SDS 322E - Project 1

Overview

The goal of this project is to apply the skills you have developed in this course so far to a topic of your interest. The 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 Markdown file) containing your code and your results, similar to what you have done previously with Labs and Homeworks.

Working in a Group

For the project, you must work in a group of **4 students** (including yourself). Although you will be working in groups, you will produce one unique report per group. The groups are assigned on Canvas and it is a different group than Lab groups.

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 (indicated in the `roundTypeId` column), 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:

```
usethis::create_from_github("SDS322E-2025Fall/Project1")
```

The data files all have the file extension ".tsv.bz2". These are bzip2 compressed tab-separated-value files that can be read into R using the `read_tsv()` 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 an R Markdown template that you can use to complete the report for this project. The template has some code for reading in the data into R. General 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.

Report

The text of your final report will provide a narrative structure around your code and outputs with R Markdown. Answers without supporting code will not receive credit and code 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 (knit early, knit often)!

Remember that each group only submits one report.

Questions

NOTE: This dataset was downloaded in 2024 and therefore may not reflect the most recent information posted on the WCA web site. Please use **this dataset** to answer all of the questions below and do not use the web site to answer the questions.

Active Speed Cubers

How many active (3x3x3) speedcubers are there registered with the WCA? For this question an *active speeedcuber* 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. Who holds the current world record single? On what date was this record set?
2. Who *previously* held the world record single? 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 plot of their solve times vs. the date of the solve, with date on the x-axis and solve time on the y-axis.

NOTES

- Each round of a competition has 5 solves that should be considered separately when counting the number of solves.
- Consider all attempts before a sub-5 time is achieved, even attempts in the same or previous rounds of a competition. The ordering of the rounds is indicated in the 'rank' column of the `RoundTypes` table where the smaller numbers come before the larger numbers.

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.

Region Rivalries

Europe and North America are both regions with strong speedcubers in the WCA. Which region has the faster group of speedcubers on average?

To answer this question, characterize each person using their best *average* score according to their listing in the `ranksaverage` table. In the `persons` table the `countryId` indicates each person's country affiliation. The `countries` table lists the region that each country is in via the `continentId` column (Europe is "_Europe" and North America is "_North America").

Before attempting to answer the question, state what you expect the answer to be below. What do you conclude about speedcubers in Europe vs. North America?

Alternative Explanations

Develop an alternative explanation/hypothesis regarding speedcubers from Europe and North America that is

1. Consistent with the results you produced in the previous question; but
2. Provides a different interpretation or explanation for what is going on.

If the results from the previous question were unexpected, make use of systems thinking to develop an alternative hypothesis. If the results were consistent with your expectations, then use skeptical thinking. In either case, you should present an analysis that shows evidence for or against this alternative explanation relative to the conclusion that you made in the previous question.

What is your alternative explanation?

Summarize the evidence in the data for/against your alternative explanation.

Format of the Final Report

Answer All Questions

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

Discussion

Putting it all together, what did you learn from your data?

- Reflect on the process of conducting this project. What was challenging, what have you learned from the process itself?
- Was there anything unexpected that you found in your analysis? Or was everything essentially as you expected?
- Include acknowledgements for any help received
- Report the contribution of each member (i.e. who did what).

Formatting

- Create the report using R Markdown knitted to a PDF file, with headers for each section and each question answered;
- Include comments to the 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);

Submission on Gradescope

Remember to add all group members to the final report on Gradescope. If you do not do this the other members of the group will not get credit!