IDS project report: Analysis of cubing competition data

Karl Oskar Kuuse

Link to the project repository: https://github.com/oskarkuuse/IDSproject

Business understanding:

Background: My project is on the topic of speedcubing competitions. In a general sense, speedcubing is the act of solving various twisty puzzles, of which different types of Rubik's cubes are most well known, as fast as possible. Speedcubing also has a relatively big competitive scene, which is the focus of this project. Speedcubing competitions are in-person competitions hosted all over the world where speedcubers compete in various different events (cube types) where obviously the faster solver is considered the winner. It's important to add that the governing body of cubing competitions is the World Cube Association (WCA), who regulates the rules of the competitions.

The motivation for this project mainly stems from the fact that speedcubing is a competitive activity where any bit of extra information can help a person become a better competitor and achieve better times. This also includes knowing the trends and statistics from different competitions, which this project aims to describe. It is worth mentioning that speedcubing really isn't a business and this project isn't meant directly to support any business, instead, the main benefactors and the target audience are the people that appreciate speedcubing and compete in speedcubing, so this report is mostly written from this perspective.

Business goals: The main goal of this project is to analyze the data from various speedcubing competitions and outline information and findings that might be interesting or helpful to people that are actively competing in speedcubing or just appreciate the activity. For example, a question I hope to answer could be in the style of 'Does the number of competition participations affect a competitors average times and if so, how exactly?'.

Business success criteria: Since the goals are currently somewhat subjective, mainly because there's no certainty about what kind of trends will be discovered from the competition data, then I can only describe what I will consider as a success for this project. The most important success criteria for me would be to discover some new trends/statistics from the data that I (and so hopefully also others) haven't seen before and that they would offer deeper insight into cubing competitions. Secondly, I would consider this project successful if in the end, it forms a logical and comprehensive whole, which besides describing the findings, offers supporting visuals/plots and explanations.

Inventory of resources: The single most important resource for this project is the cubing competition database provided by WCA that has data about different cubing competitions. Besides that, the main software I plan on using for this project will be Python/Jupyter notebooks and also MySQL for working with the database, which should be more than enough to achieve the results. It's also worth mentioning that I have personally been involved with speedcubing for quite some years, so I have knowledge about the field.

Requirements, assumptions, and constraints: There aren't really any constraints besides the fact that the project's deadline is around the middle of December (posters 12th of December). There are also no constraints when it comes to the database used, since it's publicly provided by WCA and they also state that it can be used freely for statistical analysis.

Risks and contingencies: Since I'm working on this project alone, there aren't really any problems that I couldn't overcome. There's a possibility that I'm not able to find any interesting trends like I am hoping to, but it's unlikely and in that case I have to make do with what I am able to put together.

Terminology: Some speedcubing terminology that might come up in the project:
speedcuber - term for people that engage with speedcubing
WCA - World Cube Association, the governing body of cubing competitions
event - the specific puzzle which is being solved, most well known is the regular 3x3x3 Rubik's cube, but in total there are 17 events recognized by the WCA
AO(N) - average of N, where N is some number, most common being AO5 (average of 5). Standard way of measuring the solving time of a speedcuber: they do N solves, the best and the worst times are left out, and the mean is taken from the remaining times, which is the competitors final result. In competitions, AO5 is most commonly used.
scramble - sequence of puzzle moves in order to scramble the puzzle. These are randomly generated in order to ensure fair competition and to have the ability to reconstruct the scrambled cube if needed

Costs and benefits: Not really relevant for this project, since there are no extra costs nor are there any monetary benefits.

Data-mining goals: Mostly already described under business goals, namely that I aim to report on the trends that I discover from the competition database, and create supporting plots/visuals that assist describing the findings. As a side-goal, I'd like to create a simple model that would attempt to predict the performance of some competitor in future competitions.

Data-mining success criteria: I would consider data-mining successful if I was able to find interesting trends in the data that I'm working with and visualize a minimum of 10 different interesting findings (rough estimate). When it comes to the model then I would be happy if it could accurately predict the general time range that the person would achieve in future tournaments.

Data understanding:

Outline data requirements: It's clear that in order to gain insight into cubing competitions, I need a lot of data about them. There should mainly be data about different competitions that have been held around the world, and also information about the participants and their results. It's also beneficial for the data to span a long time, since this would create an opportunity to also analyze the evolution of cubing competitions. It would be ideal if the data format would be easy to import into Python, since that's what I'm planning on using for creating the necessary plots and visuals.

Verify data availability: Like mentioned earlier, there's an amazing resource available in the form of the WCA database. This covers all the requirements mentioned in the previous point, since it encompasses all of the official WCA competition data dating back to 2003. So the required data does indeed exist.

Define selection criteria: The only data source I'm going to be using is the WCA database as mentioned earlier, which I will be importing into a local MySQL database that allows me to easily explore the data. For this project, the most useful tables from the database are persons, competitions and results. This is because these are the topics that my project is mainly based on. The database has 13 tables in total and even though the others aren't as relevant as the three mentioned, I still might find use for them, just right now I will focus on the three mentioned.

There were no problems with importing the database contents into a local database.

Describing data: In the previous point I mentioned the most useful tables of the database, so here I will describe the data in them more in depth. It's also worth mentioning that there is a short description for the database's tables on the page linked earlier.

The persons table has information about the competitors that have participated in WCA competitions. There's information about a person's name, country and gender, which all could be useful, if we want to do some comparison between competitors. There is sadly no direct birth-date/age field for a person, but there is the WCA id which is assigned to every competitor and it contains the year when a person was registered to the WCA database, so if there's a need to compare 'age', then using that would be one possibility. In total there are 174 660 different entries in the persons table.

The competitions table has information about the official WCA competitions that have been held. There's information about the location of the competition (country, city, venue) and the time it took place. Also, additional information like what events were available at the competition and who were the organizers. There's not really anything missing from this table. In total there are 8681 entries for competitions.

The results table is probably the most important, since it has information about all the times achieved at competitions by competitors. Each entry represents a single competitor's performance in a single event round at some competition, so usually it has information about the AO5 the person achieved (separate solve times are also all listed as table attributes). There's also additional information about what was the competitor's position in that round and whether the result was a new world/national record at that time. There are 3 108 759 entries in total.

In conclusion, the most important data required for this analysis is present in the database.

Exploring the data: Having done some initial data exploration, I haven't found any major issues with data quality, and the data seems to be quite uniform from start to finish. Under this I mean that data from, say, 2004 and 2022 is formatted similarly and is comparable. Using some simple sql queries I found out that speedcubing is clearly a growing and evolving activity, for example, in 2004 there were only 12 competitions held, but in 2019 that number was 1331, but there is a considerable drop in activity during the ‘covid years’ of 2020 and 2021. Similarly, it is easy to check that the average times of solves have gotten better as time goes on. It’s also worth mentioning that I found out that the times in the database are mainly represented with centiseconds. I won’t go more in depth here and save it for the project, but it is clear that the data has good quality and many interesting aspects to analyze.

Verifying data quality: Based on the previous points, I can say that the data has the necessary quality and information for this project. The database has been quite professionally maintained and because of that the data is uniform for the most part. There is some ‘leftover’ data from the earlier years of the database (like some countries that don’t exist anymore and some event formats that aren’t used anymore), but that data shouldn’t interfere with the information that I plan on extracting and working with.

Project plan:

Tasks:

1. Exploring the data in order to find interesting trends: For this I’m planning on using the MySQL copy of the WCA competition database. I intend on making queries about potential topics that I want to study and make some rough plots of the data to see if there’s anything interesting to report.
2. Compiling useful datasets: If I find some topic/data that I want to analyze more in depth and work with it using Python then it is probably better to export it as a separate CSV file. The database IDE I’m using (DataGrip) offers an easy way to export queries as formatted CSV files, but maybe I need to do some manual adjustments.
3. In-depth analysis of selected data and creating plots: After finding the valuable information, I intend on working with it using Python (/Jupyter notebooks) to create the necessary plots and do some statistical analysis if required.
4. Combining my findings and adding commentary to it: Besides just statistics and plots I want to comment on my findings and talk about the potential implications. This is for tying the whole project together.
5. Training a simple model to predict competition performance: Like mentioned earlier, I want to try using the skills learned during this course (so using Python and algorithm covered) to train a simple machine learning model, but this is not the main goal of the project.

Since I’m working on this project alone and the time spent should be around 60 hours, then the time split should roughly be around 15-5-15-15-10 hours (in the task order). This is, of course, a rough estimate and is subject to change. Also the task order definitely won’t be as linear as I described it.