DISC GOLF POPULARITY IN ESTONIA ANALYSIS

Malin Püttsepp
Karmo Saviauk

Repository: https://github.com/Karmosav/discgolf-data.git

Task2

Disc golf is a rising outdoor sport where players throw disc at a target, and is played using rules very similar to regular golf. Modern disc golf started in the early 1960s. The game is played in about 40 countries and Estonia is one of them. In Estonia there are 202 disc golf courses and has the highest density of disc golf courses per km2 of dry land. From that can be concluded that disc golf is pretty popular in Estonia. And can not forget our own Estonian disc golf double world champion - Kristin Tattar.

Our first goal is to analyze the growth of player base in Estonia. Second goal is to analyze the competition/training frequency by season. Final goal is to analyze the correlation of player's ranking and the number of training courses and competitions played.

The outcome will be measured and judged by the Data Science course supervisors.

CEO of Disc Golf Metrix is a relative of one of us. He gave us all the data we need and is a very good expert in disc golf. Our resources include three datasets from Disc Golf Metrix. The First one contains information about players, the second and the third contain information about user training and competitions where the user has competed. We have two laptops and for software we use a jupyter notebook and VSCode.

For security reasons we are not able to see people's personal information like name, e-mail, address and personal identification number. Only the submission date is given by the Data Science course by that time we have it ready. The final product will also not show people's personal information and will not be oddly specific.

All data we have is in Git and also in our personal computers and automatic save is enabled. All technical errors that may occur are brought to minimum. We can continue our work with small power outages and internet outage is no harm for us.

Used terms:
Rating/Ranking - Indicator how well the player is playing (the higher the ranking the better player is)
PDGA (rating)- Professional Disc Golf Association, also rating given to disc golf player by PDGA.

Our project does not have any costs or gain us any money.
Data we got was mostly cleaned so we did not have to process the datasets so thoroughly.
Data-mining goals are mainly graphs that show us relations and correlations.Also we will calculate correlation coefficient to see how strong the correlation actually is. We will put them together in one presentation where all gathered information is together.
To see how correlated are attributed are we use Spearman and Pearson correlation.

Task3
In this Analysis we need at least three datasets to address the data mining goals. Like mentioned above the project involves three datasets: user information, user competitions, and user training in disc golf competitions. Each dataset serves a specific purpose in understanding user engagement and performance in disc golf.
All three datasets have been obtained from Disc Golf Metrix who were kind enough to give us access to these datasets. But we do not get access to confidential info such as names, email addresses and so on, but that is expected and we had taken that into account. We even got to have a look at all the datasets that they had, from what we chose three. The first dataset was about user information, where there was a lot of information about user's info, but to us most of that was non-important. The attributes that mattered to us were when the account was created, what ranking they have now, and personal ID to match them with competitions. The second datasheet that interested us was the competition to user dataset, where one of the columns is user ID and the other is competition ID. The fact that interests us is the number of competitions from a player. But this datasheet is so large it is almost impossible to download it to csv, so we got grouped version on this datasheet (player ID and number of competitions). The third one is basically the same as the previous datasheet but instead of competitions they had training courses. And same thing with this datasheet we had to get grouped version of this.
But what do we have? The first dataset (user.csv) contains information about every account ever created on Metrix. It has over 185 000 rows and 13 columns. From these 13 columns

some of them are important to our data-mining goal and others not so much. The useful columns are ID - to join tables with competition and training datasets, Metrix ranking, user's birth year could give us interesting facts about the popularity of this sport, CreatedOn shows what time the account was created on, from this we could see the growth of player base and CountryCode which we use to filter out Estonian accounts. We also have not-so important features and we have to think about the weather to analyze them. First of them is Area and City, which don't seem that important to us. Then there is PDGA code, attribute for PDGAMembershipStatus, LastFinishedRoundOn is attribute to when did user play his/her last round and DeletedOn shows when was the account deleted (mainly NULL). The second dataset hold's two attributes: first is user ID and the second is the number of competitions they have taken place in. The third one is similar but there is a player's ID and the number of training courses taken.

After the extraction of the player's who weren't Estonian we were left with about 30 000 rows. From now on we only look at the Estonian player base. Generally the data quality is good. We had a first look at data when we plotted every attribute distribution. Everything was as expected. Male to female ratio is about 4 to 1. From the birth year distribution I found a few incorrect values such as 0 or 3000, but overall it seemed to be a normal distribution. Ranking attribute was also as expected, overall normal distribution but a lot of people have a ranking of 0. The values vary from 0 to about 1000. Other attributes are not worth mentioning. Overall the quality is as good as expected, some anomalies occurred and there were some NaN-s but no severe data quality issues. In conclusion I would say data quality is good enough to support our plan.

Task4

Tasks:

- Get to know data - we will do it together, look through the csv files we have manually just to see what we are working with, together - 0.5 hours
- Extract Estonian players from others, Karmo - 0.5 hours
- Look over the data - Plot all attributes to see anomalies, clean if needed. Using matplotlib and visually looking for anomalies or incorrect data. Karmo - 2 hour
- Design poster - Design base for poster we are going to present. Using canvas and some google extensions to design some elements on poster, Malin - 3 hours

- Analyze Estonian player base growth with plot - see if there are any trends on graph. First we have convert time to correct type or else we can not plot the timeline. Then we use “groupby” to to get new players per month, and then plotting , Malin - 2 hours
- Analyze correlation between number of training courses and ranking - calculate the correlation with Pearson and Spearman’s correlation. We need to join tables to get the player’s number of training courses and ranking. Separate the most extreme values to get the plot look more detailed and finally plot. Karmo - 1.5 hours
- Analyze correlation between number of competition participated and ranking - calculate the correlation with Pearson and Spearman’s correlation. We need to join tables to get the player’s number of competition participated and ranking. Separate the most extreme values to get the plot look more detailed and finally plot. Karmo - 1.5 hours
- Finish poster - put all the gathered information on the poster and make it look attractive for anyone who sees it. We are editing our poster in canva. together - 4 hours