Originally wanted to calculate the “stamina” of each player, but that turned out to be too subjective. How do we define stamina?

So we decided to try and develop “team play”, correlating passes amongst team members with points made by the team.

First we needed to get the number of points scored by each team. A problem we ran into is that the game event files (simple\_markings) do not take shots from fouls into account. We decided this was ok, because foul shots aren’t in the actual gameplay. These are the only points that aren’t accounted for.

So we can’t get which team won because of this.

Some game infos in frames are not in simple\_markings, from 2017-06-07 onwards. This is just incomplete data, so we couldn’t use the extra frames information.

Some game infos in frames say that player ids are on both teams. This is just buggy input data.

We spent most of the first day trying to figure out how to scrape data out of the files given to us and into a panda data structure. We used dictionaries for the key-value pairs between events and counts of that event for each player.

{Player id: {Team : home/away, 2PM : count, 2PX : count, PASS : count, etc…} }

We practiced this using only one file on Google Colab, for easy collaboration amongst multiple computers.

Upon accomplishing this, we needed to find a way to run the program on multiple files. Colab allows us to upload files to the cloud and access them from there. However, upload speeds are awful, and space is limited. So to make this more efficient, we whittled down the frame files to just what we needed from them: a dictionary of all the player ids from the game. This significantly reduced the size of the files needed to be uploaded to the cloud.