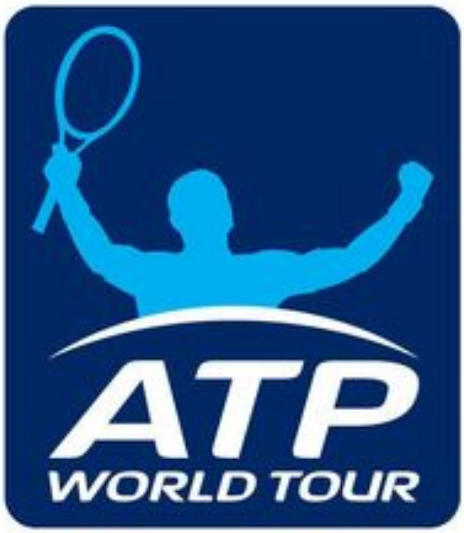
Head to Head:
ATP Matchups
From 1991 to 2016



Matthew Lin
904281426
Dis 1C

Introduction

Objective

Men's tennis has evolved tremendously in the past 25 years. The players today move better, hit harder and with more spin than players of the past, and have access to modern rackets, strings, and training. Players nowadays barely have any technical weaknesses to exploit. They have entire teams dedicated to every aspect of tennis: the fitness trainer who helps the player build his physical stamina and prevent injuries, the sports psychologist who strengthens the player's mental toughness and fortitude, and the technical coach who ensures that the player's form and technique remains razor sharp. But are they really better than the champions of previous eras?

I have been an avid tennis player and fan since the age of 9, and I grew up in the era when Roger Federer first started to dominate. Since then, I have seen Rafael Nadal, Novak Djokovic, and Andy Murray climb the ranks to world number one. Having frequented tennis fan forums on the Internet, I have seen a lot of hypothetical questions getting thrown around. "Is Roger Federer better than Pete Sampras at Wimbledon?" "How would Novak Djokovic stack up against Andre Agassi at the Australian Open?" "Could any player have stood a chance against Rafael Nadal in his prime at Roland Garros?" These are all fabulous questions that fans could only speculate about, but what if we can use mathematical modeling and statistics to provide tangible evidence to our fantasy matchups?

I sought to answer these questions using datasets found on the ATP World Tour website. Since 1991, the Association of Tennis Professionals (ATP) has been keeping track of statistics on various metrics such as the serve, return, and performance under pressure. Since there was so much data to be analyzed from this dataset alone, I decided to limit my search to players who have been ranked number one in the world since the start of 1991. The players I will be analyzing are listed below:

1. Boris Becker
2. Stefan Edberg
3. Jim Courier
4. Pete Sampras
5. Andre Agassi
6. Thomas Muster
7. Marcelo Rios
8. Carlos Moya
9. Yevgeny Kafelnikov
10. Patrick Rafter

11. Marat Safin
12. Gustavo Kuerten
13. Lleyton Hewitt
14. Juan Carlos Ferrero
15. Andy Roddick
16. Roger Federer
17. Rafael Nadal
18. Novak Djokovic
19. Andy Murray

List 1. World number ones since 1991

I will be analyzing the datasets to provide more comprehensive evidence to predict head to head matchups between players from different eras.

About the Dataset

All of the data that I have compiled is publicly available on the ATP website (http://www.atpworldtour.com/en/stats). Unfortunately, they did not have an easy method of parsing through all this data, so I had to input each player and their corresponding serve, return, and under pressure statistics manually in Excel. This was tedious but it actually gave me pretty good insight on statistics that I did not know beforehand such as the average aces and double faults each player served per match.

Specifically, the categories for serve, return, and under pressure statistics are defined as follows:

- Serve
    - % 1st serve
    - % 1st serve points won
    - % 2nd serve points won
    - % service games won
- Return
    - % 1st serve return points won
    - % 2nd serve return points
    - % return games won
    - % break points converted
- Under Pressure
    - % break points converted
    - % break points saved
    - % tie breaks won
    - % deciding sets won

All of these statistics combined to form a bigger picture of how players perform against everyone else in the field. The serve has actually had two more categories (average aces per match and average double faults per match), but truthfully, both categories are a subset of % 1st serve points of won and % 2nd points won, so essentially, each category has 4 subcategories to base my analysis off of.

In addition to these statistics, I also compiled mthe actual head to heads of all the players in my dataset. However, it was difficult to perform direct analysis on these head to head matchups using Matlab because of several factors. First, it was not reasonable to base a head to head matchup off of winning percentage because a 1-0 winning head to head is treated the same as 10-0 winning head to head (both are 100% winning records). Additionally, a losing head to head without any wins is treated the same way as a sample size of two players who have never played before i.e. 0-1 losing head to head is the same as 0-0 head to head (0% winning records). Lastly, it was difficult to work with ratios or fractions in Excel because it would keep simplify fractions such as 7/19 (meaning 7 wins in 19 matches) and round up to 3/8, so this data would just be an approximation.

Because of these reasons, I included the original CS 170A Course Project.xlsx with the actual head to heads and used decimal values of the head to heads in my actual .csv and .mat files in order to perform analysis techniques with the serve, return, and under pressure statistics. An important thing to note is that some Davis Cup matches are omitted on the ATP website when I looked at court surface specific head to heads, so the head to heads of clay, grass, and hard court matches may not add up to the overall head to head. Additionally, -0.2 represents the player in order of the legend, -0.1 means the players have never played each other, 1 means a perfect winning record, and 0 means a perfect losing record.

In the interest of concerning space on my lab report, I have all my code inside a GitHub repo that is public domain. The link to this can be found at https://github.com/Darthpwner/CS-170A-Project. Inside this repo are the corresponding CS 170A Course Project.xlsx, .csv, and .mat files that I used to compile and store the data so that Matlab could access them. The original .csv datasets had the corresponding players on the vertical axis and the associated statistics on the horizontal access, but since Matlab could only interpret the numerical values for .csv files, I had to start of at row 1, column 1 for each .csv file.

Modifications to the Dataset

In order to parse this dataset successfully, I had to do convert all the percentages into decimal because .csv files can only interpret numerical values. This was pretty easy to do in Excel, and I changed this for all the percentages in my dataset.

The other challenge was that on grass, three of the players were missing information on the ATP website: (1) Thomas Muster, (2) Marcelo Rios, and (3) Gustavo Kuerten. The reason for this was that these players were not as proficient on grass and failed to progress to far enough in grass court tournaments, so their sample size was too small to be aggregated together. For their subcategories, I just marked all of them as 0 because .csv files could not use N/A as an option.

Log

Getting the data was the straightforward part, the main challenge was generating heuristics in order to predict the keys to winning a match. I was initially inspired by IBM SlamTracker's Keys to a Match, but I tried doing more research on how they generate their models and it appeared to me that Watson Analytics was extremely cumbersome to use even though it had a plethora of data.

As a result, I had to generate my own heuristics to come up with keys to a match. I came up with these using several mathematical modeling techniques learned in class such as correlation, covariance, and analysis of distributions. I also tried to use another technique called the k-nearest neighbor algorithm to come up with some heuristics to predict matchup outcomes.

My GitHub repo contains all the files used for loading the .csv data, creating different bar, line, and scatter plots, and analyzing the serve, return, and under pressure statistics as well as the head to heads of all the world number ones on clay, grass, and hard courts in addition to all surfaces.

Plots

My report employs a frequent use of plots ranging from bar graph plots, dotted line plots (head to head log), and scatter plots (k-nearest neighbor analyses). For the sake of convenience and organization, these plots are spread out across various files in my MATLAB project, but most of them follow a similar structure of organization, which involves setting up the matrix and then plotting either using bar(x) for bar graphs, plot(x) for dotted line plots, and scatter(x, y) for scatter plots, where x and y are matrices.

In my /plots folder, I have MATLAB code that generates bar graph plots for the serve, return, and under pressure subcategories for all 19 players for the different court surfaces as well as the average on all surfaces. Each plot shows the subcategories on the x-axis (4 on each plot) as well as the percentage represented in decimal form on the y-axis. In addition, each plot has a corresponding legend that shows which links the player to the bar graph representation. There are a total of 12 bar graph plots that I generated shown below.

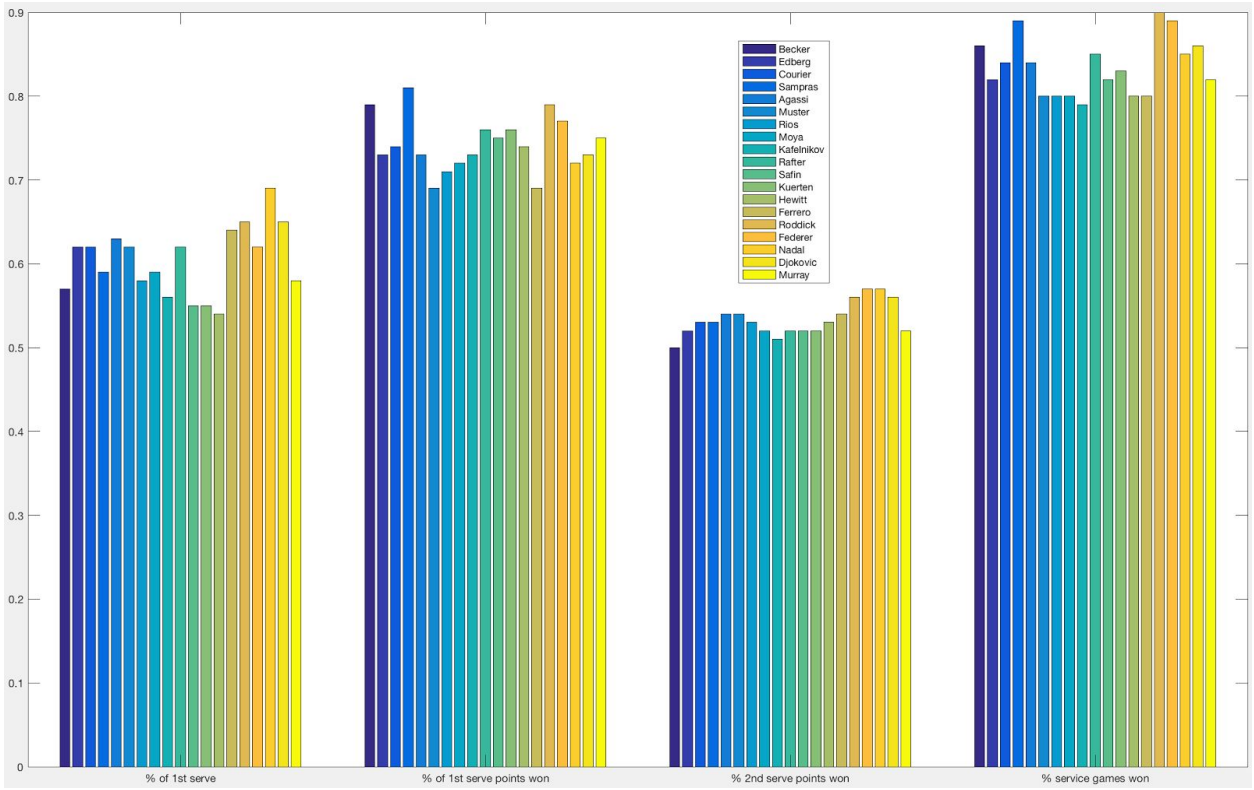


Figure 1. Serve Stats All

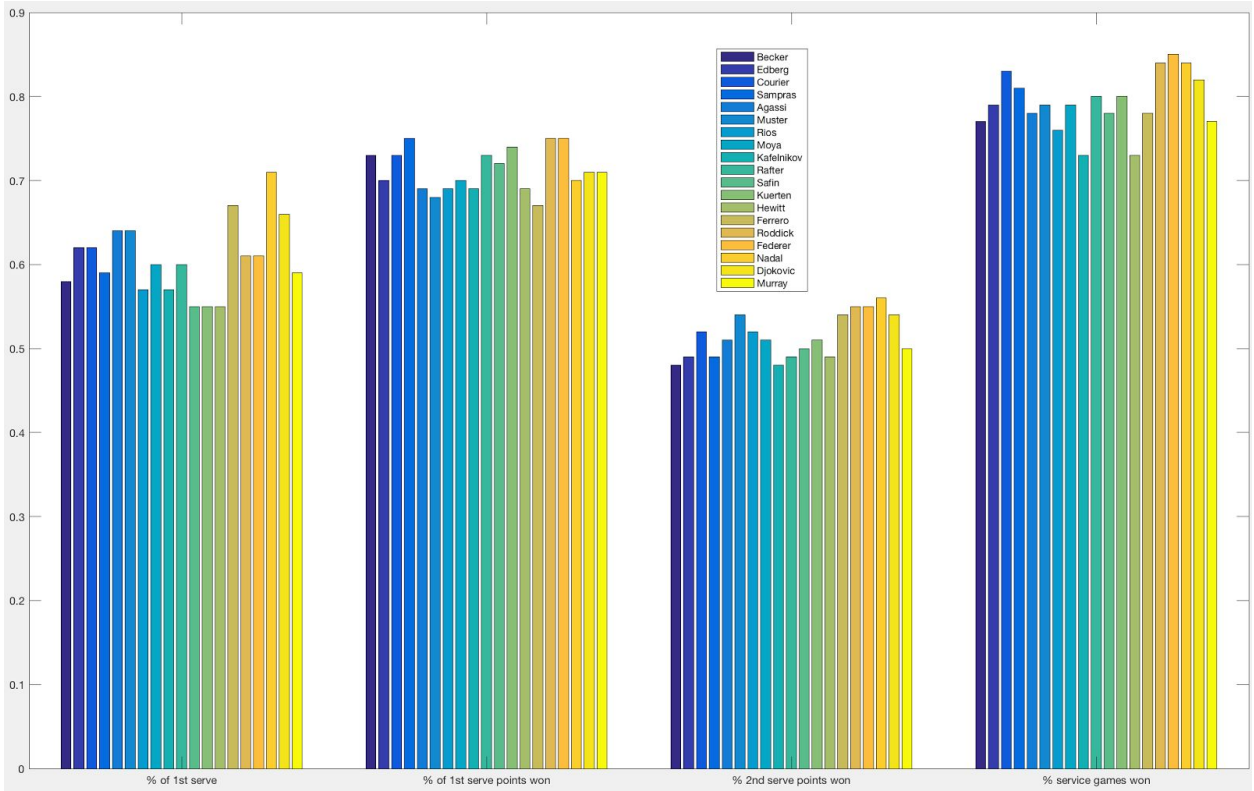


Figure 2. Serve Stats Clay

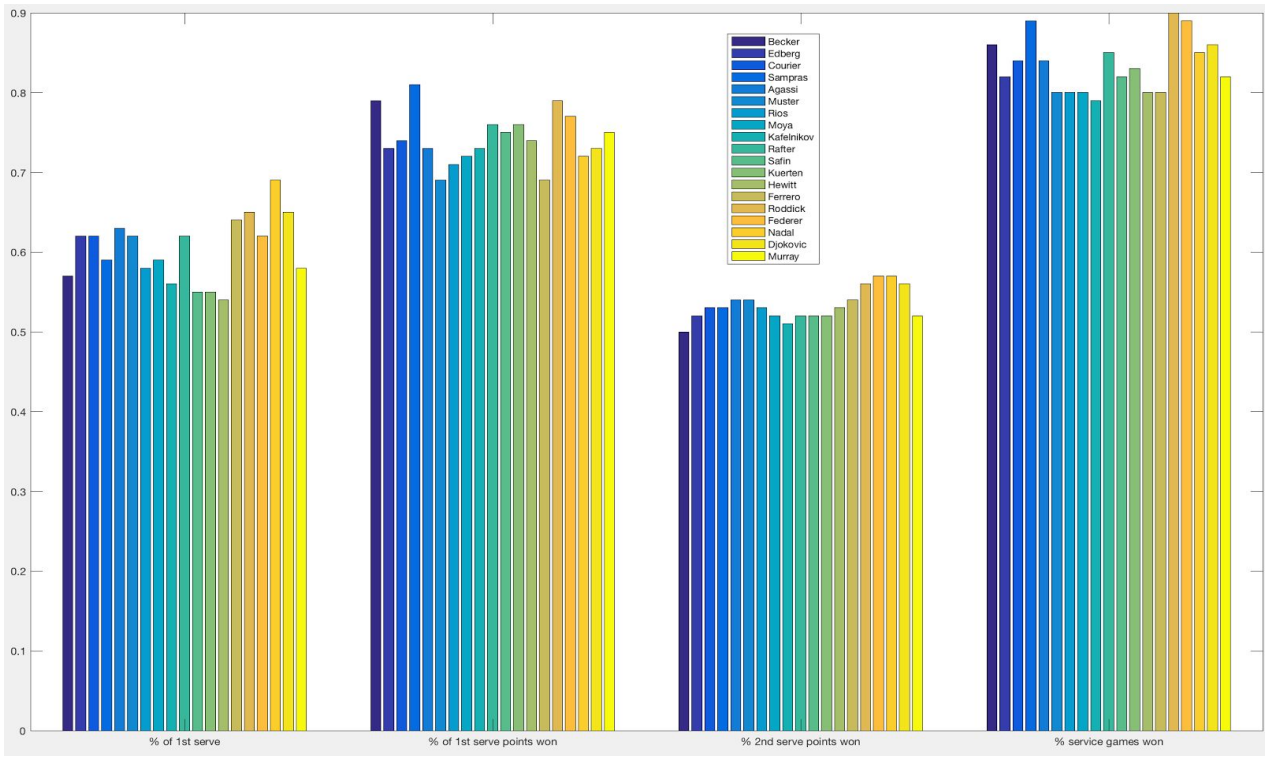


Figure 3. Serve Stats Grass

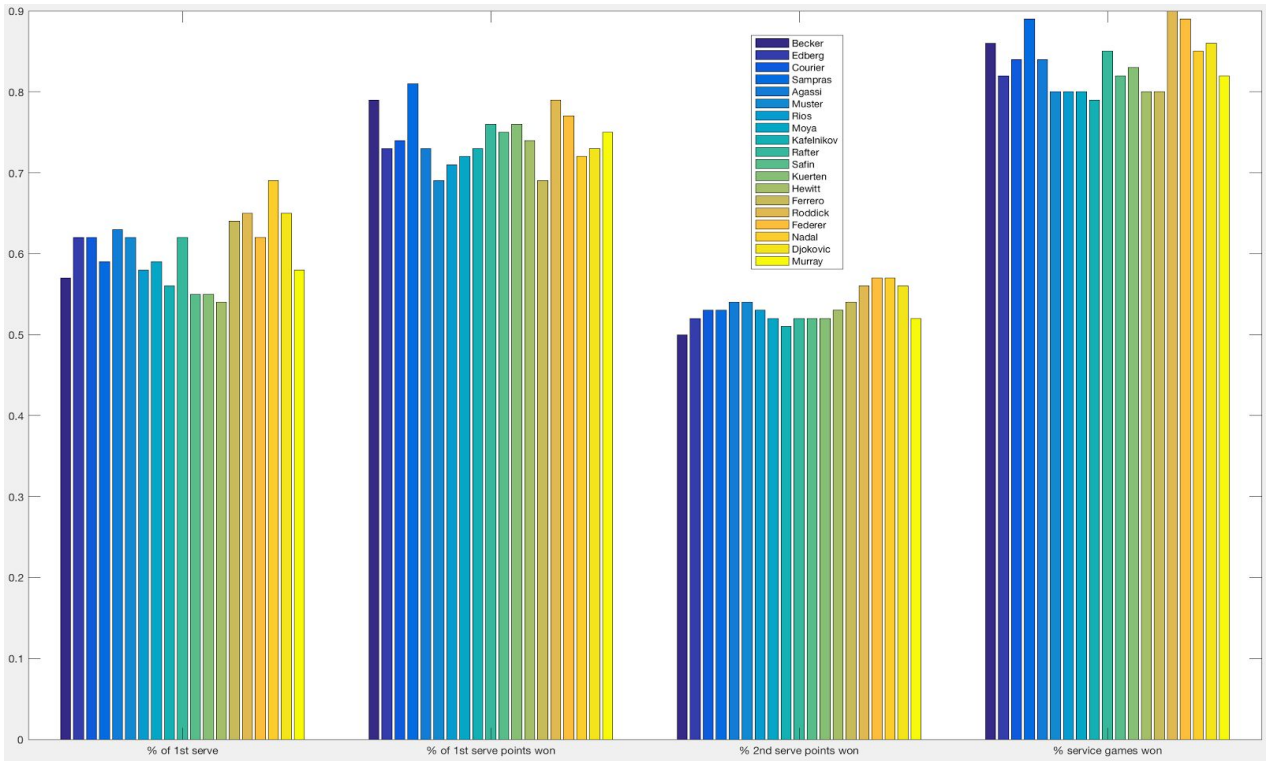


Figure 4. Serve Stats Hard

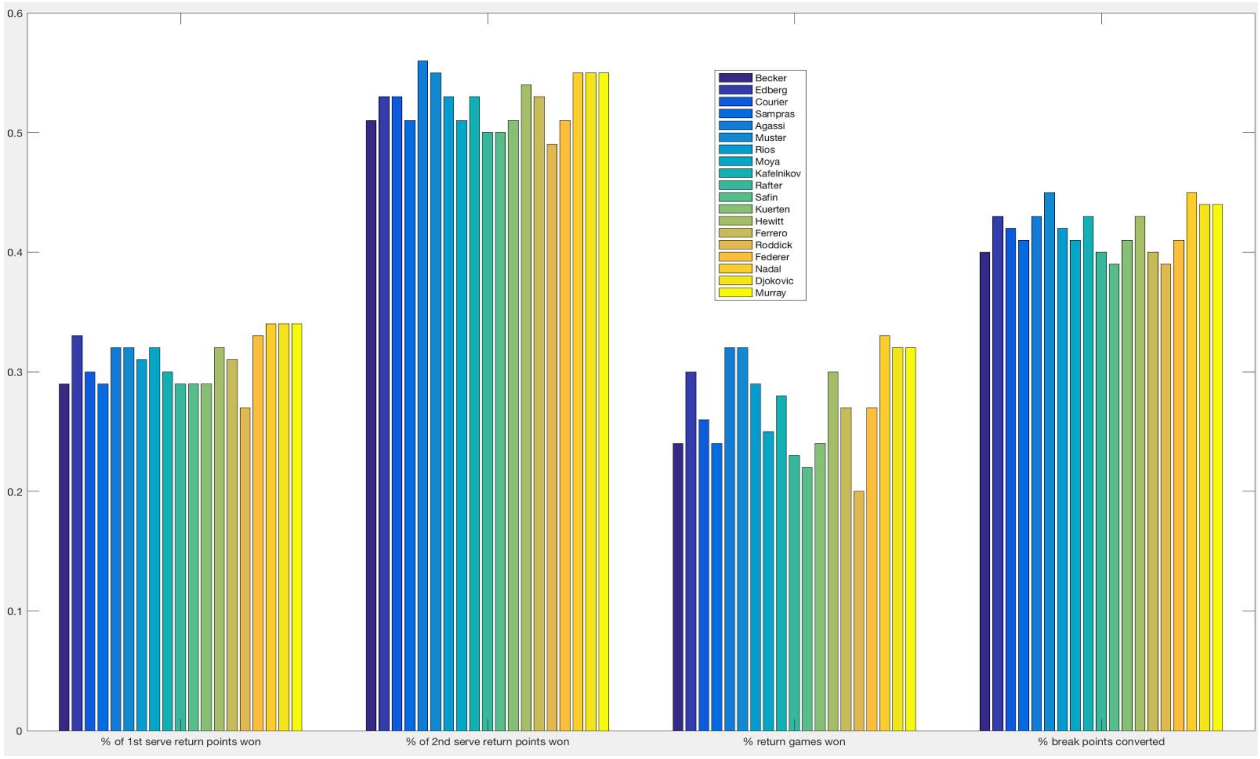


Figure 5. Return Stats All

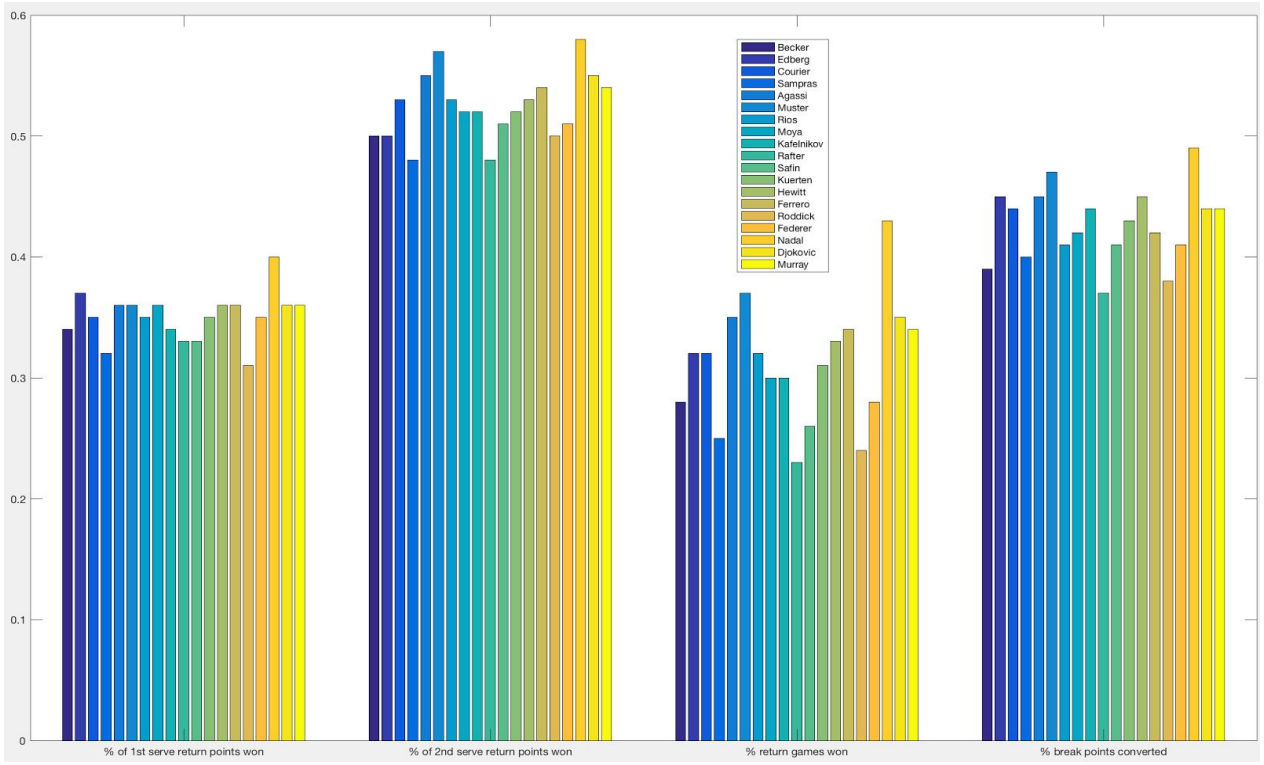


Figure 6. Return Stats Clay

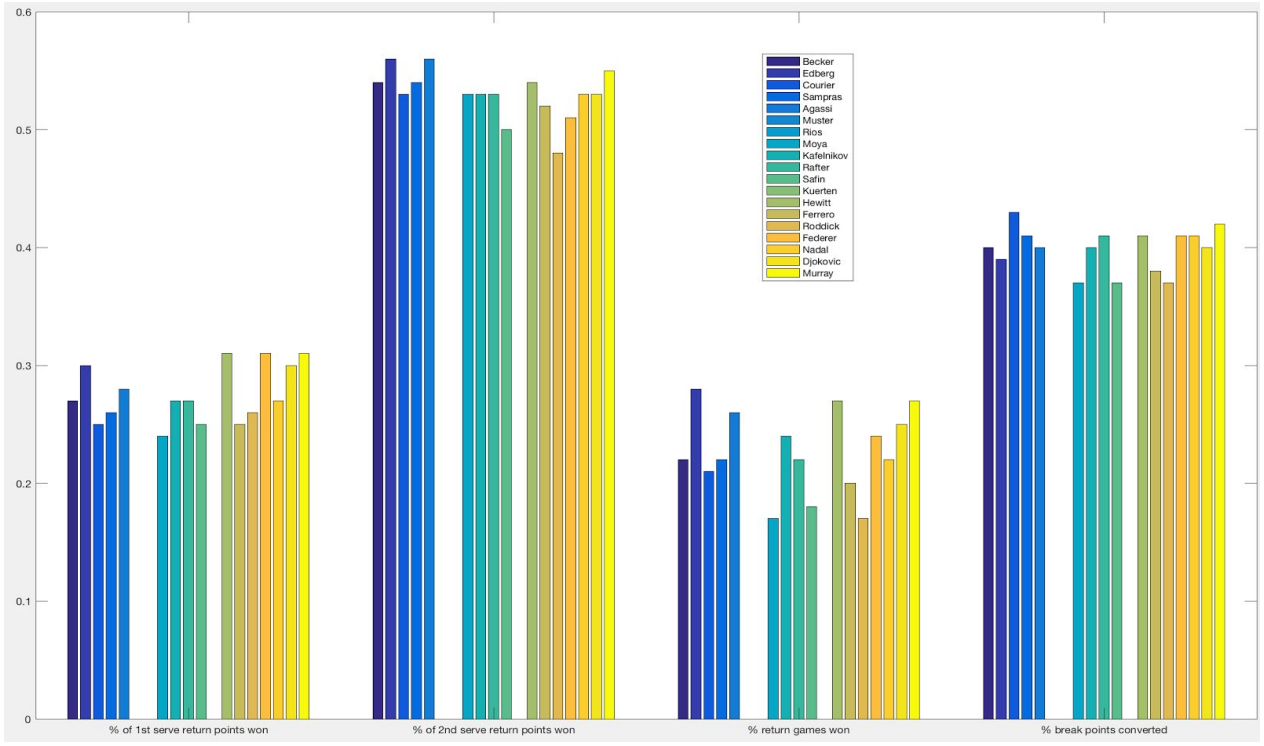


Figure 7. Return Stats Grass

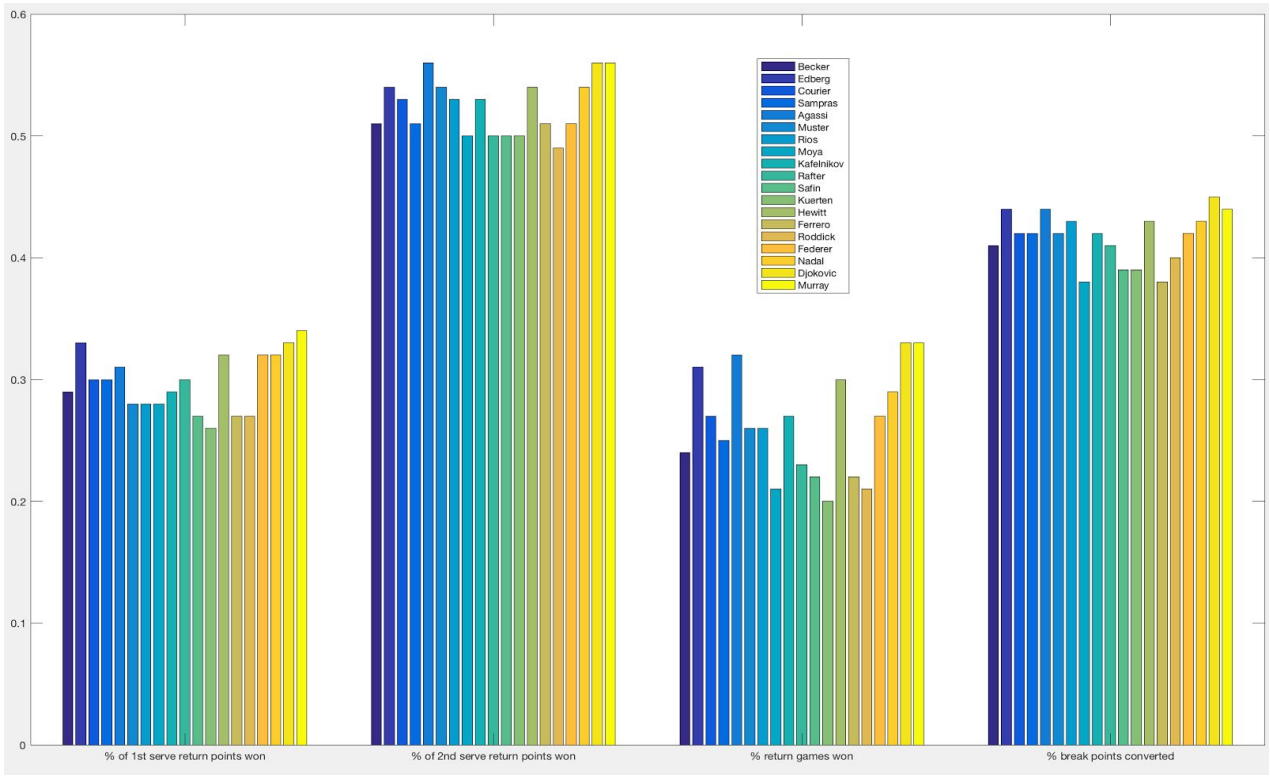


Figure 8. Return Stats Hard

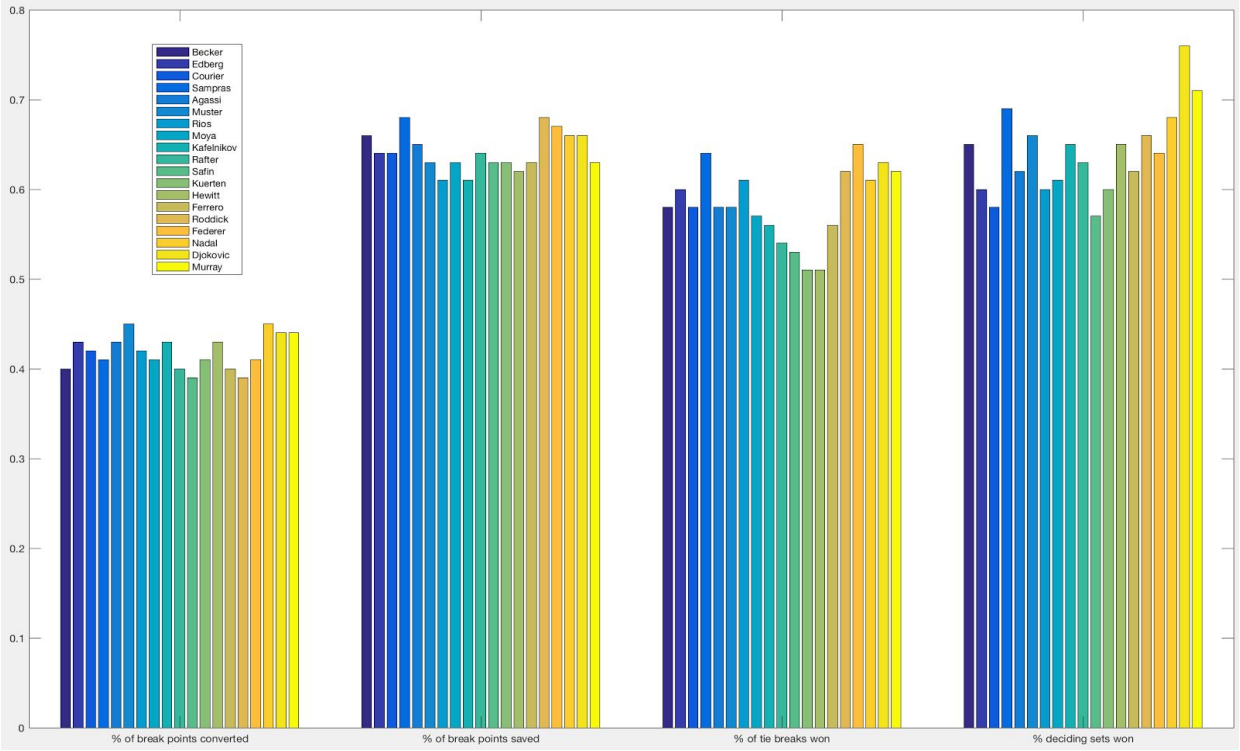


Figure 9. Under Pressure Stats All

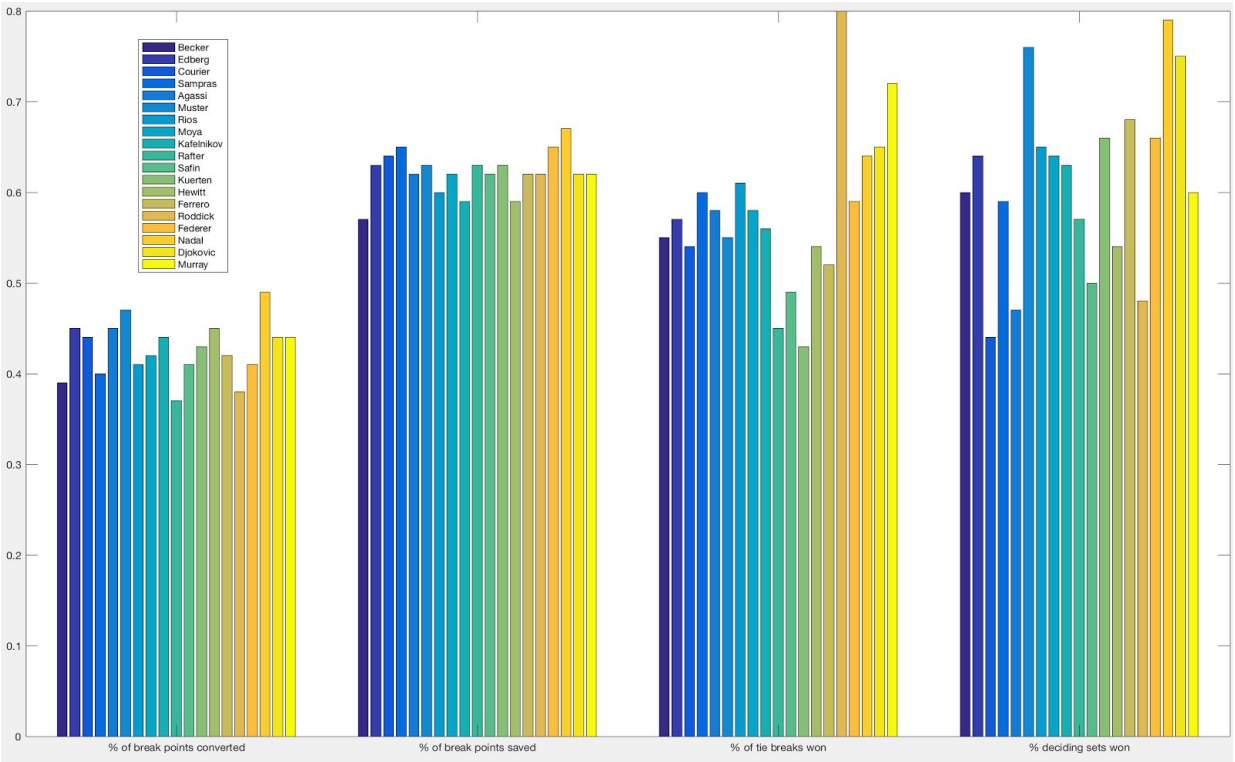


Figure 10. Under Pressure Stats Clay

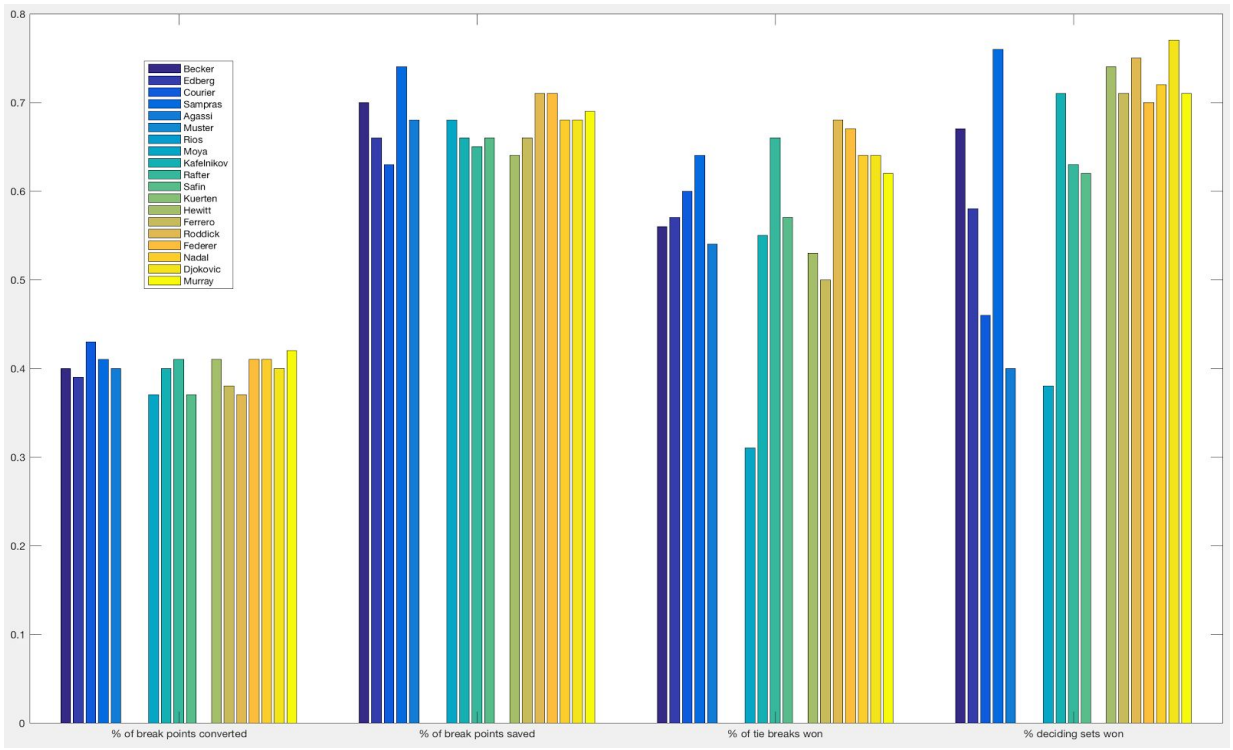


Figure 11. Under Pressure Stats Grass

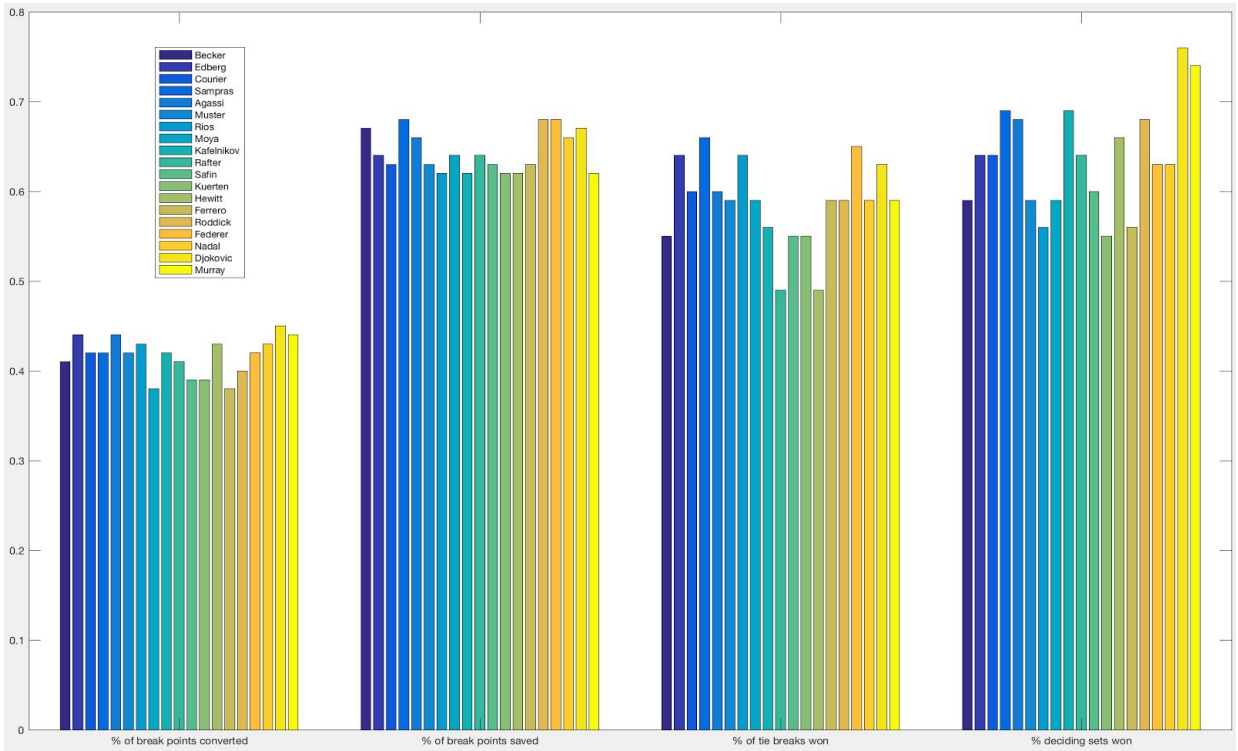


Figure 12. Under Pressure Stats Hard

My /plots folder also contains MATLAB code to produce point plots for head to head matchups for all surfaces as well as clay, grass, and hard courts. As mentioned earlier in the **About the Dataset** section, -0.2 represents the players in the order of the legend, -0.1 means the players have never played each other, 0 means that the player has a perfect losing record against another player, and 1 means that the player has a perfect winning record against the other player. Note that these plots are missing quite a few data points; this is because the plot shows the last player in the legend that corresponds to the percentage value. This is why most of the plots for the earlier players like Becker and Edberg only show Murray on their plots for players who have never faced off, even though other players like Federer, Nadal, and Djokovic have not played Becker and Edberg. These 4 plots are represented below.

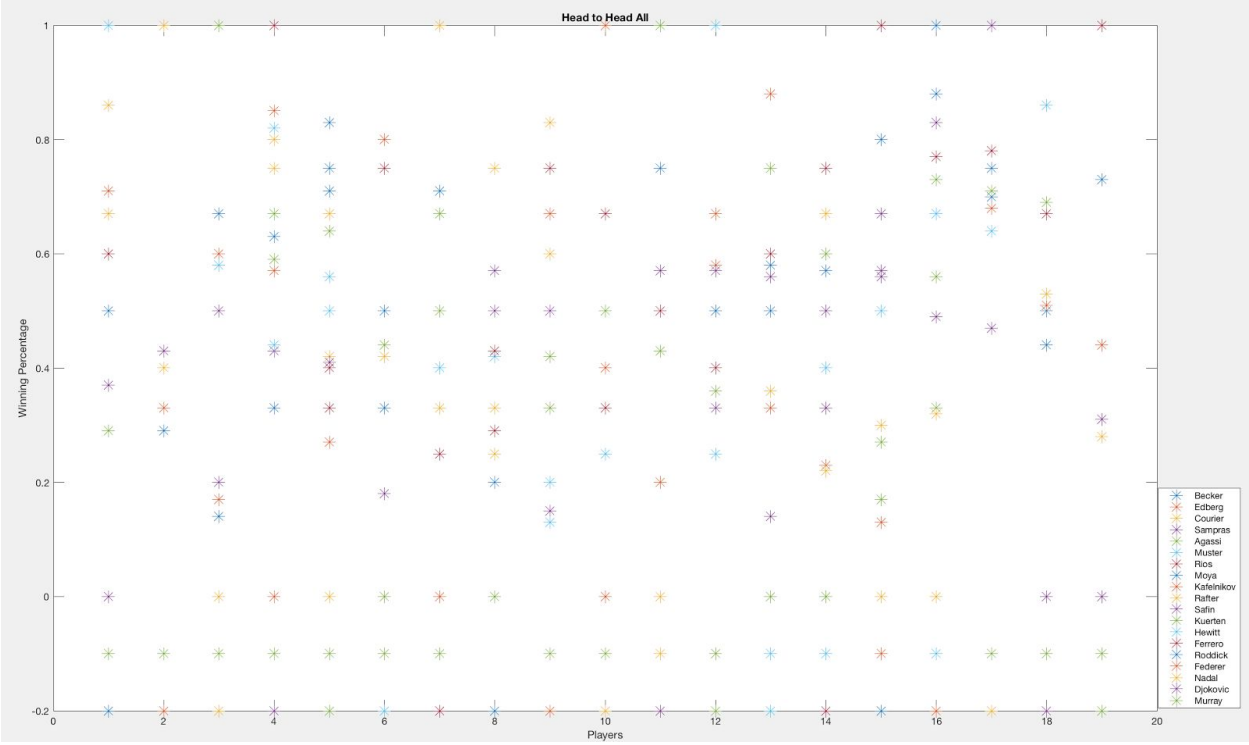


Figure 13. Head to Head All

The script best_head_to_head_all.m creates a 1x19 vector of doubles that shows all the positive head to heads of each player. I analyzed this and saw that Pete Sampras (10), Roger Federer (9), Andre Agassi (8), Rafael Nadal (8), Boris Becker (7), Marat Safin (7), and Lleyton Hewitt (7) had the best head to heads against the rest of the field. For some cases like Sampras, Agassi, Federer, and Nadal, they really were the dominant players of that era. Others like Safin and Hewitt played in a more

transitional era and had particularly good records against world number ones from the past.

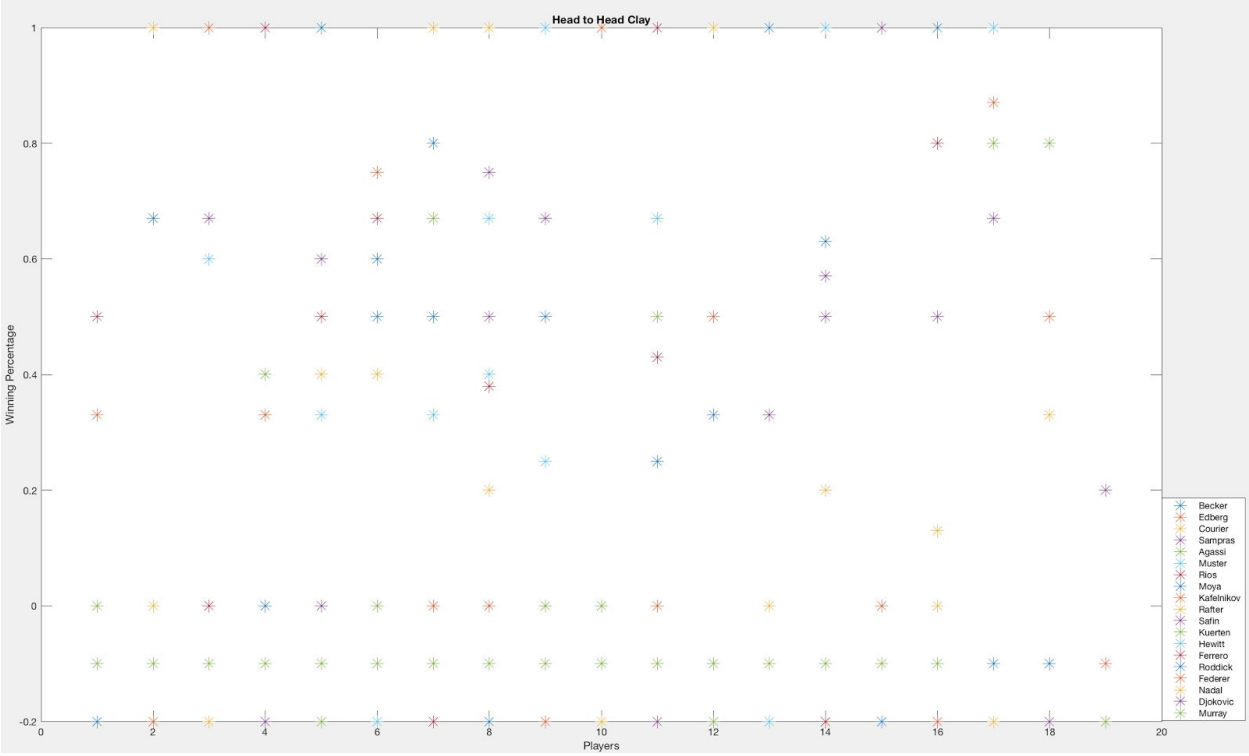


Figure 14. Head to Head Clay

The script best_head_to_head_clay.m creates a 1x19 vector of doubles that shows all the positive head to heads on clay of each player. Roger Federer (6) and Rafael Nadal (6) were at the top of this list. Stefan Edberg (5), Jim Courier (5), and Andre Agassi (5) were all tied for 2nd, and Thomas Muster (4), Carlos Moya (4), Yevgeny Kafelnikov (4), Patrick Rafter (4), Marat Safin (4), and Juan Carlos Ferrero were all tied for 3rd. The interesting thing is that Edberg, Rafter, and Safin had never won the French Open (the clay court major), but they were all rated fairly highly on this list. This suggests that many of their wins on clay against the other world number ones have come at smaller tournaments such as those at the Masters 1000 level or 500 level, which are less prestigious. The clay court season lasts 3 months from April to early June.

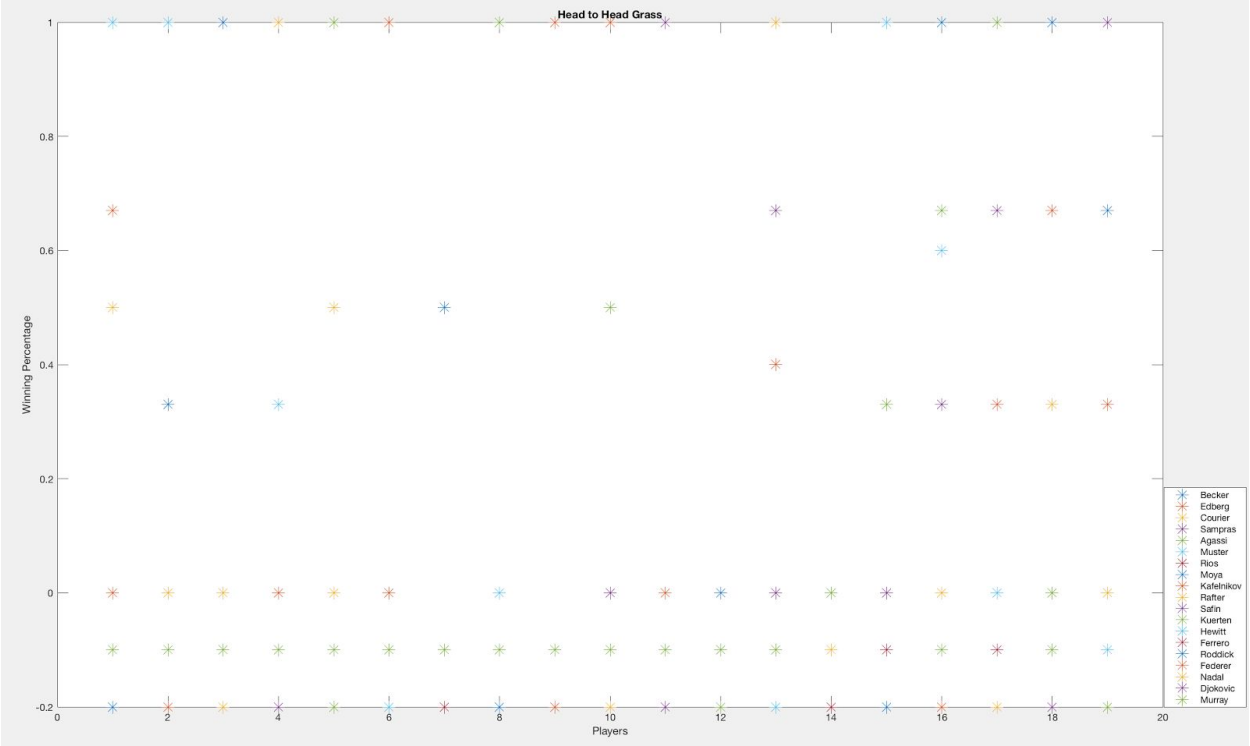


Figure 15. Head to Head Grass

The script best_head_to_head_grass.m creates a 1x19 vector of doubles that shows all the positive head to heads on grass of each player. Roger Federer (7) as at the top of this list with Pete Sampras (4), Rafael Nadal (4) and Novak Djokovic (4) coming in at number 2. Yevgeny Kafelnikov (3), Lleyton Hewitt (3), and Andy Murray (3) are all tied for 3. All of these players except Kafelnikov have won Wimbledon (the grass court major) at least once. For Kafelnikov's case, he took advantage of a small sample size of grass court matches against world number ones that were either past their prime (Becker, 1-0), not natural grass courters (Courier, 1-0), or inexperienced up-and-comers (Federer, 1-0), so I do not rate his grass court prowess above the others. This head to head sampling is a much stronger indicator of grass court success than clay because there are far fewer grass court tournaments on the ATP World Tour, so most of the matches will be played at Wimbledon. The grass court season is usually a short 3 week period from mid-June to early July.

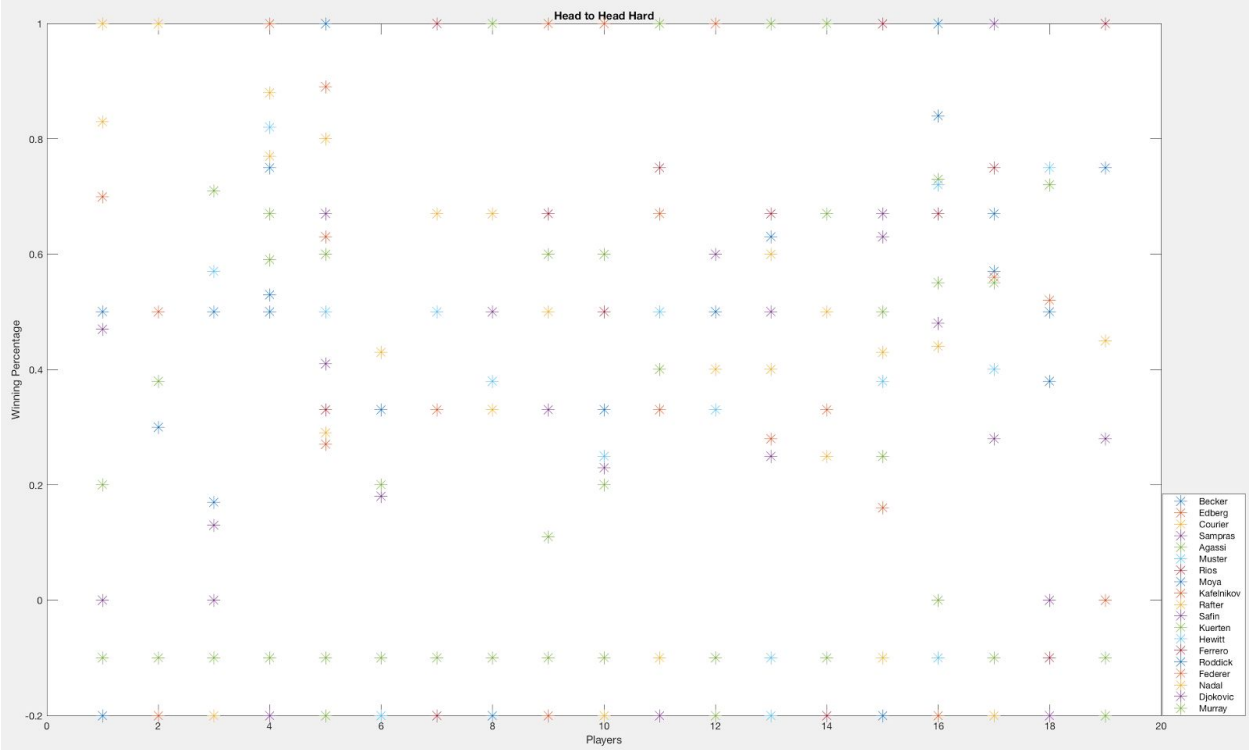


Figure 16. Head to Head Hard

The script best_head_to_head_hard.m creates a 1x19 vector of doubles that shows all the positive head to heads on hard courts of each player. Pete Sampras (9) and Andre Agassi (9) posted the best records on hard courts with Marat Safin (8) and Roger Federer (8) in 2nd place and Lleyton Hewitt (7) and Rafael Nadal (7) in 3rd place. All of these players have won a hard court major (either the Australian Open or the US Open), and all of them except Hewitt have won both hard court majors at least once. This dataset is the most indicative of hard court success compared to clay and grass because there are more hard court majors than clay or grass majors, and also because hard courts are the most common courts on the ATP World Tour since January - March and July - November are the hard court seasons.

Upon analyzing the winning head to head plots, it was immediately clear to me that just using these metrics would be flawed reasoning that is skewed towards the previous era. The Big Four today (Federer, Nadal, Djokovic, and Murray) are the only players who are still actively playing who have been ranked world number one. Many of the previous legends were already retired by the time they turned professional, so of course they will not have as good of a head to head against the other players since the sample size is skewed. However, the main purpose of these datasets was geared towards providing a baseline to make my predictive head to heads, so I am not looking into these data points as deeply as I am the serve, return, and under pressure stats.

Analysis

Before we can start comparing the players against each other, we need to establish some baselines from the plotted data. I calculated the max, min, and mean for serves, returns, and under pressure statistics on all surfaces combined. I chose not to do an analysis for every individual surface (clay, grass, hard) because that would be too many images to include and the total average is the most important piece of data to draw assumptions from. The findings are shown below. The files can be found in the /analysis folder. This section uses concepts learned in **Analysis of Distributions** such as mean and standard deviation.

Serve Stats All

```
max_percentage_first_serve =

    0.6900


min_percentage_first_serve =

    0.5400


mean_percentage_first_serve =

    0.6037


std_percentage_first_serve =

    0.0404
```

Figure 17. max % 1st serve is Rafael Nadal, min % 1st serve is Lleyton Hewitt.

First serve percentage is very important to maintain in tennis because it enables you to set up the point in a more offensive matter. It also means you hit less second serves, which are generally weaker and more exploitable. Not surprisingly, the more conservative servers lead this list with Rafael Nadal being at the top, while the relatively short Lleyton Hewitt (5”11) is at the bottom of this list. The bigger servers tend to have lower first serve percentages with the exception of Andy Roddick who is at 64.80%.

```
max_percentage_first_serve_points_won =

    0.8100


min_percentage_first_serve_points_won =

    0.6900


mean_percentage_first_serve_points_won =

    0.7426


std_percentage_first_serve_points_won =

    0.0323
```

Figure 18. max % first serve points won is Pete Sampras, min % first serve points won is Thomas Muster.

The big servers stand tall in this category with Pete Sampras leading the way followed by Boris Becker at 79.40%, Andy Roddick at 79.30%, and Roger Federer at 77.10%. Only two players had below a 70% average: (1) Thomas Muster, who had the lowest percentage of first serve points won and (2) Juan Carlos Ferrero. Both dominated on clay but were weaker on other surfaces, which makes sense that they did not have to rely on their first serves as much.

```
max_percentage_second_serve_points_won =

    0.5700


min_percentage_second_serve_points_won =

    0.5000


mean_percentage_second_serve_points_won =

    0.5332


std_percentage_second_serve_points_won =

    0.0197
```

Figure 19. max % second serve points won is Rafael Nadal, min % first serve points won is Boris Becker.

This was one of the most surprising stats to me; intuitively, I thought at first that the big servers would again dominate this metric, but ironically, the player who hits the least second serves (Nadal due to his high first serve percentage) dominates the second serve points won category. Other great baseline players without huge serves like Novak Djokovic, Andre Agassi, and Juan Carlos Ferrero also were extremely high on this list. Surprisingly, many of the big servers like Sampras and Becker ranked closer to the bottom in this category, which indicates to me that once the point gets started on more level terms, they are more at the mercy of their opponents than players like Nadal, Agassi, Djokovic, and Ferrero. The exceptions were Roddick and Federer, who won 55.90% and 56.50% of second serve points, respectively. Roddick does not have a superb baseline game, but it apparently got the job done and he has one of the best kick second serves of all time. Federer has a great kick second serve and a dazzling repertoire of strokes from the baseline, so that made more sense to me that he dominated this category.

```
max_percentage_service_games_won =

    0.9000


min_percentage_service_games_won =

    0.7900


mean_percentage_service_games_won =

    0.8347


std_percentage_service_games_won =

    0.0341
```

Figure 20. max % service games won is Andy Roddick, min % first serve points won is Yevgeny Kafelnikov.

The big servers dominated this category with Roddick being the highest and others like Sampras, Federer, and Becker rounding out the top four spots at 88.70%, 88.50%, and 85.80%, respectively. Novak Djokovic was 5th on this list at 85.60% and Rafael Nadal was 6th on this list at 85.20%; It is a testament to their overall baseline games because their serve is not his strong suit and more of a way to get the point started. The only players to win less than 80% of their service games were Thomas Muster (who won 1 French Open), Marcelo Rios (who won no majors), Carlos Moya (who won 1 French Open), Yevgeny Kafelnikov (who had the lowest percentage of service games won but won 1 French Open and 1 Australian Open), and Juan Carlos Ferrero (who won 1 French Open). This is the most important service stat because winning service games is a necessity to maintaining a winning position in matches because as soon as you lose serve, you have to come back. This is why all of the Top 6 except Roddick have at least 6 major titles (Roddick has 1, Sampras has 14, Federer has 17, Becker has 6, Djokovic has 12, and Nadal has 14).

Return Stats All

```
max_percentage_first_serve_return_points_won =

    0.3400


min_percentage_first_serve_return_points_won =

    0.2700


mean_percentage_first_serve_return_points_won =

    0.3105


std_percentage_first_serve_return_points_won =

    0.0207
```

Figure 21. max % first serve return points won is Rafael Nadal, min % first serve return points won is Andy Roddick.

Rafael Nadal holds the top spot winning a whopping 34% of 1st serve return points. Novak Djokovic and Andy Murray are tied at 33.60%, Stefan Edberg is at 33%, and Roger Federer is at 32.60%. The Big Four of Men's Tennis today (Nadal, Djokovic, Murray, Federer) are among the best returners of all time and given that the surfaces are much slower today, it makes sense how much more important the return is today than it was 10-20 years ago. Surprisingly, Stefan Edberg also was at the top of this list even though I don't normally associate him as a great returner. He played more serve and volley, but he did have a great return that was not super aggressive but allowed him to get into a more advanced position at the net to knock off a volley. The players under the 30% mark included Boris Becker, Jim Courier, Pete Sampras, Patrick Rafter, Gustavo Kuerten, and Andy Roddick. This made sense because Becker, Sampras, and Roddick were huge servers so they could rely on that instead of breaking serve. Rafter was a serve and volleyer so rely on that over returning well. Both Courier and Kuerten were most successful on slow clay, so their returns could be exposed on faster surfaces due to their extreme grips (both Courier and Kuerten) or long backswings (Kuerten).

```
max_percentage_second_serve_return_points_won =

    0.5600


min_percentage_second_serve_return_points_won =

    0.4900


mean_percentage_second_serve_return_points_won =

    0.5258


std_percentage_second_serve_return_points_won =

    0.0206
```

Figure 22. max % second serve return points won is Andre Agassi, min % second serve return points won is Andy Roddick

Andre Agassi is considered to be one of the best returners of all time, and this stat certainly indicates it. Agassi was very aggressive off the return, and most notably off the second serve, which is where most return points are won. The rest of the Top 5 in this category are Nadal, Murray, and Thomas Muster at 55.20% and Djokovic at 55.10%. All of them had great success on slower surfaces and clay, and their returns played a big role in this. At the bottom of the list are most of the big servers like Becker, Sampras, and Roddick. I was most surprised at how low Federer was in this category at 51.10%, but he is not known for being a very aggressive returner, so it made sense that quite a few players were ahead of him here. It is worthy of note to see that Jim Courier is a significantly better 2nd serve returner than he is a 1st server returner, winning 52.70% of 2nd serve return points which is the highest of the players who win less than 30% of 1st serve return points. Unfortunately for the other players who win less than 30% of 1st serve return points, they have similarly poor 2nd serve returning stats and rank among the lowest with Andy Roddick coming in at last place.

```
max_percentage_return_games_won =

    0.3300


min_percentage_return_games_won =

    0.2000


mean_percentage_return_games_won =

    0.2737


std_percentage_return_games_won =

    0.0393
```

Figure 23. max % return games won is Rafael Nadal, min % return games won is Andy Roddick.

Just like how percentage of service games won is the most important statistic for serves, the percentage of return games won is the most important statistic for returns. Rafael Nadal leads this category just like how he led the percentage of 1st serve return points won, leading me to believe there is some correlation between these two statistics. A lot of this can be attributed to his success on clay, where he is without a doubt the best player and returner on that particular surface. This does skew a bit of his returning stats though since he is a weaker returner on grass and hard courts. Djokovic comes in 2nd place at 32.20% and he is a much more versatile returner on all surfaces. Murray is in 3rd place at 32%, while Agassi, Muster, and Edberg are in 4th, 5th, and 6th with 31.70%, 31.60%, and 30.20%, respectively. The main thing that surprises me is that Federer is at 26.90%, which is under the mean percentage of return games won. I was not surprised that he was below the other three in the Big Four (Nadal, Djokovic, Murray), but I would have expected him to be above average in returning games won. Not surprisingly, the big servers like Roddick, Sampras, and Becker fared poorly.

```
max_percentage_break_points_converted =

    0.4500


min_percentage_break_points_converted =

    0.3900


mean_percentage_break_points_converted =

    0.4189


std_percentage_break_points_converted =

    0.0188
```

Figure 24. max % break points converted is Rafael Nadal, min % break points converted is Andy Roddick.

This statistic goes hand in hand with return games won because you need to convert break points to win return games. A low break points converted metric means that a player is good at creating opportunities but fails to convert them. Not surprisingly, Nadal is at the top given his proficiency at breaking serve. The great returners like Muster, Agassi, Djokovic, and Murray also rate highly on this list. This metric does not have a whole lot of deviation, but players under the 40% mark include Boris Becker, Patrick Rafter, Marat Safin, and Andy Roddick, indicating that they have trouble converting break points due to a combination of factors such as performing under pressure or having weaker return games.

Under Pressure Stats All

Note: the break points converted stat is duplicated from the return stats, so I chose not to re-include the max, min, and mean percentage of break points converted.

```
max_percentage_break_points_saved =

    0.6800


min_percentage_break_points_saved =

    0.6100


mean_percentage_break_points_saved =

    0.6421


std_percentage_break_points_saved =

    0.0212
```

Figure 25. max % break points saved is Pete Sampras, min % break points saved is Marcelo Rios.

The ability to save break points is extremely important and stems from two key factors: (1) the player's serving ability and (2) the player's mental game. Pete Sampras excelled at both of these categories and leads the field in break points saved. Andy Roddick, Roger Federer, Boris Becker, Rafael Nadal, Boris Becker, and Novak Djokovic are the only players above 65%, which indicates their ability to serve well under pressure. This statistic is what separates the number one players who have had more success from the ones who were less successful, which is very apparent because Marcelo Rios was the only world number one without a major title and he is at the bottom of this list. It is not surprising to see players who are considered to be less mentally tough at the bottom of this list such as Marat Safin (62.70%) and Andy Murray (62.90%).

```
max_percentage_of_tiebreaks_won =

    0.6500


min_percentage_of_tiebreaks_won =

    0.5100


mean_percentage_of_tiebreaks_won =

    0.5832


std_percentage_of_tiebreaks_won =

    0.0415
```

Figure 26. max % of tiebreaks won is Roger Federer, min % of tiebreaks won is Gustavo Kuerten.

The biggest servers with the exception of Boris Becker shine in this category with Roger Federer in 1st place. Pete Sampras comes in 2nd at 63.80% and surprisingly, Novak Djokovic is in 3rd place at 63.20%. Roddick is 4th at 62.20%, while Murray rounds out the top 5 with 61.80%. Tiebreaks are important because you only need one mini break (one point won on the opponent's serve) to win the tiebreak, so the serve and mental toughness is heavily emphasized here. What I found most unusual was that Marcelo Rios was one of the few players who had a tiebreak winning percentage over 60%, but he was neither a big server nor mentally tough! Although Nadal has fantastic mental toughness, he lacks a big serve and is always vulnerable to losing a mini break to an on-fire opponent. Stefan Edberg is the only other player on the list who wins over 60% of tiebreaks (60.20%). The rest of the players are have closer to 50/50 odds with Gustavo Kuerten being the lowest at 50.60% odds.

```
max_percentage_deciding_sets_won =

    0.7600


min_percentage_deciding_sets_won =

    0.5700


mean_percentage_deciding_sets_won =

    0.6411


std_percentage_deciding_sets_won =

    0.0469
```

Figure 27. max % deciding sets won is Novak Djokovic, min % deciding sets won is Marat Safin.

The percentage of deciding sets statistic is arguably one of the most important metrics, if not the most important metric in determining under pressure ratings. Since 2011, Novak Djokovic has been considered one of the most mentally tough champions on the circuit, and he is on a different level to the rest of the players on this list at 75.60%. Andy Murray is the next closest at 70.90%, while Pete Sampras (68.50%), Rafael Nadal (68.30%), and Thomas Muster (66.10%) make up the remaining of the Top 5. The most notable player missing from the Top 5 is Roger Federer, who is at 64.40% in terms of deciding sets won. One of the knocks on Federer throughout his career was that he was considered mentally fragile, and this statistic does show that he is relatively weaker under pressure than some of the other players on this list. The only players under 60% were Jim Courier at 57.50% and Marat Safin at 56.8%. I found it strange that Courier was rated so low because his mental tenacity was considered to be extremely high in his prime, but admittedly, his prime was a brief period from 1991 to 1993 and he was burnt out after that, so that's probably why he ranks so low on this list. Safin had always been mentally weak, so I was not surprised he was very low here.

Narrowing Down the Data

While it was definitely cool to see all these different metrics for serve, return, and under pressure statistics, I ultimately concluded that the four main categories I would use to predict head to head matchups would be (1) % service games won, (2) % return games won, (3) % of tiebreaks won, and (4) % of deciding sets won. The other stats were definitely interesting, but I felt these four in particular encompassed many of the other metrics and would give me the data to really conclude how players would fare against each other just based on the numbers.

K-nearest Neighbor Search

A useful analysis technique I read about online was the **k-nearest neighbor search algorithm**. This method is primarily used for classifying data, and in my case, I would be able to use it to help me to do my prediction analysis. I selected my k to be 19 because my dataset has 19 players, so this parameter looks for the 19 nearest neighbors. Additionally, I just chose to use Euclidean distance by default. I analyzed and plotted my findings in the three graphs below for a comparison between head to heads on all surfaces against head to heads on clay, grass, and hard courts, respectively. The x-axis represents each player numbered from 1 to 19 according to **List 1**. The y-axis represents the standard deviation from the mean.

I had some difficulties at first understanding how to use this algorithm to help my predictions, but after some experimentation and help from Professor Parker, I was able to run the k-nearest neighbor algorithm on the head to head matrices corresponding to those four main categories by surface. This generated a total of 16 scatter plots that I will discuss in more detail below. This will form the core foundation of my prediction algorithm that I will use to draw conclusions in the end. An interesting thing to note is that the k-nearest neighbor model is not fully accurate, so I cannot verify that my predictions are the most mathematically sound; however, this is my foray into it.

K-nearest Neighbor Search All

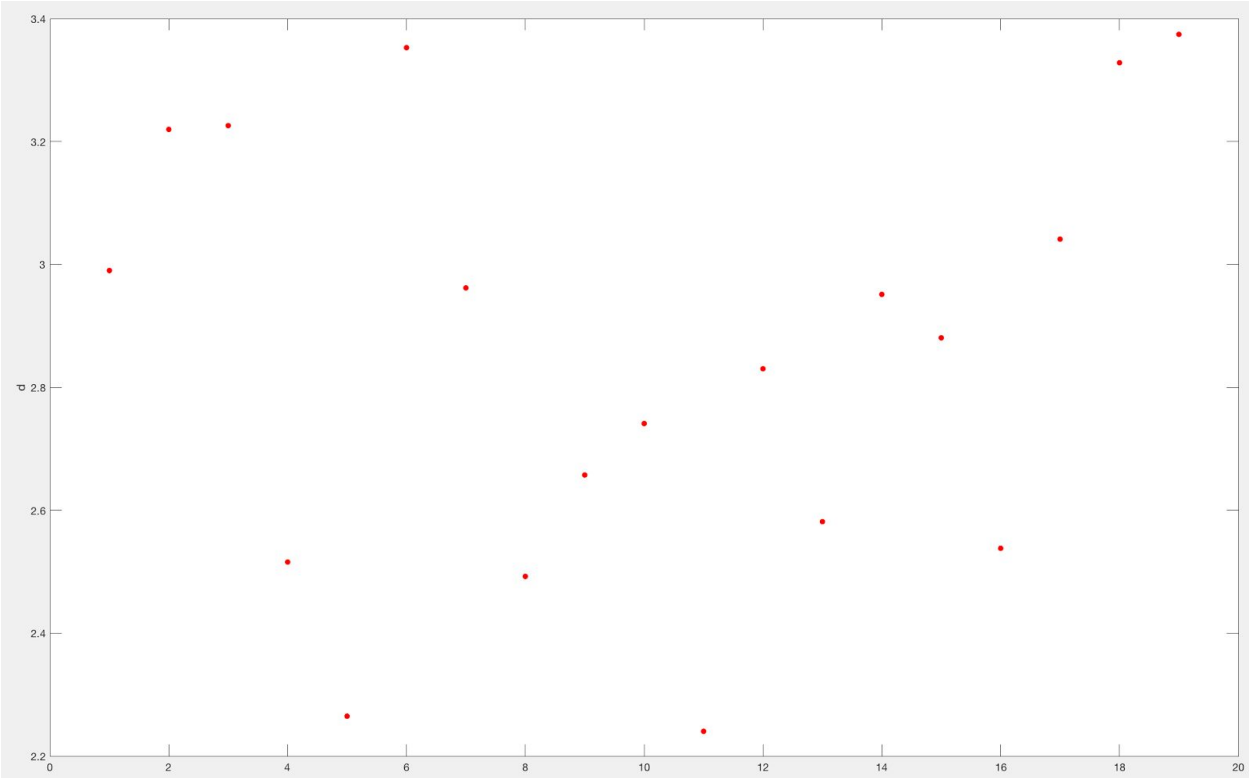


Figure 28. K-nearest neighbor search between head to head all and % of service games won.

Across all surfaces, the standard deviation was highest for Andy Murray (6), which makes sense because Murray's first serve and second serve have a huge disparity in quality. Stefan Edberg and Pete Sampras were (2) and (3) on the x-axis, respectively, and it made a lot more sense that there standard deviation was higher as a result since their success was heavily dependent on having easy holds. Other big servers like Boris Becker (1), Andy Roddick (15), and Roger Federer (16) had lower standard deviations, so this suggested they were either more consistent in their service games (Roddick, Becker) or had a stronger overall game to rely on if their serves were off (Roger Federer). Djokovic and Muster are also high on this list, which is interesting because their serves are not their big weapons.

The players with lower standard deviations did not rely as much on serve (Andre Agassi (5) and Marat Safin (11)). Agassi had one of the best returns of serve and thus it made sense that his standard deviation was lower because he did not rely on his serve as much. Safin had a pretty big serve, but I guess his baseline game carried him far more than his serve.

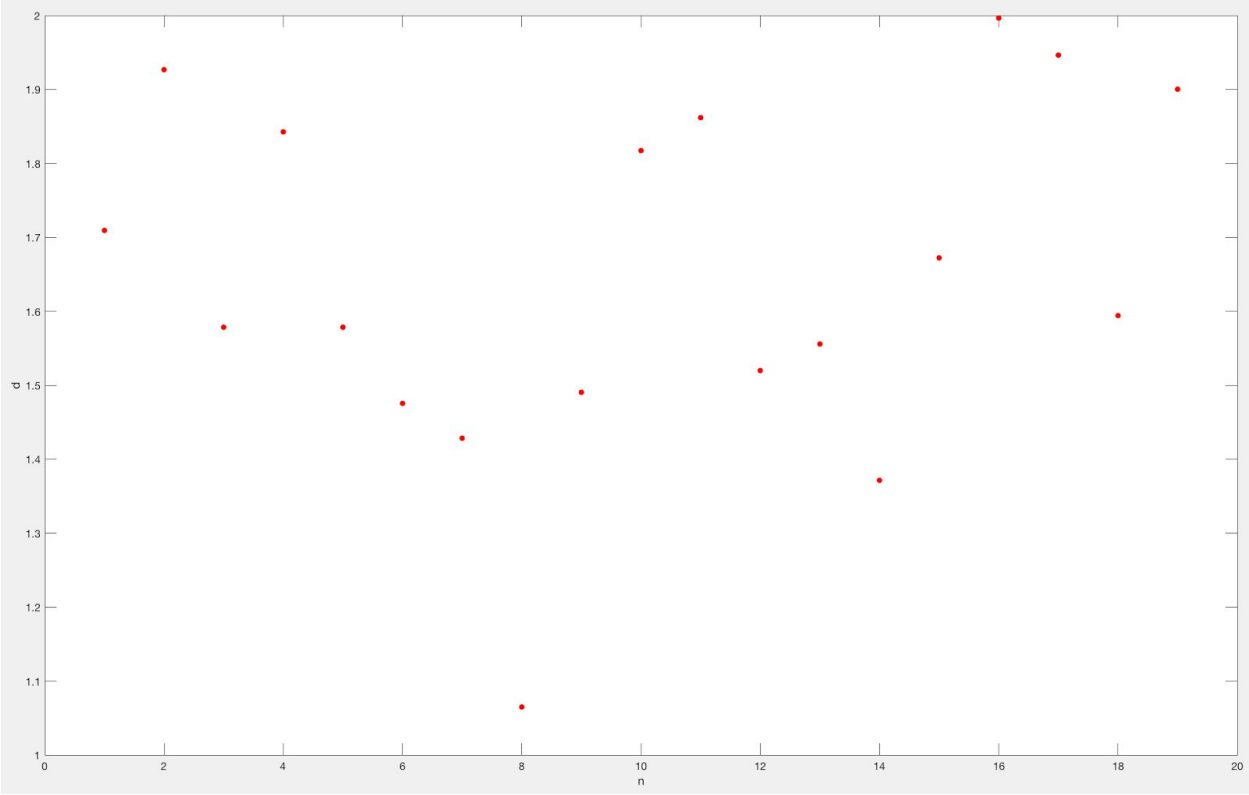


Figure 29. K-nearest neighbor search between head to head all and % of return games won.

Breaking serve is a key statistic in every match; the person who breaks more will generally win the match! In this case, the k-nearest neighbor search algorithm indicates that Federer (16) has the highest standard deviation at 2, while Carlos Moya (8) has the lowest standard deviation here. The interesting thing to note here is that the best returners like Andre Agassi (5) and Novak Djokovic (18) sit closer to the 1.5 standard deviation range. This could possibly mean that the overall quality of their returns were far more consistent than a guy like Federer, who is not as good of a returner in general but returns particularly well against big servers like Roddick. Carlos Moya's standard deviation seems like an abnormality to me; he was not really a great returner, but it seems like it did not deviate much from a norm.

A lot of the players higher on the list like Edberg (2) and Sampras (4) are more aggressive players, but Edberg had a particularly good return for an attacking player. Sampras' return was weaker, which makes sense why his standard deviation was so high, but Edberg's return quality makes me question why his standard deviation is astronomical. The same can be said for Nadal's and Murray's returns at (17) and (19), respectively. Given more knowledge of the k-nearest neighbor algorithm, I will have to check these abnormalities again and make sure this is a reasonable result.

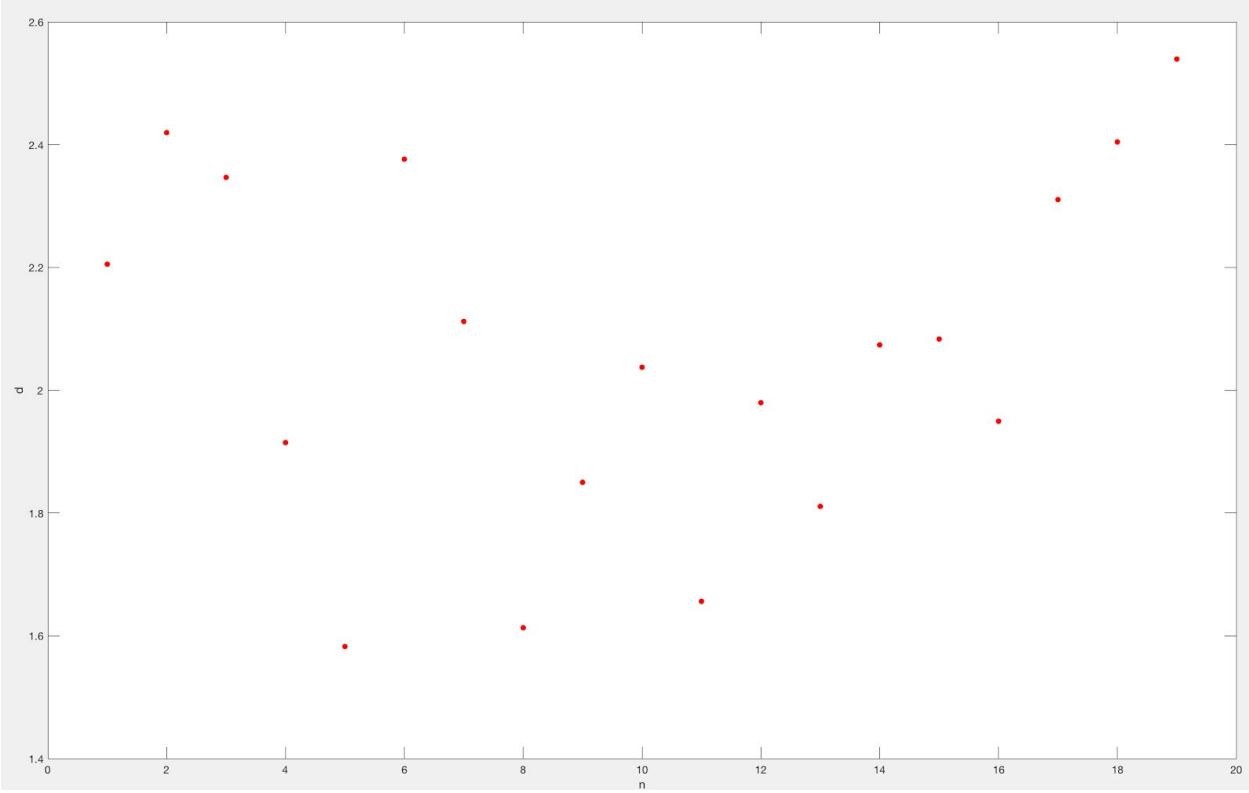


Figure 30. K-nearest neighbor search between head to head all and % of tiebreakers won.

In the case that you do not break serve, you have to win a tiebreaker to win a set. In men's tennis, the players play best of 5 sets in the major tournaments and Davis Cup and best of 3 sets everywhere else. The highest standard deviation here belongs to Andy Murray (19) and the lowest appears to be Andre Agassi (5). Andy Murray has an exceptional tiebreak record, so I find this strange, while Agassi has a below average tiebreak record, so this is not as surprising. Other players with fantastic tiebreak records like Djokovic (18) and Nadal (17) have high standard deviations, so I am not sure what is going on here. A possible explanation is that this indicates that this statistic is important in determining the outcome of their head to head matches and is therefore more volatile. Meanwhile, others like Sampras (4), Roddick (15), and Federer (16) have lower standard deviations, which could suggest that their big serves give them far more stability when it comes to getting the mini break and winning tiebreakers.

The lowest standard deviations belong to Agassi (5), Moya (8), and Safin (11). All three are less reliant on serves than the big servers and have better overall groundstrokes. Moya did not have a great backhand, while the other two did, but they had the lower standard deviation in common, so I feel this metric is more heavily emphasized to favor the great returners and baseline players.

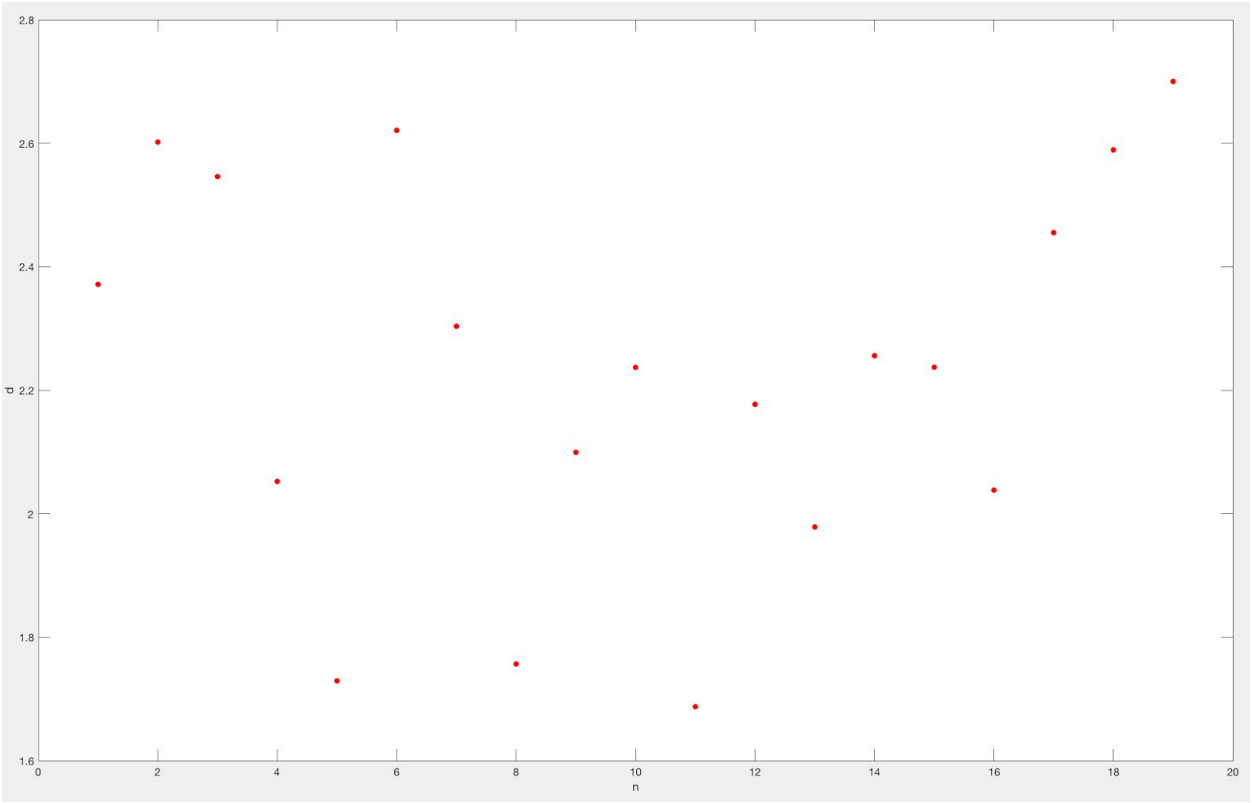


Figure 31. K-nearest neighbor search between head to head all and % of deciding sets won.

Deciding sets are the last set to decide a match; in a best of 3 set match, it is the 3rd set and in a best of 5 set match, it is the 5th set. Obviously, the deciding set is a test of stamina, both physically and mentally, as much as it is a test of tennis ability. Thus, it is not surprising to see Djokovic (18) and Murray (19) at the top of this list. Their standard deviation indicates that their winning percentages in deciding sets is an important factor for their matches. This can clearly be shown in their exceptional deciding set records mentioned earlier. Edberg (2) and Muster (6) are also high up here, and they did have some marathon matches in their careers (especially Muster).

Agassi (5), Moya (8), and Safin (11) have low standard deviations here. Just like for their tiebreak records, this suggests that when matches went to a deciding set, they were not as reliable as some of the other players. This is actually the case, since all three of them have fairly poor deciding set records compared to other world number ones.

In terms of the other Big Four, Nadal (17) has a pretty good deciding set record, but Federer's (16) is the worst of the Big Four. This is reflected here since Nadal has a higher standard deviation, while Federer's is closer to the norm of 1.5.

K-nearest Neighbor Search Clay

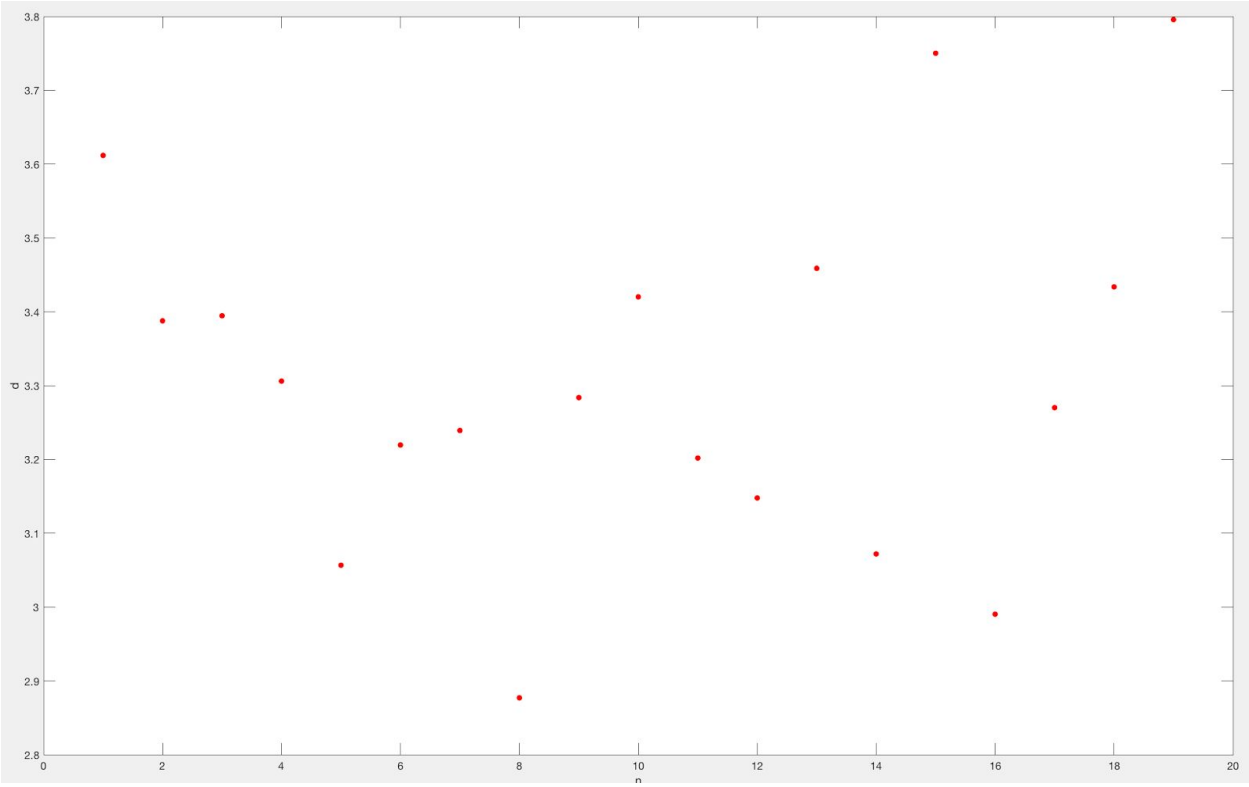


Figure 32. K-nearest neighbor search between head to head clay and % of service games won.

The two Andys (Roddick at 15 and Murray at 19) are the highest on this list, while the lowest on this list is Carlos Moya (8). First off, it is really weird that Carlos Moya has a very low standard deviation on most of the lists I have seen so far. This suggests that my k-nearest neighbor search algorithm could be a flawed analysis. Secondly, the two Andys are not as good on clay (Roddick in particular was quite awful on clay for a world number one). Thus, I am not sure how to interpret these very high standard deviations.

Another thing to note is just how high this range for standard deviation is! It ranges from 2.8 to 3.8, which is a lot higher than the ranges I have seen before. A benchmark for quality on clay is always Rafael Nadal, who sits a bit under the mean at 3.3. This could mean that he consistently plays at the same high level on clay in terms of serving. Other great clay courters like Muster (6) and Kuerten (12) have lower standard deviation values here, so this could mean that they follow the same pattern as Nadal and serve at a consistent level on clay.

Federer is very low on this list and Djokovic is pretty high up. This could mean Federer maintains his serve quality well on the dirt, but Djokovic can be prone to lapses of inconsistency while serving.

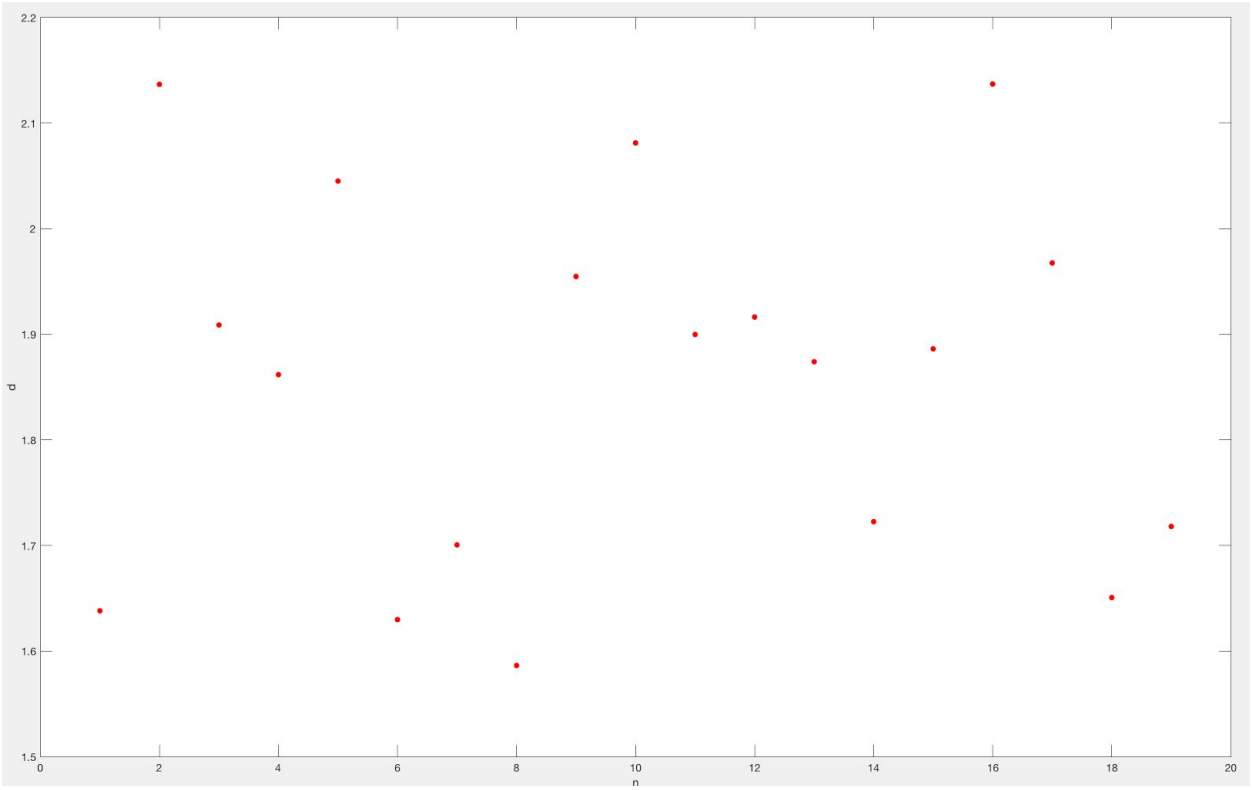


Figure 33. K-nearest neighbor search between head to head clay and % of return games won.

Once again, Carlos Moya (8) has the lowest standard deviation, while Federer's standard deviation mirrors his return on all surfaces with a very high standard deviation. Other high values include Stefan Edberg (2) and Patrick Rafter (10), which is interesting because neither were clay court players. This could suggest that lapses of inconsistency on returns determined their results.

The low values belong to Becker (1), Muster (6), Rios (7), Ferrero (14), Djokovic (18), and Murray (19). With the exception of Becker, most of these players have had some pretty decent success on clay. Even Becker was not too shabby, reaching the finals of the Italian Open, Monte Carlo Masters 1000 as well as the semifinals of the French Open. This could imply that returning was either less emphasized for them (Becker) or more consistent on average (everyone else basically).

Nadal is slightly above average but not as high as Federer's. Nadal has the best returning stats on clay, but again, this could suggest that there is something flawed with my reasoning or that his return could fluctuate a bit and affect his head to head matchups.

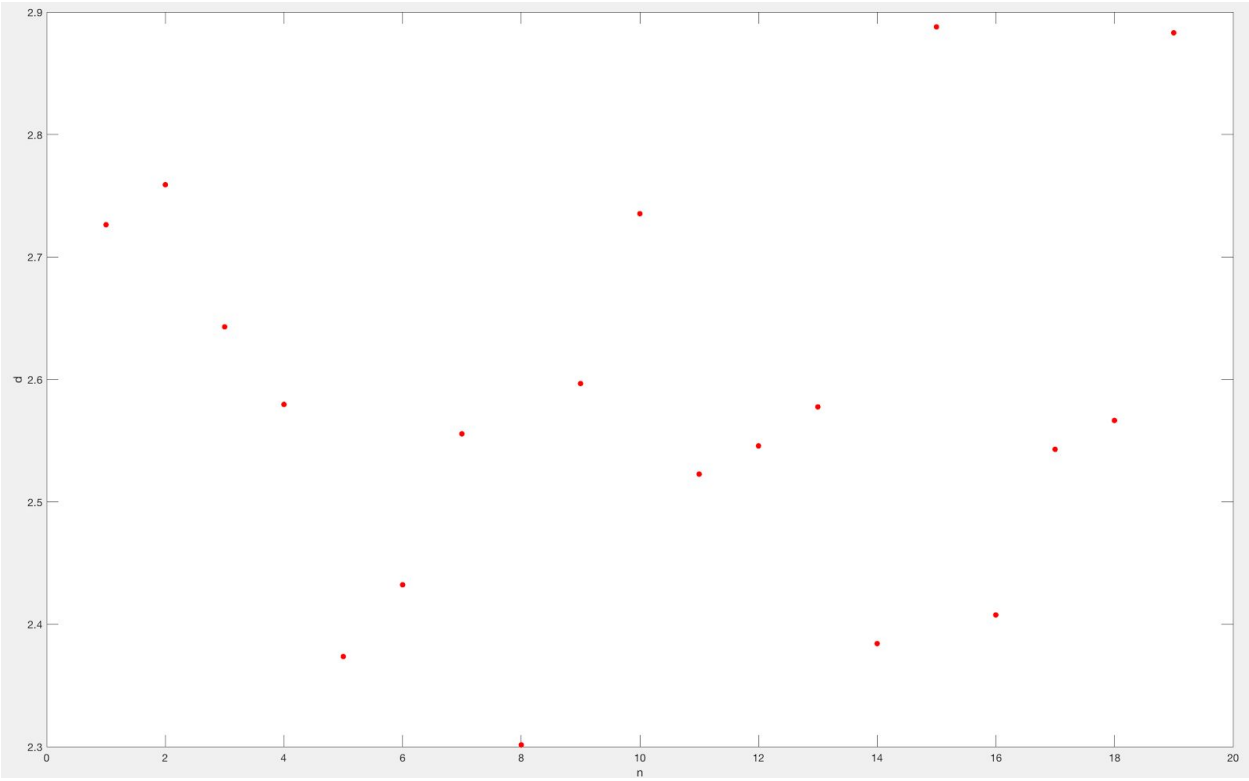


Figure 34. K-nearest neighbor search between head to head clay and % of tiebreakers won.

Big surprise! Carlos Moya (8) ranks at the very bottom of this list once again. Roddick and Murray are also high on this list once again like on the k-nearest neighbor search vs serve list for clay. What does this suggest? Serving quality is related to the chance of winning a tiebreakers, and apparently, there is greater standard deviation for the Andys.

It is interesting to note that clay is the slowest of the 3 surfaces, so it seems like tiebreakers are not as important on clay as they are on hard courts or especially grass. Other players who have lower standard deviations in tiebreakers on clay are Andre Agassi (5), Thomas Muster (6), Juan Carlos Ferrero (14), and Roger Federer (16). This could suggest that their tiebreak results are less likely to fluctuate on clay, due in large part to a strong baseline game to rely upon.

Big servers and attacking players like Becker, Edberg, and Rafter are quite high on this list. Interestingly, Sampras (4) is under the average standard deviation, which is strange to me. Sampras did have better movement and better groundstrokes than some of the other attacking players, but still nothing special compared to the clay courters of his day.

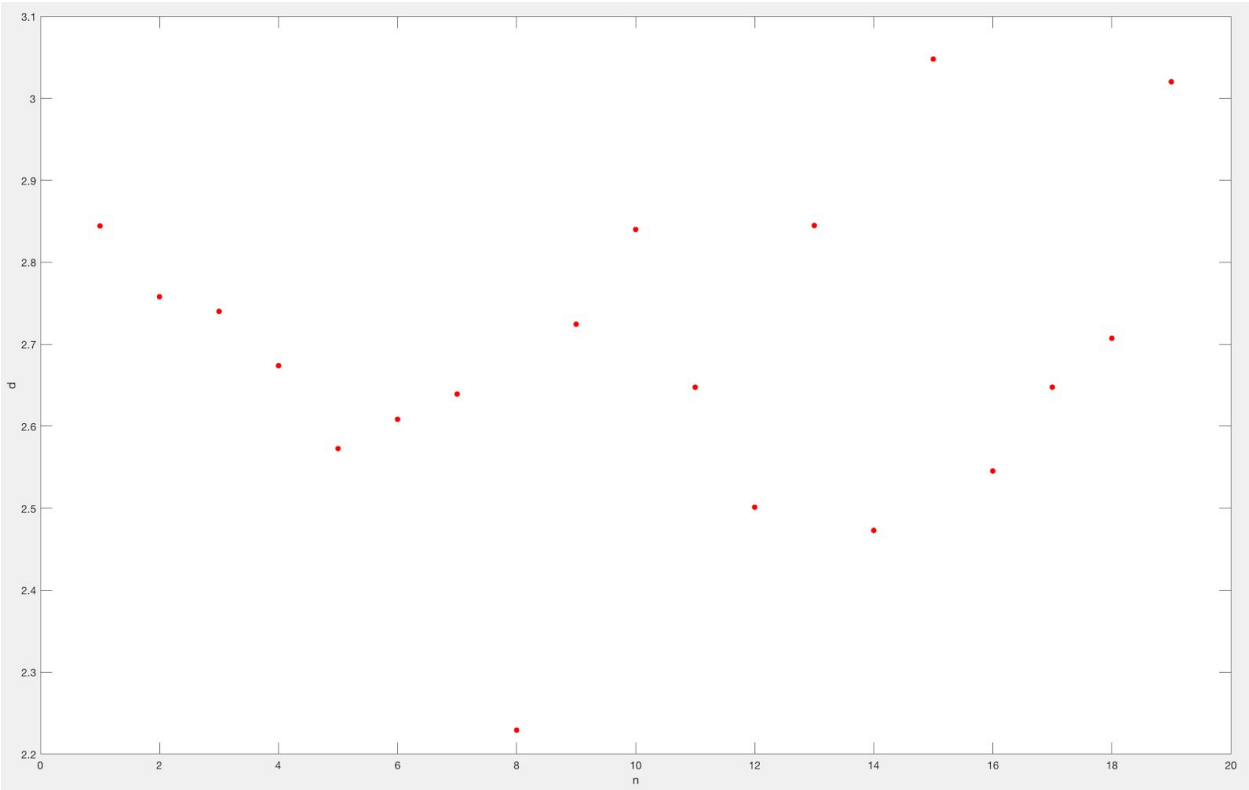


Figure 35. K-nearest neighbor search between head to head clay and % of deciding sets won.

Andy Roddick (15) leads this list, while Moya once again has the smallest standard deviation. Just like for the all section of deciding sets won, there is a lot of correlation between percentage of tiebreaks won and percentage of deciding sets won. The Andys again rank highly in this mark for clay.

The attacking players i.e. Becker (1), Edberg (2), and Rafter (10) are above the average standard deviation. Lleyton Hewitt (13) is also very high on this list, although he was a counterpuncher who did not have as much success on clay.

The players who were lower on this list were Andre Agassi (5), Thomas Muster (6), Gustavo Kuerten (12), Roger Federer (16), and Rafael Nadal (17). All of these players have at least one French Open title and were great clay courters, so I think it makes sense they would have a lower standard deviation since they would consistently win clay court deciding sets.

K-nearest Neighbor Search Grass

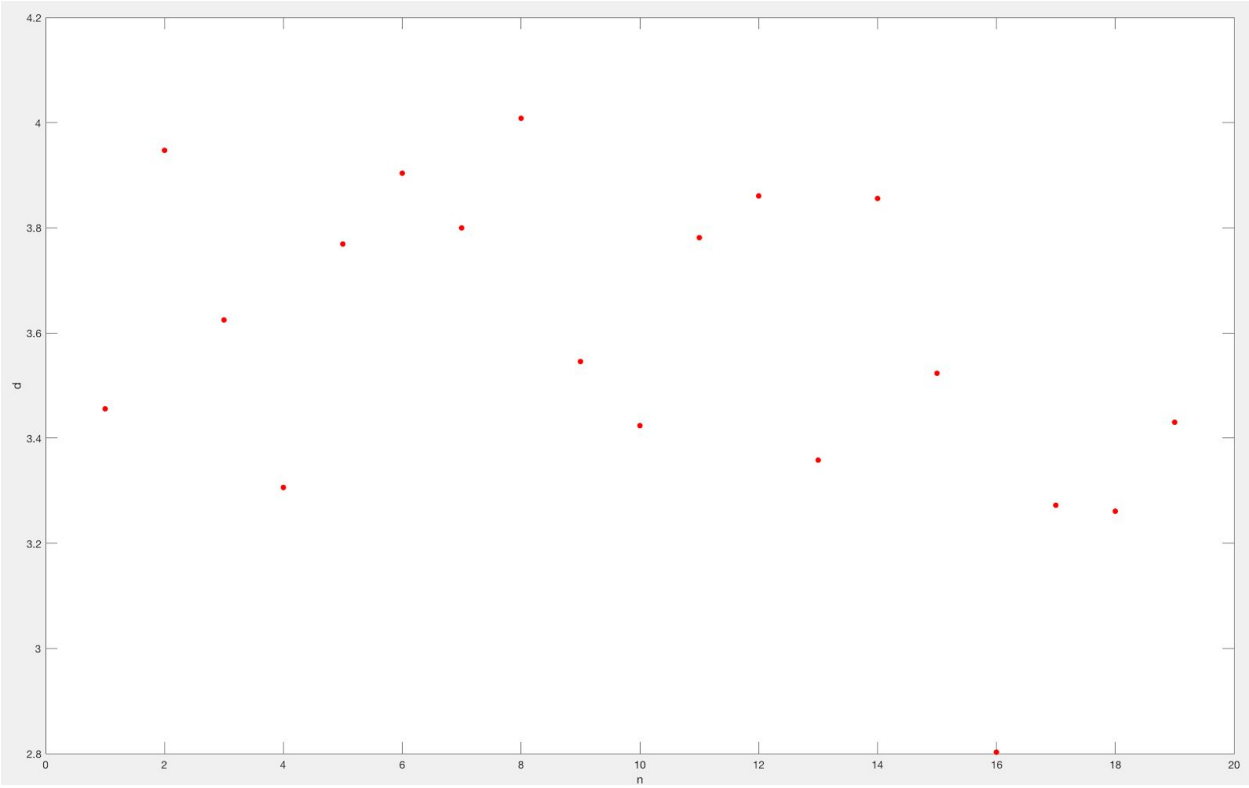


Figure 36. K-nearest neighbor search between head to head grass and % of service games won.

For once, Carlos Moya at 8 is at the higher end of the standard deviation spectrum, while Roger Federer (16) is the lowest on this list. Moya was not a grass court player and his serve was not a fantastic weapon. The opposite is true for Federer; he has an excellent service game and grass is his best surface.

Becker (1), Sampras (4), Rafer (10), Hewitt (13), Roddick (15), Nadal (17), Djokovic (18), and Murray (19) all rank among the lower standard deviations than normal, and all of them have won Wimbledon at least once except Rafter and Roddick. Even Rafter made two consecutive Wimbledon finals in 2000 and 2001, while Roddick made three Wimbledon finals from 2004, 2005, and 2009, so it seems like a lower standard deviation indicates a higher quality grass court hold game.

The players higher on this list include Agassi (5), Muster (6), Rios (7), Safin (11), Kuerten (12), and Ferrero (14). All of them had less relative success on grass and their games were not as well-suited to grass compared to clay or hard courts.

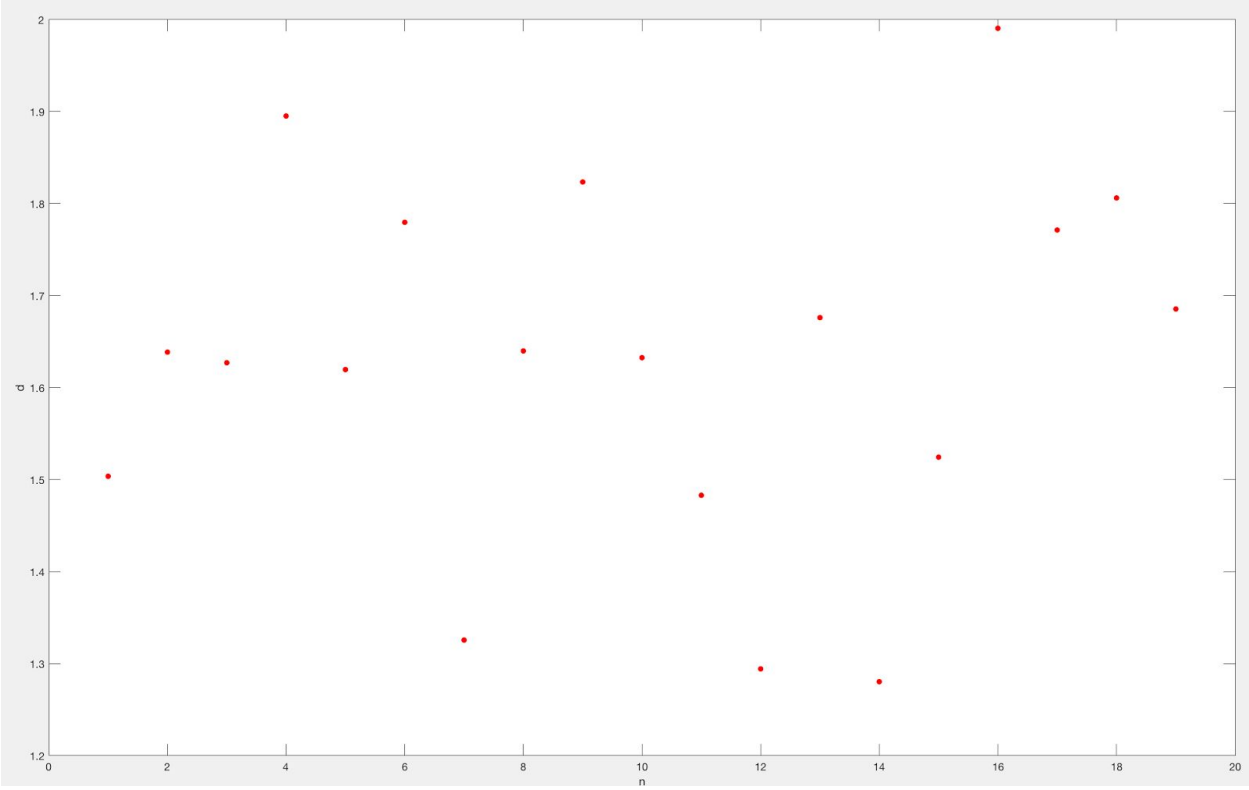


Figure 37. K-nearest neighbor search between head to head grass and % of return games won.

Federer (16) leads the standard deviation in return once again just like on all surfaces and clay. The lowest standard deviation appears to be Ferrero (14). The lower standard deviations appear to be Becker (1), Rios (7), Safin (11), Kuerten (12), and Roddick (15). The middle three players were less successful on grass, but Becker won Wimbledon three times and Roddick made three Wimbledon finals, so I am not sure why their standard deviations are like this.

Sampras (4), Muster (6), Kafelnikov (9), Nadal (17), and Djokovic (18) are quite high on this list. This is quite a motley crew because it includes arguably the best grass court player (Sampras), the worst grass courter on this list (Muster), a fairly poor grass courter (Kafelnikov), and two multi-Wimbledon champions (Nadal and Djokovic). What they do have in common is that a lot of their matches could have been deciding by breaks of serve, so this is arguably the greatest contributor to their respective placements on this plot.

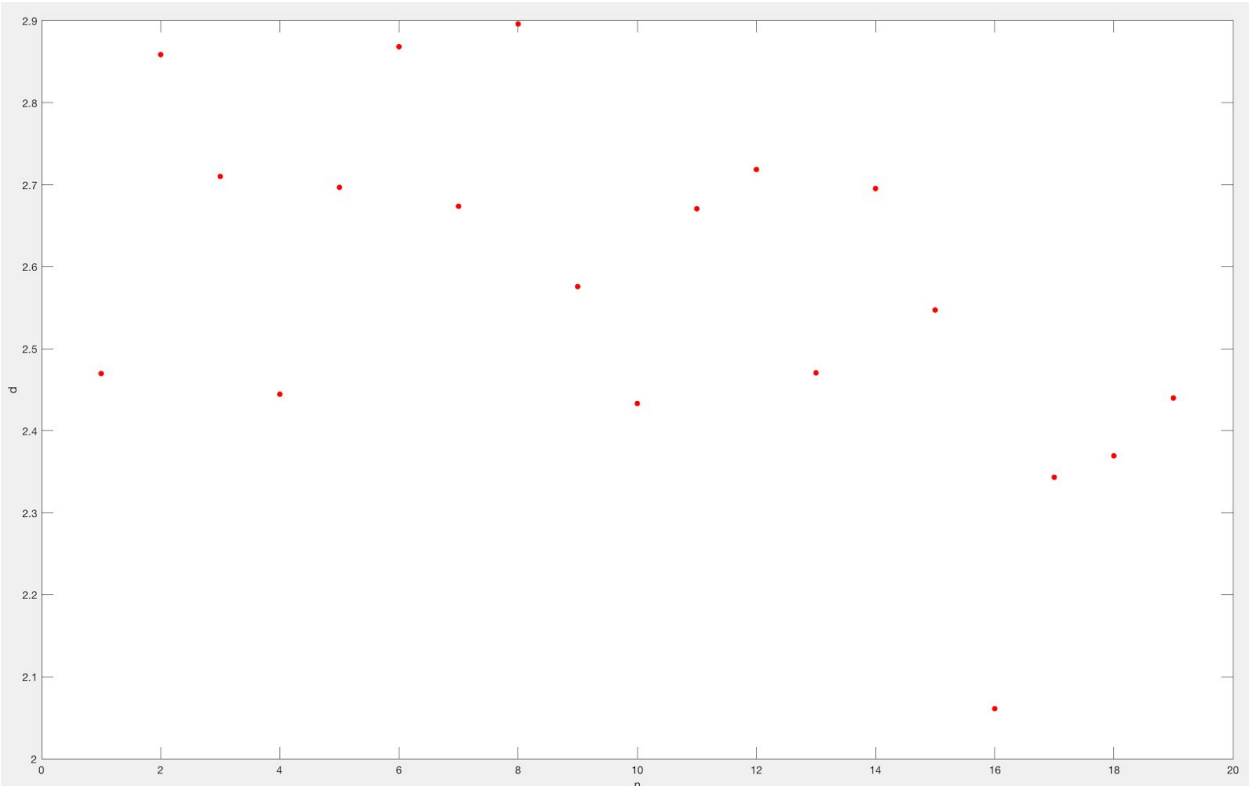


Figure 38. K-nearest neighbor search between head to head grass and % of tiebreakers won.

The one thing I have noticed across all these k-nearest neighbor plots is that Moya and Federer are always on opposite ends of the spectrum. In this case, Moya has the highest standard deviation, while Federer has the lowest standard deviation. Other high standard deviations include Stefan Edberg (2), Jim Courier (3), Andre Agassi (5), Thomas Muster (6), Marcelo Rios (7), Marat Safin (11), Gustavo Kuerten (12), Juan Carlos Ferrero (14), and Andy Roddick (15). The lower standard deviations belong to Boris Becker (1), Pete Sampras (4), Patrick Rafter (10), Lleyton Hewitt (13), Rafael Nadal (17), Novak Djokovic (18), and Andy Murray (18).

Just looking at these two lists, it is not immediately obvious what the relationship is between this k-nearest neighbor search algorithm result and the quality of the players on grass. Both Edberg and Agassi had won Wimbledon, while Roddick made three Wimbledon finals and Courier made a Wimbledon final. The lower standard deviation players all did win Wimbledon though. This could mean that tiebreakers could have be associated with more consistent winning results on grass.

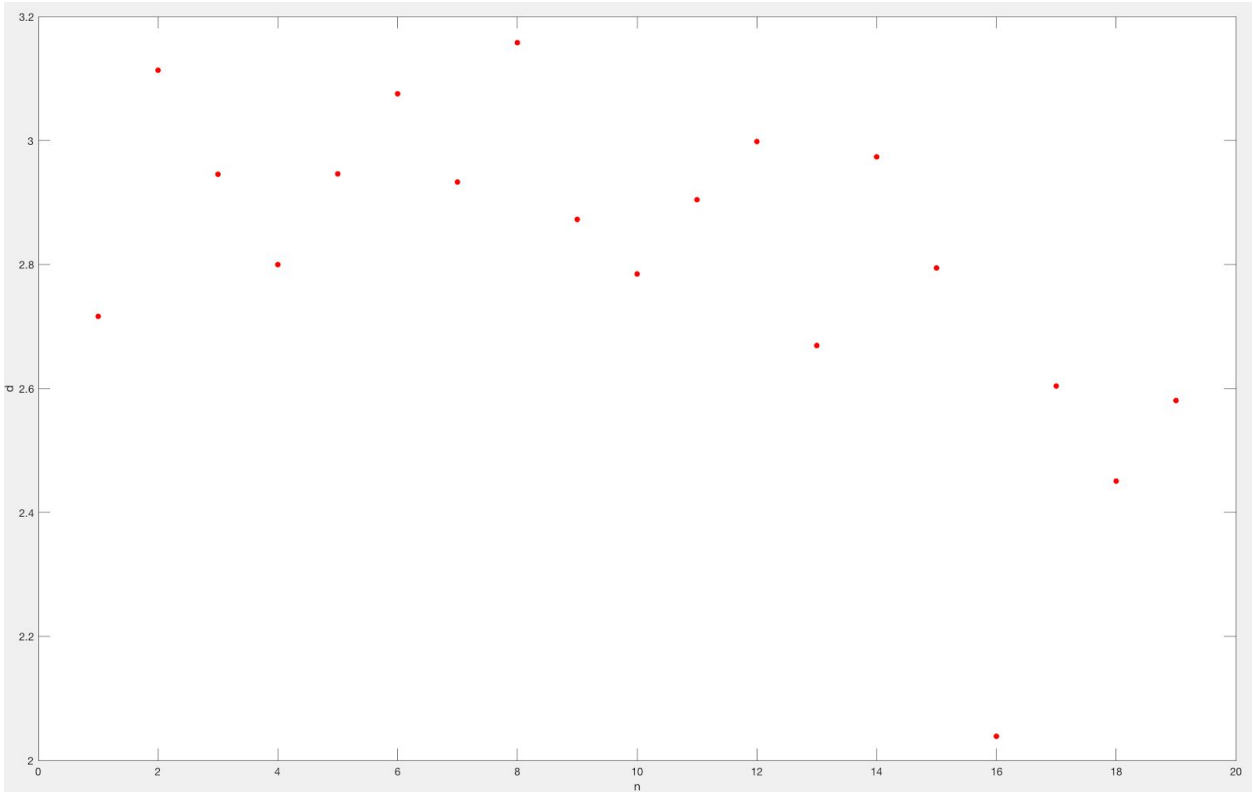


Figure 39. K-nearest neighbor search between head to head grass and % of deciding sets won.

This virtually mirrors **Figure 38**. Both extremes of the spectrum are identical. There are far fewer players under the mean for the standard deviation though. It appears that Djokovic is the only other player besides Federer below the standard deviation, although fellow Big Four members Murray and Nadal hover near the 2.6 mark for standard deviation. Everyone else is quite high on this list.

An important thing to be reminded of is that grass is an extreme surface that few people grow up on. Wimbledon is the only major tournament on grass, so it does make sense for the standard deviation to fluctuate more on grass given a smaller sample size than clay. At the French Open and the many clay court tournaments on the ATP World Tour, we are far more likely to see a deciding set since more matches are played on those surfaces.

K-nearest Neighbor Search Hard

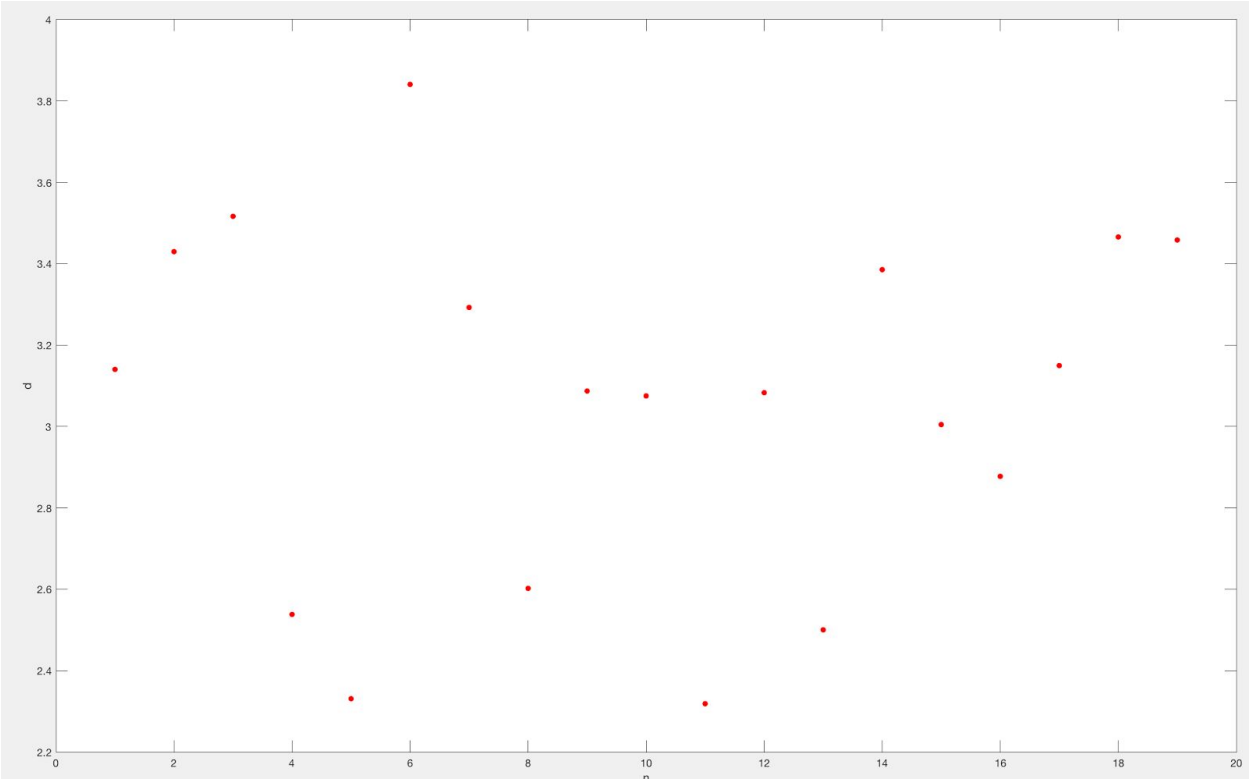


Figure 40. K-nearest neighbor search between head to head hard and % of service games won.

Andre Agassi (5) and Marat Safin (11) have the lowest standard deviations on this list, while Thomas Muster (6) has the highest standard deviation on this list. Muster did not rely too much on serve on any surface, so it makes sense that his serve quality could have varied quite a bit. Agassi and Safin also did not rely on their serves as much, but it should not have mattered because their baseline games were so strong that they can rely on that.

Other players with low standard deviations were Pete Sampras (4), Carlos Moya (8), and Lleyton Hewitt (13). Players with high standard deviations were Stefan Edberg (2), Jim Courier (3), Marcelo Rios (7), Juan Carlos Ferrero (14), Novak Djokovic (18), and Andy Murray (19).

Rafael Nadal (17) was pretty close to the norm for percentage of service games won on hard courts. This made sense because his game is not heavily predicted on serve, so it makes sense that he is more in the middle of the spectrum.

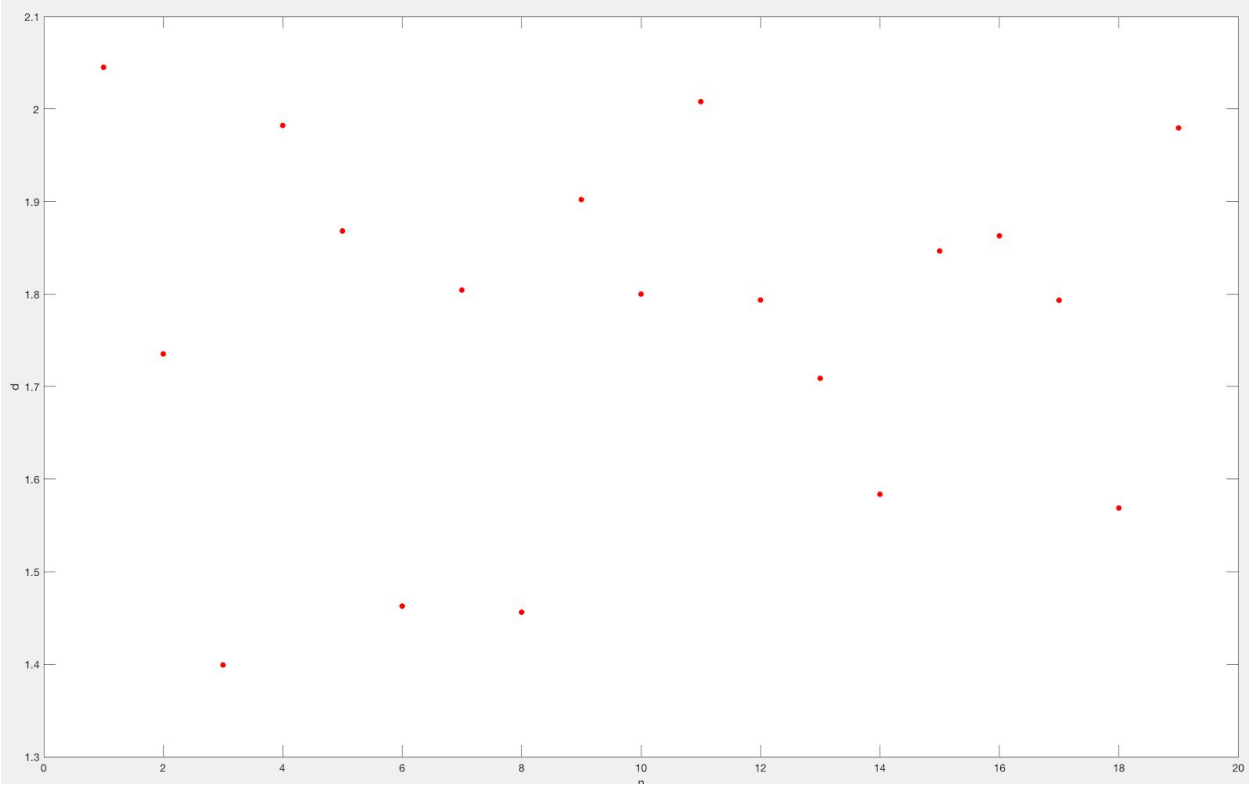


Figure 41. K-nearest neighbor search between head to head hard and % of return games won.

Boris Becker (1) has the highest standard deviation on this list, while Jim Courier (3) has the lowest standard deviation here. This is interesting because they did not account for any of the extremes in any of the previous results for k-nearest neighbor search. Becker was pretty solid on hard courts, winning a US Open and two Australian Opens, but he was a pretty up and down player, so that could account for his high standard deviation. Courier was also good on hard courts, winning two Australian Opens, but he was much more mentally tough and focused than Becker, so this could have allowed him to capitalize on breakpoints a bit better.

On the low end of the list are guys like Muster (6), Moya (8), Hewitt (13), Ferrero (14), and Djokovic (18). All of whom had pretty strong mental toughness, which could have helped them convert break point opportunities.

The higher end of the spectrum includes Sampras (4), Agassi (5), Kafelnikov (9), and Murray (18). Agassi was the odd one out of this list because his return was out of this world and his mental game was not too shabby (later in his career, he squandered most of his earlier career). The rest either did not care too much of their return game (Sampras), or lacked something extra (Kafelnikov's overall game and Murray's mental strength).

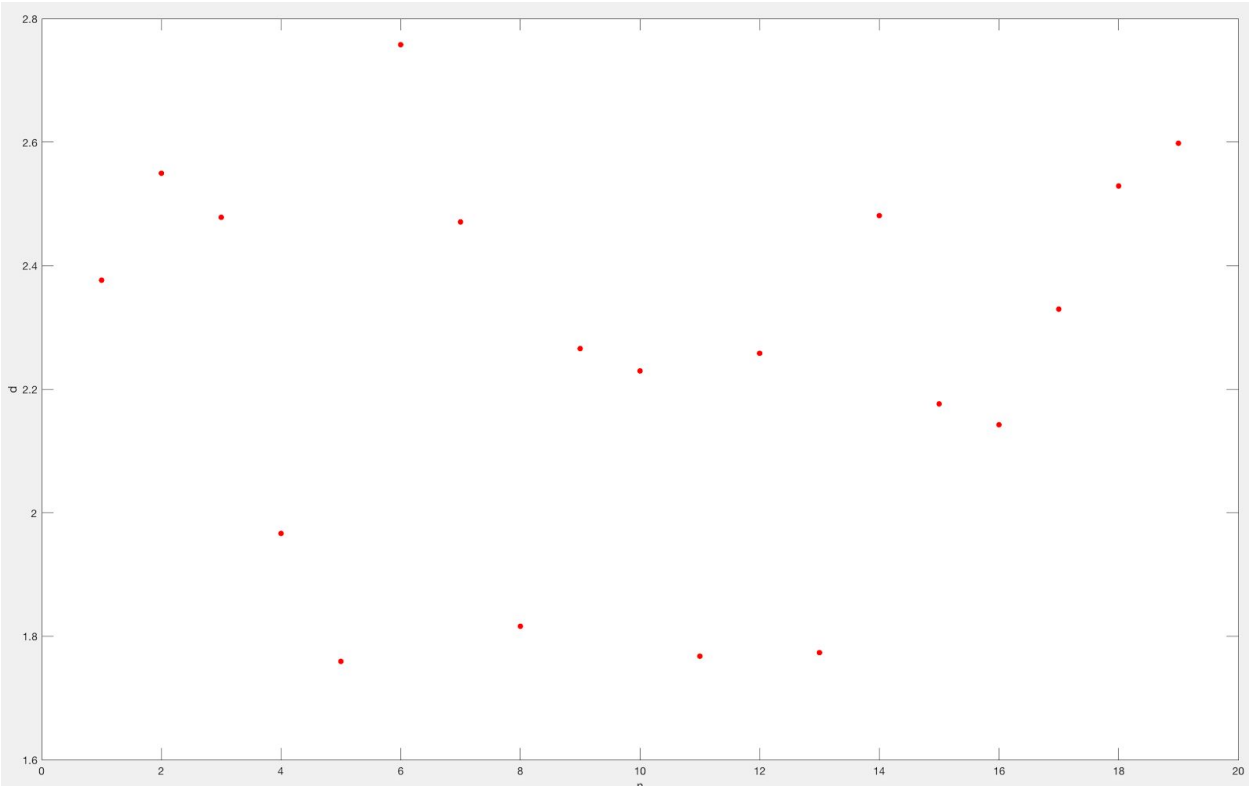


Figure 42. K-nearest neighbor search between head to head hard and % of tiebreakers won.

Muster (6) leads this list, while Agassi (5) has the smallest standard deviation of any of the world number ones. Other players with notably low standard deviations were Pete Sampras (4), Carlos Moya (8), Marat Safin (11), and Lleyton Hewitt (13), while players with notably high standard deviations were Stefan Edberg (2), Jim Courier (3), Juan Carlos Ferrero (14), Novak Djokovic (17), and Andy Murray (18).

Hard courts are in the middle in terms of speed of the court compared to clay and grass. Thus, I would expect to see more tiebreakers than clay, but less tiebreakers than grass. The results from k-nearest neighbor search suggest that players who have had their greatest success on hard courts like Kafelnikov (9), Rafter (10), and Roddick (15) are in the middle of this standard deviation spectrum. The odd person out seems to be Gustavo Kuerten (12) who is slightly above the mean of the standard deviation but had his greatest success on clay. He did have a good serve for a clay courter though, and he won the Year End Championships in 2000 to finish as world number one on indoor hard courts.

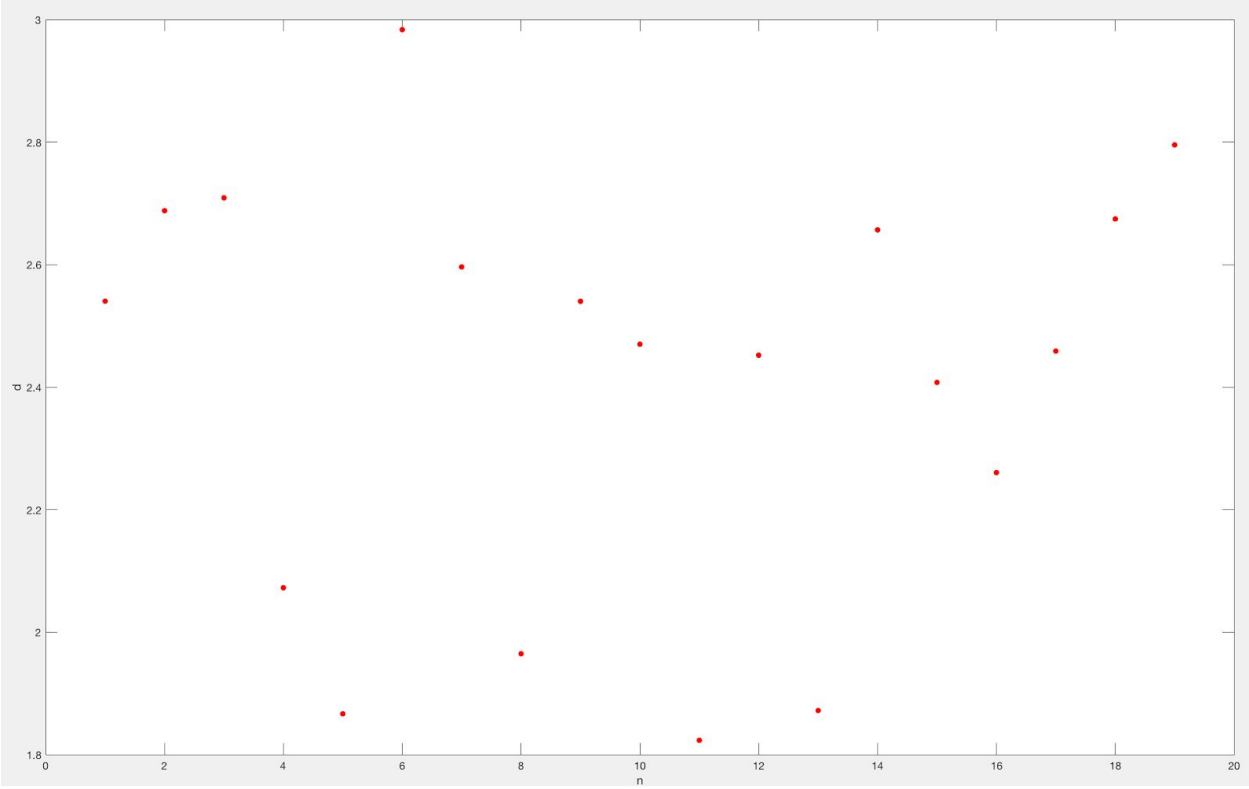


Figure 43. K-nearest neighbor search between head to head hard and % of deciding sets won.

Just like for the other surfaces, there is a lot of overlap between percentage of tiebreakers won and percentage of deciding sets won. Thomas Muster (6) has the highest standard deviation on this list, while Marat Safin (11) has the lowest standard deviation for deciding sets won. Players with lower standard deviations include Pete Sampras (4), Andre Agassi (5), Carlos Moya (8), Lleyton Hewitt (13), and Roger Federer (16), while players with higher standard deviations are Boris Becker (1), Stefan Edberg (2), Jim Courier (3), Marcelo Rios (7), Yevgeny Kafelnikov (9), Patrick Rafter (10), Juan Carlos Ferrero (14), Novak Djokovic (18), and Andy Murray (19).

With the exception of Moya, all of the players with lower standard deviations have won hard court majors (either the Australian Open or the US Open), suggesting that they performed well under pressure in the deciding sets on hard courts. In general, the players with higher standard deviations did not have as much success in terms of deciding sets on hard courts, but the oddities here are Djokovic and Murray, who lead the tour in winning percentage in deciding sets won. I would have to analyze this deeper, but their success on hard courts is directly related to their performance under pressure in the deciding set, so it would make sense if they represent the positive extreme of the standard deviation spectrum.

Covariance for Head to Head Among Court Surfaces

Another useful technique I learned in class that I applied in my project was **covariance**, which is defined to be a measure of the joint variability of two random variables. A positive covariance indicates that two variables show similar behavior, while a negative covariance indicates that two variables indicate dissimilar behaviors. Covariance could only be run on two matrices with identical dimensions, so I ran covariance on the head to head of all surfaces matrix against each of the head to heads on clay, grass, and hard courts. This generated 3 plots that I will describe more below. This also meant I could not use covariance (or correlation, which I will talk about more later) to compare the serve, return, and under pressure statistics to the head to head matchups.

For all the covariance plots, the percentage was a positive value from 0 to slightly under 0.20. The four bars on each bar plot represent the obtained results from the covariance of the head_to_head_all matrix against the head_to_head_clay, head_to_head_grass, and head_to_head_hard matrices. The right yellow bar is highest on hard courts, while the left purple bar appears to be highest on grass. The two inner bars are the highest on hard courts in terms of scale and actual height. These results suggest that grass and hard courts have a higher covariance than clay, so the head to heads indicate more similar behavior to the overall head to head matchups.

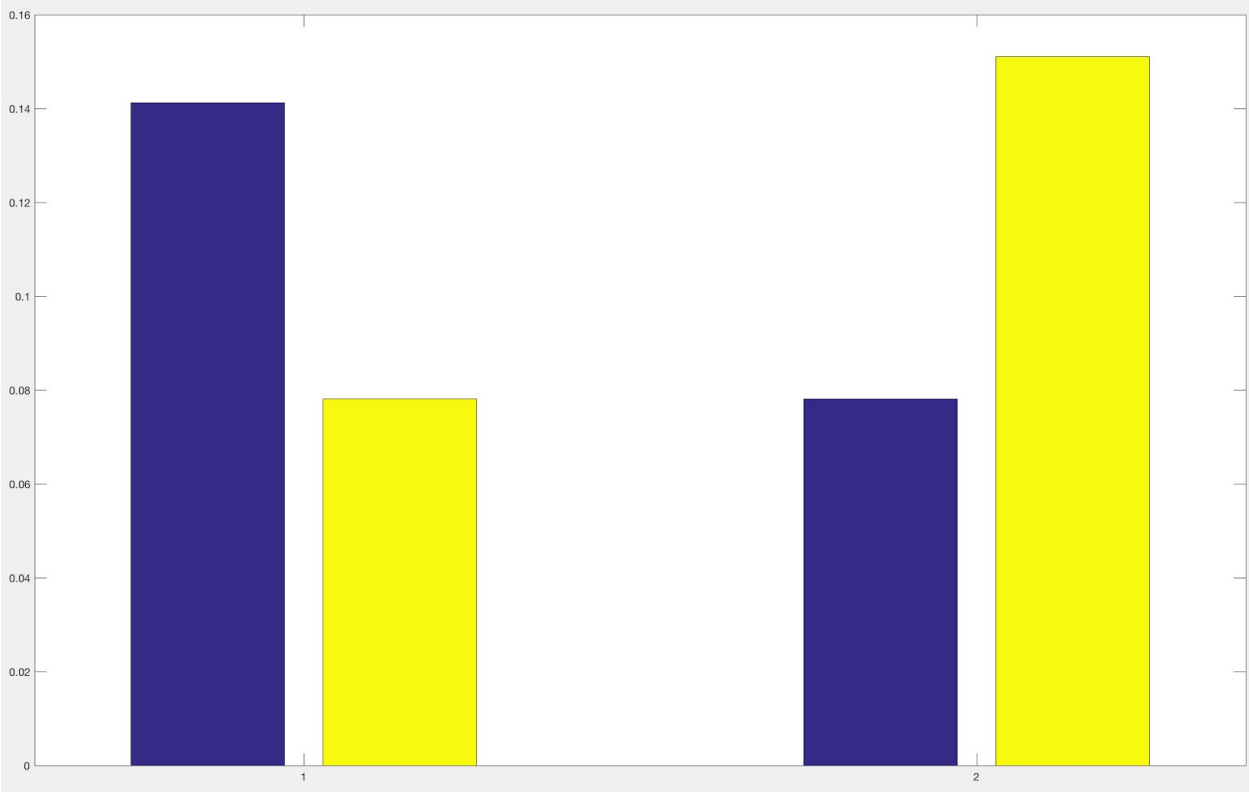


Figure 44. Covariance between head to head all and head to head clay.

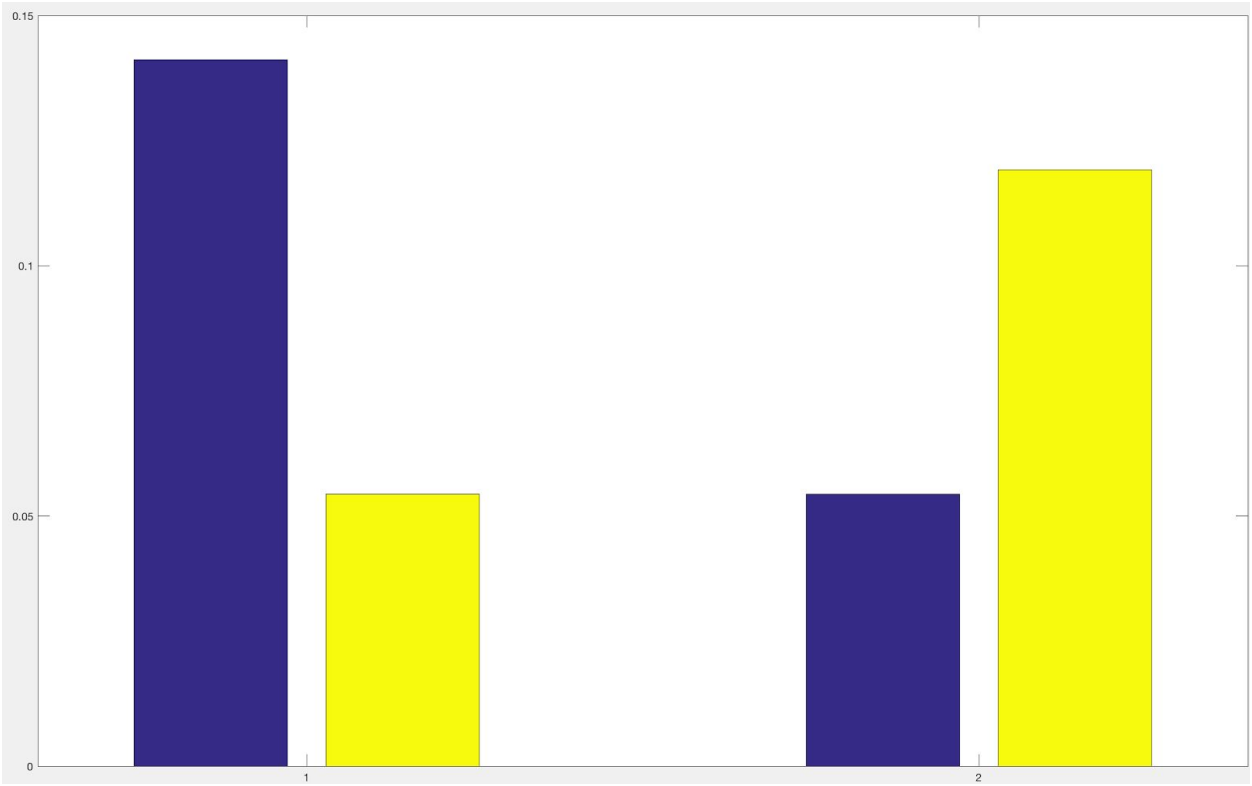


Figure 45. Covariance between head to head all and head to head grass.

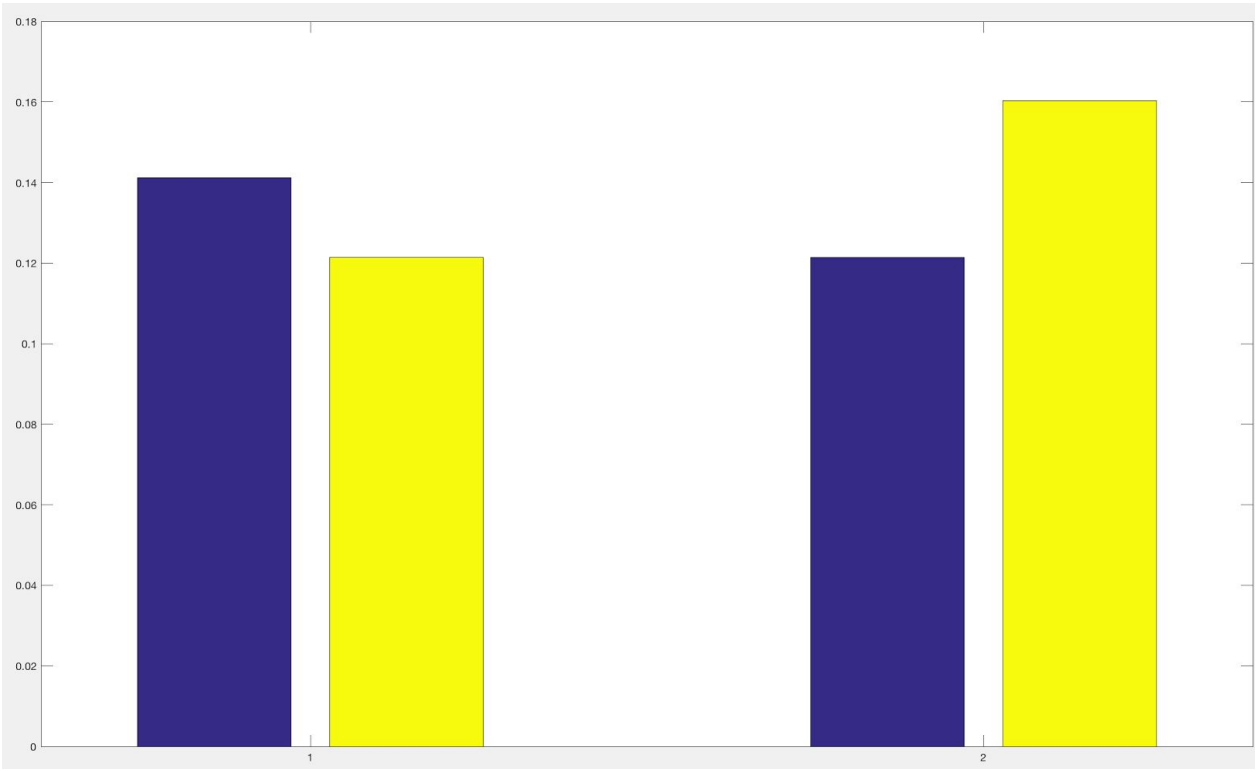


Figure 46. Covariance between head to head all and head to head hard.

Correlation for Head to Head Among Court Surfaces

In addition to covariance, I used **correlation** for the different court surfaces as well as the average on all surfaces. Although correlation is defined to be normalized covariance, I noticed that running correlation on my two 19 x 19 matrices for the court surfaces gave me another 19 x 19 matrix in return in stark contrast to running covariance on my two 19 x 19 matrices, which gave me 2 x 2 matrices. My findings are displayed below. A greater positive correlation means that there is for a positive increase in one variable, there is also a positive increase in the second variable. A greater negative correlation means that the variables move in opposite directions; for a positive increase in one variable, there is a decrease in the second variable. The positive correlations indicate that the players have met on the respective surfaces, and the negative correlations indicate that the players have never played on that surface. The higher the positive correlation, the greater the increased weight against that one's head to head overall yields in regard to the surface compared against.

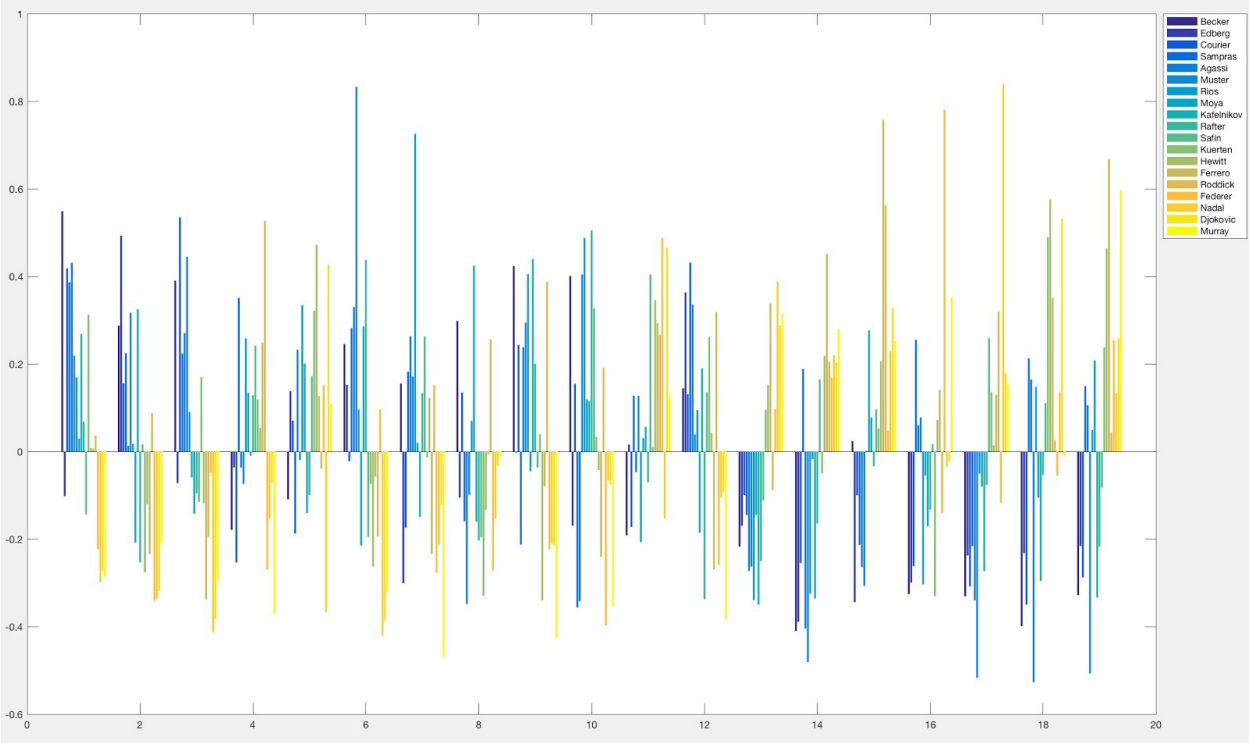


Figure 47. Correlation between head to head all and head to head clay. The leaders on this list are 6 (Muster), 7 (Rios), 15 (Roddick), 16 (Federer), 17 (Nadal), and 19 (Murray). All of them except Roddick had solid results on clay, so I believe Roddick's listing here is an abnormality.

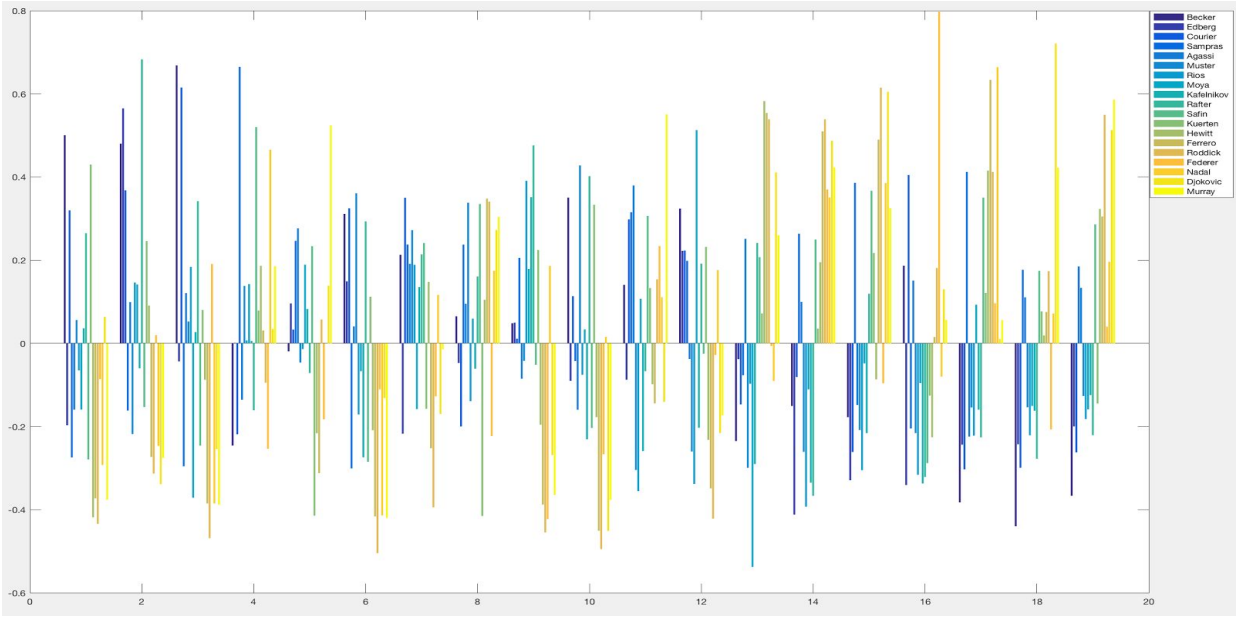


Figure 48. Correlation between head to head all and head to head grass. The leaders on this list are Stefan Edberg (2), Jim Courier (3), Pete Sampras (4), Andy Roddick (15), Roger Federer (16), Rafael Nadal (17), and Novak Djokovic (18). All of them have reached at least one Wimbledon final.

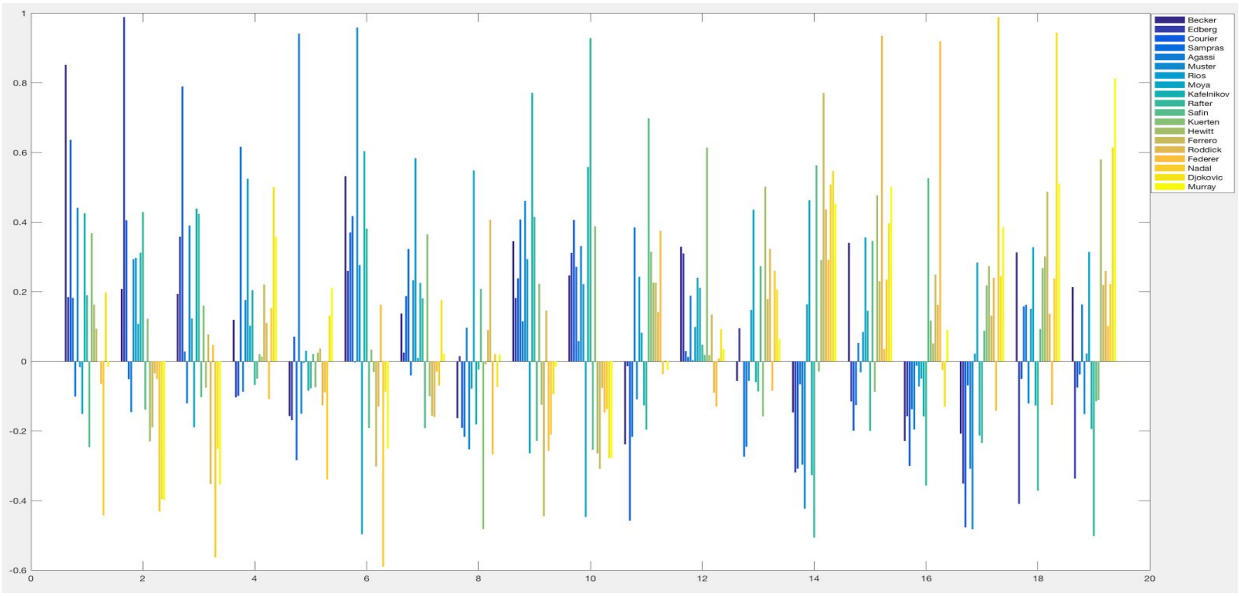


Figure 49. Correlation between head to head all and head to head hard. The leaders on this list are Boris Becker (1), Stefan Edberg (2), Andre Agassi (5), Thomas Muster (6), Patrick Rafter (10), Andy Roddick (15), Roger Federer (16), Rafael Nadal (17), Novak Djokovic (18), and Andy Murray (19). With the exception of Muster, all have won the US Open (one of the two hard court majors).

Predictions

While it was definitely interesting to see the results from k-nearest neighbor search as well as correlation, covariance, and analysis of distributions, all of these methods were rather flaky and inconsistent when coming up with a prediction algorithm. Therefore, I decided to use my own heuristic to come up with a judging criteria. This would be based off the following equation:

- $A(.30) + B(.30) + C(.30) + D(.10\%)$ where A = % of service games won, B = % of return games won, C = % of tiebreaks won, D = % of deciding sets won. I decided to weigh the first three categories more because the chance of going to a deciding set is dependent on those previous three factors.

From this equation, I would calculate a weighted average for each player and then take the standard deviation of all the players in the set. The player who has the higher weighted average will win more matches out of 10, and the heuristic is described as follows:

- $Matches\ won\ by\ player\ leading\ head\ to\ head = 5 + round(x)$
- 0 < x < 5, round(x) rounds to the nearest .5, and x is the number of standard deviations away from each other. $1\ \sigma = 2\ match\ wins$
- $Matches\ won\ by\ player\ losing\ head\ to\ head = 10 - Matches\ won\ by\ player\ leading\ head\ to\ head$

In the interest of saving space and time, I decided to only compare certain matchups that were the most speculated over the Internet. This led to a total of 12 matchups of interest: (1) Federer vs Becker, (2) Federer vs Edberg, (3) Federer vs Sampras, (4) Nadal vs Courier, (5) Nadal vs Muster, (6) Nadal vs Kuerten, (7) Djokovic vs Sampras, (8) Djokovic vs Agassi, (9) Djokovic vs Safin, (10) Murray vs Becker, (11) Murray vs Edberg, and (12) Murray vs Sampras. The MATLAB outputs are below.

Federer leads Becker (9 - 1)	Nadal leads Courier (9 - 1)
Federer leads Edberg (9 - 1)	Nadal leads Muster (8 - 2)
Federer tied with Sampras (5 - 5)	Nadal leads Kuerten (10 - 0)
Djokovic leads Sampras (6 - 4)	Murray leads Becker (7 - 3)
Djokovic leads Agassi (6 - 4)	Murray leads Edberg (7 - 3)
Djokovic leads Safin (9 - 1)	Sampras leads Murray (6 - 4)

```
>> prediction_federer
federer =
    0.6190

becker =
    0.5680

edberg =
    0.5680

sampras =
    0.6130

predict_federer =
    0.6190    0.5680    0.5680    0.6130

std_predict_federer =
    0.0278
```

```
>> prediction_nadal
nadal =
    0.6520

courier =
    0.5510

muster =
    0.5890

kuerten =
    0.5280

predict_nadal =
    0.6520    0.5510    0.5890    0.5280

std_predict_nadal =
    0.0542
```

```
>> prediction_djokovic
djokovic =
    0.6220

sampras =
    0.6120

agassi =
    0.5990

safin =
    0.5400

predict_djokovic =
    0.6220    0.6120    0.5990    0.5400

std_predict_djokovic =
    0.0367
```

```
>> prediction_murray
murray =
    0.6050

becker =
    0.5680

edberg =
    0.5680

sampras =
    0.6130

predict_murray =
    0.6050    0.5680    0.5680    0.6130

std_predict_murray =
    0.0239
```

Figure 50. Head to head predictions for the Big Four on their best surfaces.

Conclusion

In the end, I felt that the heuristic that I came up with closely approximated how these fantasy matchups would turn out because I used the most important statistics in tennis. Even though some of the other methods I learned about in class or researched online did not work as successfully, it was a fantastic experience for me to learn how to apply some of these concepts in the world of data science to learn more about a topic I love.

I definitely learned a lot more about intriguing tennis statistics that I never knew about before such as Andy Roddick's fantastic tiebreak winning percentage on clay (75%), Stefan Edberg serving more double faults than aces on clay, and Marat Safin's poor record at saving break points (62.70%). A lot of my findings also reinforced some higher-level ideas that I already knew about. I knew that clay courters like Gustavo Kuerten or Jim Courier had lower winning percentages of 1st serves because of their style of play or that Federer had the worst return of serve statistics in the Big Four, but these statistics provided empirical evidence to back up my ideas.

From the data scientist perspective, I gained a lot of experience working extensively with MATLAB and learning different types of mathematical modeling techniques such as analysis of distributions (mean and standard deviation), covariance, correlation, and k-nearest neighbors search. Additionally, I got a lot more comfortable performing basic matrix operations such as transposes, plotting different types of graphs (bar, line, and scatter plots), and loading .csv files into MATLAB projects. These skills will definitely serve me well in the long run if I want to pursue a career in data science.

Of course, future plans on how to improve this project were something I had in mind. Given more time, I would definitely try to analyze all the possible matchups using the heuristic I came up with. I would also include a much larger head to head sample size instead of just having the 19 world number ones since 1991. This idea would be quite ambitious and would take a lot of time, but I think if there were enough other tennis enthusiasts, I could collaborate with them on expanding my data collection.

In terms of improving the heuristic I came up with the predict tennis matches, an idea I had for future research could be to use more advanced mathematical modeling techniques like Markov Chains to predict matchups given the current game score as well as other factors. This research paper that I found online outlines these techniques: http://www.doc.ic.ac.uk/teaching/distinguished-projects/2015/m.sipko.pdf