Game On: The Rivalry Between Men's and Women's Tennis

Ava Scharfstein

Dartmouth College 2024

BA in Mathematical Data Science and Computer Science

Repository

Abstract

This project is motivated by gender disparities, namely the treatment and discourse of male versus female athletes in tennis' most prestigious international tournaments. After mining match-level data from the Sportradar Tennis API and historical rankings,[1] I investigated the entertainment value of tennis matches by gender through an exploratory data analysis and hypothesis testing. Findings suggested that men's tennis showcases greater competitiveness with players having similar match-level statistics, particularly excelling in service games,[2] while women's tennis leans towards risk-taking, producing more winners and longer games with more unforced errors. Thus, the data shows that neither gender's matches are statistically more entertaining, as they offer different dynamics and value to viewers.

Background

Despite the tremendous strides made by equal pay activist Billie Jean King,[3] there remains a striking disparity in the treatment of male and female players in professional tennis. In many highly prestigious international tournaments, women earn a smaller share of winnings compared to their male counterparts.[4] A glaring example from the 2022 Roland Garros tournament underscores this inequity. Amélie Mauresmo, the tournament director, allocated just one women's match to the prestigious nighttime slot, while prioritizing nine men's matches. Women also receive second billing in mixed tournaments – less desirable schedules on smaller courts.[5] These inequities hint at the systemic gender discrimination within professional sports: a perception that men's matches hold greater attraction and appeal than women's.[6]

---

[1] Rankings are merit-based methods used by the Association of Tennis Professionals (ATP) and Women's Tennis Association (WTA) that determine the qualification for entry as well as seeding in all tournaments.

[2] In tennis, players take turns serving the ball to start a point. The service game specifically refers to the period when a single player serves the ball.

[3] Billie Jean King is the former No. 1 tennis player with 39 major titles and considered one of the best players of all time. She received a Presidential Medal of Freedom for her advocacy for women in sports and LGBTQ rights. In particular, King championed the effort for equal pay for Grand Slams threatening to boycott the US Open until pay was equalized in 1973 (Jean King). In 2023, the Women's Tennis Association (WTA) announced that by 2033, WTA-ATP 1000 and 500 level tournaments – the levels directly below the Grand Slams – the prize money will be equal (Futterman, "Pay Equity Is Officially Coming to Tennis. Eventually").

[4] For example, the Western & Southern Open in Ohio pays women 63 cents on the dollar compared to men and the Italian Open 46 cents on the dollar (Futterman, "The Same Work but a Lot Less Pay for Women. Welcome to Tennis in 2023").

[5] Less desirable schedules refers to match timings or court assignments that may not be ideal for players or spectators. Smaller courts are generally less desirable because they have fewer seating capacities, limiting the number of spectators who can attend and reducing the atmosphere of the match. Further, smaller courts and early morning or late night matches may also receive less media coverage and attention compared to matches on larger, show courts like center courts or matches played at peak times.

[6] Novack Djokovic, world No. 1 men's singles player, was criticized at Indian Wells for saying men "should get awarded more" because "the stats are showing that we have much more spectators" ("ESPN"). Indeed, it's a myth that fans are always more interested in men's tennis. In 2013 and 2014, the women's US Open final had higher TV ratings than the men's final (Futterman, "The Same Work but a Lot Less Pay for Women. Welcome to Tennis in 2023").

Indeed, the chairman and CEO of the Women's Tennis Administration (WTA) claims that the market values men's tennis more than women's, particularly in the context of sponsorships and media rights.[7] This financial inequity means that aspiring players at lower levels – who are reliant on sponsorship for sustenance in their professional tennis journey – have a diminished chance of sustaining their career. This critical issue of gender inequities serves as the driving motivation for my cumulative final project, outlined below.

Project Goal

This project takes a data-driven approach to investigate, quantify, and interpret the conception of *appeal* in mens versus women's tennis. What makes a match appealing is subjective, so I chose to quantify this notion of entertainment through competitiveness. Competitiveness occurs when a match between two opponents is closely contested and evenly balanced. The outcome of a competitive match may be uncertain, with momentum shifting between players and neither having a significant advantage over the other. Spectators and fans are typically engaged and enthralled by the level of play and the excitement of not knowing who will emerge victorious. Ultimately, conducting this analysis of tennis *appeal* by sex may be used to counter aforementioned gender-based biases that prevail in media rhetoric and amongst player treatment.

Data

Data Extraction

As a former data science intern at Sportradar, I was able to obtain free access to Sportradar's Tennis application programming interface (API). This API contains detailed match information and an array of supplementary data. With help from the API documentation, I built a streamlined pipeline in Python that navigated Sportradar's complex and nested API to extract match-level data for close to 1000 singles[8] Grand Slam[9] matches. As outlined in appendix A, the data extraction process involved iterative endpoint identification, request building, data retrieval, parsing, and transformation, and required meticulous error handling and optimization for efficiency and accuracy. Since tournaments only seed the top players, I also scraped historical ATP (Association of Tennis Professionals, for Men) and WTA (Women's Tennis Association) rankings online to supplement the missing data. Using the BeautifulSoup package, I built

[7] One of the biggest barriers to garnering more sponsorships and media coverage is the fact that the WTA does not require player participation in all tournaments below the Grand Slam level. This thereby diminishes the marketing appeal of sponsoring female players, yielding smaller funding and less female player sponsorships. Consequently, the WTA brings in substantially less money than the men's circuit and therefore less money to contribute toward prizes. But, if equal prize money is important to the tournament owners, then they could choose to pay the difference (Futterman, "The Same Work but a Lot Less Pay for Women. Welcome to Tennis in 2023").

[8] Singles tennis involves a single player on each side of the court, competing head-to-head.

[9] Grand Slams are the most prestigious and watched tournaments in tennis, including the US Open, Wimbledon, the Australian Open, and Roland Garros.

time-stamped datasets of men's and women's rankings. See appendix B for more details. Despite the brevity of this section, I want to underscore the extent of time and effort that went into data extraction, as well as the other data preparation processes. See data_aquisition folder for code.

Data Cleaning

The first step in transforming the extracted data into one concise and clean dataset was merging all the data I extracted into one single dataset. After merging, I undertook a lengthy data-cleaning process, which included handling missing data, normalizing continuous variables, and aggregating, removing, or renaming variables to minimize noise and improve clarity and consistency. The process was iterative, especially as new issues arose during the initial exploratory analyses. See Appendix C for details. Ultimately, I concluded with a DataFrame in which each row is a match and each column is specific details about that match, including individual player statistics. See preprocessing folder for code.

	name_home	aces_home	breakpoints_won_home	...	name_away	aces_away	breakpoints_won_away
0	Kudermetova, Veronika	7.000000	0.500000	...	Kanepi, Kaia	1.500000	0.000000
1	Samsonova, Liudmila	2.500000	1.500000	...	Bogdan, Ana	3.000000	1.500000
2	Parks, Alycia	2.500000	2.000000	...	Friedsam, Anna-Lena	1.000000	0.500000
3	Rakhimova, Kamilla	0.666667	1.666667	...	Bucsa, Cristina	2.000000	2.000000
4	Pegula, Jessica	1.000000	1.666667	...	Davis, Lauren	0.666667	0.666667

Figure I: A snapshot of the cleaned, normalized data

Methods

In order to evaluate the competitiveness between two players, I considered the difference between their individual statistics in a match (eg. aces, unforced errors, service points won). Indeed, when a certain player's stat is equivalent to their opponent – yielding a net zero difference – the players may be considered well-matched and competitive. Conversely, when one player's stat is much larger than their opponent – yielding a net non-zero difference – the match may be considered to be dominated by one player and less competitive – in regards to that specific statistic. For simplicity, I will denote this difference between players' stats as *stat-difference* where the stat is a feature in the aforementioned dataset, such as number of aces[10] or number of service points won. The dataset is defined such that the home-player is always the higher-ranked player and the away player is always the lower-ranked; thus, a negative stat-difference means the lower-ranked player has surpassed their opponent for that stat. Read more details on feature engineering and other methods in Appendix D.

For twenty four different features (stat-differences and other overall match stats),[11] I conducted four hypothesis tests that compared the male and female distributions. I also leveraged

[10] An ace is a serve that goes untouched by the receiver, winning the point for the server.

[11] A list of all features: aces, breakpoints_won, double_faults, first_serve_points_won, first_serve_successful, games_won, max_games_in_a_row, max_points_in_a_row, points_won, points_won_from_last_10, second_serve_points_won, second_serve_successful, service_games_won, service_points_lost, service_points_won, tiebreaks_won, total_breakpoints, errors, unforced_errors, winners, avg_set_length, avg_game_length, avg_set_diff, avg_points_per_game,

empirical cumulative distribution functions, qq-plots, and histograms to aid my analyses. I conducted a similar analysis for the distribution of *stat-sums* (stat totals for both players) to aid in the interpretation of the stat-differences. See Appendix D's feature engineering section for a further explanation of stat-sums.

A. Kolmogorov–Smirnov test: a nonparametric test which assesses the whether two distributions are derived from the same underlying distribution
B. T-test: assess the equality of the means between two groups, assuming normativity
C. Mann Whitney U test: a non-parametric test used to compare two independent samples, generally as assessing equality of the medians (McClenaghan).
D. Levene test: assess the equality of the variances between two or more groups ("1.3.5.10. Levene Test for Equality of Variances").[12]

Results

The Levene tests indicated statistically significant differences in variance between male and female distributions for 21 out of 24 features, with the variance being consistently smaller for men. Given that variance quantifies the dispersion of player similarity, a larger spread suggests less similarity among players and consequently, less competitive opponents for a particular statistic. For smaller variance, similarity scores are clustered closely around the mean, indicating a more consistent performance or style of play among opponents. Therefore, these findings imply that, generally, men's matches may exhibit higher competitiveness compared to women's, as reflected by the narrower variance in male player statistics.

Notable features include unforced errors (Levene p-value = 6.7449e-16) and winners (Levene p-value = 2.54959e-10), where men exhibit significantly smaller variance. Interestingly, the only feature in which men's variance exceeds women's is aces (Levene p-value = 1.82336e-10), suggesting that there is more variability in men's serving. See *figures II,III, and IV* on page 7.

Conducting T-tests, statistically significant differences in means between male and female stat-difference distributions were observed for certain features. The mean, which represents the average similarity between opponents, is crucial in understanding the dynamics of competitiveness within matches. Positive mean values suggest greater overall dominance of the home/higher ranked players, whereas negative values indicate dominance for away/lower ranked players for a specific statistic.

T-tests for unforced errors (p = 0.00281319) and winners (p = 0.000221226) revealed statistically significant differences across gender. Unforced errors generally occur when a player fails to execute a shot successfully without immediate pressure or interference from their opponent. While they may be attributed to a player's inconsistency, unforced errors may also be an indication of risk taking. In fact, there is a higher degree of risk involved with hitting a quality shot, such as a winner – a shot unreachable by an opponent. Thus, the results may suggest that

[12] I normally would have used the F-test but only the Levene test was implemented in the stats.scipy package.

higher ranked female players are more willing to take risks compared to high ranked male players. On the other hand, lower ranked male players are more willing to take risks compared to lower ranked female players. The analysis also revealed a statistically significant difference in winners-difference between men and women. Women exhibited a greater mean in winners-difference compared to men (T-test p-value = .00022126), suggesting that higher ranked female players are more likely to hit winners than their opponents compared to men. See *figures II and III* on page 7.

To supplement these results, I also compared stat-sums and overall match stats between men and women. Across all of the serving-related features (eg. first serve points won, second serve points won, aces, service games, etc), the Mann Whitney U tests revealed statistically significant differences between the medians of the male and female distributions, with the median higher for men's stat-sum (see *Table 1)*. This is an unsurprising result, since men are known to be stronger servers, more likely to hold their service games.[13] Holding one's service game is essential for maintaining a lead or staying even with your opponent in a match. See *Figure V*.

On the other hand, the Mann Whitney U tests revealed significant differences between the medians of the male and female distributions, with the median higher for women for several key features. First, as seen in *figure VI*, the median for total break points is greater for women (p-value = 1.9145e-32). Breakpoints are crucial moments in a game where one player is on the verge of winning a game served by their opponent. They are significant because they can lead to momentum shifts and often play a decisive role in determining the outcome of a set or match. The higher median for total break points suggests that service games in women's matches are less predictable, with players facing more challenges to hold their serves, compared to men. This could be attributed to the aggressive return games of female players or their ability to capitalize on opponents' weaknesses. Next, there were also greater unforced errors (p-value = 4.4393e-15) and winner totals (p-value = .01198), a further indication of higher risk taking among female players (see *figure VIII*). Finally, there was a statistically significant difference in the medians for the distribution of average game length (p-value = .00238) and average points per game (p-value = 2.648e-12) by sex. Higher median points per game and game length highlights a greater intensity and competitiveness of women's matches, as players engage in longer and more closely contested games. Ultimately, these findings underscore that there is more unpredictability in service games, more risk taking, more impressive shots, and longer, more intense games for women's matches on the whole. See figures below and Appendix E for more details. For space sake, I did not include all of the figures I produced, since there were close to two hundred.

[13] In tennis, "holding your service game" refers to successfully winning the game when you are serving.

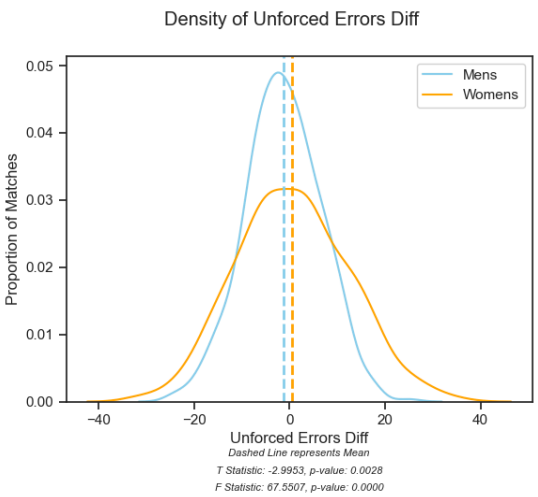


Figure II

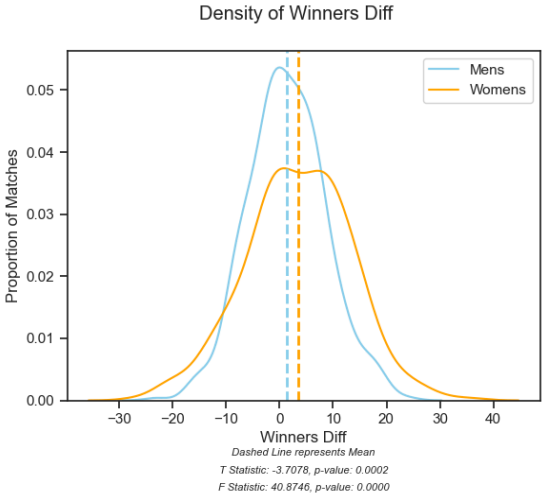


Figure III

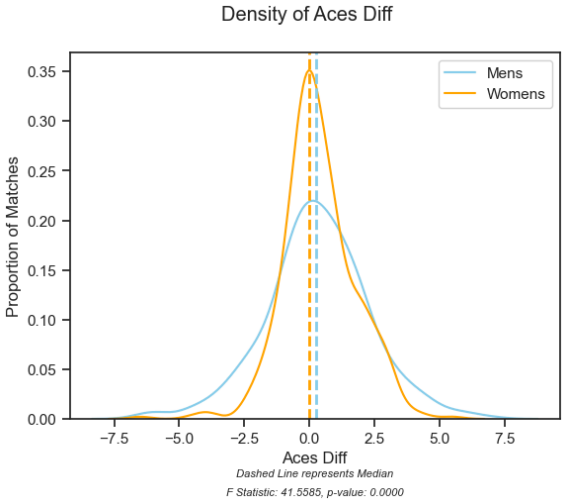


Figure IV

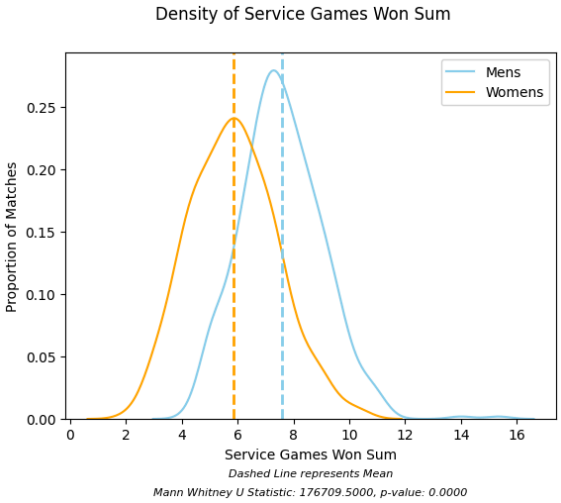


Figure V

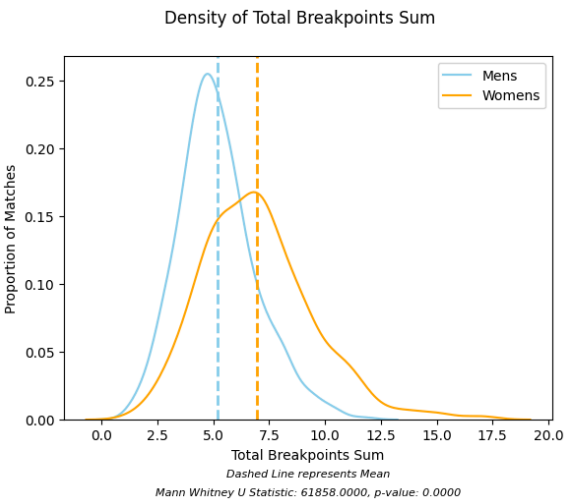


Figure VI

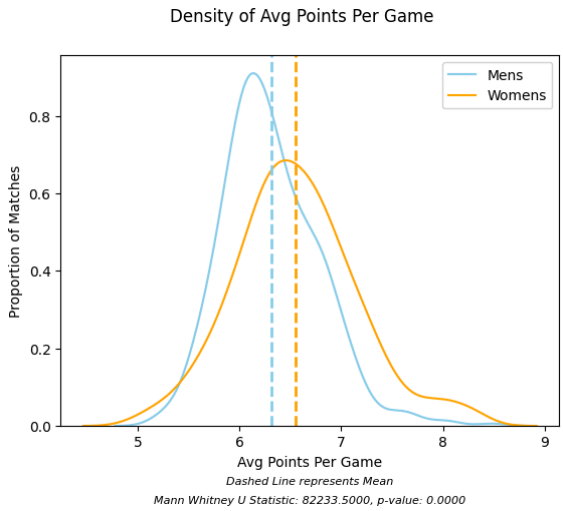


Figure VII

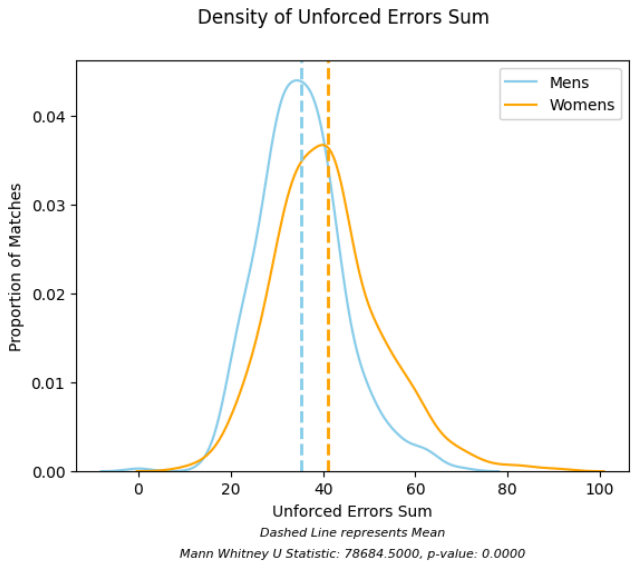


Figure VIII

Stat-sum/Avg Feature	P-value	Men > Women
1st serve points won	9.6482e-23	True
2nd serve points won	1.1289e-12	True
Aces	8.3515e-65	True
Service games won	1.6928e-54	True
Total break points	1.9145e-32	False
Unforced errors	4.4393e-15	False
Winners	.01198	False
Avg game length	.00238	False
Avg points per game	2.648e-12	False

Table 1: Mann Whitney U P-Value Summary Table

Future Work

The breadth of data analysis opportunities provided by the Sportradar API is truly impressive. There are numerous avenues I would love to explore given additional time and resources. Firstly, incorporating data on viewership, match attendance, and prize money could have certainly enhanced my analysis and further connected my project motivation to my analysis. I also would have liked to explore additional statistics such as the number of deuces, matchpoints, percentage of points won on the opponent's serve, and number of consecutive

games won. They could have provided further depth to my analysis, capturing nuances in match dynamics across genders.

Next, examining the frequency of upsets – when lower ranked players defeat higher ranked players – may uncover intriguing patterns regarding the level of competition and unpredictability in men's and women's Grand Slam tournaments. In addition, variability across different Grand Slam tournaments and differences in early versus later rounds of tournaments could shed light on the dynamics of tournament play and the factors influencing match outcomes.

Furthermore, instead of taking a match-level approach, it would be interesting to analyze the playing styles, strengths, and weaknesses of the top overall 20 players from each sex. The results might offer valuable insights into the factors contributing to their success and entertainment value as well as identify differences between male and female players. Even more, analyzing player consistency across tournaments and assessing the variance in individual performance could provide valuable insights into the consistency and predictability of top players.

Conclusion

While the results show that men's tennis tends to exhibit greater competitiveness in terms of players having similar statistics, it's worth noting that dominance can manifest in various ways, and one player may excel in certain aspects while another excels in others. Men typically outperform in service games and exhibit greater consistency, but are more prone to forced errors. Conversely, women tend to outperform in risk-taking, producing more winners and longer, more unpredictable and closer games, albeit with more unforced errors. Ultimately, I claim that neither gender's matches are inherently more entertaining; they simply offer different dynamics and value to viewers. Therefore, the entertainment value of men's and women's tennis matches cannot be definitively ranked against each other, as they represent distinct and equally compelling experiences.

Acknowledgements

Sportradar is a sports data analytics company that I interned for from January-March (23W) and June to August 2023 (23X). They kindly granted me access to their Tennis API for free. Thank you to Molly Labelle and the NY Office at Sportradar for their help in facilitating this project and helping me resolve API bugs throughout the process. Thank you also to Professor Peter Mucha for his support despite taking leave during 24W and Professor Demidenko for his feedback on my initial draft.

Data

"ATP Rankings Singles." *ATP Tour*, www.atptour.com/en/rankings/singles?RankRange=1501-5000&Region=all&DateWeek=2024-02-05. Accessed 29 May 2024.

"Coverage Matrix." *Sportradar*, coverage-matrix.sportradar.com/. Accessed 29 May 2024.

"ESPN." *ESPN.com*, 21 Mar. 2016, www.espn.com/tennis/story/_/id/15031425/novak-djokovic-questions-whether-women-deserve-equal-pay-tennis.

"Overview Tennis API." *Sportradar*, developer.sportradar.com/tennis/reference/overview. Accessed 29 May 2024.

"WTA Singles Rankings." *Women's Tennis Association*, 2023, www.wtatennis.com/rankings/singles.

"WTA Tennis Rankings." *Tennis Explorer*, www.tennisexplorer.com/ranking/wta-women/. Accessed 29 May 2024.

External Sources

"1.3.5.10. Levene Test for Equality of Variances." *Itl.nist.gov*, itl.nist.gov/div898/handbook/eda/section3/eda35a.htm.

Futterman, Matthew. "Pay Equity Is Officially Coming to Tennis. Eventually." *The New York Times*, 27 June 2023, www.nytimes.com/2023/06/27/sports/tennis/wta-women-pay-equity-prize-money.html.

---. "The Same Work but a Lot Less Pay for Women. Welcome to Tennis in 2023." *The New York Times*, 10 May 2023, www.nytimes.com/2023/05/10/sports/tennis/women-men-prize-money.html.

amoeba. "How to Reverse PCA and Reconstruct Original Variables from Several Principal Components?" Cross Validated, Apr. 2017, stats.stackexchange.com/questions/229092/how-to-reverse-pca-and-reconstruct-original-variables-from-several-principal-com.

Jean King, Billie. "Billie Jean King Enterprises." *Billie Jean King Enterprises*, 2014, www.billiejeanking.com/.

McClenaghan, Elliot. "Mann-Whitney U Test: Assumptions and Example." *Informatics from Technology Networks*, 6 July 2022, www.technologynetworks.com/informatics/articles/mann-whitney-u-test-assumptions-and-example-363425.

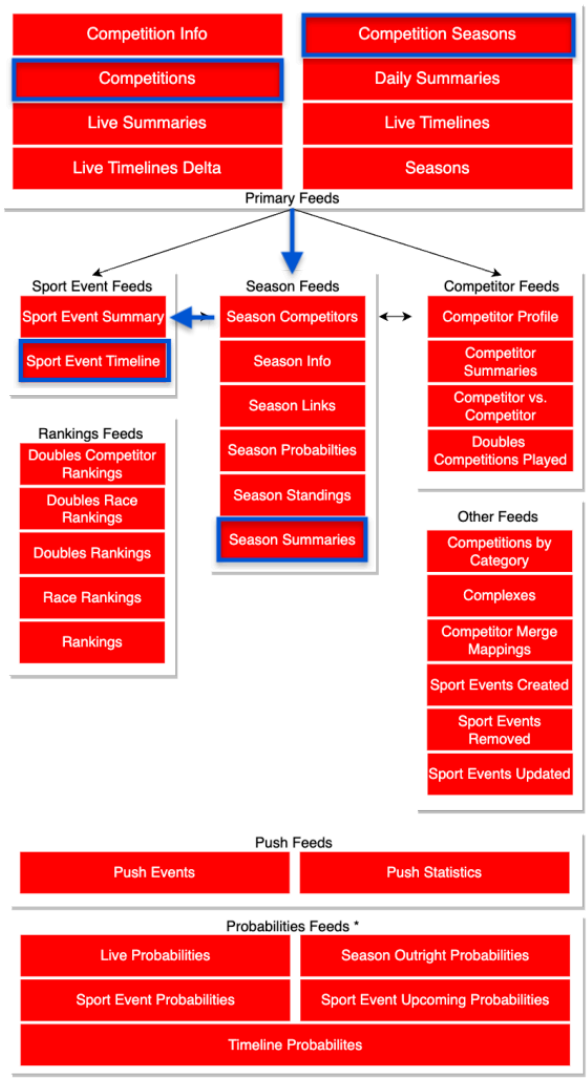
Siegrist, Kyle. "5.13: The Folded Normal Distribution." *Statistics LibreTexts*, 5 May 2020, stats.libretexts.org/Bookshelves/Probability_Theory/Probability_Mathematical_Statistics_and_Stochastic_Processes_(Siegrist)/05%3A_Special_Distributions/5.13%3A_The_Folded_Normal_Distribution. Accessed 29 May 2024.

Appendices

Appendix A: Sportradar API Data Extraction

About the API

The Tennis API offers real-time and point-by-point match scoring, along with supplementary data, covering over 4,000 competitions for both men's and women's tennis (eg. tournaments such as Wimbledon Men's Singles) ("Overview Tennis API"). It maintains consistency in structure, format, and behavior with other General Sport APIs. The breadth of the API is illustrated on the *figure IX* to the right: Primary feeds include seasons, competitions, player data, and real-time scores and additional feeds provide various stats such as win probabilities for matches, player profiles, historical results, post-match corrections, and in-depth match statistics. Note that the coverage for detailed stats and probabilities is limited to a small subsection of the competitions covered.



In order to access the detailed match-level data, I had to traverse through the API, as indicated by the blue rectangles and arrows.[14] First, I utilized the Competitions and Competition Season primary feeds to acquire the Competition Season Ids for Grand Slam tournaments by year. I used the Competition Season Ids to extract Season Summaries, a list of all of the matches (or sport events) in that competition. Finally, using the Sport-Event-Ids, I acquired the Sport Event Timeline for detailed match-level information. I built a streamlined pipeline to traverse from competitions to seasons to matches because I was extracting data for hundreds of matches, which entailed multiple thousands of requests. Details are elaborated further below.

It's worth noting that a substantial part of this project involved learning to navigate the API's structure, determining relevant endpoints, and recognizing its limitations. In fact, most of fall term and a week or two of winter term (September to January) was dedicated to experimenting with Sportradar's API sandbox.

[14] Traversing through an API involves using a set of predefined commands to communicate with a server and retrieve or send data, similar to navigating through different pages of a website using links.

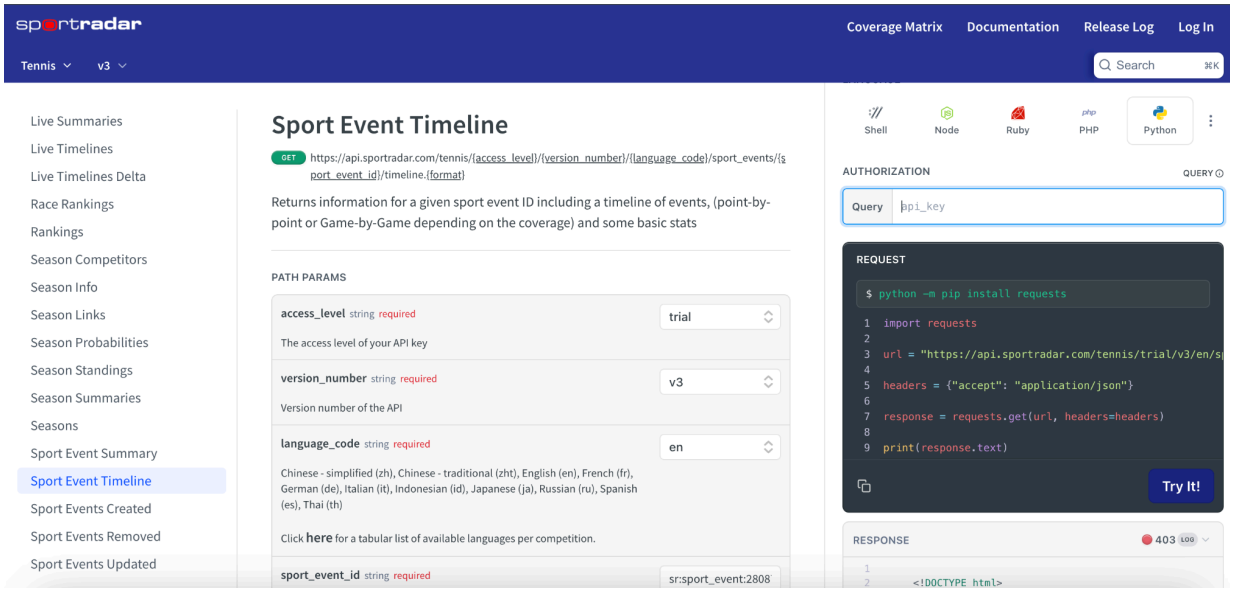


Figure X: Example of API Sandbox

Data Extraction

Initially, I had planned to extract data from a variety of types of tournaments (Grand Slams, ATP and WTA Finals, 1000-level tournaments, and 500-level tournaments). However, I found that the Sportradar API did not have play-by-play coverage of any tournament aside from the Grand Slams.[15] Furthermore, since the Tennis API was only recently released by Sportradar, the historical coverage of play-by-play data was limited to 2022 and later. These limitations significantly reduced the amount of data I initially thought I could extract from the API.

First, using the Sportradar coverage matrix,[16] I acquired a list of Competition-Ids for tournaments of interest (eg. Grand Slams) ("Coverage Matrix"). For each competition, I then queried a list of Competition-Season-Ids, which are for competitions for a given year. The historical data only goes back three years, so I only had ids for 2022-2024. For each Competition-Season, I retrieved a Competition-Summary, which included a list of every match played in that tournament and some basic statistics on that match. I filtered the matches that were play-by-play and had enhanced-stats only, and then saved their ids to a pickle file. This was important because it was computationally extensive to run and I had a limited number of queries.[17]

---

[15] Play-by-play data is timestamped point by point scoring information for a match. Matches with the play-by-play feature also have detailed match statistics for each player. The enhanced statistics and play-by-play features were the bulk of my content for data analysis

[16] A coverage matrix for the Sportradar Tennis API is a structured tool used to systematically track and analyze the extent to which endpoints for particular competitions (tournaments) are covered/captured. In other words, full coverage means there is no missing data, partial coverage means there is some missing data, and no coverage means there is no data. Each competition has a corresponding id which is used for querying data and information.

[17] Due to API request limits, I had to be very efficient and precise in my API calls, to avoid running out of requests. On two occasions I had to contact my former employer to reset my API key so that I could continue querying. Eventually, they gave me a production key which had around 3000 calls a month for three months.

I then iterated through that list of play-by-play matches. For each match, I extracted detailed player-specific stats that were pre-calculated (eg. aces, total breakpoints, breakpoints won, games won, etc). Further, since these Sport-Event-Timelines had timestamped point-by-point data, I was able to compute other valuable statistics. These included, set lengths, game lengths, average points per game, match length. These computations were tricky and took me a while to figure out because there frequently were suspensions in the middle of matches due to injuries, weather, or because the time between a game/set's end and the recording of the score was non-zero. Another roadblock included early retirement (matches that ended early), which I resolved by simply omitting those matches. I figured it wasn't fair to include the data for those matches because they're not representative of a complete whole match and they could skew or mislead the results. Further, I ran into a lot of 403 errors – access to the requested resource is forbidden – too many queries per second. I resolved by leveraging the Time package's sleep function in between many consecutive requests.

All in all, the data extraction process was not linear. I discovered multiple issues (eg. outlier set length) during exploratory data analysis and had to return and rerun the data extraction process multiple times over. With the self-computed statistics, specifically regarding time, I rigorously checked my code to guarantee accuracy.

Reflection

Despite the considerable time (four weeks of the ten week term) I invested in navigating the Sportradar API for data extraction—identifying data sources, addressing missing data, and handling inconsistencies—these aspects were not covered in the data science curriculum (nor my computer science curriculum) at Dartmouth. Yet, I firmly believe that these skills are aligned with the practical challenges encountered in real-world data science scenarios. The hurdles related to data acquisition and scarcity frequently pose significant barriers for data scientists. The ability to adeptly navigate diverse platforms for data extraction is nearly as crucial as the data analysis itself. Data science inherently demands adaptability, quick thinking, and the capacity to generate innovative ideas and alternative solutions in response to dynamic challenges.

Back to Data.

Appendix B: Scraping Historical Rankings

The ATP and WTA rankings are systems used to rank professional tennis players based on their performance in tournaments over a designated period. These rankings provide a numerical representation of a player's standing within the global tennis community, taking into account their results in various competitions. The rankings consider factors such as match wins, tournament victories, and the caliber of opponents faced, serving as a benchmark for players' competitiveness and skill levels on the ATP and WTA tours. The rankings update weekly and are used to seed players in tournaments.

Of the 128 players in Grand Slam tournaments, only the top 32 players are seeded based on their pre-tournament ATP and WTA rankings.[18] Consequently, I wanted to acquire pre-tournament rankings for all players to fill in the missing data. I scraped the official ATP rankings from their website for the players ("ATP Rankings Singles"). For women's rankings, I ran into an issue extracting data from the official WTA website because the data is loaded dynamically using JavaScript after accessing the initial page ("WTA Singles Rankings"). While debugging, I tried to use Selenium, a browser automation python package, to manually change the date and update the data on the URL. However, I encountered an 'ElementNotInteractableException', which typically arises when an element (in this case, a button to update the data) is present but cannot be interacted with, often due to factors like visibility, clickability, overlapping elements, or timing issues. After hours of debugging, I had made no progress, so I turned to alternative solutions, including trying to convert the PDF version of the website into a dataframe. After many failed attempts, I scoured the internet for another website that had WTA rankings ("WTA Tennis Rankings"). I cross-checked some of the rankings by date with the official WTA website to verify its credibility. This website was easier to scrape, using BeautifulSoup and Selenium.

Back to Data.

[18] Seedings are rankings used to separate the top players in a draw so that they will not meet in the early rounds of a tournament.

Appendix C: Data Cleaning

Merging

The first crucial set of creating a complete, clean dataset is merging the extracted four datasets – ATP rankings, WTA rankings, men's matches, and women's matches – together. This process was a lot less simple than I had expected, most notably due to discrepancies in data formatting. Indeed, while the names in the Sportradar match datasets were formatted as 'Last, First', ATP rankings were formatted as 'First Last' and WTA rankings were formatted as 'Last First' (note the lack of comma in the latter two). This small difference ended up being a huge issue for players with three or more words in their names: I had to figure out how to distinguish between a first and last name in order to place the comma to replicate the 'Last, First' format.[19] Other issues included players where part of their names were misspelled.

At a high level, I resolved the missing player names in the merged dataset by searching for similar names in the rankings datasets. I split the missing names into parts and tried to find matches for each part. If a match was found, I updated the missing name with the correct one. If not, I tried again using only the last name. The corrected names were then merged back into the original dataset, updating the missing values. This process helped to ensure the completeness and accuracy of player names in the dataset. In the end, I was able to successfully merge rankings for more than 4000 players, unsuccessfully merging only five players.

	name_home	seed_home	name_away	seed_away	date
0	Thompson, Jordan	NaN	Nakashima, Brandon	NaN	2023-06-26
1	Musetti, Lorenzo	14.0	Varillas, Juan Pablo	NaN	2023-06-26
2	Baez, Sebastian	NaN	Barrios Vera, Marcelo Tomas	NaN	2023-06-26
3	van Assche, Luca	NaN	Karatsev, Aslan	NaN	2023-06-26
4	Rublev, Andrey	7.0	Purcell, Max	NaN	2023-06-26

	player	rank	week
0	Carlos Alcaraz	1	2023-06-26
1	Novak Djokovic	2	2023-06-26
2	Daniil Medvedev	3	2023-06-26
3	Casper Ruud	4	2023-06-26
4	Stefanos Tsitsipas	5	2023-06-26

Figure XI: ATP Rankings Dataset and Men's match dataset

	name_home	seed_home	name_away	seed_away	date
0	Kudermetova, Veronika	12.0	Kanepi, Kaia	NaN	2023-06-26
1	Bogdan, Ana	NaN	Samsonova, Liudmila	15.0	2023-06-26
2	Parks, Alycia	NaN	Friedsam, Anna-Lena	NaN	2023-06-26
3	Bucsa, Cristina	NaN	Rakhimova, Kamilla	NaN	2023-06-26
4	Pegula, Jessica	4.0	Davis, Lauren	NaN	2023-06-26

	player	rank	week
0	Swiatek Iga	1.0	2023-06-26
1	Sabalenka Aryna	2.0	2023-06-26
2	Rybakina Elena	3.0	2023-06-26
3	Pegula Jessica	4.0	2023-06-26
4	Garcia Caroline	5.0	2023-06-26

Figure XII: WTA Rankings Dataset and Women's match dataset

Cleaning

Once the data sets were merged into one, it was important to clean up the dataset to ensure consistency. First, I removed rows with significant amounts of missing values (only a few rows) and removing variables that were redundant from the merge (eg. removing the original seed variable, which was replaced by ATP and WTA rankings).

[19] For example, Juan Pablo Varillas in 'Last, First' form is 'Varillas, Juan Pablo', but Luca van Assche is 'van Assche, Luca'. There is no way for the computer to know what is the correct way to reformat the names on its own

Cleaning also involved aggregating similar variables like winners, unforced errors, and forced errors into single variables, thereby going from twenty variables to three. I was hesitant at first to aggregate variables because in the process I was potentially losing information. However, given the direction of my analysis (which I outline in Methods), I decided to aggregate.

```
winners = ['backhand_winners', 'drop_shot_winners', 'forehand_winners',
'groundstroke_winners', 'lob_winners', 'overhead_stroke_winners',
'return_winners', 'volley_winners']
unforced_errors = ['backhand_unforced_errors',
'drop_shot_unforced_errors', 'forehand_unforced_errors',
'groundstroke_unforced_errors', 'lob_unforced_errors',
'overhead_stroke_unforced_errors', 'volley_unforced_errors']
errors = ['backhand_errors', 'forehand_errors', 'groundstroke_errors',
'overhead_stroke_errors', 'return_errors']
```

In other cases, I averaged variables togethers. For example, set_x_diff was a variable that represented the score differences for set x of a match. For instance, if a match score was 6-3, 2-6, 6-1, then I had three variables, set_1_diff = 6-3 = 3 and set_2_diff = 2-6 = -4, set_3_diff = 6-1 = 5.[20] Since the number of sets played in a match can vary not only by gender but in general, I thought consolidating these five variables into a single average set difference or score would provide a more meaningful representation. I also averaged the total number of games per set for each set (eg. if the match score was 6-3, 2-6, 6-1, then set1_games = 9, set2_games = 8, set3_games = 7). Finally, I computed the average set length to derive a single value representing the average duration of a set for the entire match and the average game length to represent the mean duration of games within each set.

The most important data cleaning step was normalizing all of the match statistics by dividing by the number of sets played. All columns that represented match totals were normalized by the number of sets played in that match.[21] Since men and women's Grand Slam singles matches play to a different number of sets, normalizing ensures we'll be comparing men's and women's stats on an equivalent level.

Data Relabeling

Another crucial adjustment I made in the data preprocessing pipeline was reassigning the 'home' and 'away' labels. According to Sportradar documentation, the 'home' and 'away' player labels were assigned arbitrarily. *Figure XIII* below displays the rank versus the difference in the number of 'home' and 'away' label designations across all of the matches. The figure suggests

[20] Most matches are played as best-of-three or best-of-five set contests. A player must win at least six games with a margin of at least two games over the other side, to win a set. A final set score of 6-3 indicates the home player won 6 games while the away player won 3.

[21] In Grand Slams, Women's Singles play best of three sets and Men's Singles play best of five sets. In other words, Women's matches can have two or three sets and men's matches can have three, four, or five sets per match. This is what distinguishes Grand Slam tournaments from other lower-level tournaments, which play best of three for men and women.

that players with higher rank (closer to zero) are more likely to consistently be labeled as home or away players, compared to lower ranked players.[22] Across all matches, however, there is roughly an equal number of matches (48.47%) where higher ranked players are designated as 'home' vs 'away'.

As a result, I decided to relabel the cleaned dataset to ensure that the home-player is always the higher ranked player and the away-player is lower ranked. This way, the sign of the stat-difference would be universally interpretable. For example, a negative stat-difference indicates an unexpected outcome where the lower ranked player surpasses the higher ranked player (expected winner) for a given feature.

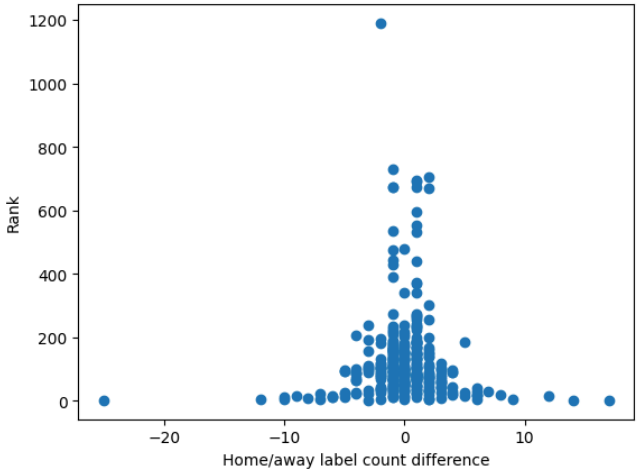


Figure XIII

Back to Data

[22] As outlined in Appendix B, a ranking is a metric that scores professional tennis players based on their performance. A rank of 1 is the highest rank, while a rank of 100 is much lower.

Appendix D: Feature Engineering and Failed Dimensionality Reduction

Preface

At the outset of my project, I initially planned to conduct an unsupervised clustering analysis to group matches by competitiveness and use visualization among other techniques to interpret the clusters and how they differed by sex. Dimensionality reduction is an appropriate initial step given the high-dimensionality of my data set. However, most techniques such as principal component analysis (PCA) and linear discriminant analysis (LDA) contribute to a loss of interpretability. Thus, even if distinct clusters emerged, it would be difficult for me to assess what features contributed to a discrepancy. Potential options for interpretation include back-projection or examining the correlation between the original features and the principal components or cluster assignments (amoeba). In other words, the unsupervised clustering approach paired with dimensionality reduction was a poorly thought out idea. Given the scope of the project and the time it took to conduct my data extraction and cleaning, I quickly pivoted to a more manageable approach, exploratory data analysis and hypothesis testing. However, this appendix outlines some of my initial efforts and analyses.

Feature Engineering: Absolute of Stat-Differences

An important initial step of handling high-dimensional data involves choosing the most pertinent and insightful features that reflect the inherent structure and trends within the data. This process can diminish any unnecessary noise, sparsity, or redundancy within the dataset, consequently enhancing the efficacy and interpretability of any analysis. To encapsulate a measure of player similarity or competitiveness, I spent a lot of time trying to figure out how to combine the per-player stats (eg. aces_home and aces_away). You may lose a lot of information when you consolidate variables, so it was important that I thought carefully about each option. I considered additive aggregations, differences, ratios, and using absolute values. The stat-difference represents how similar the players are on the basis of that stat (eg. aces_home - aces_away represents how much the home player beats the away player in number of aces; a negative suggests that the away player has more aces). The stat-sum represents the degree to which the match demonstrates high skill levels (eg. aces_home + aces_away reflects total aces in the match). Stat-ratio would also represent the relationship between the two values in terms of their proportion (eg. aces_home/aces_away compares the magnitude of aces_home to aces_away). I felt that stat-difference would best reflect the competitiveness between two players, while stat-sum would not be able to distinguish between the two competitors and stat-ratio would not sufficient measure the magnitude (eg. 60/30=2 & 20/10=2 but 60-30=30 & 20-10=10).

First, I considered the absolute value of the stat-differences, ignoring which player was dominating (indicated by the sign). I started with this approach because initially the home and away player labels were designated arbitrarily by the Sportradar API. That is, the ranking of the player was unrelated to the ‘home’ or ‘away’ label designation and I didn’t want to make

conclusions based off of these arbitrary designations.[23] Further, I wanted to initially focus on whether the players were evenly matched and didn't care which player (higher or lower ranked) was better than the other. The absolute value of features in a dataset created a set of folded distributions, where all data points to the left of x=0 (See *figure XIV* below).

I conducted some preliminary analyses, such as looking at the correlation matrix of a subset of stat-difference variables (see *figure XV* below). Note that these results took place pre-aggregation of some of the variables, as outlined in Appendix C). There was very little indication of correlation. Further, I plotted a pairwise scatterplot between a subset of variables, to see if any linear relationships between variables existed. Only a few variables had linear relationships, though they were all pretty obvious and not insightful. For example, points won and service games won are positively correlated and exhibit a linear relationship because winning a higher percentage of points typically leads to winning more games. In service games, consistently winning points, especially on serve, directly contributes to holding serve and thus winning those games. Since the majority of the variables do not have linear relationships with each other, this ruled out the possibility of using PCA with absolute value of stat-differences. I created Q-Q plots to check for normativity, an assumption of LDA (see *figure XIV*). The left tail in the Q-Q plots are consistently above the line, suggesting heavier tails or left-skewness. This is confirmed by the histograms of many of the variables, in which distributions appear to be asymmetric and folded.

Given my lack of experience working with folded distributions, I found them difficult to work with and interpret. Namely, I had trouble reconciling folded normal distributions, skewed right distributions, and half-normal distributions (Siegrist). Ultimately, I decided to keep the stat-differences, but remove the absolute value.

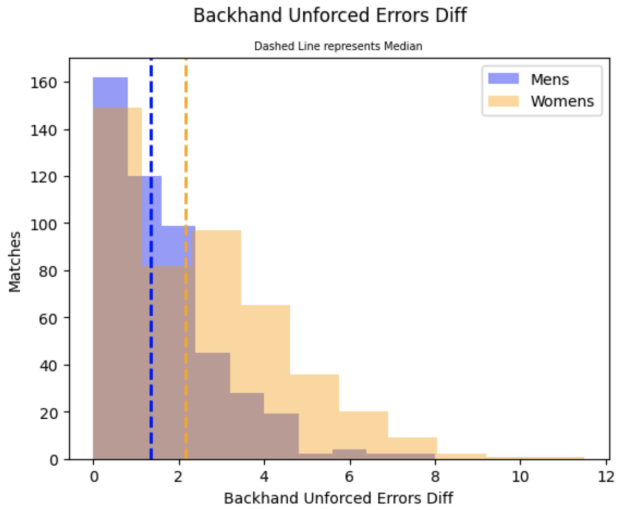


Figure XIV: Folded distribution caused by the absolute value of stat-difference

[23] Later in the analysis process, as outlined in Data Relabeling, 'home' and 'away' labels were reassigned such that the home player is the higher ranked player and the away player is the lower ranked player.

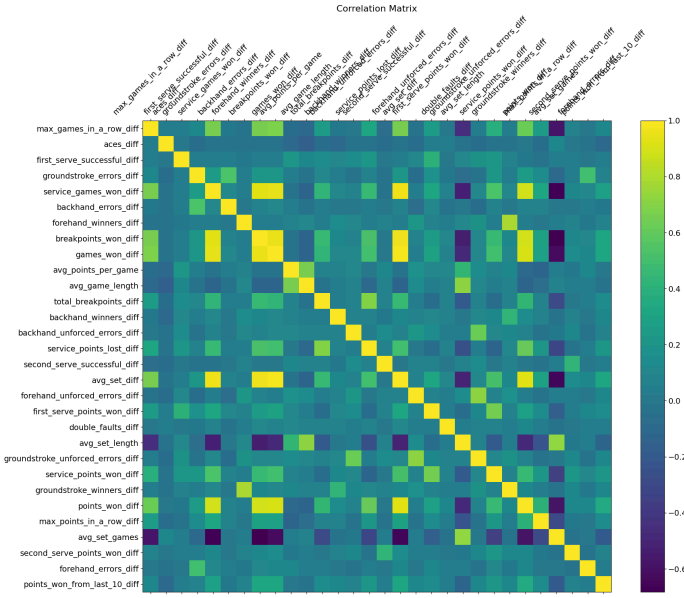


Figure XV

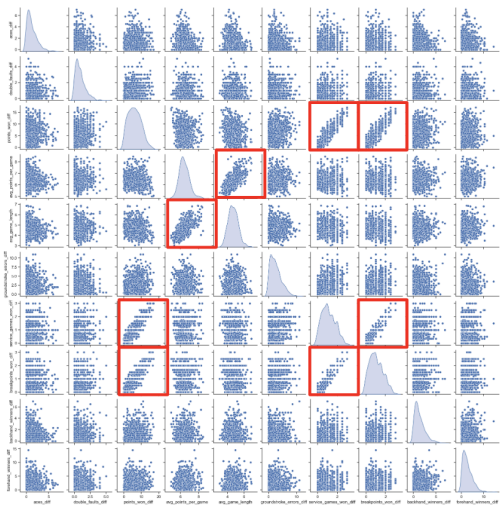


Figure XVI

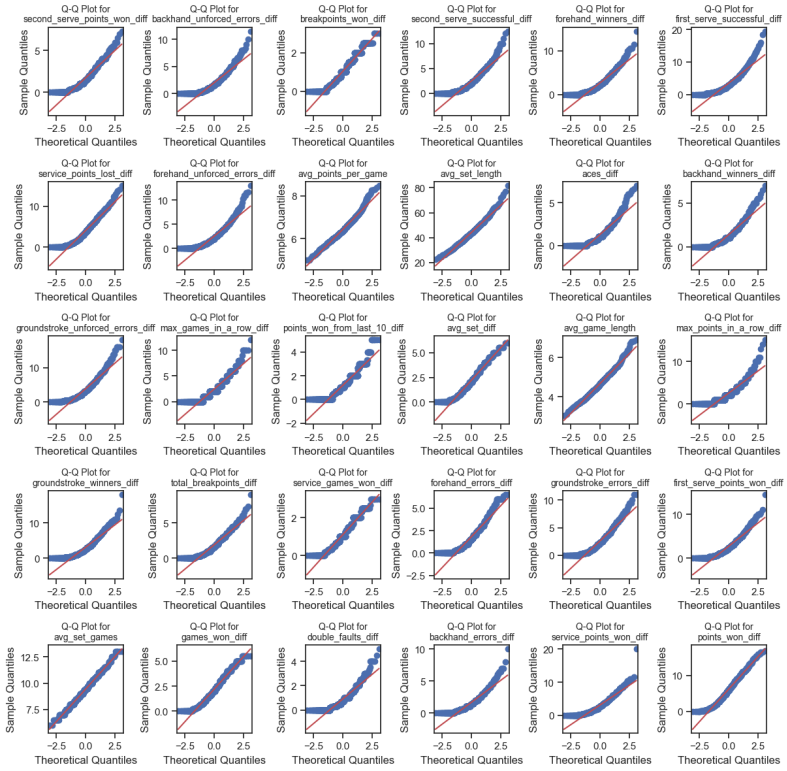


Figure XVII

Dimensionality Reduction and Clustering

After the data was relabeled, as outlined in Appendix C, I went through the same process of checking for linearity and normativity across the stat-difference variables. I plotted a correlation matrix, a pairwise scatter plot, and Q-Q plots for large subsets of the variables.[24] The results indicated much more correlation across a greater variety of variables and a much high proportion of variables with linear relationships (see *Figures XVIII and XIX*). This meant my new variables met the criteria for PCA. Furthermore, the Q-Q plots indicated normativity, suggesting that LDA was also an option (see *Figure XX* on page 22).

I conducted PCA but unfortunately the results were not revealing (see *Figure XXI* on page 23). The explained variation for the first three principal components were as follows [0.33451301 0.11527081 0.09087206], which make up a little more than 54 % of the variance, not very significant. I plotted the first two principal components and each match is represented by a point on the scatter plot and colored by sex. Aside from perhaps greater variance, there is a lot of overlap and very little revealing information from PCA and very little room for interpretation. Reiterating my discussion from the preface, I decided to move on from dimensionality reduction and clustering and move towards exploratory data analysis and hypothesis testing.

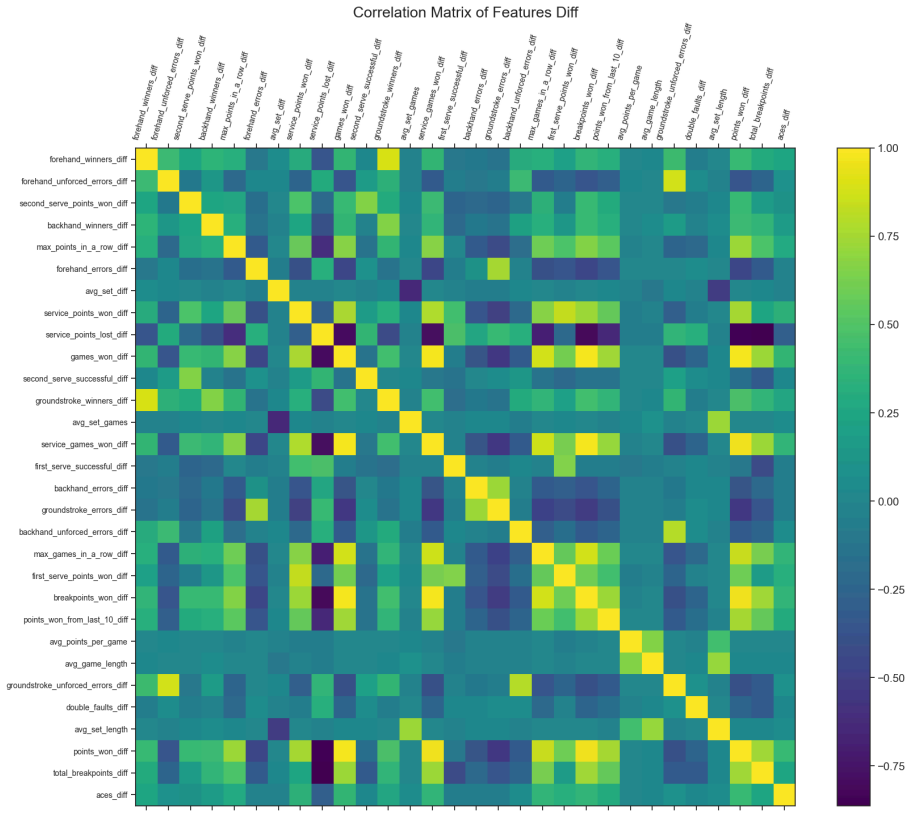


Figure XVIII

[24] I didn't include all of the variables in the scatter plot, since at the time there were more than thirty variables (30 x 30 = 900 scatter plots).

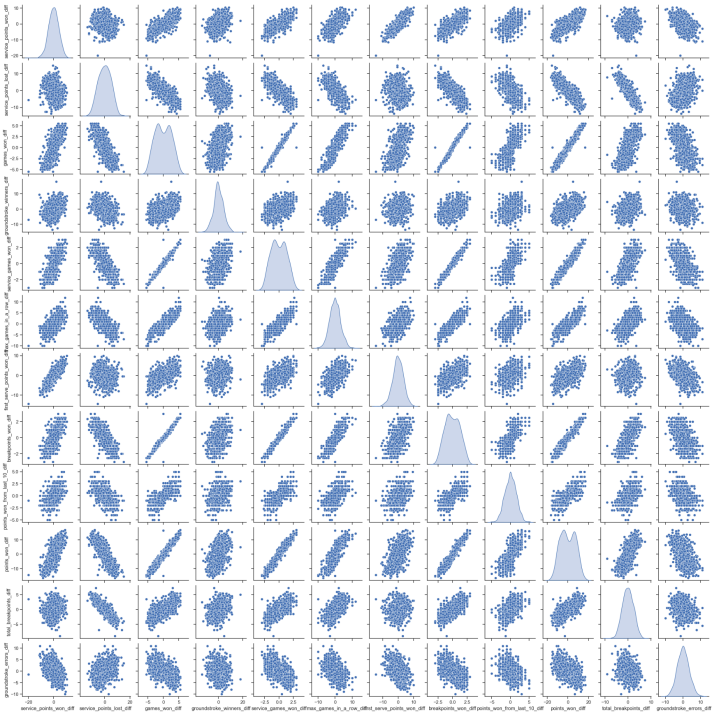


Figure XIX

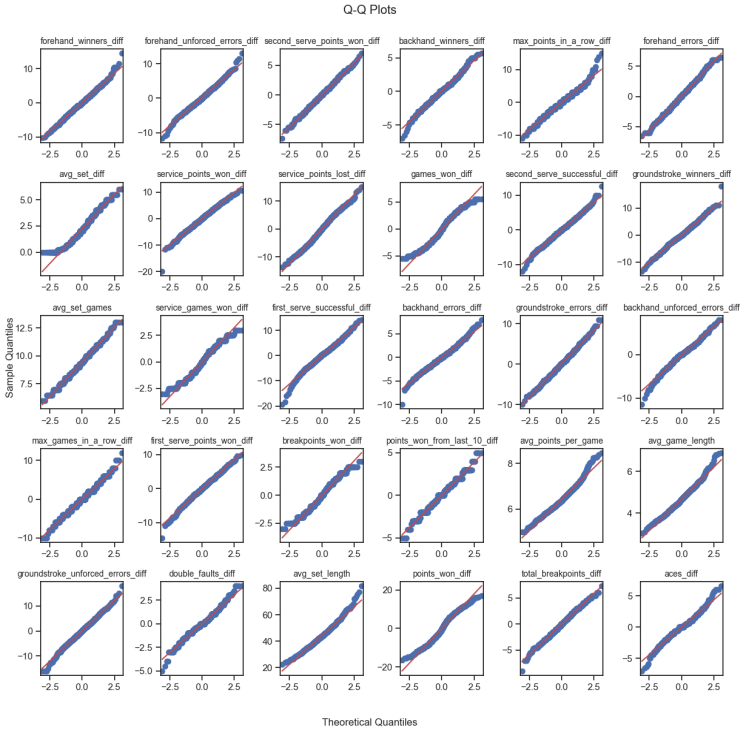


Figure XX

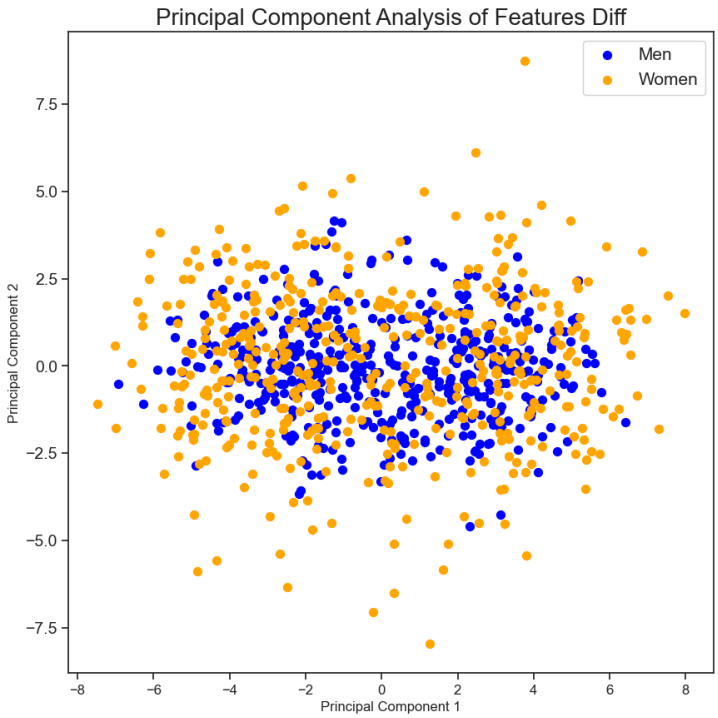


Figure XXI

Back to Methods

Appendix E: Results Elaboration

I began my analysis of distributions by conducting Kolmogorov-Smirnov (KS) tests for my stat-differences. For the KS-Test, the null hypothesis is defined as two identical distributions and the alternative as two non-identical distributions. A rejection of the null hypothesis would indicate that the cumulative distribution functions of the two samples are unlikely to come from the same distribution. The computed KS tests indicated that 22 out of 24 features had a p-value < .05. I plotted empirical cumulative distribution functions for visual confirmation of these results. Since KS Tests are not specific to certain qualities of distributions, this was more a preliminary test to indicate whether I was heading in the right direction. The below figures are examples of ECDFs for Aces Diff and First Serve Points Won Diff, with p-values of .0015 and .0022 respectively.

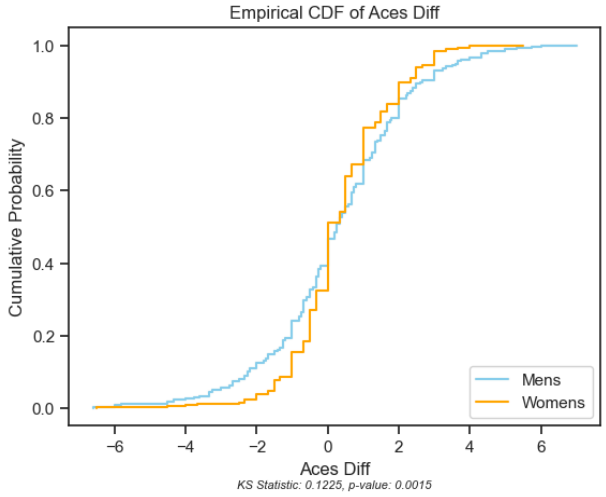


Figure XXIII

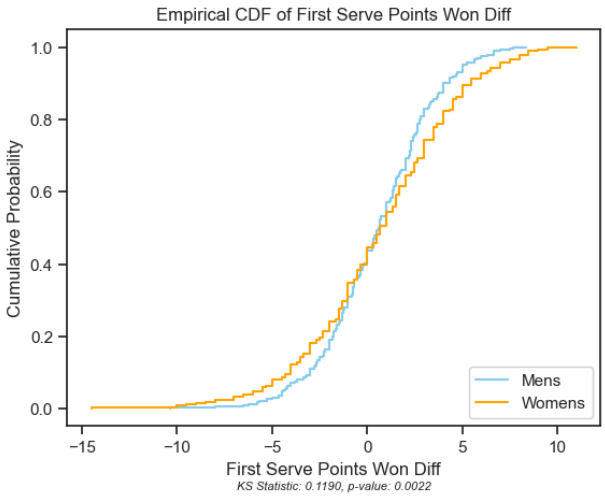


Figure XXIV

In addition, T-tests operate under the assumption of normativity. I conducted Q-Q plots to identify which variables were normal. Below are the Q-Q Plots for unforced errors diff and winners diff. Most of my other stat-difference variables indicated similar Q-Q plot results.

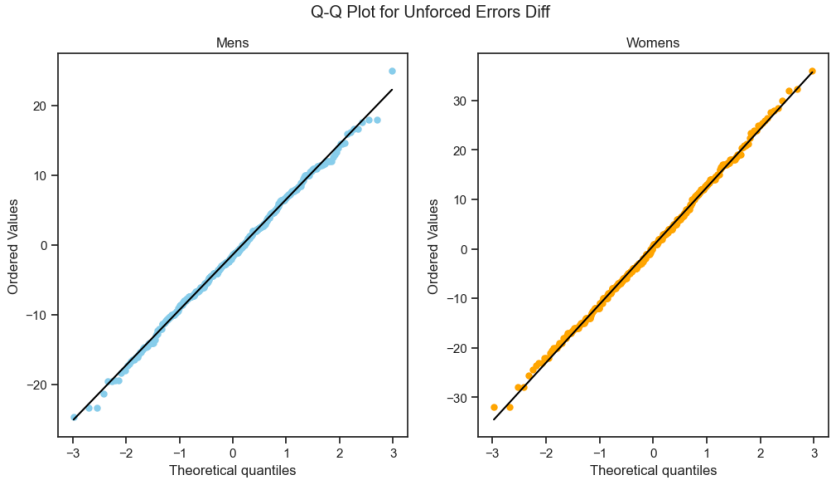


Figure XXV

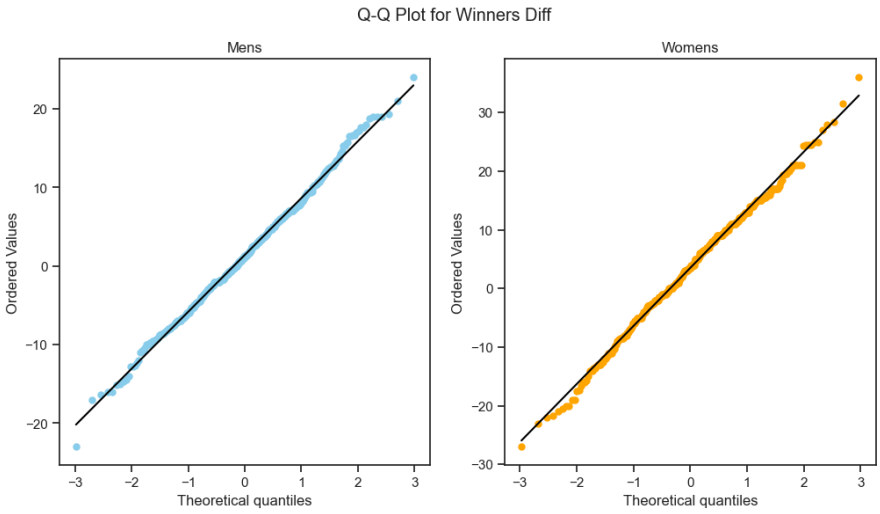


Figure XXVI

Back to Results