Classifying WTA Playing Styles and Identifying Pattern Weaknesses

Emily Zhang & Beatriz Avila-Rimer

1. Introduction

Tennis is a highly competitive and physically demanding sport that requires a combination of skill, strategy, and mental focus. With the increasing availability of technology and data analytics, statistics and analysis have become a valuable tool for coaches, players, and fans. Tennis analysis involves using granular data points and a variety of statistical models to gain insights into player performance, match strategy, and overall trends in the sport. In recent years, tennis analysis has become an increasingly popular topic of research, with many studies exploring the effectiveness of different playing styles using a multitude of analysis techniques, such as HawkEye and match predictions, and their impact on player performance. In this paper, we aim to explore and classify various playing styles on the WTA (Women's Tennis Association) tour and identify specific shot pattern weaknesses for each playing style.

This research question is important for several reasons. Firstly, there is a limited amount of research surrounding WTA players as opposed to the ATP (Association of Tennis Professionals, representing the men's tour). Addressing this research question can pave the way for other researchers to build upon what we have discovered. Secondly, WTA players tend to have higher volatility in terms of results; since 2018, the WTA has seen 12 different players win a grand slam (out of the Australian Open, US Open, Wimbledon, and French Open) whereas the ATP tour has only seen 6 different players. This reason may tie into why there is limited research

in the WTA space since patterns become harder to identify with a lower consistency of match predictions. We want to compare and contrast the differences and/or similarities between both WTA and ATP styles and discuss the different developments of the men's and women's games. Additionally, the results of this research can provide a quantitative measure to aid strategic decisions, meaning that these outcomes can give tangible and actionable ideas to implement in practice and matches when facing a specific playing style. Fourthly, this research will contribute to the development of women's tennis, applying to not only the professional tour but also women's college tennis and the junior circuit. Lastly, classifying playing styles can improve the quality of individualized coaching. Creating drills that practice against specific shot patterns can improve the productivity of lessons and strengthen a player's game to adapt to different playing styles. By providing valuable insights into specific tactics, improving decision-making, and enhancing overall performance, this question can contribute to the development of the sport and benefit coaches and players of all ages.

2. Literature Review

There is currently only one published paper and two online articles that address a similar question. The first article explores a similar research question pertaining to the top 100 ATP players, quantifying their playing styles and how effective they can be in predicting outcomes [1]. Based on the players' average ace percentage, serve speed, points played at the net, net points won, unforced error rate, and their forehand backhand strokes, Ontiveros used K-means clustering, an unsupervised machine learning algorithm that groups data points together based on similarities, to determine that there were four main playing styles: left counter-puncher, right counter-puncher, all-court, and big server. Furthermore, he uses regression analysis to determine

which playing style has a higher predicted win percentage over another; for example, when a left-handed counter-puncher plays against a big server, the counter-puncher's winning odds are increased by 1.95. We are looking to expand upon this article's research by applying K-means clustering to WTA players with a larger set of attributes as well as pinpoint specific shot patterns to play against different playing styles.

The second article similarly discusses player style classification on the WTA tour and whether or not there are associations between playing style and success on tour [2]. The research article engineers two features, aggression and consistency, based on two separate equations:

$$(1)\ aggression = \frac{(total\ winners + total\ unforced\ errors)}{total\ games\ played}$$

$$(2)\ consistency = \frac{total\ winners}{total\ unforced\ errors}$$

By imposing specific cutoffs using the averages in that year, Devin separated the players into discrete categories of aggressive or defensive and consistent and inconsistent. The results demonstrated the importance of consistency in previous Grand Slam winners and overall win percentage and that aggressive consistent players have a slight edge over defensive consistent players in the past years. Additionally, he examined the correlation between playing styles and win percentages across four other figures: aggression, consistency, average number of unforced errors per game, and average number of winners per game. One of the more significant findings is that defensive players generally saw an increase in win percentage with higher aggression. To build upon these ideas, we would like to explore a higher dimensional model, considering more data points, such as shot type, rally length and direction, instead of only looking at winners and unforced errors to analyze playing styles. We believe that our findings can be impactful in determining weaknesses against specific playing styles that can expand upon this article's conclusions in improving consistency and aggression.

In our last piece of literature, the published paper aims to find the optimal shot sequences from different court positions for different player types. Liyanage first classifies both ATP and WTA players through anthropometric data (e.g. height, weight, BMI), grip types, handedness, and backhand type using the K-Prototypes Algorithm with an optimal number of clusters for each handedness in both ATP and WTA, determined by the Elbow method. In contrast to the previous two articles, this paper utilizes data from HawkEye, providing a spatiotemporal data perspective in calculating the shot sequences. A sequential market basket analysis was conducted to determine the most used shot sequence, most successful shot sequence, hidden gem sequence (selected as most surprising by domain expert), and a band-aid sequence (best sequence from a bad segment of the court) for each player cluster. One of the research's findings is that deeper court positions are preferred by shorter males and extreme-grip players in order to create more time to neutralize the ball or hit with their full swing path. Another finding is that the forehand crosscourt to forehand down-the-line was a common pattern used by most of the player types. We plan to draw ideas from the sequential market basket analysis approach to detect shot sequences between two to three shots.

3. Hypothesis

Based on our research, we hypothesize around three distinct playing styles across the WTA tour: all-court player, counter-puncher, and aggressive baseliner. In the past, many had considered serve-and-volley a predominant playing style; however, as the technology of racquets, the physicality of players, and the speed of the game have increased, this style has no longer become as effective. Nevertheless, we still want to acknowledge that there are currently

some outliers on the WTA tour who perform serve and volley quite efficiently. We expect this niche of players to be classified under all-court players.

All-court players are versatile and well-rounded, adept at mixing different strategies and adapting to various game situations. They possess a combination of power, consistency, and tactical awareness, allowing them to be effective in both offensive and defensive aspects of the game. For example, Ons Jabeur is a model representation of an all-court player who is creative and unpredictable with an array of shot-making abilities that keep her opponents off balance. In contrast, aggressive baseliners are characterized by their ability to generate significant pace and aggression in their shots. They often rely on their strength and explosive strokes to overpower opponents and finish points quickly. Serena Williams is an excellent representation of an aggressive baseliner who uses her serve and her overwhelming power to dominate points right at the start without the need to lengthen rallies. Lastly, counter-punchers excel at retrieving and neutralizing their opponents' shots with exceptional defensive skills. They prioritize consistency, movement, and shot placement, relying on patience and the ability to turn defensive positions into offensive opportunities. Simona Halep has shown exceptional footwork and speed to cover the court and a remarkable ability to redirect the pace of the ball against bigger hitters as a counter-puncher. However, we also acknowledge that players may exhibit elements of multiple playing styles; we plan to predict the playing style that dominates their average play during matches.

After performing our classification, we expect all-court players to have a higher average net effectiveness and the win percentage based on rally length to be normally distributed (i.e. we expect the highest win percentage to be for medium-length points while a relatively lower percentage for short and long points). We also expect all-court players to show the most diverse

shot selection. On the other hand, we predict aggressive baseliners to hit mostly groundstrokes and have a relatively high number of both winners and unforced errors. These players are likely to display low net effectiveness and have the highest win percentage for shorter-length points. Finally, we expect counter-puncher to play long points and thus are more likely to have a high win percentage for both long and medium-length points. Counter-puncher tend to be crafty, so players in this classification are likely to slice more than the average player on tour. They are also likely to have the most well-defined shot direction patterns in order to neutralize their opponent's aggressiveness to stay in the point.

Instead of considering handedness as previous papers have mentioned, we believe that handedness does not have a large impact on differences between playing styles; for example, the consistency and the amount of spin on the ball will stay relatively the same as well as the movement around the court across these players. Additionally, we will not be considering physical features such as height and weight although we are aware that certain physical characteristics may skew high in the different classifications. For example, we expect aggressive baselines to be taller on average than the players in the other classifications because taller individuals tend to generate more power. We also plan to limit our data to between 2017 and the most recent 2023 matches in order to acknowledge and account for advances in racket technology, surface changes, and the evolution of playing styles over time.

In terms of shot pattern analysis for each player type, we have the following expectations motivated by the USTA Player Development Journal [6]. We anticipate counter-punchers to keep the ball deep, moving the opponent side to side, likely in an “X” pattern. We also expect long cross-court to cross-court rallies until a short ball is recognized and redirected down the line to be a predominant pattern in their game due to their patience and consistency. In contrast,

all-court players are expected to use a wider variety of shot directions and shot types. We predict similarity in patterns between counter-puncher and all-court players; for example, they are also likely to engage in long cross-court to cross-court rallies until an opportunity to redirect or change the opponent's rhythm arises (e.g. with a slice or a drop shot). All-court players are also likely to use a chip shot (in other words, a short slice ball with backspin such that the opponent must move into the court) and then hit a deep ball to the opposite side to mix up the depth of their shots. In general, we expect these players to exhibit greater control of all areas of the court and for their patterns to illustrate that. We believe that the process of pattern recognition might be the most difficult for players in the all-court classification given the variety that characterizes their game. Finally, for aggressive baseliners, we expect these players to dominate the court more heavily with their forehand, following patterns that rely on multiple inside-out forehands (i.e. forehands hit on the backhand side of the court to the opponent's backhand) until they see an opportunity to attack. Aggressive baseliners are more likely to have an effective and precise serve that they use to set the point up. A common pattern we expect to see is a serve out wide and the next ball to the open court or a serve to the T with the next ball hit behind them (i.e. to the same serving side).

In terms of predicting shot pattern weaknesses, we expect aggressive baseliners to struggle against patterns that keep the ball deep and with lots of topspin, such as the "X" pattern explained earlier. These strategies make it harder for aggressive baseliners to attack a ball from behind the baseline; if they choose to attack in spite of this situation, the resulting shot will likely be less effective and become a liability. Another pattern that is likely to be a weakness is the use of a chip slice followed by a deep drive to the other side of the court. This pattern can be troublesome for these players to handle because of the differences in spin which can be difficult

to adjust to, especially when the aggressive baseliners have preferences for flat and fast balls. For counter-punchers, we expect these players to have trouble dealing with patterns that mix up changes in the speed, spin, height, and/or depth of the ball. Counter-punchers are patient and experienced in playing long points with the same rhythm and such changes are likely to throw them off their game. An example of such a pattern would be a deep rally that keeps the counter-puncher behind the baseline then a drop shot to move the counter-puncher forward. Another pattern would be a crosscourt rally then a change in direction down the line with a different spin and height. Finally, all-court players are likely to struggle with patterns that keep the ball deep and do not allow for many angles. These types of patterns are likely difficult for all-court players because there are fewer opportunities to vary the shot selection and placement. Examples of such patterns are long crosscourt exchanges with high topspin balls, or the "X" pattern mentioned above.

4. Data and Methods

The data used in this research comes from the public GitHub repository created by Jeff Sackman as part of the Match Charting Project, which aims to amass detailed records of professional matches [5]. We worked with the WTA match data from January 2017 to April 2023 (the latest available data). This data comes from individuals who volunteer to watch and record match statistics for every shot hit and is subject to human reporting mistakes. This issue is especially prevalent when recording shot direction and depth as the reporter must make judgment calls. Furthermore, since there are a multitude of different users recording matches, the judgment calls might not be consistent across all matches. There is no method to assess the accuracy of these observations, so we must trust and acknowledge that there may be some inconsistencies in

some of the observational data. Each match is recorded by a single observer, in our analysis we use data from a wide variety of observers using the available aggregate stats computed for each match and we also computed new statistics from the available data.

From the comprehensive records available on the website, we chose to utilize the data sets that contained information about general match statistics, shot types, rally lengths, and serve influence per match in order to create features that would help us classify players into distinct playing styles. In total, we collected a set of 353 WTA players from the match IDs from 1306 matches. Considering that there are over 50 WTA events in addition to the four Grand Slams and that tournament sizes range from a draw size of 64 to 128, we estimate that there are approximately 5200 WTA matches played each year. Therefore, we acknowledge that we are analyzing a very small dataset of about 4% of the matches played in the last 6 years. We also recognize that out of a total of 1650 WTA players, we are only analyzing about 20% of the population with a bias toward the top players, this is due to the fact that the users who record the matches are likely more interested in recording the top players.

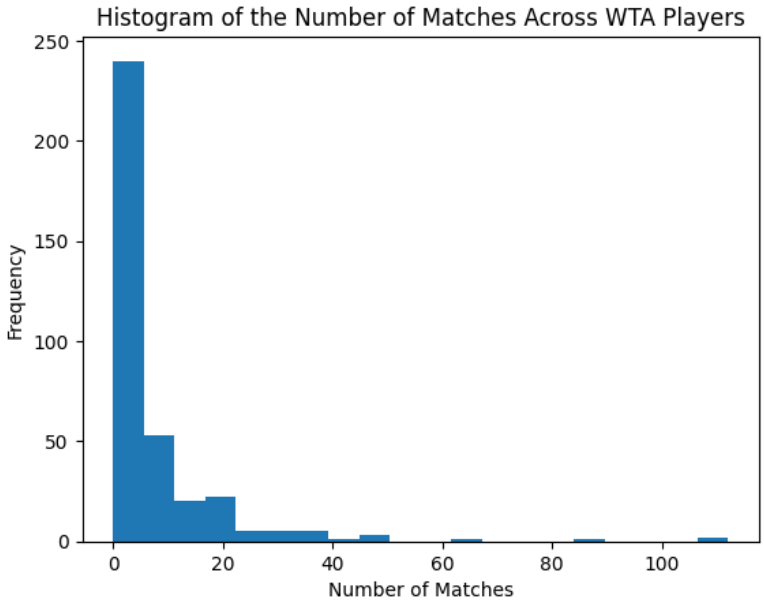


Figure 1: *A histogram of the number of matches recorded across the WTA players.*

From Figure 1, we can see that there is a large disparity between the number of matches recorded for each WTA player. Kiki Bertens, Biana Andreescu, and Iga Swiatek were among the few with the highest number of matches recorded with 112, 109, and 88 matches respectively, while 123 players only had 1 match recorded and 54 players had 2 matches recorded.

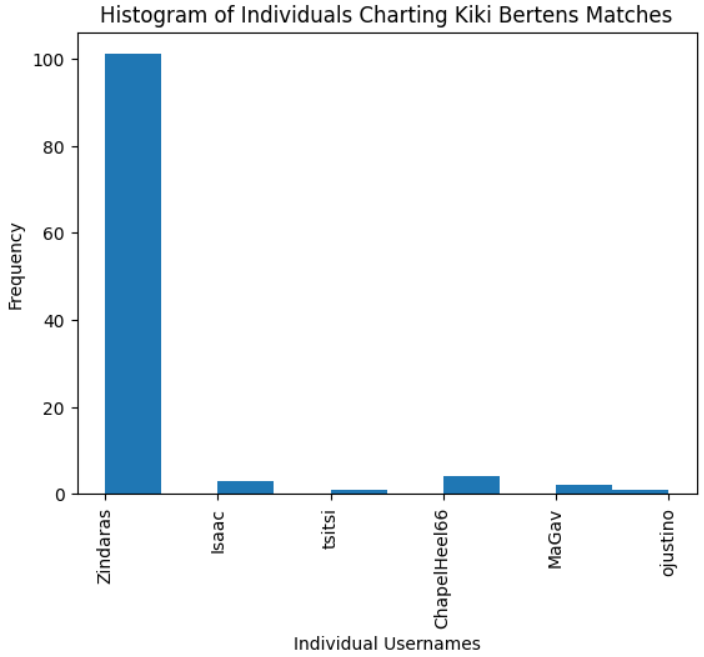


Figure 2: *A histogram of the list of individuals charting Kiki Bertens's matches.*

After looking more closely at the data, some of the top players such as Kiki Bertens and Biana Andreescu had a skewed proportion with only one individual (i.e. Zindaras in Figure 2) charting the majority of their matches. This result implies that these individuals were charting the matches of their favorite players (e.g. Kiki Bertens or Biana Andreescu). As the number of matches recorded for a WTA player decreased, the list of individuals became more evenly distributed and the WTA matches were skewed towards top-ranked players such as Iga Swiatek (88 matches), Ashleigh Barty (49), and Naomi Osaka (47). These observations clearly demonstrate bias in the uneven distribution of the recorded matches for players. We then decided

to remove the players who had less than 3 matches recorded in the database as part of the preprocessing stage which left us with 181 WTA players and 766 matches. The players who had a high number of matches thus had a more accurate representation of their game with lower variability. In addition to the players, the majority of the matches also come from high stakes, competitive tournaments such as the Grand Slams (Australian Open, US Open, Wimbledon, and Roland Garros) as the top four match locations respectively, followed by the Indian Wells (commonly referred to as the fifth Grand Slam) and Miami Open, Madrid Open, and Italian Open, which are all top WTA 1000 events (i.e. events with $1,000,000 in prize money). This observation also promotes bias in the player-recorded matches as higher-ranking players are more likely to play in these events and make it further in these tournaments. However, these biases cannot be avoided with such a small dataset, and more matches that include the same few players do not negatively affect the data of the other WTA players.

After loading the data and filtering out the matches and players, we proceeded to narrow down the wide range of data included per match. Additionally, we had to clean the match IDs to identify whether or not a player was defined by ID number 1 or number 2 in order to obtain the correct corresponding data for each. In order to extract relevant information from the available raw data, we created new features to better capture patterns and player characteristics. From the rally dataset, we constructed six new features: “Short Point Win Percentage”, which consists of points that are won within 1-3 shots, “Medium Point Win Percentage”, consisting of points that are won within 4-6 shots, and “Long Point Win Percentage”, consisting of points that are won after 7+ shots. For each match per player, we found the ratio of short/medium/long points won to the total number of short/medium/long points and averaged the win percentage across all matches. This information can help to project how aggressive or patient a player is; for example,

an offensive player would have a higher winning percentage with shorter points than longer rallies compared to a more defensive player. We also looked at the raw short point, medium point, and long point percentages. The motivation to derive both of these metrics was to gain a more comprehensive understanding of a player's performance and to offset a bias towards more successful players.

Extending this idea, we additionally took data from the shot types table: number of forehand groundstrokes, backhand groundstrokes, slices, and net shots played per person per match. The total number of forehand and backhand slices was combined into a single category since we believe that the difference on which side the slice was hit on does not drastically affect their playing style and how the ball is received. Half-volleys, overheads, and volleys were also categorized into net shots because we wanted to examine the general patterns of how consistently a player moves into the net. This data allows us to see the overall shot-making decisions and how much each player utilizes various shots across their matches.

Furthermore, general match statistics were taken from the WTA dataset, specifically, average service win percentage, calculated by taking the number of points won while a player is serving divided by the total number of points played while this same player is serving. We similarly calculated the average return win percentage, number of winners, and number of unforced errors over all matches played. The average serve and return percentages can give a sense of the player's offensive and defensive abilities while the number of winners and unforced errors examines a player's balance between consistency and aggression.

Lastly, we studied data from the serve influence table. This data reveals how far into rallies each server retains the advantage of serving and whether or not a player's serve is a strength that can help to dominate and win points easily. The serve influence is calculated by

obtaining the percentage of points won by the server for a certain range of rally lengths. We chose to analyze serve influence for both first and second serves in the 1-5 rally length range.

Having derived comprehensive player features we combined attributes into a single data frame organized by player and ran the K-means clustering algorithm. We use the K-means clustering algorithm instead of the K-prototypes or K-modes algorithms mentioned in previous papers because we are only working with numerical data and not categorical. Through the elbow method, which is a graphical technique used to determine the optimal number of clusters by running K-means on a range of cluster values, we decided to use 3 groups for K-means clustering. Initial runs of the K-means clustering provided results that contradicted our intuition of the game and seemed to be more accurately quantifying degrees of success rather than characterizing playing styles. In order to obtain more pertinent results we reevaluated the attributes that were being fed into the algorithm and ultimately decided to use the following: “Serve Win Percentage”, “Return Win Percentage”, “Winners”, “Unforced Errors”, “Short Point Percentage”, “Medium Point Percentage”, “Long Point Percentage”, “Percentage of Forehand Groundstrokes”, “Percentage of Backhand Groundstrokes”, “Percentage of Slices”, “Percentage of Net Shots”, “1st Serve Influence 1+”, “1st Serve Influence 2+”, and “1st Serve Influence 3+”.

Once the groups were finalized by K-means, we labeled each match ID (using code) with the corresponding winner by using another of Jeff Sackman’s CSV files which included the winners from previous WTA matches. In order to study effective pattern strategies against specific playing styles, we labeled each of the 353 WTA players with their corresponding K-means group and sorted the players into a Python dictionary based on the match results (e.g. 1 v. 2 represented matches where players of style #1 won against style #2). We decided to create three match groupings, we categorized the matches based on the outcome, specifically focusing

on instances where one group had lost to the two other groups. This resulted in 3 different groups to analyze the shot sequencing patterns (2 & 3 def. 1, 3 & 1 def. 2, and 1 & 2 def. 3). We performed analysis across all losing matches for each player type to determine the most significant patterns weakness. From now on, we will be making a distinction between the player of interest and the opponent. For example, if we were looking at Group 1 (2 & 3 def. 1), the player of interest is the player with playing style 1 and the opponent is either categorized 2 or 3.

In order to perform shot sequencing analysis, we first determined a valid shot sequence criterion. A shot pattern is considered relevant if it results in a point lost by the player of interest. We analyzed a subset of points, specifically focusing on points that were lost during matches that were lost by the playing style of interest. The motivation behind this filtering is to attempt to normalize the context of gameplay across matches and points. Furthermore, a shot sequence is valid if it contains at least two shots hit by the opponent and it occurs in a point that leads to a disfavorable outcome for the player of interest. Additionally, shot patterns are independent of what the player of interest hit. For example, consider a point where the opponent served out wide, the player of interest hit a forehand cross-court, and the opponent forced an error by hitting a cross-court forehand. In terms of shot sequencing, we only focused on the opponent's actions. The rationale behind this metric is to help the player of interest recognize what patterns they are most susceptible to.

After the appropriate filtering, we were left examining approximately 53,000 points across all matches for all player types. In order to extract the most common patterns, we performed subsequencing on the points where we considered three or more shots at a time and then calculated the frequencies of each pattern in its respective dataset. If the points were shorter than 3 shots, they were removed from the set; given that a significant portion of points lost is a

result of a serve or return error (which is the case for most matches), we knew that the size of our dataset would dramatically decrease. After evaluating how many of these points met the prescribed criteria, we were left with around 5,000 subsequences in total and took the top three from each group listed in our results below. This shot sequencing analysis was performed with the point-by-point data from the Match Charting Project, characterized by a list of symbols representing serve direction, shot type, how the point ended (i.e. unforced error, winner, or forced error), and the type of error (i.e. in the net, out wide, or deep). Specific samples of this point-by-point data can be found in Appendix A and a link to our Google Colab which consists of our code and more graphs can be found in Appendix B.

5. Results

Before diving into the results of the K-means clustering, we first did an initial investigation of any similarities between the features we selected.

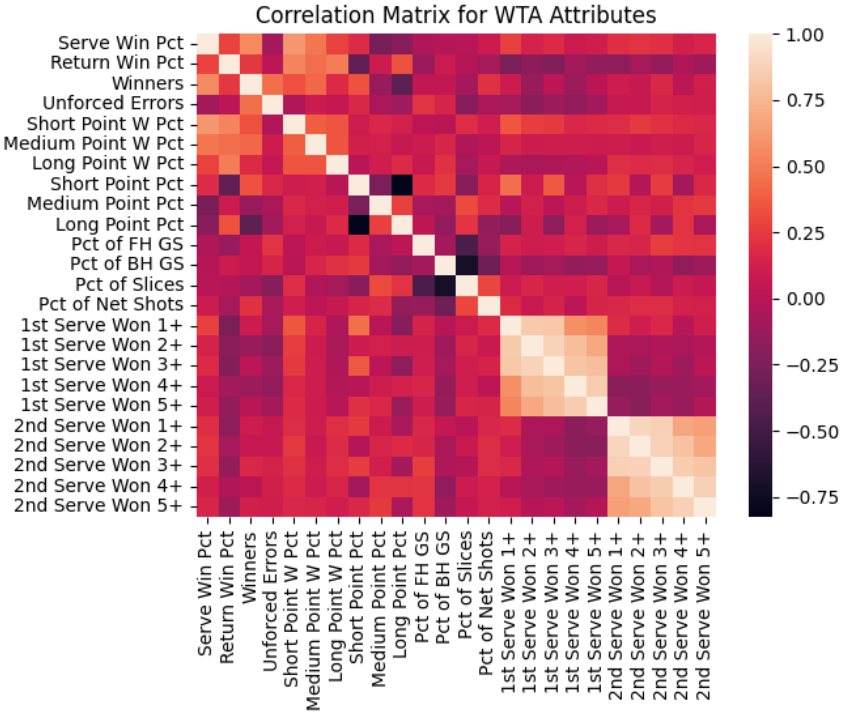
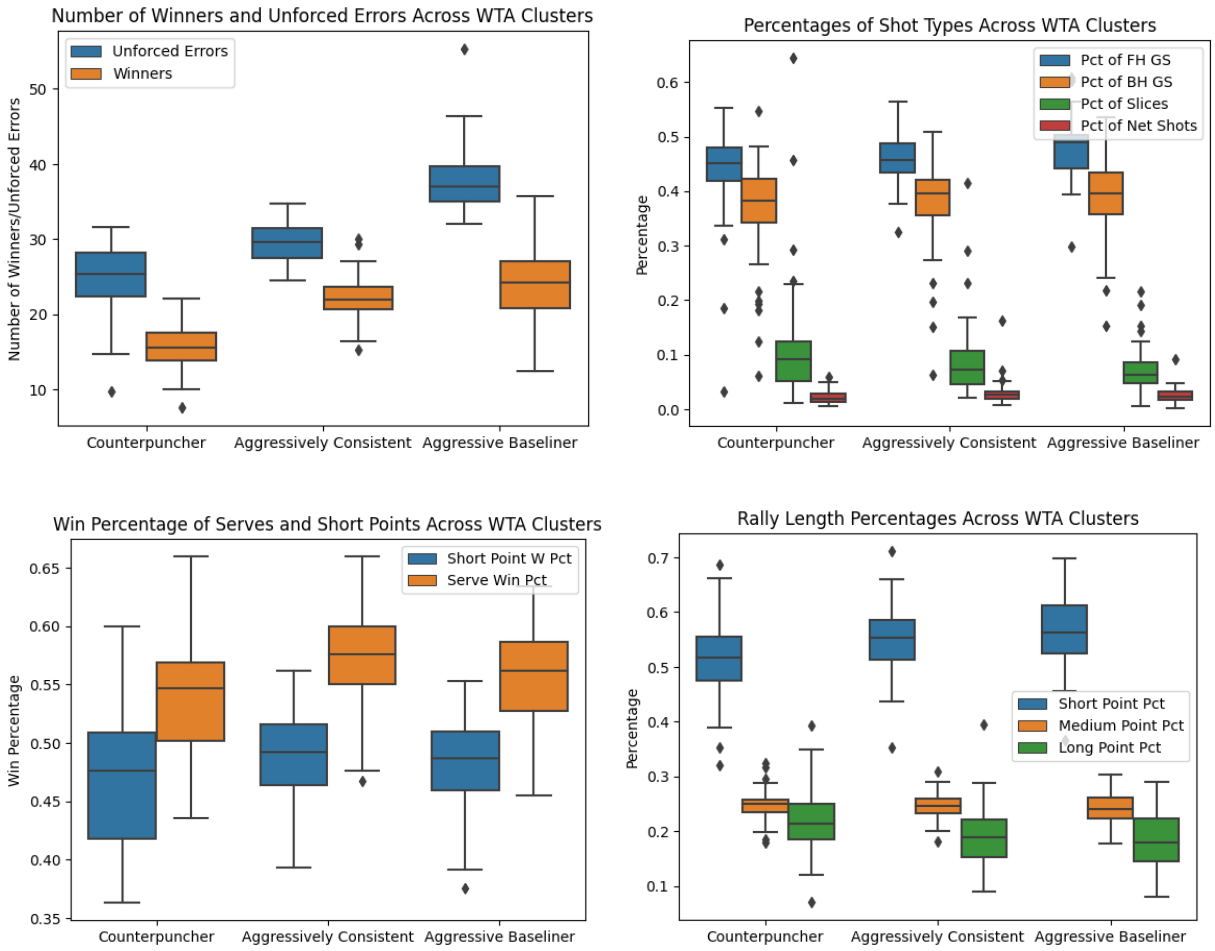


Figure 3: *A correlation matrix of all selected features from the WTA Match Charting Project data. The lighter colors of the gradient represent a high positive correlation between features while the darker colors represent a high negative correlation.*

First, it is important to note some trivial correlations from Figure 3. For example, for the serve influence metric, it follows that if a player has a high 1st serve 4+ win percentage, then she will also have a high 1st serve 3+ win percentage, and so on. In contrast, we see almost no correlation with the 2nd serve percentages, demonstrating that 2nd serves may be less effective in setting up points on the WTA tour. We also see a positive correlation between winners and unforced errors. This result is as expected as it follows that players who are producing a lot of winners are pushing for more difficult shots. This, in turn, means that they are more likely to make unforced errors while attempting riskier shots. Another result shown in the table is the positive correlation between serve win percentage and short point win percentage. A high win percentage on the serve indicates that the player's serve is very effective in leading to a weak or short return and ending the point in 1-2 shots after. We also want to note the negative correlation between slices and backhand groundstrokes, illustrating that most players prefer to slice with their backhand.

From our observations on the K-means results, we labeled the three player types as (1) counter-punchers, (2) aggressively consistent players, and (3) aggressive baseliners.



Figures 4, 5, 6, and 7: *Boxplots of selected attributes from the WTA dataset divided by the K-means groups.*

Based on our graphs, we observed some distinct characteristics between each group. The aggressive baseliner has the highest average number of winners and unforced errors across the three groups, followed by the aggressively consistent players and the counter-punchers. This observation aligns with our initial hypothesis that this group of players takes riskier shots. The aggressive baseliners also have the lowest average percentage of slices and net shots and the highest average short point percentage. This agrees with our prediction that this group of players takes the first strike at the ball and continues building the point on the baseline with their pace. Some top-ranked players that were categorized in this group include Madison Keys and Petra

Kvitova, both known for their flat and hard-hitting balls. These players helped to ensure as a sanity check that the results determined by K-means were valid and sensible.

Our second group, the counter-punchers, saw the highest average long point percentage and lower serve and short point win percentage. This group also aligned with our expectation that these players are less offensive with a weaker serve but perform better in longer rallies. Interestingly, we also saw the highest average number of slices from this group; this result was not in our initial hypothesis but it follows that these defensive players retrieve and neutralize the ball in order to make time for themselves to return to the middle of the court. Top-ranked players that were grouped into this playing style include Caroline Wozniacki and Angelique Kerber, known for their precise movement and their grittiness to fight for every ball.

Our last group was labeled as aggressively consistent players, different from our initial hypothesis to be all-court players. This group appeared to be the most balanced across all categories, having the second-most winners and unforced errors and the second-highest percentages across all rally lengths and shot types. This group does not have as high of a variety of shot types as we predicted for all-court players but still possesses the toolset and capabilities to craft the point with a mix of aggression and consistency. These players also had the highest success rate with serves and short points, a characteristic that we expected to hold for the aggressive baseliners instead. However, their balance with aggression and consistency proved to have a higher win percentage than the aggressive baseliners in these categories. Players that were categorized into this playing type include Naomi Osaka, Iga Swiatek, and Ashleigh Barty, who are all former #1 ranked players with strong abilities in both offense and defense. Some of the most successful players were grouped into this playing style, emphasizing how this type of game style proves to be effective against top WTA players.

However, upon closer inspection, this group was observed to represent a mixture of playing styles. Based on the match statistics of these players, K-means categorized players such as Serena Williams, Ashleigh Barty, and Simona Halep into the group of aggressively consistent players. From the provided data, K-means did not accurately separate these players into different groups but instead, seemed to provide insight as to what makes these former #1 players successful with their wide array of shot types and ability to stay offensive while minimizing errors in their games. Another outlier that we observed was Ons Jabeur, who was placed into the aggressive baseliners category with Jelena Ostapenko and Madison Keys, but was a player who we expected to be characterized as an "all-court" player. Ons is amongst the craftier players currently playing on the WTA tour, known for her creative point construction and use of drop shots and slices. We believe that this misclassification is due to the fact that not many people have a similar playing style as her, especially at such high efficiency. Ons, on average, has a high number of winners and unforced errors per match, similar to the overall statistics of aggressive baseliners. Barty is also observed to have a similar style to Ons but has seen more success with her style, which is why she was grouped with the aggressively consistent players. Another prominent outlier we observed was Serena Williams, who was expected to be classified amongst the aggressive baseliners but instead was grouped with the aggressively consistent players. However, with her high win percentage in service points, a great balance between consistency and winners, and the lack of physical attributes used in this study, she was grouped with the most successful group of players.

The results of the shot sequencing analysis with the top three most common pattern weaknesses are listed below each group. A pattern is considered valid if it occurred more than 10 times in the dataset. The number of times a pattern occurred is listed next to the pattern.

Group 1 (Counter-punchers):

Top 5: {'b2Xf1': 512, 'f1Xf3': 379, '5Xb2': 295, '5Xf2': 285, 'f2Xf1': 277}

265 patterns were found out of the 1082 points considered.

Group 2 (Aggressively Consistent Players):

Top 5: {'f1Xf3': 668, 'b2Xf1': 573, '5Xb2': 523, '5Xf2': 472, 'b29Xf1': 457}

500 patterns were found out of the 2240 points in this dataset.

Group 3 (Aggressive Baseliners):

Top 5: {'b2Xf1': 721, 'b3Xb3': 666, 'f1Xf3': 451, '5Xb2': 367, '5Xf2': 322}

382 patterns were found out of the 1615 points considered.

We can contextualize the top results for each category following the key provided in the appendix. Additionally, as a reminder, we are only interested in the opponent's shots; thus the 'X' is simply a placeholder for the player of interest's shot. For counter-punchers, 47% of the points lost during a match contained the sequence `'b2Xf1'`. This can be understood as the opponent hitting a backhand down the middle followed by a forehand crosscourt. This sequence is a neutralizing sequence and is often used to regain control of the court using the forehand. This strategy likely plays to the opponents' strengths (aggressive baseliners' and aggressively consistent players' forehands) to regain control of the point against counter-punchers and also signifies that points against counter-punchers tend to be longer, characterized by attack and reset cycles. For the top 5 most frequent patterns, we see points similar to the top result where these patterns are likely neutralizing or opening up the court following the attack and reset cycles.

For aggressively consistent players, `'f1Xf3'` was the most prominent pattern utilized by opponents with a frequency of 30%. This can be understood as the opponent hitting a forehand crosscourt followed by an inside-out forehand to the ad side of the court. If this pattern does not force an error immediately, it puts the opponent in a heavily favorable position. This pattern is likely very effective for the opponents because it forces the player of interest into a defensive position. Opponents likely use this pattern to exploit the fact that aggressive consistent players tend to prefer closer to the baseline. For the top 5 most frequent patterns, we see a similar structure to the top 5 patterns observed by counter-punchers since these players display a strong defensive game; however, we can see the more aggressive nature of these players given that their opponents are forced to hit more backhands while they are on defense.

Finally, the most common shot pattern weakness found for aggressive baseliners was `'b2Xf1'` occurring in 45% of the points. This is consistent with their opponents' strengths and follows from the logic provided for counter-punchers. When playing against aggressive baseliners, opponents often need to neutralize their attacks in order to force an error. For the top 5 sequences, opponents tended to target the backhand side which is commonly a weaker side for aggressive baseliners.

Note that for all playing styles, cross-court rallies, especially forehand cross-court rallies were among the most common patterns. We chose to disregard these patterns because cross-court rallies tend to be neutral and do not give a clear advantage to any player, so we considered these irrelevant in recognizing pattern weaknesses. Additionally, we found that for all playing styles, the common pattern weaknesses recognized were consistent with the strengths of the opponents. Aggressive baseliners and aggressively consistent players tended to lean towards using their forehand and set up forehand patterns while counter-punchers relied on redirecting the ball.

Nonetheless, from our results, players can gain insight into what patterns to recognize during points to gain a competitive advantage.

6. Discussion

In this study, we investigated how to classify WTA playing styles and determine pattern weaknesses for each type. As stated before, the limited dataset size and the bias towards top players may influence the generalizability of our findings to the broader population of WTA players. Nonetheless, we believe that our findings reveal several important insights about how the WTA game has developed and evoke an examination of the differences between the WTA and ATP tours. First, our analysis resulted in three different playing styles for the WTA players. Playing styles are built both from nature and nurture, but we believe that nurture has a greater impact on the variety of playing styles on the WTA tour. We chose to not take into account the physical attributes of the players in this study to show that physicality is not a deterministic factor when it comes to the development of a playing style for WTA players. For example, Angelique Kerber and Karolina Pliskova share similarities in terms of height and build, as they both stand around 6 feet tall. However, their playing styles differ dramatically. Kerber is known for her exceptional defensive abilities, quickness, and counter-attacking skills, and she relies on her ability to retrieve and redirect shots to frustrate opponents. In contrast, Pliskova possesses a powerful serve and aggressive baseline game, often looking to dictate points with her flat groundstrokes and imposing presence from the baseline.

On the other hand, the conditions and the available resources in which one grows up can have a larger impact on how a player's game style develops. For instance, a player who grows up on clay courts may develop a baseline-oriented playing style, focusing on consistent

groundstrokes and defensive skills. Alternatively, players who train on fast hard courts may emphasize aggressive play, utilizing their natural power and offensive shots. Additionally, coaching and training methods play a crucial role in honing technique, footwork, shot selection, and tactical awareness. Different coaches may emphasize specific strategies or techniques, leading to variations in playing styles. Some players may be coached to adopt an aggressive and offensive style, while others may be trained to excel in defensive skills or rely on versatility and shot variety. Furthermore, the influence of nurture extends to the player's mental and psychological development. Factors such as discipline, work ethic, resilience, tactical awareness, and strategic decision-making are often nurtured through proper coaching and training, affecting a player's style and competitiveness on the court.

Comparing our results with the number of playing styles on the ATP tour, we observe that there are fewer playing styles. One of the reasons may be due to physical differences. On average, male tennis players tend to have more power and speed due to physiological differences such as muscle mass and testosterone levels. This can lead to greater diversity in playing styles on the ATP tour and a greater emphasis on physicality, with players showcasing a wider range of aggressive and power-based approaches. Coaching methods and development programs for women's tennis also may focus more on consistency, movement, and defensive skills rather than encouraging a wide variety of playing styles. Coaches may prioritize developing players who excel in specific areas, leading to a narrower range of playing styles on the WTA tour compared to the ATP tour. Because there are fewer playing styles on the WTA tour, it may explain why the results are less consistent on the women’s side where no single style or player is able to dominate on the tour (where 14 different women have won a Grand Slam title in the last six years which is a significant difference compared to the men’s side where there have only been 9 different men).

In terms of shot pattern weakness, we recognize that it is difficult to generalize a group of tennis players as having the same weakness, given that each match-up might be subjective and weaknesses are likely to vary player-to-player despite players within the same playing style. A possible objection to the relevancy of this study might be that if certain effective patterns become too prominent in the WTA tours, this will take away from the entertainment value that brings fans to stadiums. Having common patterns can discourage creative point construction and hinder the development of unique styles at the individual level. Additionally, this could force the redefinition of playing styles on tour. Nonetheless, we believe that shot pattern weakness as an aggregate statistic provides an initial framework for quantifying strengths and weaknesses in the game of tennis and opens the door for a more comprehensive analysis of the game and its evolution.

As discussed before, one of the limitations of this research project is that we worked with a very small dataset of only 4% of the matches played in the past 6 years. Because the dataset was so small, there was little room for flexibility in removing outliers, such as when the surface type or opponent playing style may be affecting a player's game or if the player is having a really good or bad day which may skew the data. There was also very little match data for WTA players who are ranked outside the top 200, in which we only covered 181 players out of a total of 1650 players on the tour. However, the issue with expanding the dataset across a longer period of time is that it would not account for technological improvements with rackets or new developments in playing styles. Another limitation of this study is that people's playing styles are multifaceted and not clear-cut as a single type. These WTA players have had very diverse backgrounds and grew up with different resources to develop them into the games they possess

now, making each player a unique aggregation of playing styles. This assumption was made as a simplification for the project but is not entirely representative of these players' games.

7. Conclusion

In conclusion, this research paper aimed to classify the playing styles on the WTA tour and explore pattern weaknesses for each type. This analysis revealed that there are three prominent playing styles on the WTA tour: aggressive baseliners, counter-punchers, and aggressively consistent players. However, more match data is needed to accurately categorize these WTA players. For each of these playing styles, we were also able to identify their most common shot pattern weaknesses, focusing on the top five most common sequences. We learned how most of these player types tend to cater toward using their strengths to pinpoint their opponents' weaknesses. These shot patterns can become useful for players to recognize during matches how their opponents are setting up the point and develop methods to combat these patterns.

Understanding the future implications of playing styles in the WTA can have several practical applications. Coaches can leverage these insights to tailor training programs that maximize players' strengths and address potential weaknesses. Additionally, analyzing playing styles can aid in the development of effective game strategies, enabling players to compete more successfully at both individual and team levels. Moving forward, further studies could explore various types of categorical data that could assist in quantifying these styles and implementing a matchup tool that can identify the most successful shot patterns to use against specific opponents. Additionally, investigating the influence of external factors like the environment these players grew up in and the conditions they played on through case studies would provide a

more comprehensive understanding of the development of styles within the WTA. Ultimately, by diving deeper into the analysis of WTA playing styles and their future implications, we can enhance our understanding of the sport, support player development, and contribute to the ongoing evolution of women's tennis.

8. Acknowledgements

We would like to thank all the individuals who contributed to the Match Charting Project. A special thanks to Jeff Sackman for creating this online community to record match statistics. Furthermore, we would like to thank Mike Alvarez and Oliver Eslinger for creating this research class and for their guidance and feedback throughout the term.

9. References

[1] Ontiveros, Johnattan. “Sorting Strokes: Classifying Tennis Players Based on Stats and Style.” *The Harvard Sports Analysis Collective*, 28 July 2021, https://harvardsportsanalysis.org/2021/07/sorting-strokes-classifying-tennis-players-based-on-stats-and-style/.

[2] Devin, Nicholas. “Data Analysis of WTA Player Styles, 2012-2020.” *NYC Data Science Academy*, 15 Dec. 2020, https://nycdatascience.com/blog/student-works/analysis-of-wta-player-styles-2012-2020/.

[3] Liyanage, Shane. “Optimal Shot Sequences by Court Position and Player Types.” *SportRxiv*, 13 Dec. 2022, https://doi.org/10.51224/srxiv.233.

[4] Sackmann, Jeff. “The Match Charting Project: Quick Start Guide.” *Heavy Topspin*, 23 Sept. 2015, www.tennisabstract.com/blog/2015/09/23/the-match-charting-project-quick-start-guide/.

[5] Sackmann, Jeff. “Match Charting Project.” *GitHub*, 2014, www.github.com/JeffSackmann/tennis_MatchChartingProject.

[6] United States Tennis Association Incorporated. “Player Development Journal.” *USTA*, 2013, http://assets.usta.com/assets/1/15/8086_Player_Development_Journal.pdf.

10. Appendix A

The following demonstrates examples of how the point-by-point data is represented and references the Tennis Abstract blog [4].

Symbol Key:

- Serve direction (4 = wide, 5 = body, 6 = down the T)
- Shot codes (f = forehand, b = backhand, s = backhand slice, r = forehand slice, v = forehand volley, z = backhand volley, o = overhead, p = backhand overhead, u = forehand drop shot, y = backhand drop shot, l = forehand lob, m = backhand lob, h = forehand half-volley, i = backhand half-volley, j = forehand swinging volley, k = backhand swinging volley)
- Shot direction (1 = right-hander's forehand side/left-hander's backhand side, 2 = down the middle, 3 = right-hander's backhand side/left-hander's forehand side)
- Point ending codes (@ = unforced error, # = forced error, * = winner)
- Type of error (n = net, w = wide, d = deep, x = wide and deep)

*4ffbbf**

This array of codes means (without including shot direction): serve out wide, forehand return, forehand, backhand, backhand, forehand winner. The table will also indicate the server (i.e. player 1 or 2) for each point charted.

6svlon@

This means: serve down the T, forehand volley, lob, overhead into the net for an unforced error.

b3s3b1f1v2

This means (without including the serve and assuming both players are righties): backhand crosscourt, backhand slice crosscourt, backhand down the line, forehand crosscourt, forehand volley down the middle).

6f2d@

This means: serve down the T, forehand return down the middle missing deep for an unforced error.

11.Appendix B

Our code is available at:

https://colab.research.google.com/drive/1X2UU0O-t9apvZ4UDBjVGrgk7mGOxC9tk?usp=sharing