STATS 306 Final Project

This is a general guideline for your final project. The total number of points is 100 points. You are to create and answer 4 different scientific questions on your data set. Please be sure to come up with different types of questions related to your data sets that will use a wide range of tools that are covered in the class. Each scientific question is worth 20 points, yielding a total of 80 points. Other than the scientific questions, 20 points will be allocated generally to your report on the formatting, packages, data preparation, and tidy data aspects. This gives a total of 100 points. Detailed guidelines on points assignments are shown in the following table.

In particular, you should articulate your scientific questions clearly. Each question can have multiple subproblems where you explore, transform, visualize your data, and finally come up with a conclusion for your scientific question. The report should be written like an essay, with markdown text interspersed with code and plots. For instance, you should explain what you are trying to do before writing a line of code, and then explain how your code will achieve what you are trying to do using words.
Articulate clearly what you have found for each of the questions.

Clearly indicate in the very beginning of your final project the contributions for each of your team members. If all team members contribute roughly equally, then you can include a phrase ``all team members contribute equally in this project". By default, every member in your team will receive the same score since this is a joint project. However, if a certain team member contributes significantly less compared to the others, then the corresponding team member will likely receive a lower grade. If your team falls into this situation, do not hesitate to contact me via email.

Read through the following table very carefully and post questions in Piazza if you need any clarifications.

Section	**Details**	**Points Possible**
Formatting	**-** Report will be written in Jupyter notebook, but submitted as a standalone **.html** - The report should be written like an essay, with markdown text interspersed with code and plots. Comment text is not acceptable. - All output must be shown; the code in your notebook will not be run by the graders. - Report **will be given a zero** if submitted in a different format **without prior consent from the professor or your GSI** - Report is organized into four parts so that there	**5**

	is: 1. An introduction at the beginning which motivates the question(s) of the dataset (NO CODE) 2. A section which discusses: a. Package loading (with the code to load libraries) b. Data import; description of variables c. Data tidying (and contains that code) 3. One or more sections which contain a transform and/or a visual which answers an aspect of the question. You should clearly state and answer (at least) four questions about the data in this section. 4. A conclusion at the end which gives an answer to the question(s), supported by the your data analyses (NO CODE)	
Packages	- The tidyverse package is properly loaded. - All packages other than tidyverse are explained in words - Load all packages in the beginning of your report, and add a comment explaining the use of the package	**5**
Data Preparation/Import	- The provided dataset is loaded correctly using read_csv - The authors give a description of the dataset, its context, and what each of the variables they use means. - If you use other datasets, it is imported correctly, the source is cited, and you explain why you're using it	**5**
Data Tidying	- The dataset is cleaned into a **tidy format,** as defined in R4DS 12 Tidy data a. Each variable must have its own column. b. Each observation must have its own row. c. Each value must have its own cell - The tidying should be done in a way so that they clean the data to a tidy form, then use that form multiple times in their EDA. - If the data are already tidy, then explain clearly *why* the data are tidy	**5**

Individual Questions:

This is the rubric for each of the four questions you will propose and answer

Section	**Details**	**Points Possible**
EDA: Transform Data	- The data is properly filtered, subsetted, joined, and reshaped to answer each specific question. - We will judge on creativity; you should try to come up with new variables using mutate(), create interesting groups to then compare in the data, etc. - You should use a variety of techniques from the different units we have covered in class	**5**
EDA: Data Analysis	- Interesting visuals are created that show an emergent pattern in the data. - Tables showing important summary statistics for variables of interest should be constructed as part of your answer to each question - Plots are professional-looking and easy-to-understand a. Descriptive title b. Good x- and y-labels c. Clear labels in legend d. Created using ggplot - You should use at least two plots for each question you propose a. Across all questions, you should use at least three different *types* of geoms/plots (e.g., histogram, scatter plot, etc.) - You should include a linear regression model as part of your answer for at least one of the four questions	**10**
Communicate	- The questions the students are trying to answer are clearly articulated and motivated. - Each step of the analysis, whether it's transforming the data or visualizing the data, is accompanied by a plain English description a. We don't want line-by-line commenting on the code. Rather, we want a high-level explanation of why a particular data transformation or visualization is needed to answer the question of interest.	**5**