replication of the case study helps the student hone their data science skills and ensures that the project can be replicated.

**What am I going to do?** Access the GitHub repository from a previous DS 4002 student to replicate a project completed during the course. Follow the instructions from this rubric and use the materials from the folder and GitHub repository to replicate the project and learn more about the topic.

**Tips for Success:**

- Read instructions carefully
- Follow the steps in the order they are given and as closely as possible
- Talk to instructors and fellow students if any issues or questions arise

**How will I know I succeeded?** You will meet expectations for the Case Study Replication Rubric if you follow the steps below:

| Spec Category | Spec Details |
|---|---|
| Data Collection | ● Goal: download the dataset necessary for replication of the project<br>● Navigate to the GitHub repository: https://github.com/jillianhaig/CS3<br>● From the "DATA" folder, the student may download the dataset of interest, named "Disney_Movie_Dataset.csv"<br>● The data includes the variable, "recent?", which signifies if the movie review is recent (1- under a year from release) versus non-recent (0- a year or more from release) |
| Statistical Analysis | ● Goal: gather sentiment scores and use hypothesis testing, confidence intervals, and linear regression to confirm or deny the presence of recency bias in the movies<br>● In the "CODE" folder, the "Case_Study_Replication_Code.ipynb" file contains steps and preliminary code to successfully complete the analysis portion of the case study<br>● Using any coding environment (Google Colab recommended), obtain the average compounded sentiment scores for all the movie reviews and compare the recent movie review sentiments to non-recent movie |

| | |
|---|---|
| | review sentiments<br>○ Use the VADER package to gather compounded sentiment scores from the text data from the review variable<br>● Use hypothesis testing (T test) to obtain a p-value; reject or fail to reject the null hypothesis that recent movie reviews have higher average sentiment scores<br>● Build a confidence interval to determine if there is any overlap between the average sentiment scores for recent and non-recent reviews<br>● Use linear regression to see the impact recency has on sentiment, if any<br>○ For the regression, the Y variable is the compound sentiment score variable and the singular X variable is the "recency?" variable |
| Data Visualization | ● Goal: produce visualizations that explain the results from statistical analysis<br>● Produce both a a boxplot and a violin plot to highlight the relationship between review timing (recency) and the compounded sentiment score<br>● Produce a bar graph to show the relationship between review timing (recency) and average compounded sentiment scores<br>● After building the confidence interval, show it in the form of a bar graph<br>● Produce the output from the OLS regression |
| Results and Conclusion | ● Goal: from the results and data provided, come to a conclusion about recency bias in animated Disney movies<br>● The student should find that recency bias does, in fact, appear to be present in this scenario<br>○ Find that the compound sentiment scores for recent reviews, on average, are higher than non-recent review sentiments<br>○ Reject the null hypothesis in favor of alternative hypothesis (in other words, recency bias does appear to be present)<br>■ Reject the null if the p-value is less than 0.05<br>● Based on statistical analysis, especially the hypothesis testing and confidence interval, the student should find that the sentiment scores for movie reviews posted recently after the release of a movie tend to be higher than those posted a year or more after the movie's release date, indicating that recency bias is present in this scenario |