DATASCI 530 – Assignment 2

For this assignment you will using the Formula 1 dataset introduced in Lecture 3 in the Github notes. Formula 1 is an international car racing competition, where drivers from all over the world compete for a title, and where individual races take place across different countries. Each driver is part of a team called a "Constructor", typically sponsored by a specific car manufacturer. You can find more information about the sport and the dataset here:

https://en.wikipedia.org/wiki/Formula_One

https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020

Use the datasets to answer the following questions:

a) Which three constructors had the highest number of total points between 1981 and 2020? How many total points did each of them get? How do the total number of points for each constructor compare to the average across constructors?
b) Which three constructors had the highest number of total points between 2001 and 2020? How many total points did each of them get? How do the total number of points for each constructor compare to the average across constructors?
c) How did the rankings change across the two time periods?
d) How many different drivers did Ferrari have between 1981 and 2020?
e) What was the best year for Ferrari between 1981 and 2020?

For this assignment you will be expected for merge multiple datasets (for instance names and points may be on different datasets), check for missing values, practice chaining operations, and create appropriate visualizations.

- Start by checking the codebook to figure out which information table contains the relevant information.
- You can come to office hours if you have any questions.

Deliverable

Submit an HTML file with you results on Canvas. Include markdown chunks to write comments on your findings and explain your methodology.

When submitting your work imagine that you are sharing this with colleagues at work, who have asked you to explain and justify key decisions and findings of your analysis. The file should be professionally presented, self-explanatory, and have clearly labeled sections.

Rubric:

You will be graded both on the code and the writing using the following point system:

Code - Code runs without issues - Code is appropriate to answer question - File is appropriately organized, with different selections and comments clarifying the definition of variables and dataset names.	4 points
Description of methodology - Explain key coding decisions via markdown chunks - Clarify any choices about how to clean data - Shows evidence of conducting quality checks (count sample size of merged datasets, check for the presence/absence of missing values, and/or compute any relevant descriptive statistics)	2 points
Results and Visualization - Write a short summary of the findings. - Correct interpretation of results and easy to follow - Includes graphs to display results from different groups - All graphs look professional (have axis and title labels, suitable color scheme, font sizes, legend, etc.)	4 points