ANALYSIS OF CUBING COMPETITION DATA

Karl Oskar Kuuse

Introduction and motivation

My project is based on the topic of (speed)cubing, which simply put is the sport of competitively solving combination puzzles, like the 3x3x3 Rubik's cube, as fast as possible. There are two main points that motivated me to choose this topic. Firstly, I have personal interest in cubing, and I believe that there are interesting aspects to be analyzed about the topic that others might also appreciate. Secondly, I was interested in whether it was possible to use data analysis in order to gain more insight into the statistics and trends of cubing competitions, and by doing so find useful tips for people that want to become better at the activity.

Data and goals

Cubing competitions are regulated by an organization named the World Cube Association (WCA), which also maintains a public database that contains information about roughly 170 000 competitors, their results and much more. For this reason, the database was an obvious choice as my data source for the project and since it was professionally maintained and documented, there were no problems with the data quality.

Based on the available data I formulated more concrete goals:
1.Use data analysis methods to discover interesting trends and relations in the cubing competition data from the WCA database
2.Visualize and describe my findings in an understandable way and create a coherent resource for possible future use

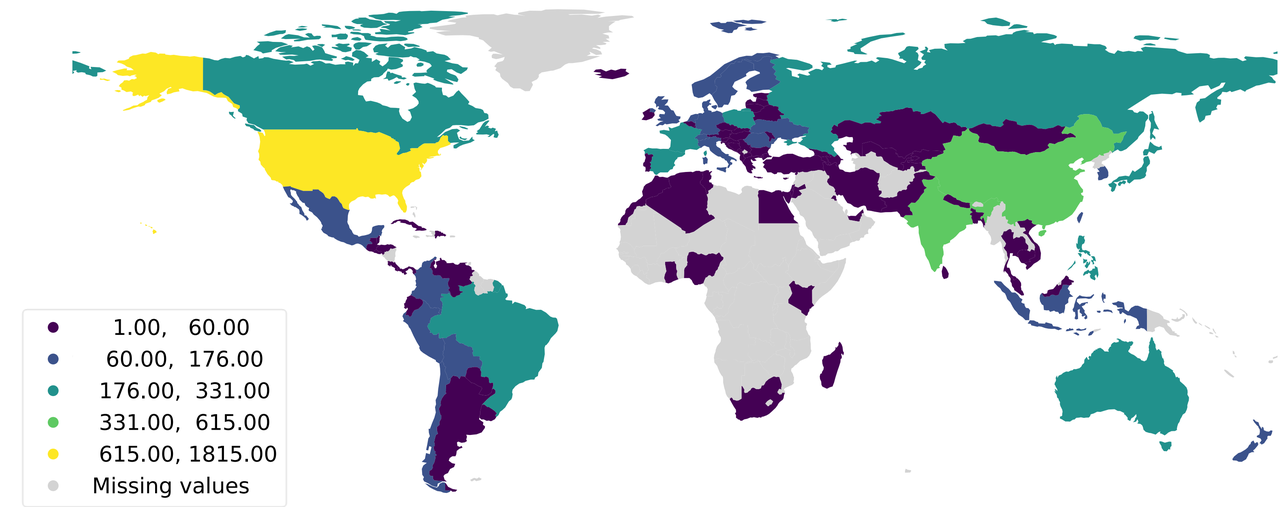
Methods

The main tools used for the project on the software side of things were Python in the form of Jupyter Notebook and MySQL for storing a copy of the WCA database on my local computer, which made making the necessary queries much faster and more convenient.

My general workflow was to first produce an idea about what aspect or trend I'd like to investigate, after which I would gather the necessary data from the database using SQL queries. This also allowed me to export the created datasets as CSV files for easy use. This was followed up with the actual data analysis part, which was a bit different every time, since the type of data I was working with and the plots I aimed to produce were different. This was also the part where the methods learned during the course came in handy.

Truly a global activity

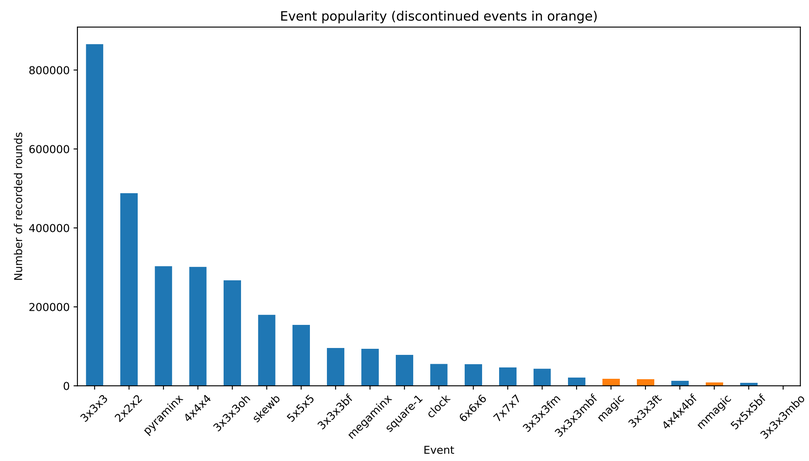
Cubing competitions are held as in-person events all over the world with a total of 110 unique countries having hosted at least one competition. This chart is topped by the US where a total of 1815 competition have been held, which also helps to explain the fact that US competitors have in total achieved 308 world records, which is the most out of any country.

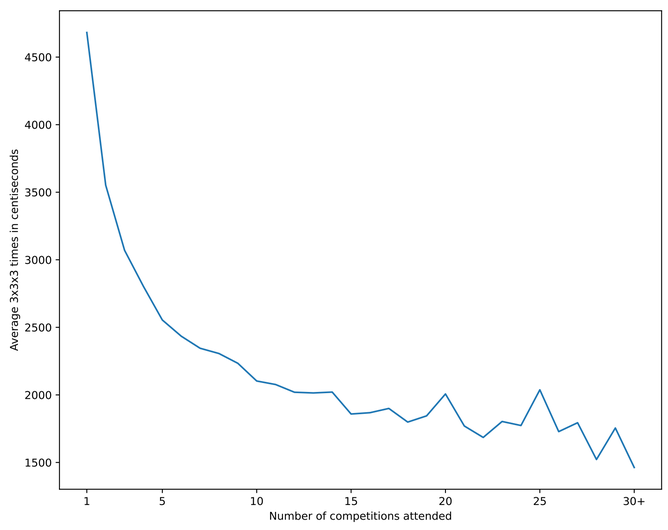


People love the classics

There are currently 17 'official' puzzles, which are used for competition, usually referred to as events. Even with this variety, the classic 3x3x3 Rubik's cube is by far the most popular event with over 800 000 rounds of competition recorder for the cube.

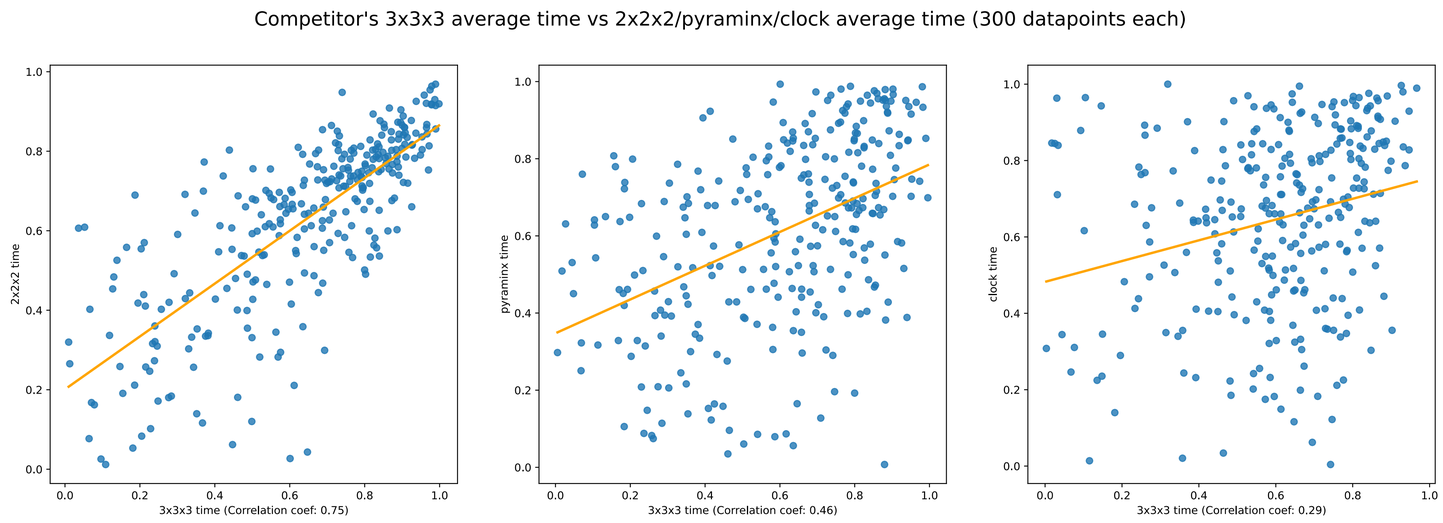
This difference in popularity also meant that at times during the analysis, when comparing all the 17 events wasn't feasible, I excluded some of the less popular events.





Experience matters

A natural question with this sort of data is do people with more competition experience achieve faster times. Here I have grouped people by the number of competitions they have attended and found the average 3x3x3 solve time for each group.
At first, the average drastically decreases as the number of competitions increases, but eventually the averages get more volatile. One reason for this is that the sample of people with the specific number of competitions gets quite small, but this also demonstrate that at some point, the 'natural' improvement stops, and dedicated work is required to improve further. Still, experience does make a difference!



Some events complement each other

Here I have explored correlation between competitor's performance in different events. For this I collected the times of competitor's who had results for both events I was comparing. I then normalized the data to a scale of 0 to 1, where 0 is the worst time out of the extracted times and 1 is the best.

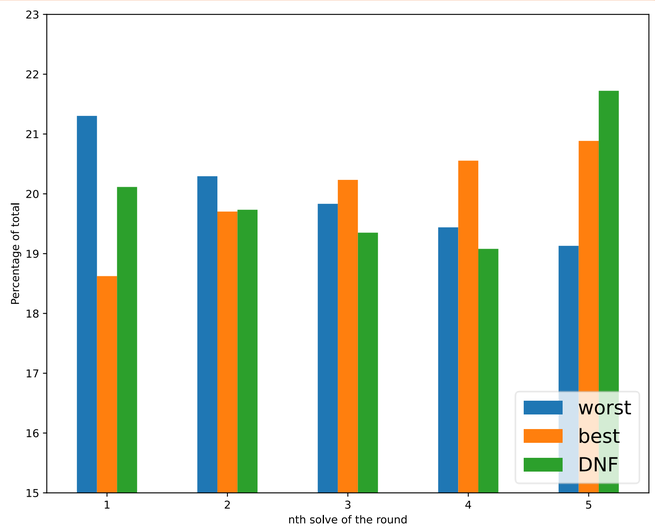
It turns out there is quite strong correlation between the results for 3x3x3 cube and 2x2x2 cube: if a person has good results in one of them, they tend to have good results also in the other one. But when we compare events that are not so similar, the correlation starts to vary. Currently we see that 3x3x3 cube and the pyraminx puzzle results are less correlated. This is further demonstrated when comparing the 3x3x3 cube and the clock puzzle, which are drastically different puzzles and there is no strong correlation left.

Based on this, if a person was looking to improve at a specific event, it might be beneficial to also practice events that are similar to the desired event.

Don't forget to warm up

A typical competition round consists of five separate solves, after which the average is calculated. Here I have explored which solve out of the five tends to be the worst or the best. The results are mostly uniform, but as we can see, the worst solves tend to occur a bit more often towards the start of the round and the best solves towards the end of the round.

I also checked when do DNFs (stands for 'did not finish', usually caused when the timer is stopped prematurely) occur. Surprisingly, there is an uptick of DNFs during the last solve, perhaps this is because the competitors are a bit too comfortable by that time and are prone to making mistakes.



Project repository

As a final note, I'd like to add that all the code for this project can be found freely in the project repository: https://github.com/oskarkuuse/IDSproject

So, if you would like to read more about the topic or explore the data yourself then I encourage you to check it out.