DSC 385 - Project 1 Exploratory Data Analysis

Overview

The goal of this project is to apply the skills you have developed in this course so far to conduct an exploratory data analysis. This first project focuses on data wrangling, exploration, and visualization. You will need to combine and transform data, tidy the resulting datasets, and create summary statistics and visualizations.

The final product of this project will be a PDF report (a knitted R Quarto file) containing your code and your results.

Topic and Data

For this project you will be working with data from the World Cube Association, which organizes speedcubing competitions around the world (the Netflix documentary The Speedcubers provides a nice introduction to the topic). The WCA collects all data on speedcubing times for all registered members in every official competition.

In general, the goal of speedcubing is to solve the puzzle in the fastest amount of time. There are a number of different puzzles but the most popular one is the well-known 3x3x3 puzzle originally developed by Erno Rubik. At competitions, there are typically multiple rounds, and during each round, a competitor gets to solve 5 different scrambles. The times for each scrambles are recorded in columns `value1`, `value2`,…,`value5` of the `results` table. For each scramble, the competitor is allowed up to 15 seconds to inspect the scrambled cube before beginning the solve. Special timers are used to detect the beginning and end of the solve (the 15 seconds of inspection are not included in the solve time). You can watch a video of speedcuber Dana Yi solving a 3x3x3 puzzle in under 5 seconds. At the end of a round, the best and worst times for a cuber are deleted and then the middle 3 are averaged to get the final average score recorded in the `average` column of the `results` table.

Sometimes a cuber finishes but does not actually solve the puzzle. This is known as "DNF" and is denoted in the dataset as a `-1` time. For example, if a cuber's second solve in a round was a DNF, then the value in the `value2` column would be `-1`.

Getting the Data

The data are organized into 13 separate tables that are stored as compressed CSV files. The data can be obtained from the following GitHub repository:

https://github.com/Principles-of-Data-Science/Project1

You can pull the GitHub repository into RStudio as follows:

1. Click on File > New Project…
2. Click on "Version Control"
3. Click on "Git"
4. Paste in the URL **https://github.com/Principles-of-Data-Science/Project1** into the "Repository URL" field
5. Type in the "Project directory name" if needed
6. Set the directory if you don't want to use the default
7. Click "Create Project"

The data files all have the file extension ".tsv.bz2". These are bzip2 compressed files that can be read into R using the `read_csv()` function from the Tidyverse. **You do not need to decompress the files before reading them in**.

In addition to the data, the GitHub repository also provides a Quarto template that you can use to complete the report for this project. The template has some code for reading in the data into R. More information about the dataset can be found at the WCA web site.

For this project you will focus on *one* speedcubing event, which is the 3x3x3 event. There are other events that people can compete in but they have been filtered out of the dataset that has been provided to you. The key tables in the dataset are:

- Persons: This table has one row per person and has some basic information about them, including their WCA ID, name, gender, and country.
- RanksSingle: This table shows everyone's ranking in the world, continent, and country, based on their best time for a single solve (not an average).
- Competitions: This table lists every competition with one row per competition. There is information about the name of the competition, location, and the date when it occurred.

- Results_333: This table contains all of the times for the 3x3x3 rounds in each competition. (NOTE: it is a *large* table). The format for each round is "average of 5", so five different scores will be recorded in columns `value1`, `value2`, `value3`, `value4`, and `value5`. The average time is recorded by dropping the highest and lowest times and averaging the middle three. Times are recorded in 1/100th of a second. So a time of 498 is interpreted as 4.98 seconds.

The Report

The text of your report will provide a narrative structure around your code and outputs with R Quarto. Answers without supporting code will not receive credit and outputs without comments will not receive credit either. Write full sentences to describe your findings. All code contained in your final project document must work correctly (render early, render often)!

Required Questions

Active Speed Cubers

How many active (3x3x3) speedcubers are there registered with the WCA? For this question an *active speedcuber* is defined as any person registered in the WCA who has competed in at least two competitions in the years 2022–2024.

World Records

This question has two parts:

1. According to the data, who holds the world record single best solve? On what date was this record set?
2. According to the data, who previously held the world record best single solve? On what date was this previous record set?

NOTE: For these questions, consider all speedcubers (not just active ones) and define "best" as the fastest time for a single solve (not for an average).

Regional Rankings

This question has two parts:

1. Amongst all speedcubers, who is the top ranked male speedcuber (for best single solve) in Australia?
2. Amongst all speedcubers, who is the top ranked female speedcuber (for best single solve time) in Europe?

Time Until Sub-5

Having a time below 5 seconds is considered an elite achievement and most speedcubers have to complete a large number of solves before they can obtain a sub-5 second solve.

1. For the current top 10 speedcubers in the world (as recorded in the `RanksSingle` table), on average, how many solves did they have to do before achieving a sub-5 second solve?
2. For one of the top 10 speedcubers make a scatterplot of their individual single solve times vs. the date of the solve, with date on the x-axis and solve time on the y-axis.

NOTE: Each round of a competition has 5 solves that should be considered separately when counting the number of solves.

Up-and-Coming Speed Cubers

Which speed cubers **not** in the top 10,000 (worldwide for single best time) should we keep an eye on for the near future?

The idea here is to identify "up-and-coming" speedcubers who are not yet achieving elite times. Come up with a list of **five** speedcubers (provide their names and WCA IDs) that you have identified as "up-and-coming". There is no one way to answer this question and the goal is to provide an analysis of the data that justifies the selection of your five names.

Format of the Final Report

Answer Required Questions

You should have a section header for each of the required questions, followed by your analysis answering the question.

Discussion

In the Discussion section, you should address all of the following:

- Reflect on the process of conducting this project. What was challenging, what have you learned from the process itself?
- Did you receive help from someone else? Include acknowledgements for any help received.

Template

The template Quarto file provided in the Git repository can serve as a starting point for your report.

Formatting

- Create the report using R Quarto knitted to a PDF file, with headers for each section and each question answered;
- Include comments describing your R code;
- Include any references (datasets, context), if needed.
- The final report should be no more than 20 pages including all code/graphics/output (the number of pages can vary greatly depending on the cleaning process).
- It is extremely important that you **select pages** when submitting on Gradescope (see more below). Points will be taken off if you do not select the appropriate pages for each question in the Gradescope outline.

Submission of the Report to Gradescope

This project report will be submitted on Gradescope for grading. Gradescope is a tool that enables the teaching staff to efficiently grade assignments like this one according to a defined rubric. **You will not be submitting this project on Canvas**. Anything submitted to Canvas will be ignored.

If you have never submitted anything to the Gradescope web site, please watch this video demonstration of how to do so.

To submit your project report, please follow these steps:

1. First render your project report into a PDF file. This can be done by either rendering directly to PDF in RStudio or by rendering to an HTML file and then "Printing" to a PDF file. Either way, **you must have a PDF file to submit to Gradescope**.

2. Go to the course Canvas page and click on the "Gradescope" link on the navigation bar on the left hand side.
3. When the Gradescope page loads, click on the assignment titled "Project 1: Exploratory Data Analysis".
4. You should be prompted with a window allowing you to submit a PDF file of your assignment.
5. After uploading your PDF file, you will be prompted to select pages of your PDF file that correspond to questions in the "Question Outline". Please make sure to do this carefully, as it essential for allowing us to grade your project efficiently.
6. After selecting the pages, submit the assignment.

Late Policy

As per the Syllabus, projects will not be accepted late. There are no exceptions; please do not contact the instructor or TA to request an exception.