QTM 151 - Introduction to Statistical Computing II

Final Project Instructions

28 October, 2024

1 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. Please submit the report as an HTML file.

2 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.
  - Both `pdf` files are the in `documentation` folder, and the data are in the `data-raw` folder.

[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.

3 Jupyter Notebook

Your Jupyter notebook should include the following sections:

- **Title and names of project members** (with section numbers)
- **Introduction**:
  - A markdown text with 1-2 paragraphs that summarise the main goals of the project. The first paragraph should briefly describe what Formula 1 is, what question you are 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 is not familiar with Formula 1 or the dataset.
- **Data description**:
  - Write a markdown chunk of 1 paragraph describing which dataset tables (amongst the 14) you will be using. State what each row represents, how many 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 and load the data.
    - Describe the data, including the number of observations, the number of variables, and the types of variables.
  - Write a paragraph in markdown describing any merging procedures:
    - Include the code used to merge the data.
  - Write a paragraph in markdown describing any cleaning procedures:
    - Include the code used to clean the data.
  - Write a paragraph describing your main columns:
    - Compute a table of descriptive statistics for the main columns of the merged dataset in which you are interested. Try to be selective. The idea is to do a *deeper analysis of a few columns* rather than a *shallow analysis of many columns.*

- **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**:
  - Write a markdown text with 1-2 paragraphs that summarise the main findings of the project. The first paragraph should describe the main results and the second paragraph should discuss the implications of the results.

Here are some potentially interesting 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?

4 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 organised 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://danilofreire.github.io/qtm151/tutorials/02-jupyter-markdown-tutorial.html#introduction-to-markdown

5 Grading Rubric

Component	Detailed Points	Total Points
Overall		2
Organisation and aesthetics	1	
Originality	1	
Introduction		2
Description of topic and question	1	
Summarise 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		1
Clarity and conciseness	1	
Total		20