#DataMadness Project Proposal Form

i-number	**Student name**
i6214757	Kristian van Kuijk
i6214168	Pavan Aakash Rangaraj
i6211839	Boris Borisov

(Provisional) Title:

Can change during the project, but make sure that you pick a title that attracts the audience

What does it take to become a legend in Tennis?

Questions and Objectives:

List at least 3 questions (high-level) of what you want to address in your analysis. More are welcome, but think of the workload

1. Does age enable you to handle pressure during break-points?

2. Does height provide an advantage during service?

3. Which is the greatest tennis nation of all time?

The Dataset

Attach the dataset or point to a web link that is available (if very big or already existing). Highlight why it can answer your questions.If the dataset does not come with datasheets, you are going to create them to best of your knowledge (see extra form)

Free ATP and WTA Results and Stats Databases – Heavy Topspin (tennisabstract.com)

GitHub:https://github.com/JeffSackmann/tennis_atp

This dataset contains all match details from 1968 -2022. It provides all relevant details about players and games such as age, height, service, nationality and key match moments such break-points and aces.

Rough Outline of the Analysis

This is a rough sketch of the analysis you want to do on the dataset and which techniques you should use to answer your questions. Not detailed at the proposal, but should act as a guide as you work on the project.

For all research questions, we start by filtering the dataset to only keep the season 2020 - 2021 - 2022 for faster analysis while providing interesting insights (since the dataset will still be large enough.

1) We look at the percentage of break points won by players of all ages (histogram with a x-axis of age, y-axis being the break points win rate between 0 and 100). Our assumption is the older a player is, more experience he has and will better manage to handle pressure. We can also add a comparison by taking into account the surface (which we expect to be always more or less equivalent between each surface).

2) To answer our 2nd research question where we assume that taller players have an advantage during service, we will compare the heights of the 2 opponents and the ratio of aces scored by each opponent. We will feature engineer a new variable *ace_ratio* which will be the ratio between the number of aces and number of legal services made by the player (as simply showing the number of aces doesn't provide any insights). We will also consider the number of double faults and 1st point won by a player. By visualizing these statistics of the taller opponent and on how many occasions the taller opponent went on to win the game we will determine whether height provides a crucial advantage during service in a tennis match.

3) An analysis on the dataset will be done. The top 20 countries will be filtered based on the winrate of each country (we do not look at the player anymore but its nationality). A dictionary with the winner and his/her country will be extracted. The dataset will be filtered between the years of season 2020 - 2021 - 2022.
Visualization techniques will be used to present the given information in order to answer the research question.