10K Analysis Take Home Question

Introduction

Thanks so much for taking the time! This open-ended take home question lets you experiment with a public dataset in a more relaxed setting. It also helps us provide a more comprehensive recruiting process. With that, we are interested in the final visualization but also your reasoning about data. So, feel free to optionally include any notes, sketches, etc that you produce during this task beyond just the final product. We hope that this is an enjoyable and interesting exercise. Let us know if you need any clarification!

Context

A large community 10 kilometer run in Boulder, Colorado attracts thousands of athletes from across the globe each year and, after the festivities conclude, they publish the results publicly for use by the athletes, the community, newspapers, running websites like splitwise, research, etc. Imagine you are working for the local university and want to create a tool to examine this data.

A. Task

1. What would be important for you to know before building this visualization? 2. How would you go about answering those questions?

B. Data

The lab provides you the attached sqlite3 database which includes the top 10,000 racers. Each row is an athlete that ran. Here's what each column means...

"time": The complete "race time" from when the athlete started running the course to when they finished the race. 43:44.71 means they completed the 10 kilometer run in 43 minutes and 44.71 seconds.
"finalPlace": The place indicates the ranking of the athlete by course time. So 1 indicates the fastest race time, 2 indicates the second fastest race time, etc.
"hometown": The city / town the athlete was from.
"division": Indicates the gender and age of the athlete. M34 indicates male of 34 years of age. F25 indicates female of 25 years of age.
"id": The ID uniquely identifies an athlete and includes their "wave" (not all athletes start racing at the same time so "waves" keep the course from getting congested). An id of AB615 means that the user was the 615th athlete registered in the AB wave. Waves start a few minutes apart in alpha order so A starts before BA starts before C.

1. Do you have any data quality concerns looking at this data? Anything you want the researchers in the lab to be aware of?
2. Any transformations you want to apply to this data to make it easier to work with?

C. Tool

The lab gives a list of example types of questions they want to answer about the top 10000 athletes... This list is not complete, just a few examples...

- What from what "age band" (like 30 - 39 years old) were most women athletes?
- How many athletes were aged 20 - 29 years old?
- What was the average race times for participants in the C wave?
- From what hometowns were the athletes in the A wave? Which town had the most A wave athletes?
- What was the fastest time of the male athletes?
- What were the final places of male athletes aged 40-49 years old from Boulder, CO?

1. Using the provided SQLite database, can you build a tool that helps answer these types of questions?

D. Evaluation

After you've built your tool...

1. How did you decide on the technologies behind your tool?
2. How would you evaluate the effectiveness of your visualization?
3. What was important to you in building the visualization?

Submission

Thank you so much again for your time! We really appreciate it. Please send along the answers to the above questions in the form of:

A1. Before building this visualization, it's important for me to know...
A2. I would go about finding out that information by...

For the tool, please send along the source code and everything needed to run that source code including build instructions. If you preformed transformations to the data provided, please describe what changes were made and send along the code / queries to do so such that we can repeat your procedure.

We are very sorry but, for security reasons, we ask that you do not send binaries or transformed sqlite databases back to us. Just queries and source code please! Only exception is a Tableau workbook (version 9.3 is preferred).