## Name: Jonathan Lawrence

## Date: 8/7/2019

## Class: DSC-530

## Assignment: Final Project

**Statistical/Hypothetical Question**

A book's rating on Goodreads can be predicted by looking at specific information about previously rated books.

**Outcome of your EDA**

I found slight correlations between some of the variables, including the page length, number of words in the title, and the number of ratings received. However, they were not significant enough to suggest causation.

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

The dataset supplied a list of genres for each book, with many books containing multiple genres. I would have liked to have performed a more extensive analysis with the genres broken up so that I could determine if genre plays a significant role in determining a book's rating. However, this would have required creating a numerous amount of datasets given that there are many books with a lot of genres. In a library, you cannot place the same book in two different places, so I would have liked to have the placement a library would have used. This would have made EDA for genre much simpler.

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

Aside from a simplified genre list, I would have liked to have a variable that tracked whether or not a book is tied to some other form of media that could have popularized it (i.e. a movie or video game).

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

I assumed that the number of ratings a book received would have meant it was more popular, and therefore drive up higher ratings. Instead, I found that it was not as clearly the case.

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

My main challenge was interpreting the custom functions provided by ThinkStats so that I could use them in conjunction with the more widely known Matplotlib and Pandas functions. There were also a few incidences where some items were deprecated with Python 3, so I needed to find a suitable replacement. A somewhat less difficult challenge was juggling all of the statistical variables in my head and trying to figure out which one would lead me to the answer I was looking for. I overcame this by looking back at the textbook to review old concepts.