QTM 151 Introduction to Statistical Computing II, Spring 2023
Final Project Instructions

I. Overview:

This project tests your ability to combine the programming concepts covered in QTM 151 and produce your own analysis for a real-world dataset. You will present a report in a **Jupyter notebook**, in groups of **3-4 students**. Submit as an HTML.

II. Dataset

We will use a publicly available dataset on Formula 1, an international car racing competition. [1] You can learn more about how this competition works in the following video:

https://www.youtube.com/watch?v=fS8Ezkxwn5g

- The dataset contains 14 tables.
    - Choose **two or more** of these datasets (you **do not** have to use all 14).
    - **"f1_codebook.pdf"** : Column attributes, types, and description
    - **"f1_entity_relationship_diagram.pdf"** : Relationships between tables

III. Jupyter Notebook:

- Title and names of project members (with section numbers)
- Introduction
    - A markdown text with 1-2 paragraph that summarize the main goals of the project. The first paragraph should briefly describe what Formula 1 is, what question you're interested in, and why it is relevant. The introduction should end with a high-level description of the results and the coming structure of the project. Try to make the text self-contained, intended for someone who isn't familiar with Formula 1 or the dataset.
- Data Description:
    - Write a markdown chunk of 1 paragraph describing which dataset tables (among the 14) you will be using. State what each row represents, how many

[1] The source of the original dataset is https://www.kaggle.com/datasets/thedevastator/formula-one-racing-a-comprehensive-data-analysis. More info on formula 1: https://en.wikipedia.org/wiki/Formula_One

observations are contained in each table, the years, and a brief overview of the of the data that is contained there:
        - Import any necessary libraries.
        - Import the data.
        - Do any calculations for counting the number of rows, etc.
    - Write a paragraph in markdown describing any merging procedures:
        - Include code for merging.
    - Write a paragraph in markdown summarizing data cleaning procedures:
        - Include code for data cleaning.
    - Write a paragraph describing your main columns:
        - Compute a table of descriptive statistics for the main columns of the merged dataset that you're interested. Try to be selective. The idea is to do a deeper analysis of a few columns rather than to do a lot.
- Results:
    - This should contain a combination of code to produce tables/plots and markdown text explaining what the findings are.
    - Be creative! The idea is to understand the relationship between different sets of columns to answer an interesting question about the data.
- Discussion:
    - Provide a brief 1 paragraph markdown chunk summarizing your findings. Describe the main things you learned from the data.

Here are some potentially exciting topics:

- Which countries produce the best drivers?
- What characteristics (including the driver, constructors team, qualifiers, and race features) are related to the success of the drivers?
- How do the results vary over time?
- How the results vary by the nationality of the drivers, or the geography of the circuits?

IV. Project Guidelines:

You can decide what question (or set of questions) to answer, but the project must include the following programming concepts:

1. Merging tables using Pandas
2. Applying multiple elements of data manipulation (recoding, renaming, transforming columns with apply, grouping, aggregating, and/or sorting)

3. Produce summary tables and plots.
4. Optional: loops and functions

- **Originality:** You can use part of the code used in lectures, quizzes, and assignments, but to get full points you should expand on what was done before
- **Running:** All the code should run properly. You will get points discounted if there are any errors.
- **Aesthetics:** The work is organized and includes all the required elements. The overall appearance is neat and professional. Use headings and other markdown formatting elements to improve the appearance of your project. See more details here:

https://notebook.community/tschinz/iPython_Workspace/00_Admin/CheatSheet/Markdown%20CheatSheet

V. Grading Rubric

Component	Detailed Points	Total Points
Overall		2
Organization and aesthetics	1	
Originality	1	
Introduction		2
Description of topic and question	1	
Summarize findings	1	
Data Description		7
Introduce your dataset	1	
Merging data	2	
Manipulating/Cleaning Data	3	
Column descriptions	1	
Results		8
Clear interpretation	2	
Formatting Tables	3	
Formatting Plots	3	
Discussion		
Clarity and conciseness	1	1
Total		20