Daniel Glus (dhg73), Ashley He (ach238), Kati Hsu (kyh24)

How to Become the World's Best Tennis Player 101

CS3300 P1 Write-Up

Our Data

We found our dataset using https://www.kaggle.com/ryanthomasallen/tennis-match-charting-project/. The dataset is comprised of detailed data on every aspect of a particular match for all existing, charted ATP (Association of Tennis Professionals) and WTA (Women's Tennis Association) matches. The dataset goes as far as delineating players' serve breakdowns, point-by-point breakdowns, shot directions, etc.

As the dataset encompasses a large number of variables, we have selectively chosen a subset of variables to take into consideration while creating our visualizations and revealing our "story."

- ***Percentage of Aces:*** In a given match, the percentage of a player's serves that resulted in an ace
- ***Percentage of First Serves Made*****:** In a given match, the percentage of a player's first serves that were made
- ***Percentage of Second Serves Made*****:** In a given match, the percentage of a player's second serves that were made
- ***Percentage of First Serve Points Won:*** In a given match, the percentage of points that were won after a successful first serve
- ***Percentage of Second Serve Points Won:*** In a given match, the percentage of points that were won after a successful second serve
- ***Percentage of Winners:*** In a given match, the percentage of shots that were not ever reached by the opponent
- ***Percentage of Unforced Errors:*** The percentage of mistakes that were made due to a player's own fault, not due to the opponent's skill
- ***Percentage of Return Points Won:*** The percentage of points won after the player successfully returned a serve
- ***Percentage of Break Points Won:*** The percentage of points won while the player was one point away from winning a game while the opponent was serving
- ***Game Score, Total Number of Sets and Points, and Server***

Aside from filtering out the categories that are not stated above, we also made the decision to filter out tiebreaker points, as they stray from the traditional scoring of tennis

matches; since they deviated from the traditional tennis scoring system, these tiebreaker points were not relevant to our visualization demonstrating point progression.

Design Rationale

- ***Tree Chart***
  - ***Marks:*** Nodes and branches of tree
  - ***Channels:***
    - Color of the nodes -
      1. “Gray” = tied score
      2. “Pink” = score is in Cilic’s favor
      3. “Green” = score is in Federer’s favor
    - Thickness of branches - Varying thickness of branch portrays varying frequency of point transitions from top node of branch to bottom node of branch (thicker means higher frequency, thinner means lower frequency)
    - Color of branches - color represents point transitions that occurred in the first, second, or last third (as specified below) of the match
      1. “Dark Blue” - First third of points in the match
      2. “Medium Blue” - Second third of points in the match
      3. “Light Blue” - Last third of points in the match

  For our first visualization, we used a modified Sankey chart to display point progression within a match. As mentioned above, the thickness of the branch was scaled by the number of transitions between the two points (nodes) on the chart. For special cases like after the score 40-40, where the score could continuously transition between 40-Ad and Ad-40, we chose to add in curved lines with arrows to better distinguish difference in those transitions. For the remaining branches, the direction is assumed to be downward as the score should be increasing at all other times.
- ***Radar/Spider Chart***
  - ***Marks:*** Plotted circle points, Area/Chart Segments/Slices
  - ***Channels:*** Axial Position - Position on the axis designates the percentage value of the axis’s category

  For our second visualization, we chose to use a radar chart with two different colored shapes overlaid to represent the statistics for two different players. A radar chart allows us to quickly and clearly compare a range of related characteristics. We initially had to experiment with which values would look appropriate for our visualization, as we didn’t want certain percentages to appear skewed (e.g. if one point was 80%, but the two points surrounding this axis were close to 0%, the area would appear skewed). We ended up using statistics important for deciding or perhaps predicting a match winner, while also close to each other within a reasonable range of about 75% between the minimum and maximum values.

Because this chart only compares two players, which we've chosen to differentiate by color, a color "scale" was not so much necessary as just choosing a pair of decipherable and recognizable colors to represent the players. As for the scaling of the chart, we created our radar chart with 5 concentric background circles equidistant from its immediately surrounding circles. From the inner circle to the outermost circle, their intersections with the axes represent 20%, 40%, 60%, 80%, and 100%, respectively, as the possible values of a percentage range from 0% to 100%.

Story Time

Oftentimes, the final score of a tennis match does not reveal anything about how close the match actually was, or how important key points were during the match. In our first visualization of point progression, we wanted to display just how closely contested the final round of the 2018 Australian Open was. With a final set score of 6-2, 6-7, 6-3, 3-6, 6-1 in Roger Federer's favor, it appears that he had a relatively easy time winning with set scores of 6-2, 6-3, 6-1, but the score doesn't reflect Marin Cilic's brief comeback in the last third of the match. Looking at the point progression visualization, it's surprisingly clear how one sided a few games were in Cilic's favor late in the match, with Cilic closing out Federer in two games. However, points transitioning in Federer's favor were more consistently spread across the entirety of the match, suggesting that although Cilic may have played well for particular games, the more consistent player will still win the overall match. Key points (break points are at 30-40 where your opponent is serving, deuce points are at 40-40) most often transitioned in favor of Federer, which may also suggest that he handled high pressure situations better than his opponent, and that he was able to perform just as consistently as he had been for the rest of the match.

The idea of the more consistent player winning a match is again reflected in the radar chart visualization of key statistics for each player in the match. Though they were close in most statistics, like percentage of winners or first serve percentage, there are a few important statistics where Federer performed significantly better than Cilic did. Players typically want to minimize unforced errors (a miss due to the player's own fault), but Cilic's percentage is above 40%, whereas Federer's percentage is almost half of that. Additionally, Federer's return point win percentage and break point win percentage are significantly higher than Cilic's. In tennis, having a high return point win percentage is significant since it's usually easier for the server to win their service game as they start off the point, placing the ball to give themselves the upper hand. The returning player, more often than not, starts off in a defensive position and will have to neutralize the serve before being able to take control of the point, which is much more difficult to do. The visualization suggests that even in disadvantageous positions, Federer outperforms Cilic and can take control of the point, giving him a better chance of winning the point. Again because it's typically easier to win a game while serving, winning break points are also an important factor in deciding the overall winner in a match. In most professional matches, it can be assumed that players will win their own service games, and winning just one break point in a set will

allow one player to take the set. So, winning break points when they are reached is crucial and should be capitalized upon. In our radar visualization, Federer dominates Cilic in break point win percentage, which again highlights his superior performance in high pressure situations. Overall, both of our visualizations convey the importance of maintaining consistent performance over the entire match, paying special attention to key points like break points and return (of service) points.

Team Contributions

- ❏ **Daniel Glus:** Initial research on building usage idea, sketches and scraping datasets for class times vs. major idea, coding for class times vs. major idea, parsed the data to be utilized in tennis visualization code, final coding for game tree and radar chart visualizations
- ❏ **Ashley He:** Initial research on gender wage gap and tennis match visualization ideas, found datasets and selected specific matches for tennis visualization, cleaned data and calculated percentages to include only relevant data points, initial sketches and code for point progression and radar chart visualizations, pair programmed with group, story and design (radar chart) sections of write-up
- ❏ **Kati Hsu:** Found chart types to use to visualize our data, executed rough visualizations of data using Tableau/Google Data Studio, cleaned up code and added comments, pair programmed with group, coordinated design details with group, write-up material