QTM (DataSci) 151 Introduction to Statistical Computing II
Fall 2025 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**.

II. Dataset

You will choose **one** of the following publicly available data collections (the .csv files are available on the repo)

1. Formula One data: https://www.kaggle.com/datasets/thedevastator/formula-one-racing-a-comprehensive-data-analysis
    a. Contains data on Formula One races that have taken place since 1950.
    b. Contains 14 .csv files
2. FIFA players data: https://www.kaggle.com/datasets/joebeachcapital/fifa-players
    a. Contains player data for the FIFA 15-22 video games.
    b. While these are video games, we can use the game developers' statistics are based on real-life performance, and can be used as a surrogate for the actual players.
    c. Contains 15 .csv files
3. NFL player and game statistics: https://www.kaggle.com/datasets/toddsteussie/nfl-play-statistics-dataset-2004-to-present
    a. Contains game data from 2004-2019 along with player, coach, and referee data.
    b. Contains 20 .csv files
    c. Two of the .csv files exceed GitHub's 100 megabyte limit. You can **download** them from the course Canvas page.
4. U.S. Election 2020 results: https://www.kaggle.com/datasets/unanimad/us-election-2020
    a. Contains state/county vote totals and results for 2020 presidential, Senate, and house of representatives races
    b. Contains 11 .csv files
5. If you find another data-set online consisting of **multiple** .csv files that share common data among them, you can choose to use that instead. Just check with me

first (mostly to save you the trouble of choosing something for which interesting data analysis will be difficult).

Once you make your choice, you must use two or more .csv files from the dataset.

III. Jupyter Notebook:

- Title and names of project members
- Introduction
    - A markdown text with 1-2 paragraphs that summarize the main goals of the project. Imagine you are giving a report to someone who is not your professor. The first paragraph should briefly describe what the topic is, what data analysis 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/FIFA/NFL/U.S. politics/other topic or the dataset being used.
- Data Description:
    - Write a markdown chunk of 1 paragraph describing which dataset tables (among the multiple options) you will be using. State what each row represents, how many observations are contained in each table, 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 (you can answer more than one!):

- (F1) Which countries produce the best drivers?
- (F1, FIFA, NFL) How do driver/player characteristics change over time?
- (F1, FIFA, NFL) What characteristics are related to the success of the drivers/players?
- (U.S. election) How strong is the relationship between a presidential candidate winning a given county vs. a Congressperson of the same party winning that county?
- (U.S. election) Are the results for Democratic Party or Republican Party politicians influenced by the number of third-party candidates?

IV. Project Guidelines:

You can decide what question (or set of questions) to answer, but the project should 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 of good quality.
4. 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. Here's a helpful Markdown cheat sheet:

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