**Project Progress Report:**

I have completed the pre-processing of the data and run a few tests on it. It was quite hard to get everything into one data frame because there are many features to combine and finding out which ones are most important proved challenging. I started by combining all the data I found and using a correlation matrix to see which features had the strongest correlation – this was interesting to see because I saw correlations or the lack there of that I was not expecting. For example, I was expecting that the winner and their rank would be correlated pretty highly but in fact this was not the case – the correlation between the winner and the number of points they had was only 40%.

However, the correlation matrix helped me eliminate some unnecessary features – the date of the tournament and location didn’t have much effect. I thought this was quite odd, because the tennis season is the whole year so I thought there was bound to be a dip in results of the higher ranked players but that did not seem to be the case.

Another challenge I had to deal with was combining the players name with their rank – the database I found with a large amount of data and features did not include the players’ names, just their rank so in order to work towards my goal of the project I needed to have a database with the players names. What I did was find a database from the same date as these results that just had the names of the players and their ranks, I then went through the “loser\_rank” and “winner\_rank” columns in my chosen database and if this rank matched a rank in the players database, I assigned the name to it.

My next step now is to narrow down these features and create meaningful visualizations of how certain features are affected by one another – this will give the reader a better insight into the game of tennis if they have a limited understanding of the sport – it will also give me the ability to narrow which features I should feed into my prediction algorithms. Then the next step is to predict how likely a player will be successful (in this case top 100) if they have certain attributes.

I will be predicting a percentage likeliness of being top 100 so this brings me to regression and looking at an Ensemble or SVR for my prediction algorithms. I still have a ways to go but I feel confident about where my data stands so the hardest part of actually gathering and assembling the database is done.