Analysis of cubing competition data

The purpose of this report is to tie together the project and give some context to the plots produced.

General Information:

The background of this project is covered in depth in the file 'C4_report' which is also available in the repository, here I will just shortly summarize the most important aspects and rather focus on the results of the project.

My project is on the topic of speedcubing, which is the sport of solving various combination puzzles (like the Rubik's cube) as fast as possible. A big part of speedcubing is competing, which is also the focus of this project.

This project is based on the data from the World Cube Association's (WCA) database which contains information about various cubing competitions, participants and their results. The database is publicly available and linked in the projects readme-file.

The goals of the project were worded as:
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

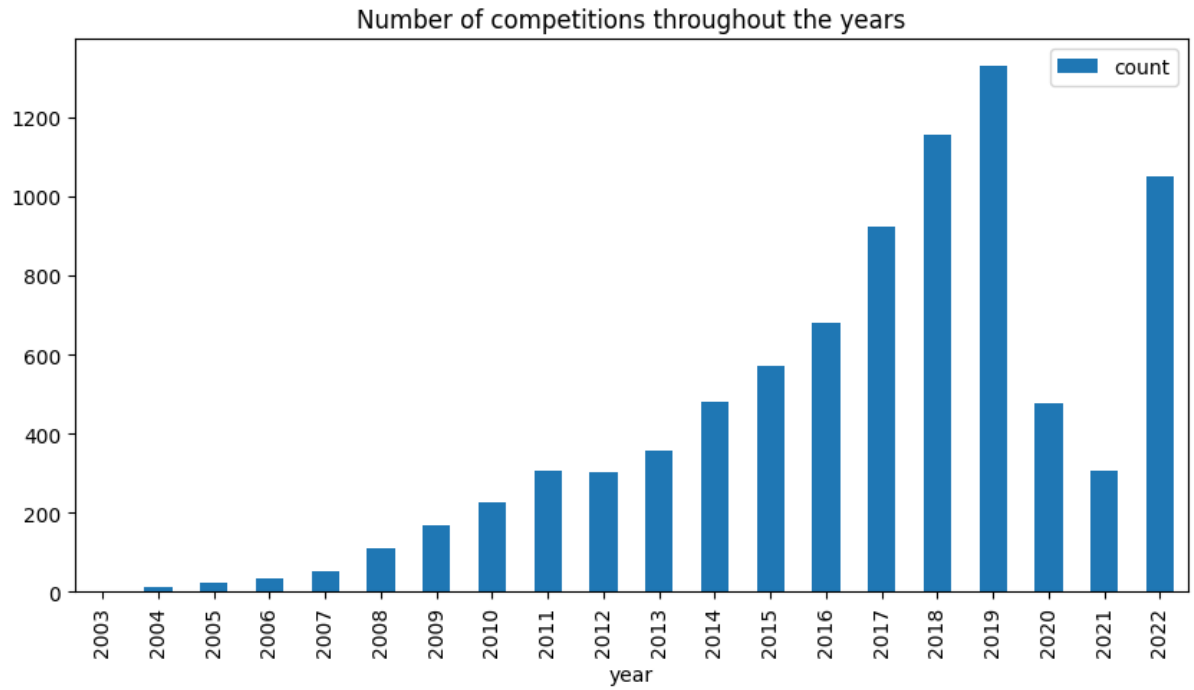
As a side note: in the file '*C4_report'* one of my side goals was to also try to create some machine learning model related to the data. While working on the project I decided to scrap that idea, since it wasn't my project's focus to start with and I found out that the other goals were more time consuming, so I decided to focus on them instead to get the best results.

Project summary:

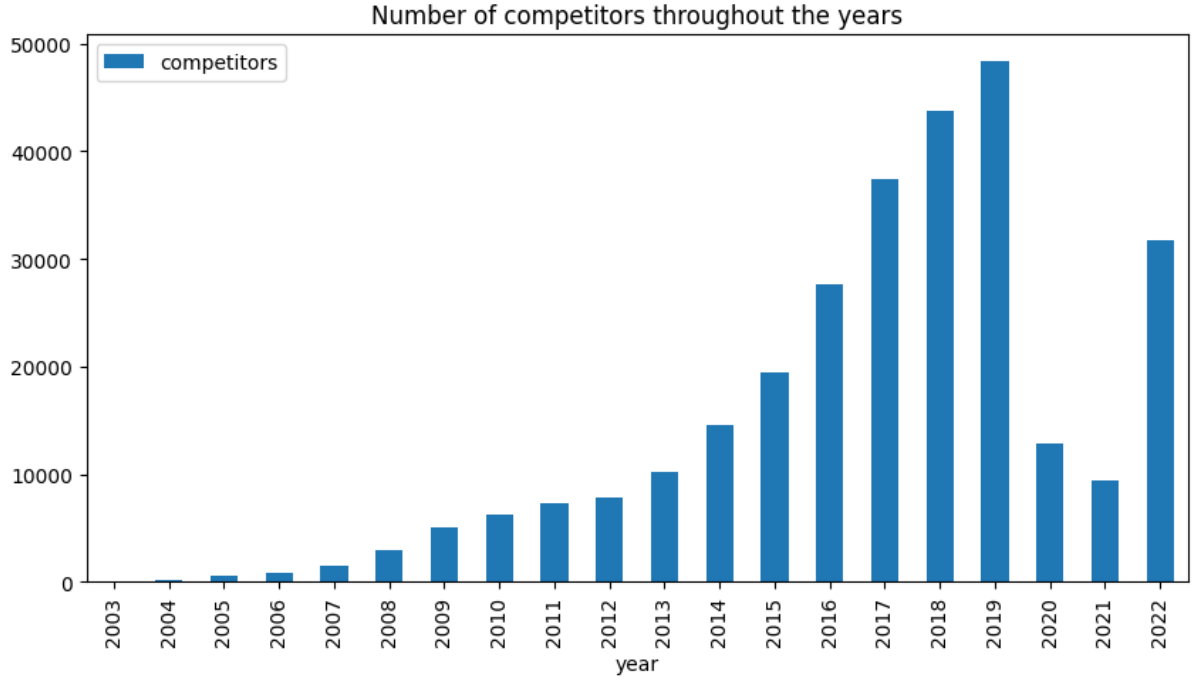
The project is thematically split into two parts: general statistics of cubing competitions and performance/result analysis. The first topic tackles more general aspects of the cubing world, for example, number of cubing competitions held throughout the years or most popular cubing events. This is useful since it gives better context about the topic and a better insight into the data sizes we are working, which is all useful when we get to analyzing the results. The performance analysis part goes more in-depth about the results of the competitor's and through data analysis attempts to find ways how one can improve at cubing.

<u>General statistics:</u>

Since a lot of the focus of this project is on the topic of cubing competitions then one of the first things I began my work with was to extract the most general information about them. The plot below displays number cubing competitions throughout the years:

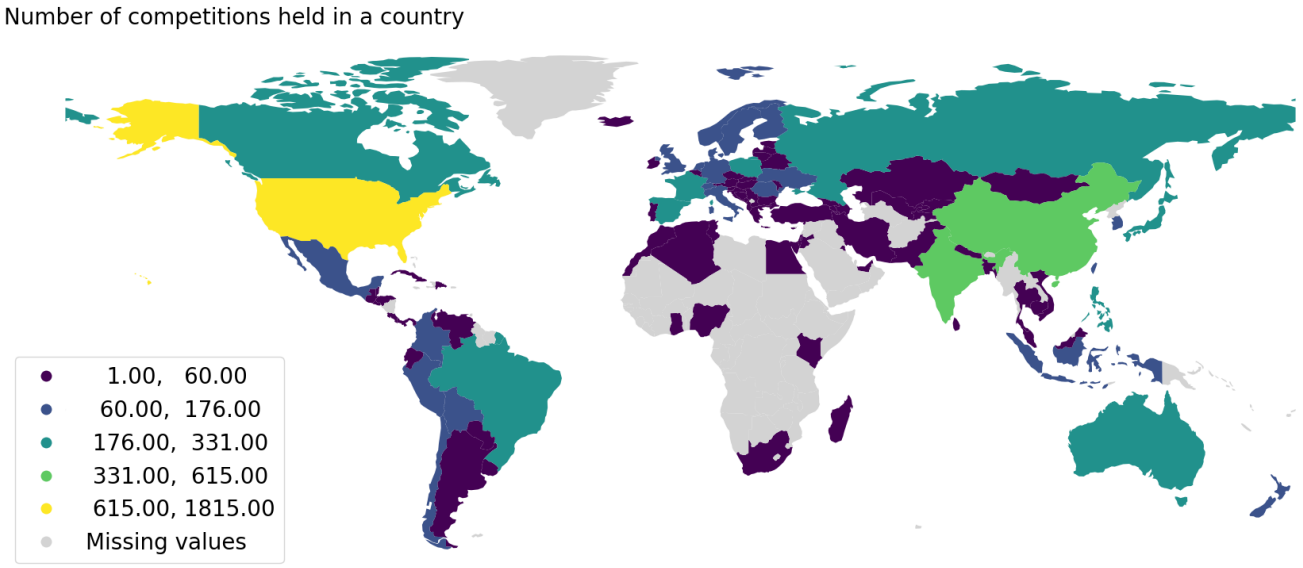


As we can see, cubing competitions have been increasing in number since the creation of WCA in the year of 2004 (few competitions have been added retroactively from 2003). This rise was sadly stopped during the years of 2020 and 2021 almost exclusively by the covid pandemic, since cubing competitions are in-person events and holding them was just impossible. Even though 2022 is not quite over as of writing this summary, we can still see that it won't likely reach the height of 2019, but hopefully the future years will do so. With the same idea, I also plotted the number of active competitor's (at least one round participation) by a given year:



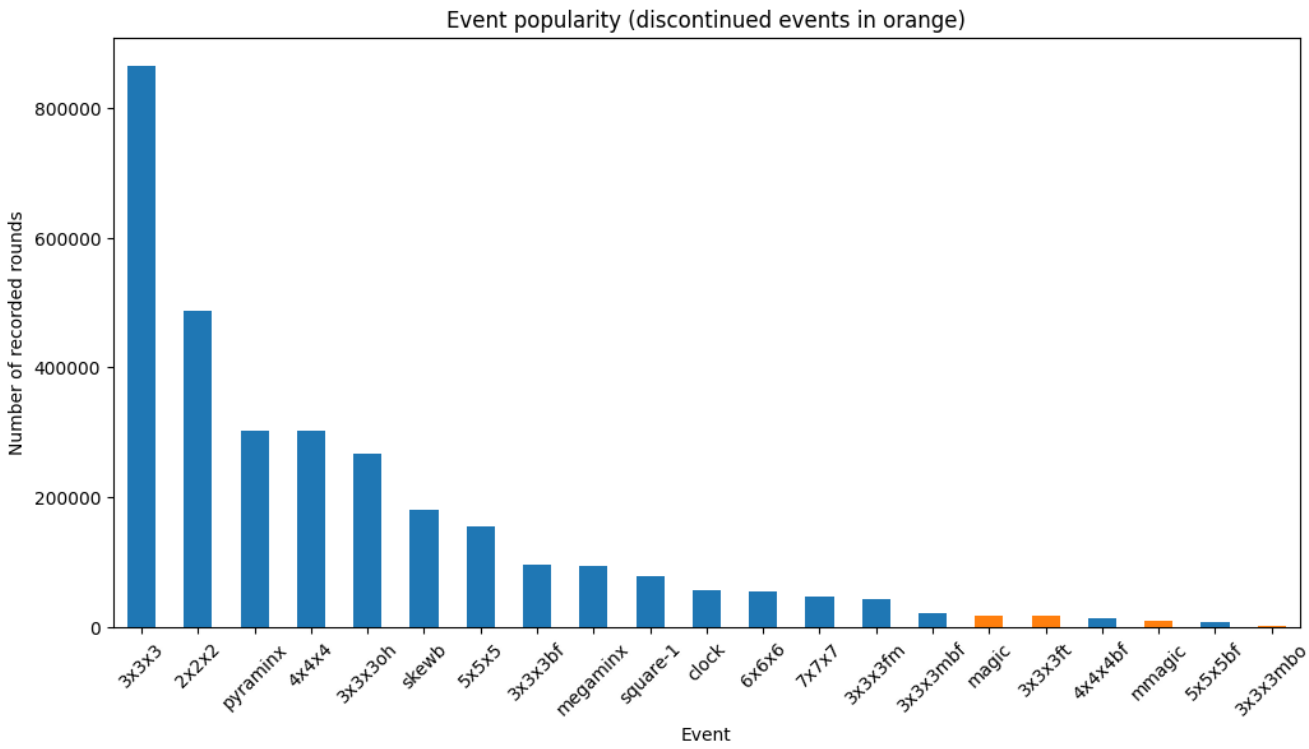
Unsurprisingly though, the shape of the barplot matches the number of competitions, so the number of competitors and competitions is quite clearly connected and it just isn't a smaller group of people participating at more events every year. Still nice to know.

Based on the competitions data, I got the idea to also visualize competitions globally:



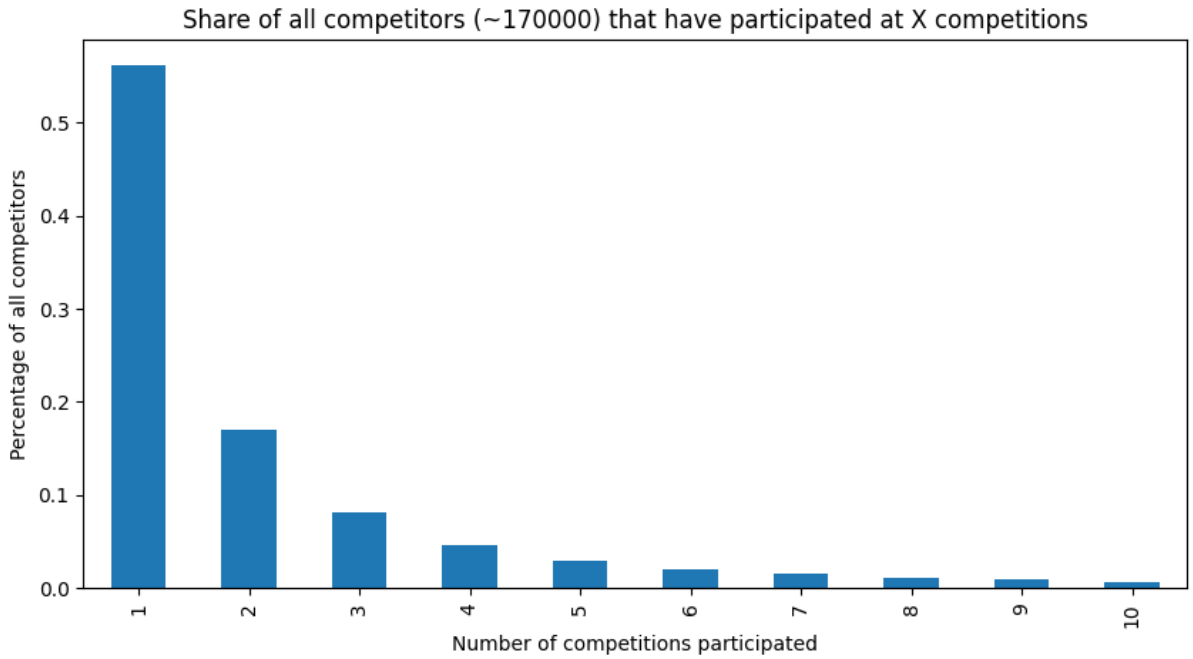
Here I have plotted the countries that have, at some point, hosted cubing events. I created the plot using geopandas library and additionally used 'Fisher Jenks' clustering method to cluster the countries by the number of cubing competitions. This is simply because the US totally dominates the cubing competition scene with 1815 competitions held, the next closest being China with 615 (~three times less), so plotting the countries without clustering then others wouldn't really stand out. This way we lose some accuracy when it comes to relative sizes, but we get the general sense where cubing is popular and most importantly, it really is a global activity. As we can see there have been quite some events even in Estonia.

Another important part of cubing competitions are the events (puzzles used for competition). There are 17 WCA approved events, but not all of them have to be a part of a specific competition. This raises the question which events are most important. This was motivation for my next plot where I visualized this:



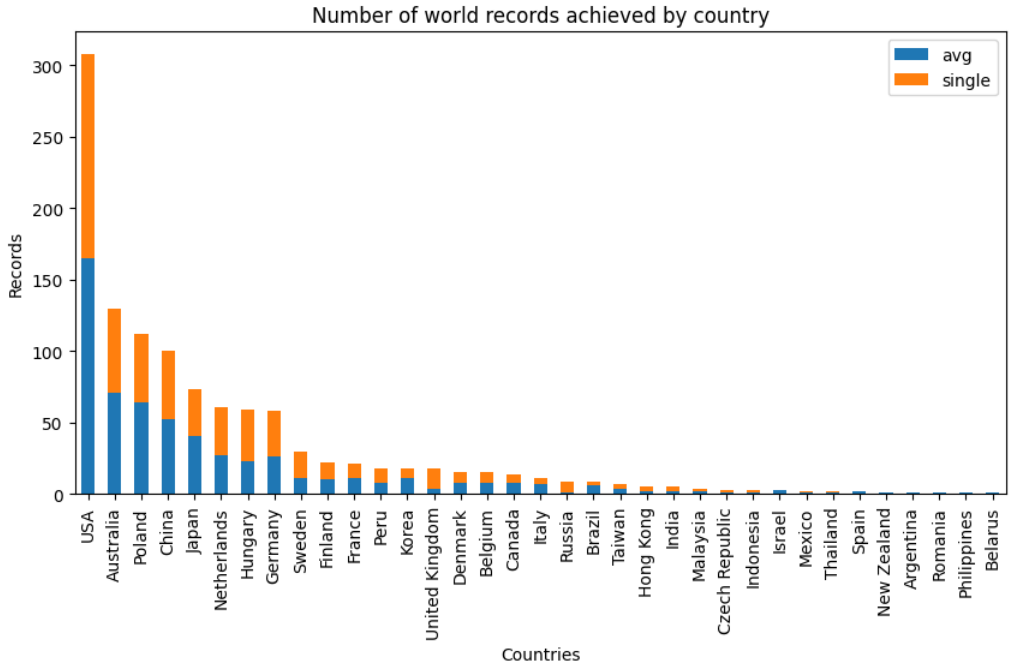
On the y-axis is the count of how many rounds have been recorded for the event (one round is a set of (usually) 5 solves done by a single competitor). We can see that the classic 3x3x3 Rubik's cube is clearly the most popular event. For this reason, in the cubing world it is considered to be the most prestigious event, but times for the cube have become so good that new records happen quite seldom. Towards the top we see other cube-shaped puzzles from sizes 2x2x2 to 5x5x5, but also some more irregular puzzles like pyraminx and skewb. There are also some variations of events, for example, bf at the end of the event name stands for blind folded (so the solved are done while the solver is blind folded) and also oh, which stands for one-handed. Even though other events don't come close to the Rubik's cube when it comes to the number of round solves, there still is a lot of variety and choosing an event to focus your efforts on is not simple.
An extra note about this plot: In the performance analysis part, some plots compare events between each-other and since comparing every event between each-other wasn't always viable computationally and time wise, I chose to focus on the more popular ones, based on this plot, to be more general with my analysis.

To get some additional overview about the competitors, I also made a plot about how many competitions have competitors actually taken part of:



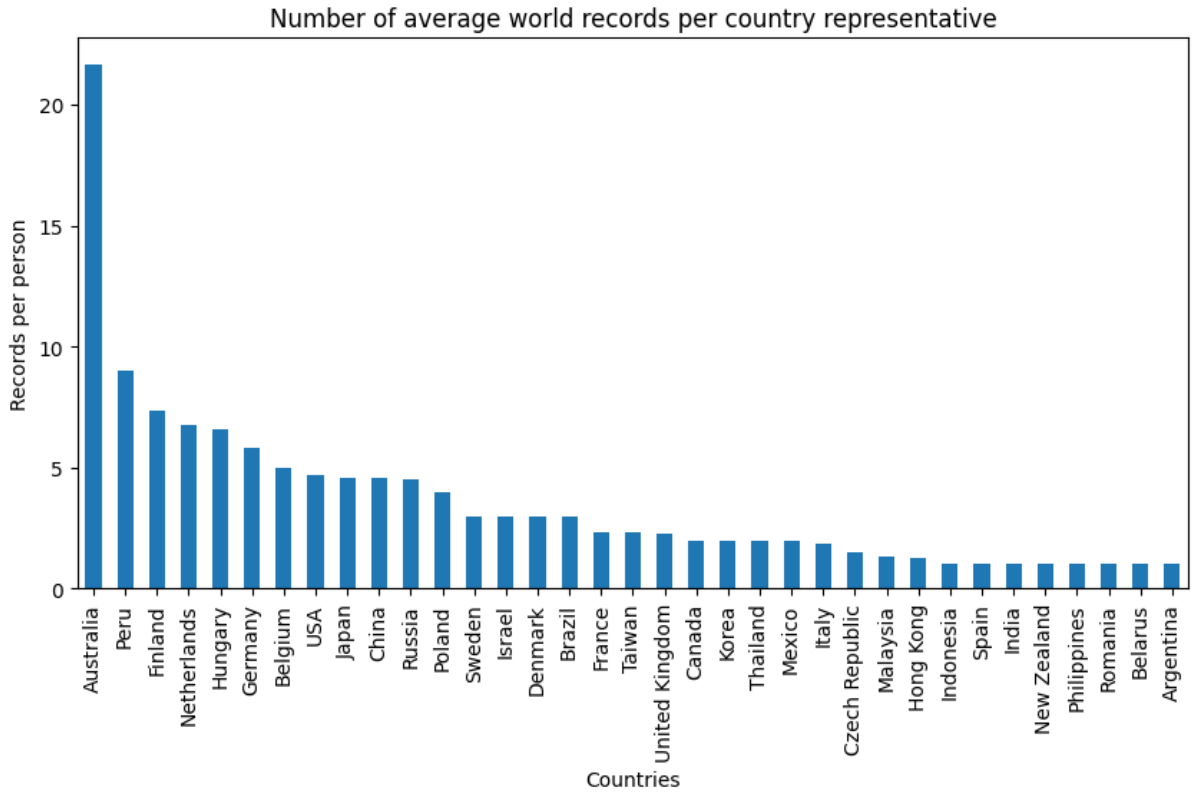
I normalized the y-axis, which is the number of people that belong to the specific group, since there are roughly 170 000 registered participants in the database, then roughly 85 000 of them have participated exactly at a single competition. As the number of competitions increases we see a clear drop in competitors (that's why I only visualized the smaller number of competitions), so not everyone sticks with cubing.

I also analyzed how many world records have certain country representatives achieved:

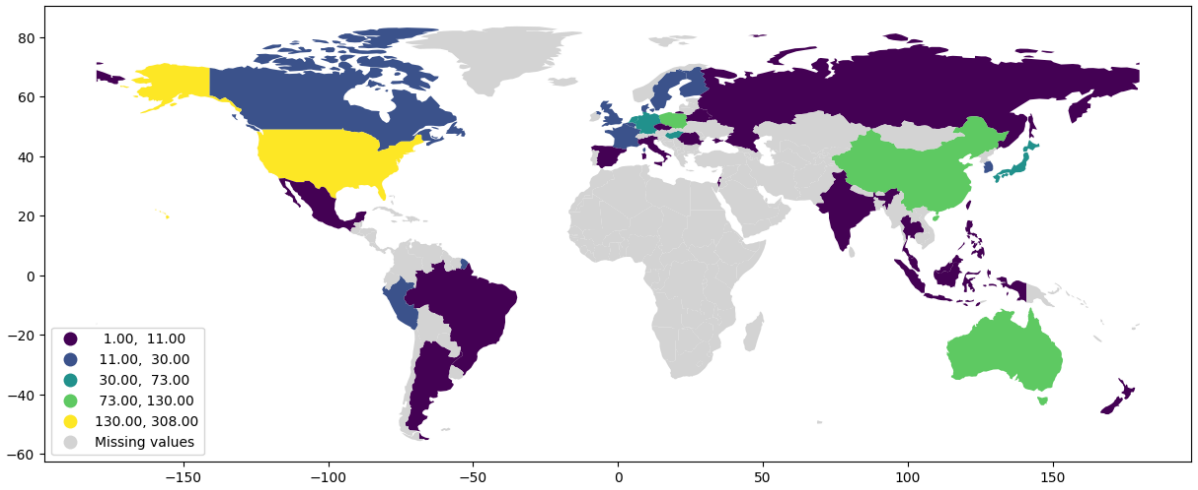


Here I have additionally split the world records into average WRs (essentially competitors round average of multiple solves) and single WRs (single fastest times achieved). We can see that once again the US dominates the field, this might be explained by the fact that they hold the most amount of competitions, which creates a bigger community and more

competition that pushes for more records. As a continuation of this graph, I also plotted how many WRs every country representative (that had achieved at least one WR) had achieved:



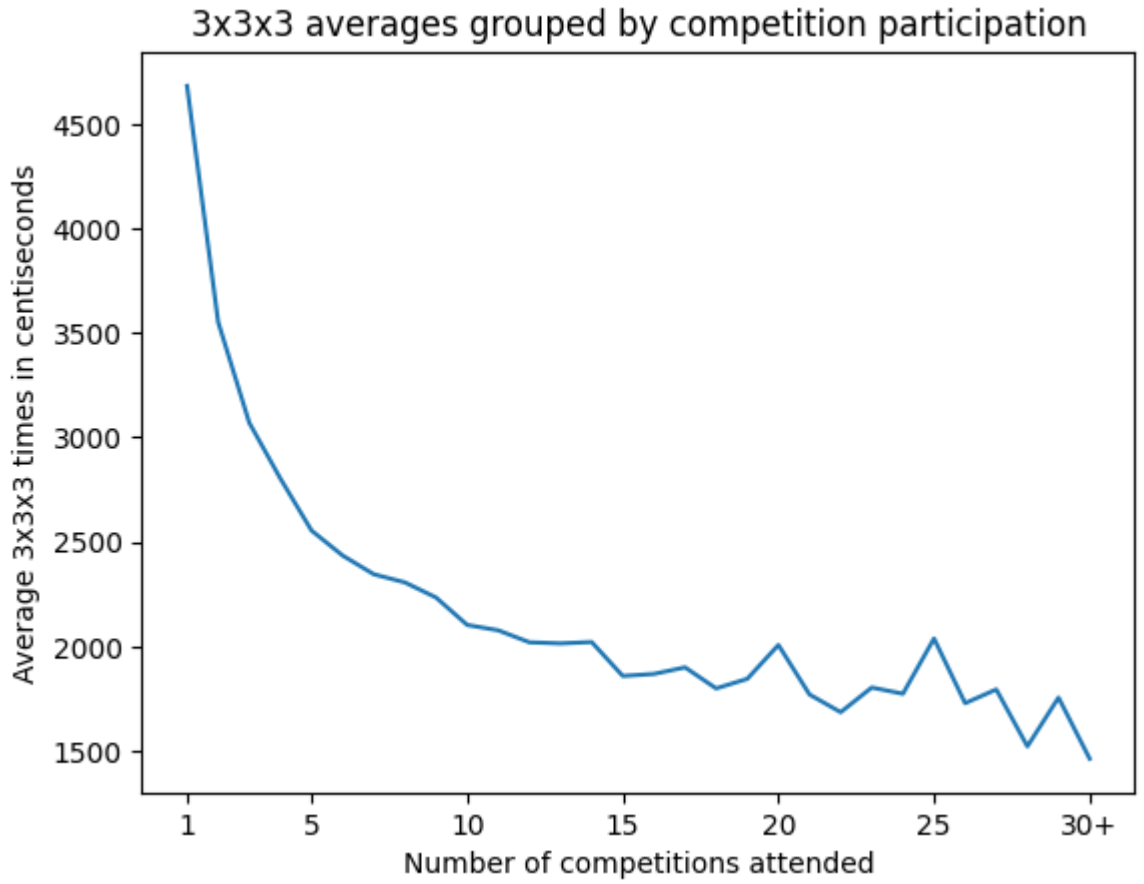
This graph shows a curious fact that Australia has achieved roughly 100 WRs, but each of their competitors has achieved around 20 WRs, so that means all of Australia's world records are achieved by around 5 people. This is actually explained by the fact that there is an Australian competitor Feliks Zemdegs who has had one of the most prolific cubing careers out of any competitor and since he has achieved countless world records, this somewhat carries the Australia's WR numbers. As a end to this topic, I also visualized the countries that have achieved world records similar to the previous world plot:



Here we can see that even though cubing competitions are held all over the world, world records are a bit more localized and there are quite some countries that haven’t achieved a single world record.

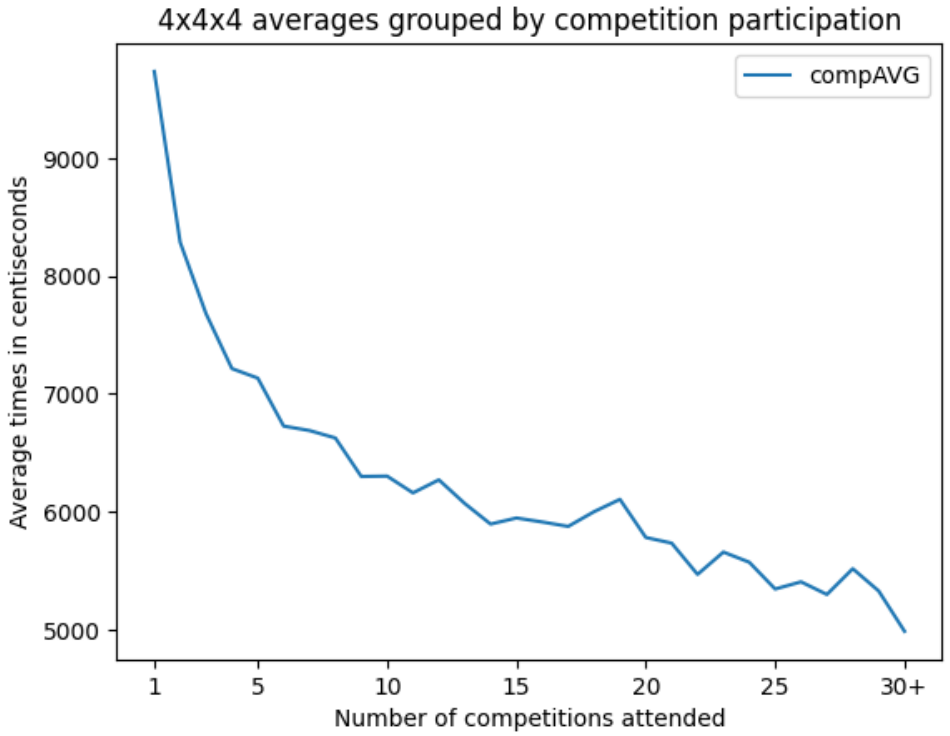
Performance analysis:

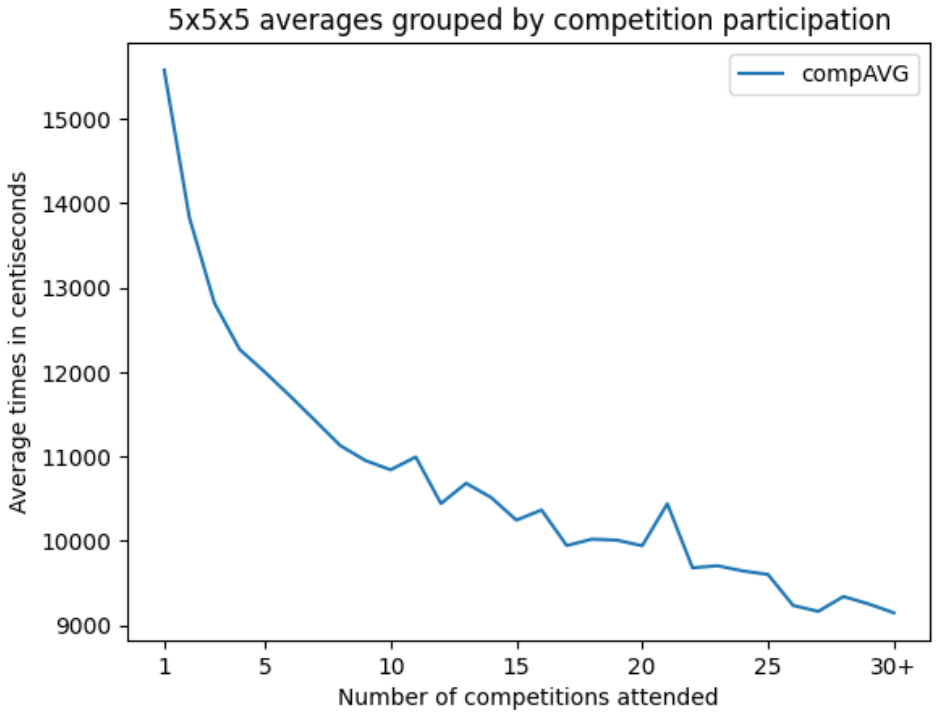
The most logical starting point to analyzing the performance of competitors was to investigate whether competition experience had some connection to the performance of the competitor. For this I made the following time series plot:



For this plot I grouped the people by the number of competitions they had participated at and found the 3x3x3 round average of each group (times are in centiseconds, so seconds * 100, since it's the format used in the database). It's not so surprising that the average decreases rapidly at first as the number of competitions attended increases, but at the same time, around 10-15 competitions the averages get more volatile and somewhat start to plateau. This might signify that at first, getting better is quite easy and it just takes some time/experience, but at some point getting better and bringing the averages down requires more specialization and concentrated effort, something which everybody doesn't do, so the averages tend to stagnate. Still, based on this I'd say it's fair to argue that competition experience does make a difference. As a side note, looking back at the plot, I think I could have approached it a bit better, since currently the groups have quite different numbers of people in them (like we saw during the general analysis), but I still believe that this gives a good general overview.

I also created a similar plot for a couple of other events, I'll outline here two more of them:

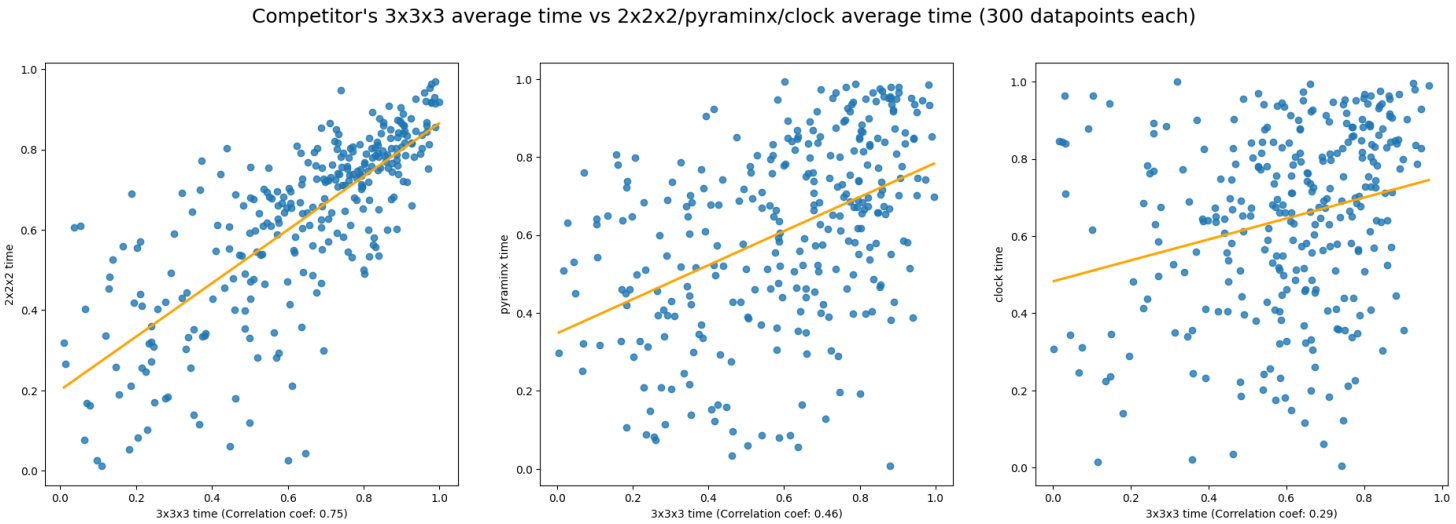




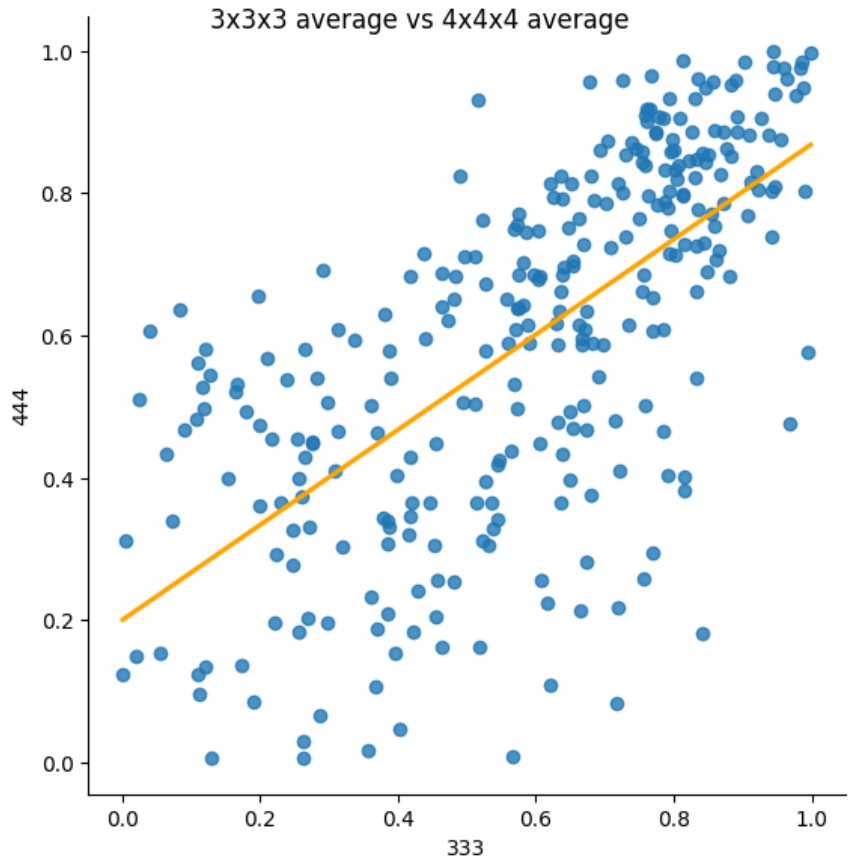
Both of these plots also show a similar trend to the 3x3x3 cube averages, but they are a bit flatter and the decrease over the first few competition groups doesn't seem to be as steep, perhaps this the 4x4x4 and 5x5x5 are just harder puzzles to get good at quickly and this causes the slower progression.

I also wanted to explore whether performance in two different events was somehow correlated. For this I chose two different events and extracted datasets of competitors, who

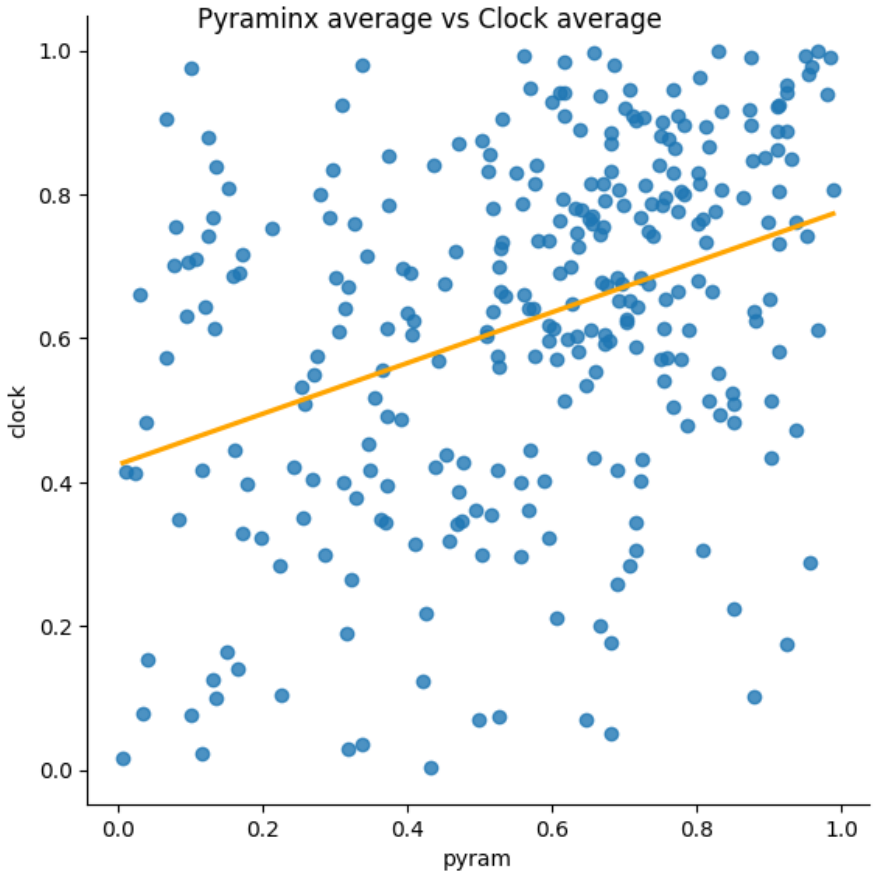
had competed in both of those events. I engineered the data by removing the most extreme times (by using the mean and standard deviation) and also standardized the data to a scale of [0,1], where 0 represents the worst time of the event in the remaining dataset and 1 the best. Then I randomly picked 300 data points from the dataset and plotted them (otherwise it would have been too cluttered) I did this for a couple of event pairs and one of the results was this (might be better viewed from the notebook):



Here the first plots in order are 3x3x3 vs 2x2x2 times, 3x3x3 vs pyraminx (puzzle) times and 3x3x3 vs clock (puzzle) times. I think it's very cool to see how some of the event results are more correlated than others. For 3x3x3 and 2x2x2 there is clear correlation between the results, especially when one of the values are over the 0.5-0.6 mark. This means that when the competitor's average in one event tends to be higher then also his other event's results tend to be higher. But now looking at the other graphs, we see that the correlation is so strong anymore. For 3x3x3 and pyraminx there still is some correlation and there don't seem to be extreme cases where a competitor is very good at one event, but bad at the other one, but when it comes to the 3x3x3 vs clock averages plot, then there seems to be even less correlation. This shows that some event results seem to be more correlated/tied than other events. Probably this stems from the fact that some puzzles are more similar to each other, for this example 3x3x3 and 2x2x2 are both cubes with different sizes. This is supported by the fact that also 3x3x3 and 4x4x4 seem to be more correlated:

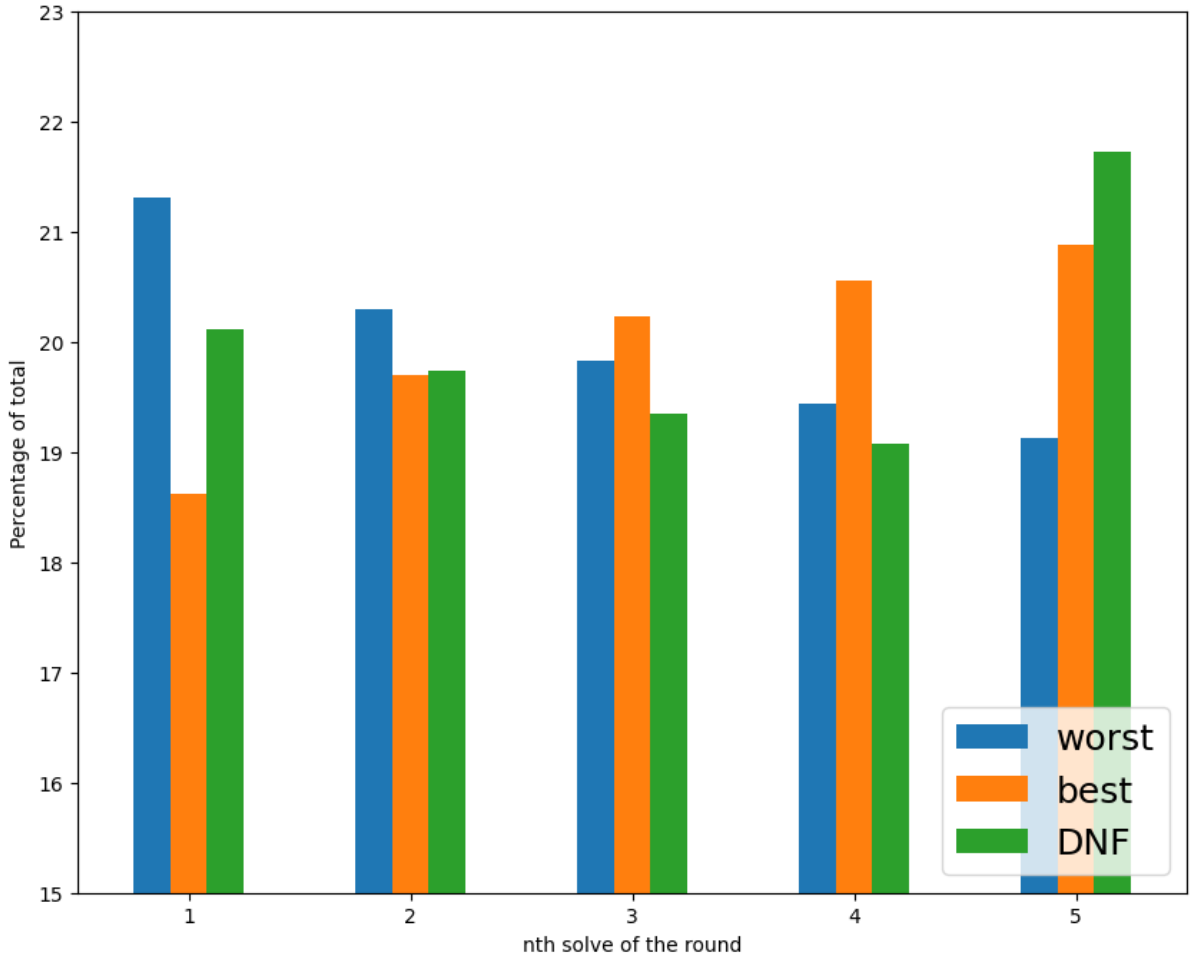


but 3x3x3 and pyraminx/clock are totally different puzzles. Also, those that know more about cubing know that many bigger sized cubes are solved by reducing them to their smaller variant (so solving a 4x4x4 can be reduced to solving a 3x3x3), which means one event benefits the other. Out of curiosity I also compared pyraminx and clock between each other:



We can see that also for them, the correlation isn't quite visible, even though the correlation coefficient is a bit higher than between 3x3x3 and clock.
As a summary for this previous part, I'd say that these plots show that it's beneficial to practice multiple different events since there really is correlation between the results of the events. Still, some events benefit each other more, so it might be good to choose your focus if you want to improve.

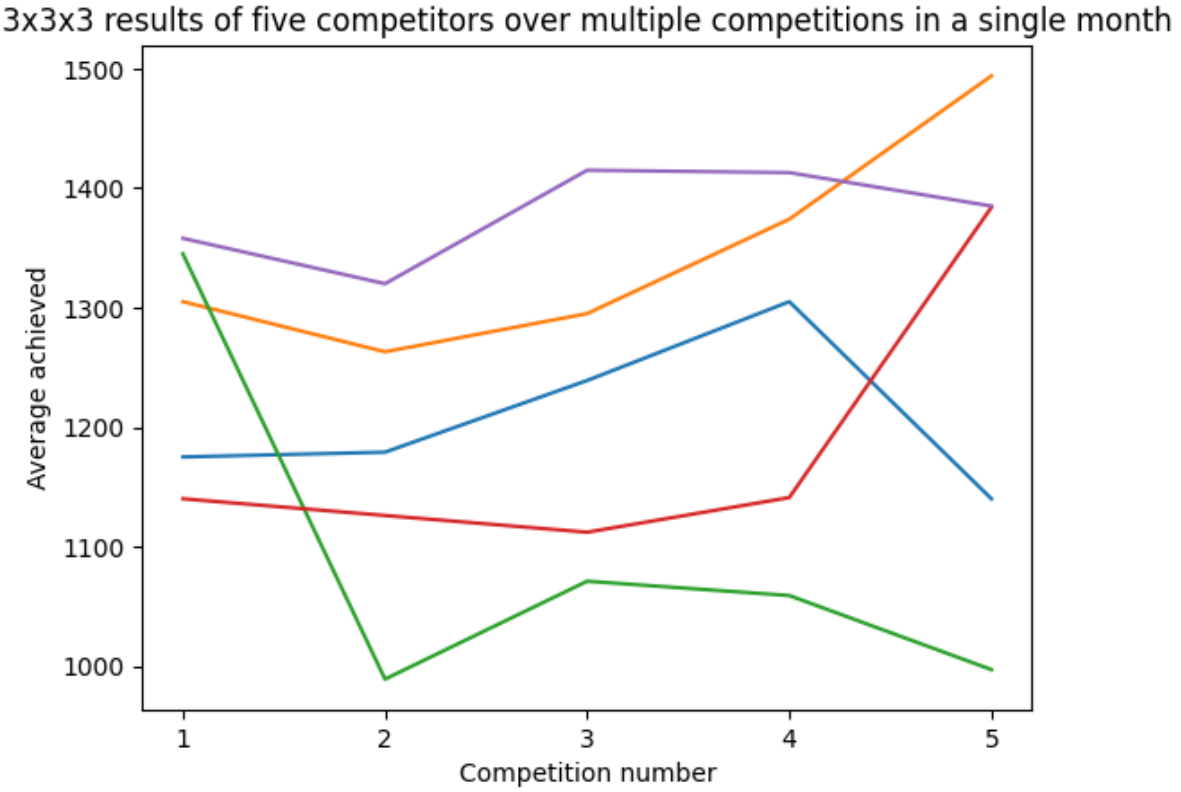
One of the things that I also investigated was the round breakdown for 3x3x3 solves. By this I mean that a typical 3x3x3 round consists of five solves and I explored which of the solves is usually the competitor's worst/best out of the five. I also checked on which solves do DNFs most usually occur (DNF=did not finish, which means that the competitor's solve wasn't counted, usually when the timer is stopped before the puzzle is solved).



Here on the y-axis is the percentage of total, so out of the 3x3x3 averages looked at, how many had the worst/best time on the n-th solve. The distribution wasn't far off from being uniform, but we can still see that the worst solves occurred a bit more frequently towards the start of the round and the best towards the end of the round. What surprised me was that DNFs spiked during the last solve of the round even though that's when getting the best solve is a bit more likely. Perhaps this signals that there is some overconfidence: the competitor is warmed up by the last round and feeling a bit faster when solving the puzzle, but this means that also mistakes are more prone to happen. Based on this, I'd say it's important for the competitor to warm up before the competition round (since worse solves

tend to happen a bit more frequently towards the start), but also be mindful at the end of the round, to avoid DNFs.

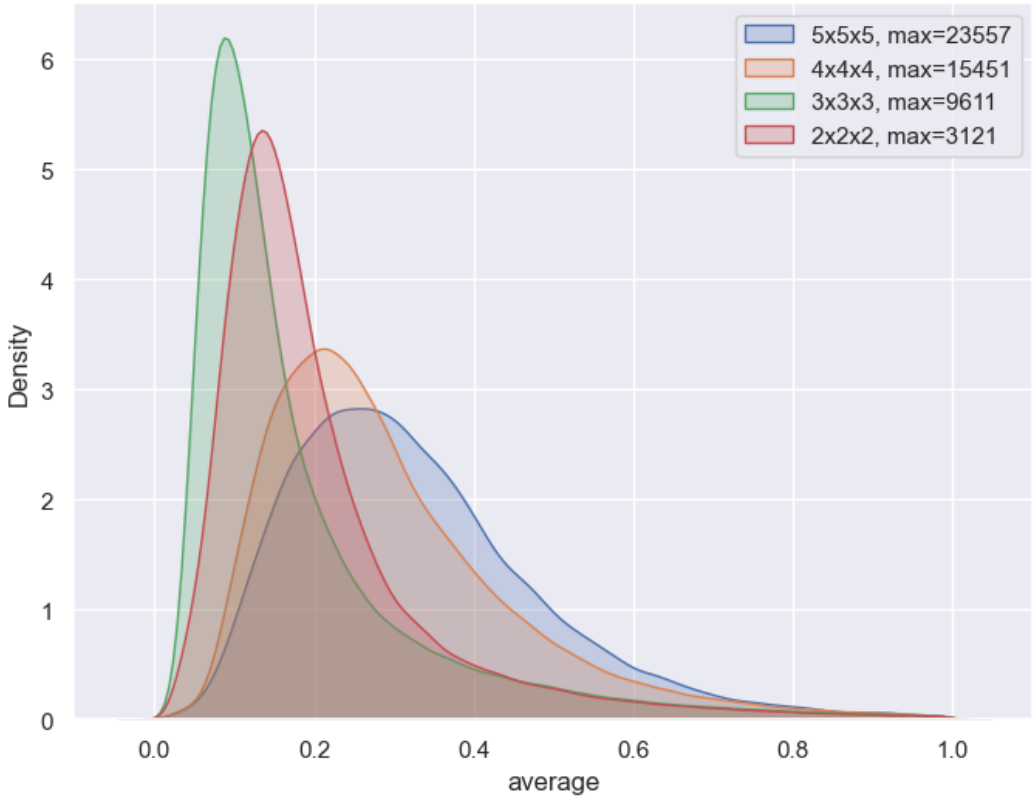
Another question I had was whether competing too frequently impacts the performance of a competitor. We know from earlier that competitive experience can be helpful, but can too much of it have a negative effect? For this I searched for people in the database that had competed in a lot of competitions in a single month, and I found five competitors that had all competed in five competitions in a single month. By ordering their competitions and extracting their average results (for the 3x3x3 event) I got the result:



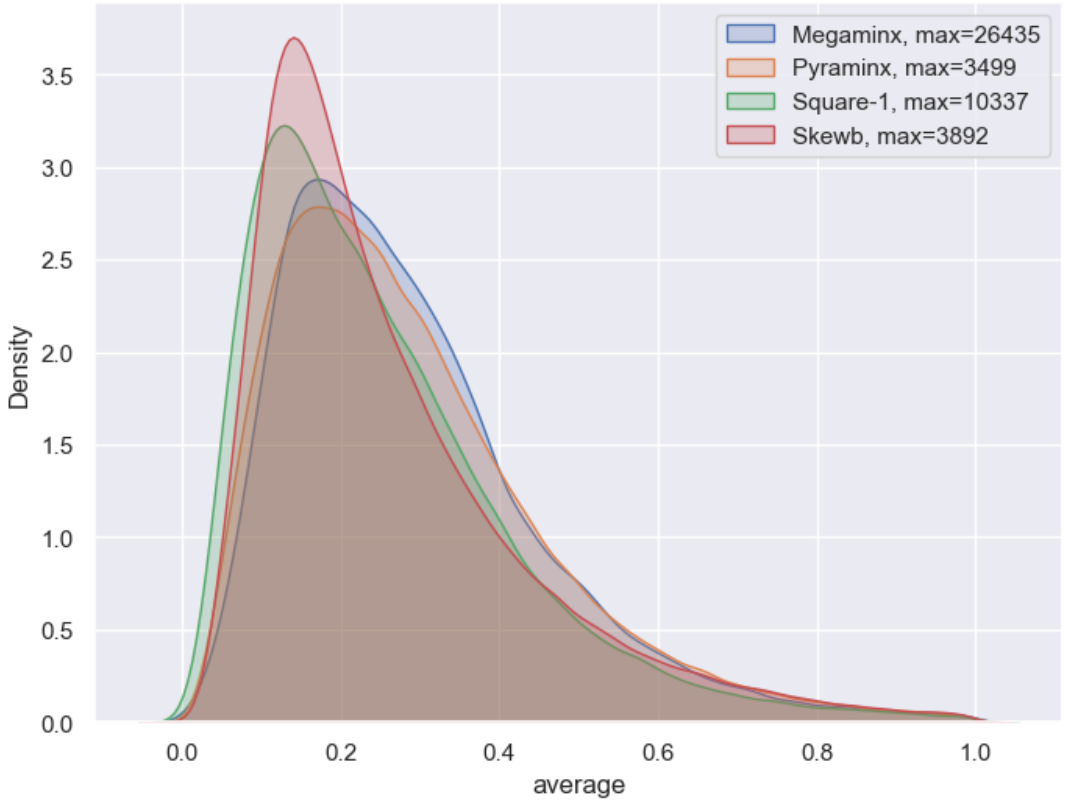
I add here that this is one of the plots that I'm not that happy with, since it currently has quite a small sample size and there are probably ways I could have improved it, but not everything is possible because of lack of time. Still, I included it since not everything always works out perfectly and I think the idea of the plot is still worth mentioning/exploring. Since the times achieved were so close to each other I didn't also normalize them. Looking at the plot there isn't really a clear trend to be seen, although one can notice that almost every competitor (different color), had a better average time at the second tournament than in the first. After that the averages go all over the place, two competitor's clearly get worse, while other two eventually get better. So in conclusion, I think that it's an interesting idea to explore with more time, but currently this wasn't my most successful investigation.

As a final part of my project, I wanted to also visualize the distribution of averages for different events. This is maybe a bit further from the competitor's point of view, but still interesting to see how similar the time distributions are. For this I used seaborn kernel density estimator plots and I picked the events I looked at were 2x2x2-5x5x5 cubes and pyraminx, megaminx, skewb and square-1:

Distribution of 2x2x2-5x5x5 cube averages using KDE



Distribution of various cube averages using KDE



Here I scaled the averages to the interval [0,1] like before, to plot the distribution on one scale. On the first plot we see that distributions differ a bit, 3x3x3 and 2x2x2 seem to be more concentrated towards the left. Perhaps since they are two of the most popular events, which means a good part of the competitors have achieved good times at those events. 4x4x4 and 5x5x5 seem to vary a bit more and the proportion of people that solve at the highest level is smaller than the first two cubes. The second plot was a bit surprising since all of the four events have quite similar distributions, they seem to be somewhere in the middle of 2x2x2/3x3x3 cube time distributions and 4x4x4/5x5xx times distribution.

Conclusion:

All in all I'm quite happy with the work done. I think I was able to analyze quite a few aspects of the cubing competition world and outline some tips that might help people get better at cubing. At the same time not everything went smoothly and I myself also realized that the topic is very deep and there is a lot more work to be done. For example, many of my plots focused on just a few events, but there are many more to be analyzed/compared (like in the correlation part), and there are topics that I didn't have time to tackle, like how do male and female competitors compare to each other or does competing abroad affect competitor's times. Still, I hope that this project can be sort of a starting point, if somebody desires to explore these topics, perhaps I'll someday explore further myself.