STAT 6012: Linear Models for Data Science
Class Activity 4

Due date: Wednesday, July 19 by 10:50 am Via Canvas.

Complete the following questions an R Markdown file and submit your compiled HTML file. If you are working in a group, list the names (last, first) of the group members in alphabetical order of last names.

The attached dataset provides comprehensive information on various aspects of books, including their publishing year, author details, ratings given by readers, sales performance data, and genre classification. Read more about the variable descriptions here:

`https://www.kaggle.com/datasets/thedevastator/books-sales-and-ratings`

As a data scientist who is looking for a fun project, you are interested in investigating:

1. [8] whether the average `gross.sales` differs across the four categories of `Author_Rating`. Conduct a full analysis to investigate your research question, including exploring the two variables through visualization, and hypothe**ses** testing. From your investigation, what are the top two `Author_Rating` categories that yields higher `gross.sales`, on average?

2. [2] whether there is a linear relationship between the response variable, `gross.sales`, and other variables in the dataset. How would you investigate the linear relationship? [Note: this is an open ended question.]