The US Open tennis tournament is fast approaching. Starting August 28th, and continuing for 2 weeks, the top tennis players in the world will come to New York City to compete on the hard courts for the final Grand Slam tournament of the year. The US Open is the highest paid tennis event in the world, with total prizes totaling $50,400,000. It is also one of the only tournaments to evenly pay both the men and women champions.

The 4 Grand Slam tournaments – Australian Open (hard), French Open (clay), Wimbledon (grass), and US Open (hard) - are the most competitive tennis events of the year and are treated similar to the four weather seasons. In fact there are many ‘feeder’ tournaments that lead up to a Grand Slam event, which give players experience on a particular court surface before the main event.

Because of the higher stakes, match predictability in Grand Slams are higher than other tournaments. I will predict the outcome of each match, and ultimately the winner of this year’s US Open. The data model is intended for recreation, and not professional gambling situations. I will personally use this model for bragging rights to family and friends.

To build the model I used three data sets:

1. Kaggle Datasets has an ATP match history, which includes fields like: tournament, date, court surface, bracket\_round, winner, loser, score, and rankings
2. World Rankings – for the predictor I would need to know current world rankings of each player, not just the historical. I scrapped this information off of the ATPworldtour website
3. Bracket – the initial bracket pairings. For training the model I used the 2016 bracket.

The Kaggle dataset surprisingly took the most amount of time to clean. I had to find the correct decoding type, reformat the player names, and change headers. I learned a good lesson on not using generic titles for headers. After I had loaded all the data and began exploring with basic statistics, I found a problem with the field ‘Round’ that wouldn’t allow me to use it. I read through the code for a couple days looking for spelling and syntax errors before a conversation with my mentor gave me the answer. It turns out ‘Round’ is a reserved name in Python and the compiler would error on calculations with this field.