UNIVERSITY OF TECHNOLOGY SYDNEY

------------------------------

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

**QUANG AN PHAM**

# DATA VISUALISATION AND VISUAL ANALYTICS

**Data Visual Assignment 1**

Sydney – Year 2023

## Executive summary

The Australian Open Champions Dataset spans from 1905 to 2023, tracing the evolution of one of the Southern Hemisphere's premier sporting events. Originating as the Australasian Championships in 1905 and later renowned as "the happy slam", the Australian Open marked a record attendance of over 902,000 in 2023. While the men's championship commenced in 1905, the women's tournament was introduced in 1922. This dataset encompasses 209 meticulously labeled records distributed across 21 columns, each detailing essential facets of the championship. It serves as a rich resource for those keen on exploring the expansive history and nuances of this prestigious Grand Slam event.

Using Excel to calculate and find the top players by finding the number of times they get the champion of the Australia Open tournament by subset function in Excel. There are 7 top players who get more than 5 times win the event, consisting of:

- 4 women: Serena William, Daphne Akhurst, Margaret Smith, Nancye Wynne Bolton
- 3 men: Novak Djokovic, Roger Federer, Roy Emerson

Key conclusion:

- Dominance of Certain Players: Novak Djokovic, Margaret Smith, and Serena Williams emerged as dominant figures in their respective categories, showcasing consistent performance across tournaments.
- Gender Disparities: Women, on average, displayed a higher win rate than men. However, the men's games were more evenly contested, signifying greater competitiveness.
- Consistent Performers: While outright victories are a clear metric of success, players like Jan Lehance and Andy Murray, despite multiple runner-up finishes, highlight the caliber of competition and the thin margins between victory and defeat.
- Countrywide Analysis: Australia prominently produced a significant number of champions, yet also showcased variability in set outcomes, indicating both dominance and closely fought matches.
- Set Analysis: The butterfly chart effectively captured the balance between sets won and lost, providing insights into player consistency and resilience. Players like Djokovic, Smith, and Bolton demonstrated remarkable skill, as evidenced by their win-loss ratios.
- Depth of Competition: The visualization underscored the depth of talent in the tennis world, where even those with fewer titles or runner-up finishes played crucial roles in making tournaments competitive and unpredictable.

Data Exploration

Data exploration is a preliminary step in data analysis where analysts and data scientists delve into the data to uncover patterns, identify anomalies, and gain an understanding of its structure. Using statistical graphics, plots, and information tables, this process provides a snapshot of the data's main characteristics and inherent relationships. By visualizing and understanding the nature of the data at hand, one can determine the appropriate analytical techniques to employ, ensure data quality, and pave the way for more intricate analyses. It's akin to a detective's initial investigation before diving deep into the mystery.

| Attribute name | Type | Description |
|---|---|---|
| Year | 4-digit data, YYYY format, interval quantitative | The year the tournament occurred |
| Gender | Binary data, categorical nominal | Either the kind of tournament or the sex of the champion |
| Champion | String format, categorical nominal | The winner's name of the tournament |
| Champion Nationality | 3 letters string format, categorical nominal | ISO Alpha-3 country code of the champion's nationality |
| Champion country | String format, categorical nominal | Typically refers to the country the champion represents, sometimes it can be different with nationality |
| Score | String format, the integers connect by comma and dash (eg: 6-3, 6-4) | The result of the whole match, each set is separated by commas, and the "tiebreak" is put inside brackets |
| Wins | Integer format, quantitative (discrete) data | The total win games in the whole match |
| Loss | Integer format, quantitative (discrete) data | The total loss games in the whole match |
| Runner-up | String format, nominal categorical data | the name of the runner-up or the name of the second-place player |
| Runner-up Nationality | 3 letters string format, Nominal categorical | 3-letter country code of runner-up's nationality |
| Runner-up country | String format, Nominal categorical | It is like champion country but it belongs to the runner-up |
| The order of sets won and lost (eg: 1 st won, 1st loss, 2nd won) | Integer format, quantitive (discrete) data | The number of games won and lost and each set in order |

Interesting findings:

After exploring the data, there are some interesting keys were discovered:

- The tournament did not take place in 1941-1945, 1916-1918, 1986
- The Australia Open was held 2 times in 1977
- In 1966, the women's champion did not need to play since the opponent walkover
- Men's tournament in 1990 and Women's tournament in 1965, they did not play the final set as their rival players were being retired.
- The women's tournament began in 1922, after 16 years since the first tournament.

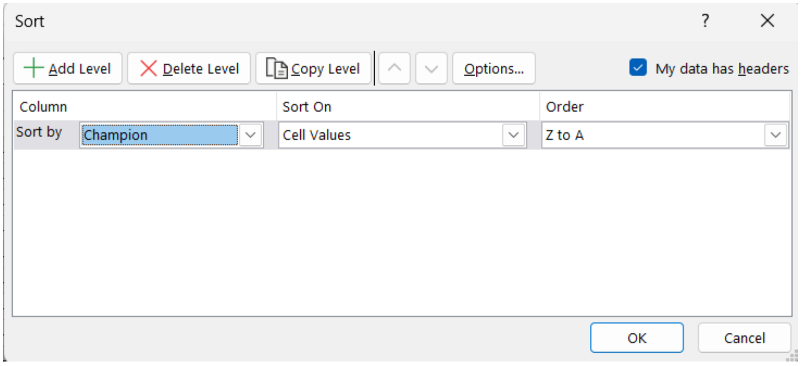
Data Preparation

Data preparation, often deemed a crucial yet time-consuming phase in the data analysis pipeline, involves cleaning, transforming, and enriching raw data to enhance its quality and structure for analysis. This step ensures that data is free from inconsistencies, inaccuracies, and missing values that could skew results or lead to inaccurate conclusions. By reshaping, standardizing, and sometimes aggregating data, analysts create a robust foundation for reliable and meaningful insights. In essence, data preparation is the unsung hero of the data science process, setting the stage for effective analytics and informed decision-making.

The dataset pertaining to the Australian Open Champion adheres to the principles of tidy data. This means that the dataset is pristine, devoid of errors, and does not contain any null values. Consequently, we can bypass the data preparation phase and utilize the original data directly for analysis. Nonetheless, there are a couple of supplementary steps required: specifically, the calculation to identify the "top players" and the formulation of a new variable named "Win rate" within Excel.

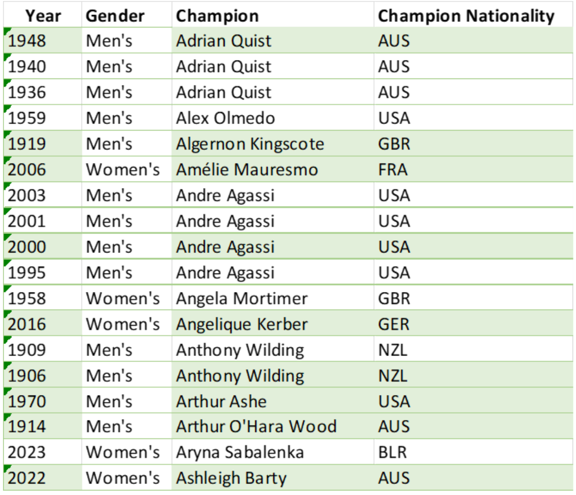
Finding top players:

First, the data needs to be sorted by the Champion. By using custom sorted as in the figure, we can result in a database with the champion column sorted by name A to Z



☑ My data has headers

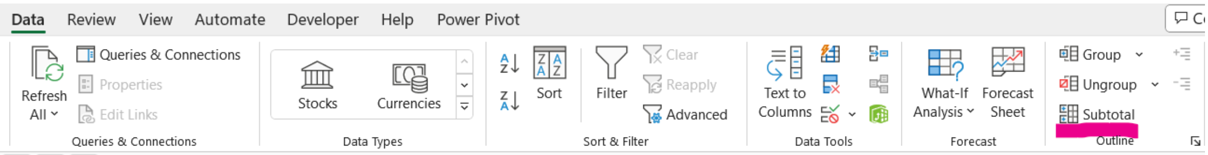
| Column | Sort On | Order |
|---|---|---|
| Sort by Champion | Cell Values | Z to A |



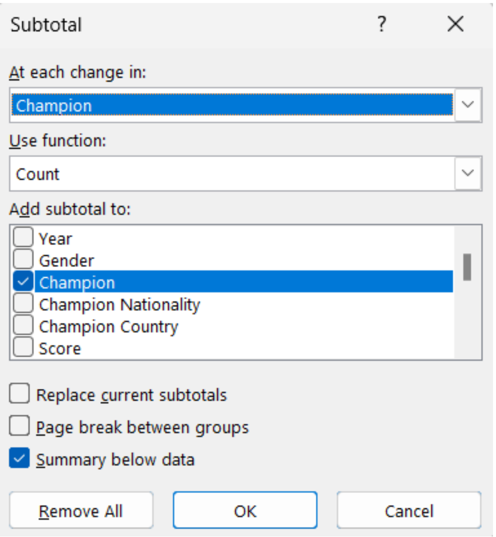
| Year | Gender | Champion | Champion Nationality |
|---|---|---|---|
| 1948 | Men's | Adrian Quist | AUS |
| 1940 | Men's | Adrian Quist | AUS |
| 1936 | Men's | Adrian Quist | AUS |
| 1959 | Men's | Alex Olmedo | USA |
| 1919 | Men's | Algernon Kingscote | GBR |
| 2006 | Women's | Amélie Mauresmo | FRA |
| 2003 | Men's | Andre Agassi | USA |
| 2001 | Men's | Andre Agassi | USA |
| 2000 | Men's | Andre Agassi | USA |
| 1995 | Men's | Andre Agassi | USA |
| 1958 | Women's | Angela Mortimer | GBR |
| 2016 | Women's | Angelique Kerber | GER |
| 1909 | Men's | Anthony Wilding | NZL |
| 1906 | Men's | Anthony Wilding | NZL |
| 1970 | Men's | Arthur Ashe | USA |
| 1914 | Men's | Arthur O'Hara Wood | AUS |
| 2023 | Women's | Aryna Sabalenka | BLR |
| 2022 | Women's | Ashleigh Barty | AUS |

The database result

In the sheet dataset Excel file, the subtotal function is to find the top players who win the champion place more than 5. Then choose all the data go to all the functions of data, and select Subtotal.



The subtotal table appears to choose what needs to be a subset. the subtotal will be the champion by counting the name of the champion, so the use function would be counted and the add subtotal to the would-be champion, following the figure:


Subtotal

At each change in: Champion

Use function: Count

Add subtotal to:

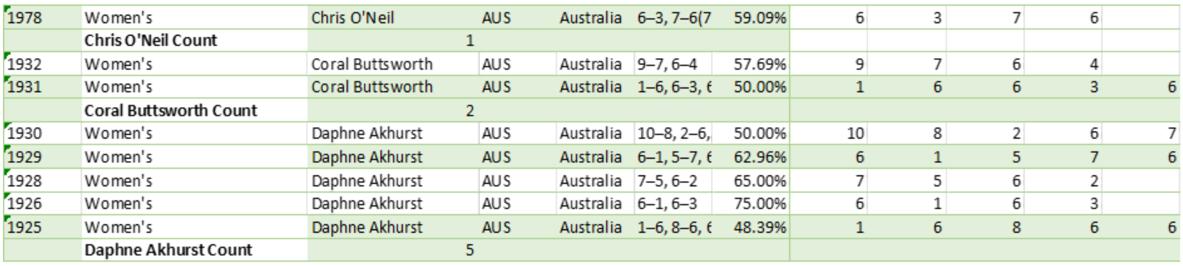
- □ Year
- □ Gender
- ☑ Champion
- □ Champion Nationality
- □ Champion Country
- □ Score

□ Replace current subtotals

□ Page break between groups

☑ Summary below data

The result will create an additional row below each subset of the champion's name like the following image:



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1978 | Women's | Chris O'Neil | AUS | Australia | 6–3, 7–6(7 | 59.09% | 6 | 3 | 7 | 6 | |
| | **Chris O'Neil Count** | 1 | | | | | | | | | |
| 1932 | Women's | Coral Buttsworth | AUS | Australia | 9–7, 6–4 | 57.69% | 9 | 7 | 6 | 4 | |
| 1931 | Women's | Coral Buttsworth | AUS | Australia | 1–6, 6–3, ( | 50.00% | 1 | 6 | 6 | 3 | 6 |
| | **Coral Buttsworth Count** | 2 | | | | | | | | | |
| 1930 | Women's | Daphne Akhurst | AUS | Australia | 10–8, 2–6, | 50.00% | 10 | 8 | 2 | 6 | 7 |
| 1929 | Women's | Daphne Akhurst | AUS | Australia | 6–1, 5–7, ( | 62.96% | 6 | 1 | 5 | 7 | 6 |
| 1928 | Women's | Daphne Akhurst | AUS | Australia | 7–5, 6–2 | 65.00% | 7 | 5 | 6 | 2 | |
| 1926 | Women's | Daphne Akhurst | AUS | Australia | 6–1, 6–3 | 75.00% | 6 | 1 | 6 | 3 | |
| 1925 | Women's | Daphne Akhurst | AUS | Australia | 1–6, 8–6, ( | 48.39% | 1 | 6 | 8 | 6 | 6 |
| | **Daphne Akhurst Count** | 5 | | | | | | | | | |

To find the top player, all the rows related to champions who have a count below 5 would be deleted.

Win rate:

To calculate the win rate of each match the total wins would be divided by the total wins and losses, the formula would be: wins/(wins+loss)
Use the function with appropriate cells to calculate automatically

## Tree Map

A treemap is a visualization tool used to represent hierarchical data using nested rectangles. Each branch of the hierarchy is given a colored rectangle, proportional in size to a particular dimension of the data, often allowing viewers to understand complex data structures and relationships at a glance. The space inside each rectangle can be further divided into smaller rectangles representing sub-categories or sub-groups. Color and size variations in these rectangles provide cues about data value differences and relative proportions, making treemaps especially useful for displaying large amounts of data without overwhelming the viewer. In essence, a treemap transforms a tree diagram's branches into colored areas, providing a spatially efficient and visually appealing way to convey multi-layered data insights.

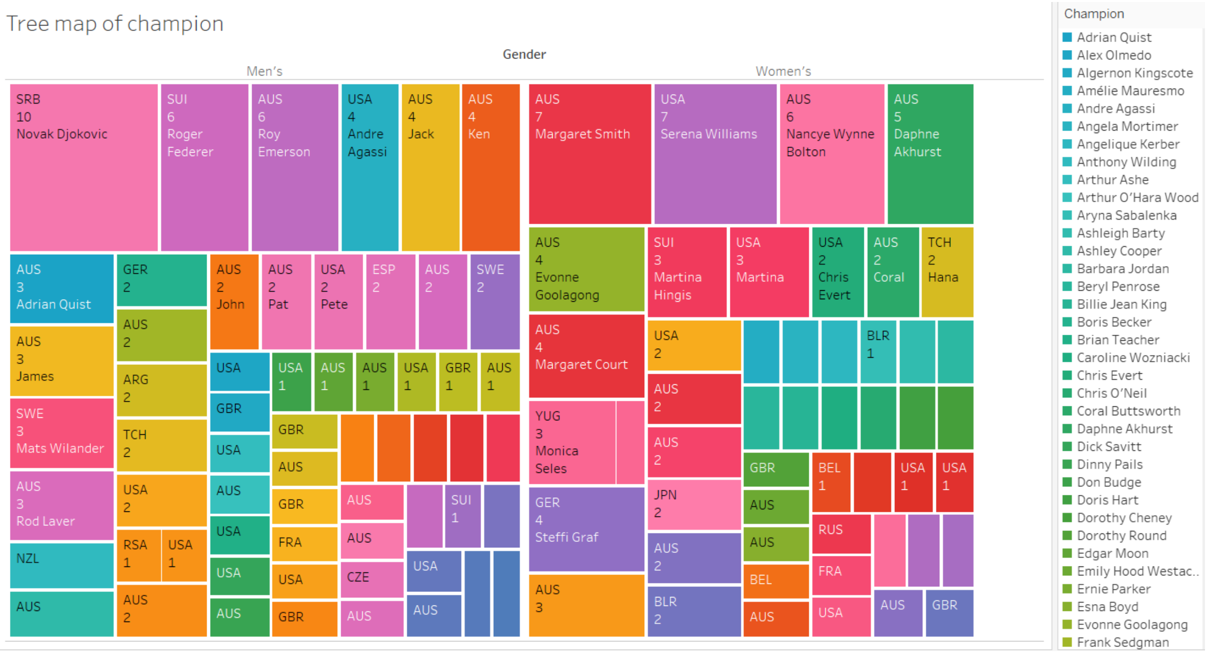


Figure 1: Champion nationality, champion name, gender, and first-place times

The design choices of this treemap were commendable. Each champion was distinctly demarcated using unique colors, facilitating quick identification. Within each cell, the treemap conveniently provided the champion's nationality, the number of victories they had achieved, and their name. Thoughtfully, the data was bifurcated into two gender-specific columns. The size of the cells directly corresponded to the number of times a player had won, allowing for an intuitive understanding of a player's dominance. Players with more titles had larger cells, and these were strategically positioned at the top right of their respective columns for immediate visibility. Notably, Novak Djokovic emerged as a significant figure in the Men's tournament with 10 titles to his name. In the Women's segment, both Margaret Court and Serena Williams were dominant forces, each with 7 titles, and they were represented in comparable positions on the treemap. This exercise underscored the power of visualization in breaking down and elucidating intricate datasets.

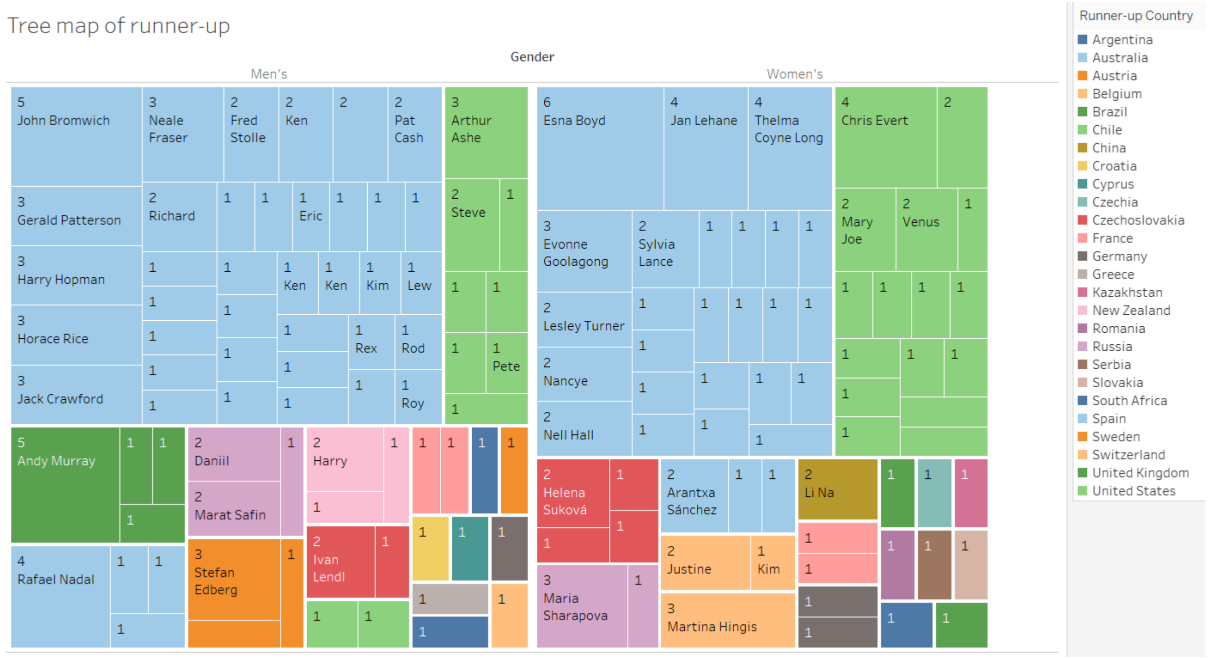


Figure 2: Runner-up country, runner-up times and gender

Unlike the champion-centric treemap, the color coding in this visualization was determined by the country of the runner-up, offering insights into geographical patterns of near-victories. Each cell was clearly marked with the number of times an individual finished as a runner-up. Maintaining consistency, the treemap was bifurcated into two gender-specific columns. For the women's division, Esna Boyd of Australia prominently stood out, having secured the runner-up position six times, a testament to her consistent performances. The men's division revealed an intriguing tie: John Bromwich of Australia and Andy Murray of Brazil both clinched the runner-up spot five times. A notable overarching trend was Australia's dominance in producing players who frequently ended up as runner-ups. This visualization underscored the importance of looking beyond just the winners and recognizing consistent performers who might not have clinched the title but displayed commendable prowess and resilience.

Advantages:
1. Space-Efficient: Treemaps utilize space effectively, allowing for the display of large quantities of hierarchically structured data in a confined area.
2. Visual Clarity: The nested rectangles provide a clear visual representation of hierarchy and category differences.
3. Immediate Insights: Differences in color and size of the rectangles allow for immediate visual comparison, aiding in the quick understanding of data patterns and outliers.
4. Flexibility: Treemaps can represent multiple dimensions of data by using color, size, and hierarchy.
5. Interactive: Modern treemaps can be interactive, allowing users to drill down into categories for more detailed insights.

Disadvantages:

1. Limited Precision: While treemaps are great for a general overview, they might not be as precise in representing exact data values.
2. Overwhelming for Large Hierarchies: If there are too many hierarchical levels or categories, the treemap can become cluttered and challenging to interpret.
3. Perceptual Issues: Due to the variation in rectangle sizes, it can sometimes be difficult for users to compare areas accurately, especially when rectangles have very similar sizes.
4. Lack of Standardization: Unlike bar charts or pie charts which have standard interpretations, treemaps might be unfamiliar to some audiences, leading to potential misinterpretations.
5. Color Limitations: If not chosen wisely, color schemes can sometimes mislead or confuse viewers, especially if they fail to consider colorblind individuals.

In essence, while treemaps are a valuable tool for representing hierarchical data, especially when space is a constraint, they need to be used judiciously and designed carefully to ensure clear and accurate communication of data insights.

## Parallel Coordinate

A parallel coordinate plot, or simply parallel coordinate, is a visualization technique used to plot multivariate, numerical data. Unlike traditional Cartesian coordinates, which plot data on perpendicular axes, parallel coordinates display each data point as a line that intersects a series of vertical, parallel axes. Each of these axes represents a different dimension or variable, and the position at which the line intersects a given axis corresponds to the value of that variable for the data point. By connecting these intersection points across all axes with lines, patterns, clusters, and relationships between variables can be discerned. The technique is especially valuable for spotting correlations and outliers in high-dimensional datasets. By enabling the simultaneous representation of multiple dimensions, parallel coordinates offer a comprehensive view of the intricate structure of complex datasets.

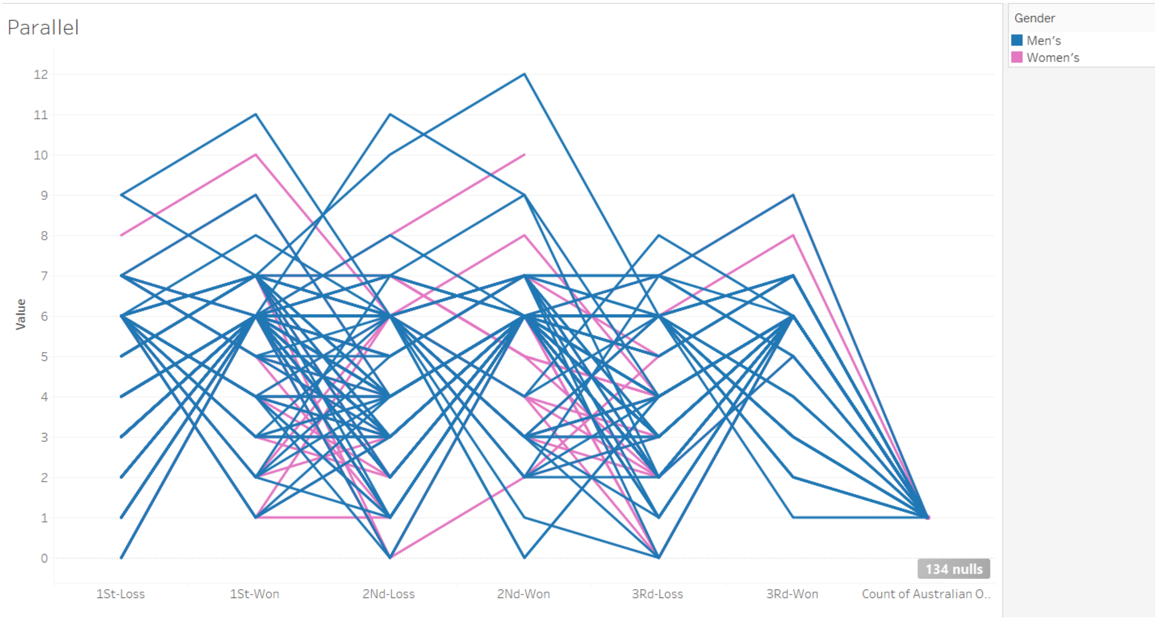


Figure 3: The parallel coordinator of the 3 first set won and loss

The parallel coordinates chart provides a clear differentiation between genders using color coding: pink lines represent women and blue signifies men. A striking observation from the chart is the resilience of women champions; very few of them experience a loss in the first set, indicating a trend where they often secure an early lead and maintain it to win the match. On the other hand, when assessing the length of play, the number of games played in each set tends to be higher for men than women. This suggests that men's matches might be more drawn-out or competitive, while women's champions often clinch their sets with more decisive victories.

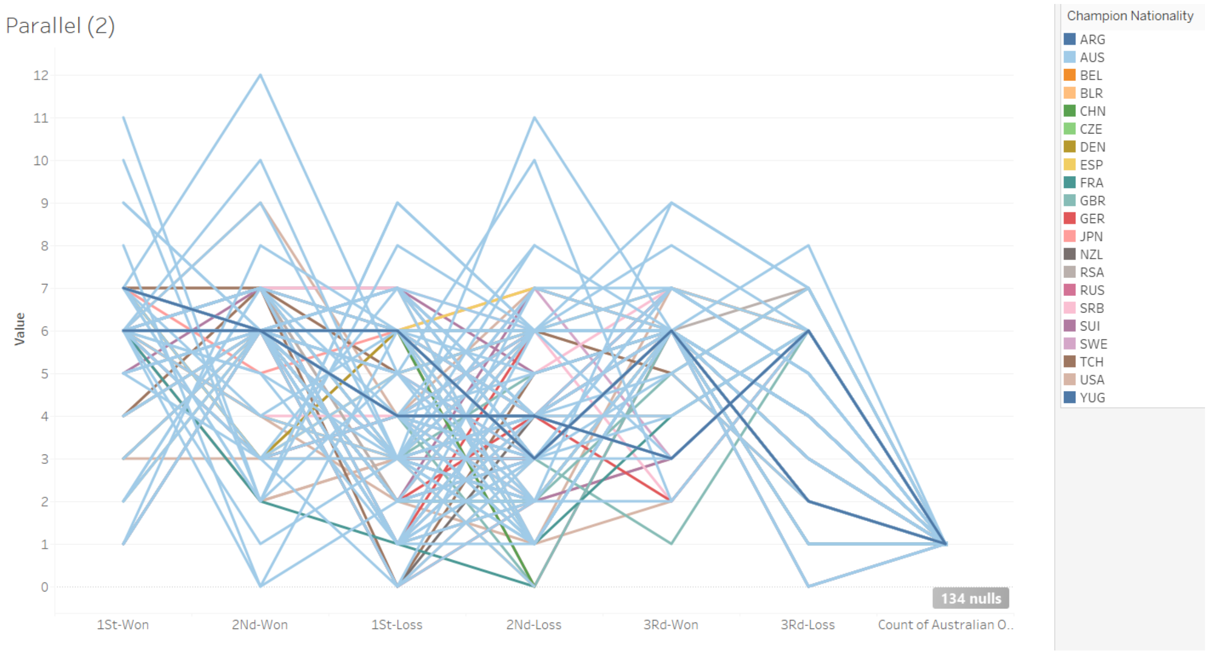


Figure 4: Parallel coordinate of first 3 sets and separated by countries

In this parallel coordinates chart, colors are distinctly assigned based on countries, providing a vivid representation of national performance. Predominantly, Australia, having the highest number of champions, is represented by a dense concentration of color lines. Additionally, the performance spread for Australian players across sets is notably vast, ranging from 0 to 12. This indicates variability in their game outcomes, suggesting that while they often dominate, there are instances where they engage in protracted battles extending to numerous games within a set.

Advantages:
1. Multidimensional Analysis: Parallel coordinates excel in visualizing high-dimensional data, providing insights into datasets that might be challenging to represent using other methods.
2. Correlation Discovery: By examining the patterns of lines (such as parallel lines or crossed lines), users can identify correlations or the lack thereof between different variables.
3. Outlier Identification: Outliers can often be quickly spotted as lines that deviate significantly from the majority.
4. Interactive Exploration: Many modern parallel coordinate tools offer interactive capabilities, allowing users to highlight, filter, or reorder dimensions to explore data more effectively.

Disadvantages:
1. Overplotting: For large datasets, lines can overlap significantly, making it challenging to discern individual data points or overall patterns.
2. Learning Curve: Parallel coordinates can be unfamiliar to many audiences, and there might be a learning curve associated with interpreting them effectively.
3. Dimension Order Sensitivity: The order of the dimensions (axes) can significantly affect the visualization's appearance and might lead to different interpretations. Selecting the best order is not always straightforward.
4. Limited to Numerical Data: Parallel coordinates are primarily suitable for continuous, numerical data. Categorical data requires additional processing or alternative visualization techniques.
5. Scale Sensitivity: The scales of different dimensions need careful consideration. If scales are vastly different, normalization or standardization might be required.

While parallel coordinates provide a powerful tool for visualizing and analyzing multivariate data, they require careful design and consideration to ensure that the insights derived are accurate and meaningful.

## Geographic map

A geographical map is a visual representation of an area, showcasing features such as landforms, bodies of water, political boundaries, cities, roads, and more. Rooted in the ancient need to understand one's surroundings, these maps serve as tools to navigate, plan, and comprehend the spatial layout of our world. Modern geographical maps are often overlaid with additional layers of data, transforming them into thematic maps that can illustrate various aspects like population density, climate regions, or economic activities. Enhanced by digital technology, today's maps are interactive, allowing users to zoom, pan, and click to access more detailed information. They bridge the abstract concept of data with the tangible reality of geography, enabling us to visualize and analyze spatial relationships and patterns in diverse contexts.

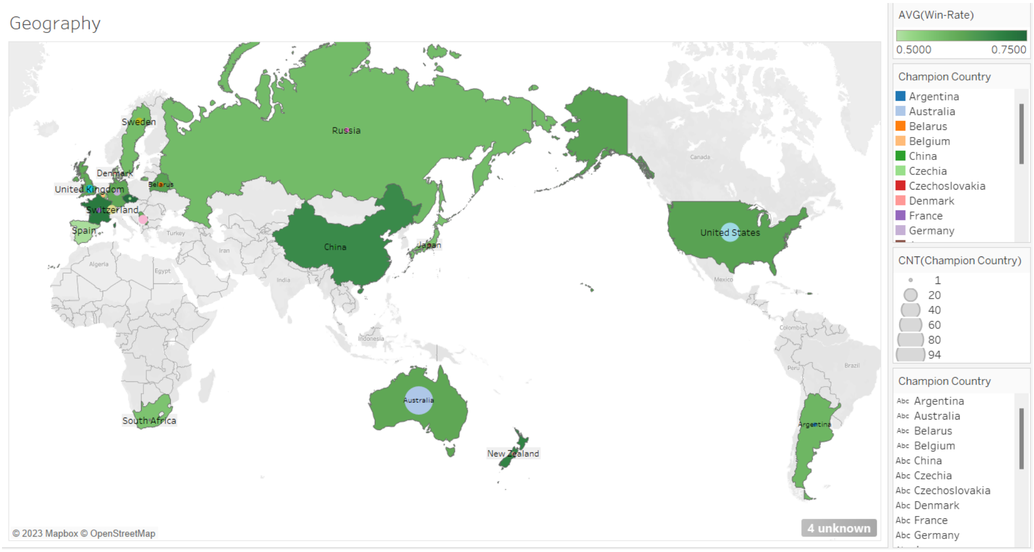


Figure5 : The geographical map of the win rate and count of champion country

On the geographical map, varying shades of color represent the win rate of each country: the darker the shade, the higher the average win rate. Overlaying this, circles pinpoint countries that have produced champions. The size of these circles corresponds to the number of champions from that country, with Australia boasting the largest circle, reflecting its impressive tally of around 94 championship wins. However, while Australia's championship count is substantial, its win rate is of medium stature. Remarkably, the top three countries with the highest win rates are China, Switzerland, and the Czech Republic. Yet, it's intriguing to note that China, despite its high win rate, has produced only one champion, emphasizing the nuance between sheer championship counts and consistent performance rates.

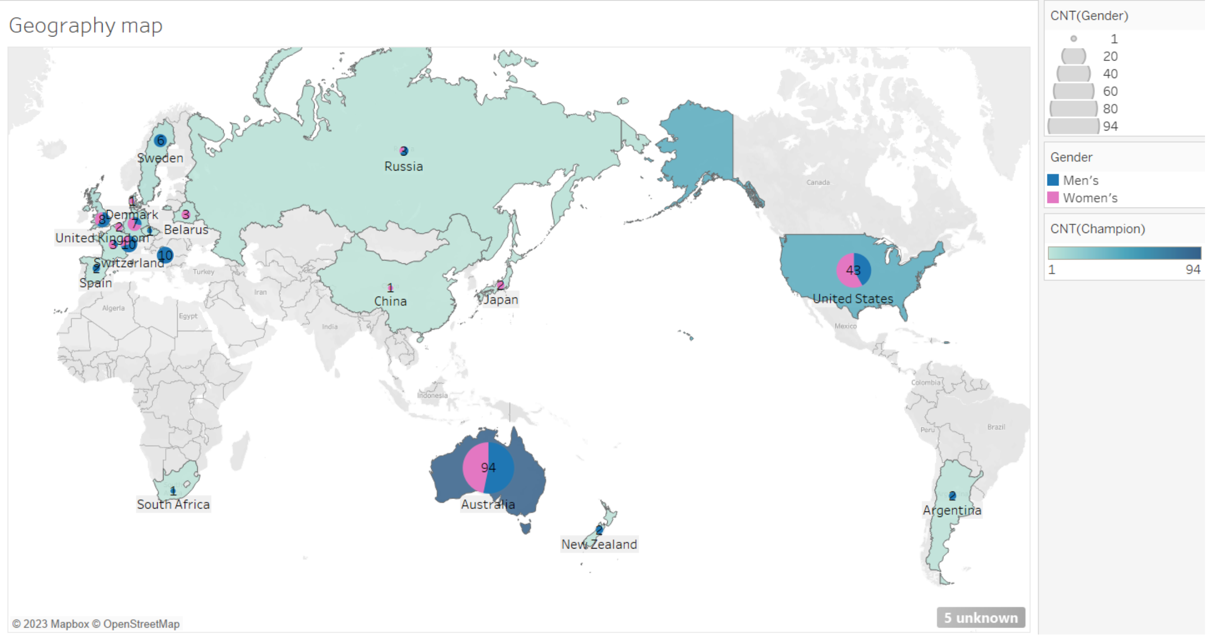


Figure 6: Geographical map of champion number and the gender rate of champion number

In this geographical map, both the circle size and the lightness of color depict the count of champions produced by each country, but with an added twist: each circle is segmented into a pie chart that provides a gender breakdown of the champions. Australia, with the most champions, has a predominant section of its circle representing male winners. Following Australia, the USA stands out not just for its number of champions but also for a greater proportion of those being women. Distinctly, countries like Belarus, Belgium, Denmark, Japan, and China are exclusively represented by female champions, as their circles lack a male segment. Conversely, Sweden, Spain, South Africa, Czechia, and Serbia have circles indicating only male champions, underscoring the gender-specific dominance in championship wins for these nations.

Advantages:

1. Intuitive Visualization: Geographical maps provide an immediate and intuitive sense of spatial relationships, offering viewers a natural way to understand data in the context of physical locations.

2. Multifaceted Analysis: Modern maps can overlay multiple layers of data, allowing users to understand and analyze complex interactions and trends across different themes within the same spatial framework.
3. Interactivity: Digital geographical maps are often interactive, enabling users to zoom, pan, click, and even incorporate real-time data, enhancing user engagement and depth of exploration.
4. Decision Making: For businesses and governments, geographical maps help in making location-based decisions, from optimizing delivery routes to selecting new store locations or understanding disease outbreaks.

Disadvantages:
1. Generalization: Maps often simplify or generalize data to make it fit or readable, which might lead to loss of detail or inaccuracies.
2. Projection Issues: Representing the Earth's curved surface on a flat map leads to distortions. Different map projections offer trade-offs between preserving area, shape, or distance, but no projection is perfect.
3. Overload and Clutter: Overlaying too much information can make a map confusing and difficult to interpret. Striking a balance between detail and clarity can be challenging.
4. Cultural and Political Biase*: Maps can sometimes reflect the biases of their creators. Borders, names, or even the data displayed can be influenced by political or cultural viewpoints.
5. Dynamic Data Challenges: While static data can be easily plotted, representing dynamic or frequently changing data on maps can be more complex and resource-intensive.

In summary, while geographical maps offer powerful tools for visualizing spatial data and relationships, it's essential to approach them critically, considering their limitations and the potential biases they may carry.

## Scatter plots

A scatter chart, also known as a scatter plot or scatter graph, is a graphical representation used to display values for typically two variables for a data set. The data is displayed as a collection of points, each with its x-y coordinates representing the values of the two variables. Scatter charts are ideal for examining the relationship between the two variables, helping in spotting trends, concentrations, or outliers within the data. When data points tend to rise or fall together and form a consistent pattern, they may indicate a correlation between the two variables. Though a simple construct, scatter charts are profoundly versatile and can offer deep insights into data distribution, density, and the underlying relationship between variables.

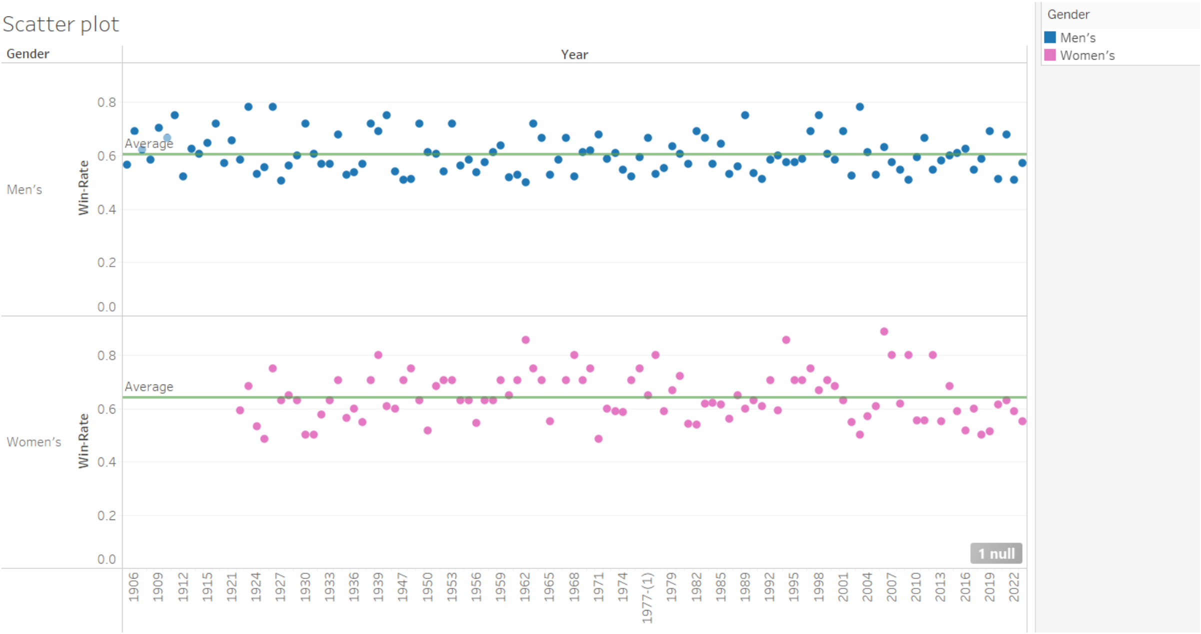


Figure 7: Scatterplot for the win rate in each year sorted with gender

In the scatter plot, blue dots symbolize the win rate of men's matches, whereas pink dots signify women's win rates. A notable observation is the elevated average win rate for women when juxtaposed with men. There's even a point representing women's tennis with a win rate nearing perfection, suggesting instances of dominant performances by female champions. In contrast, the distribution of the men's win rate is more uniform, indicating a tighter competition. This equitable spread for men suggests that their matches often involve intense battles, with closer contests between the champion and the runner-up compared to women's matches.

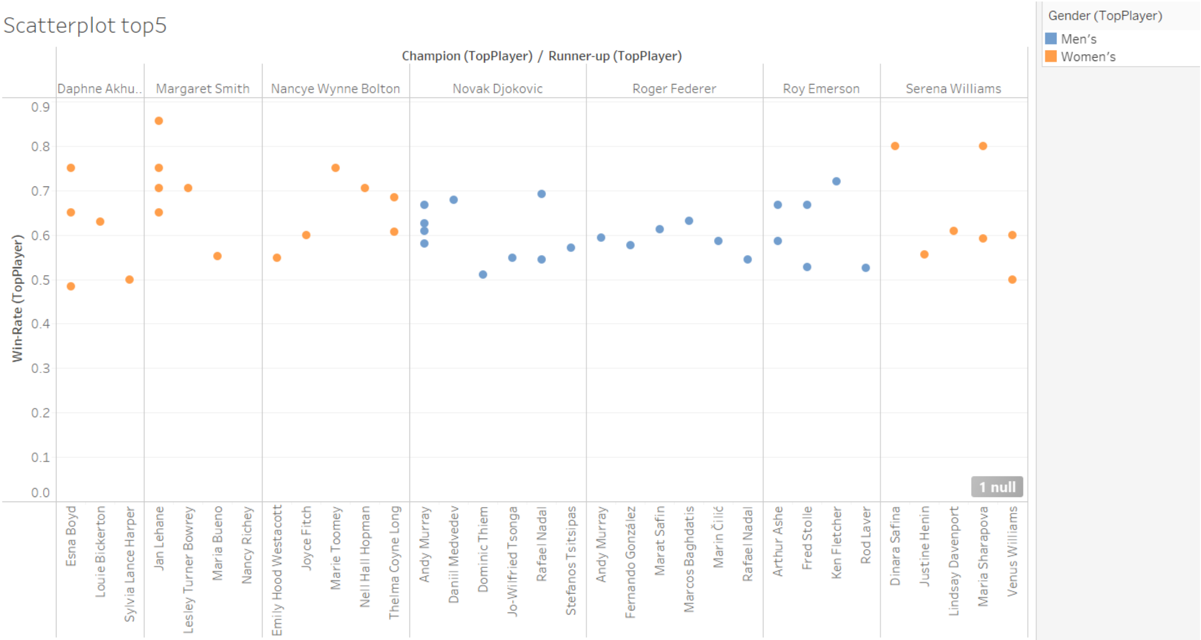


Figure 8: The win rate between the champion and the runner-up

In the analyzed data, the win rate for women tends to surpass that of men, underscoring the pronounced dominance of certain female players. It's particularly heartrending to note the fortunes of Jan Lehance and Andy Murray, with each enduring the anguish of falling just short of the title four times. In terms of unparalleled success, Margaret Smith's win rate stands out, clearly eclipsing her contemporaries and firmly establishing her legacy. Serena Williams, another titan of the sport, closely follows Margaret in terms of achievements. When observing the men's segment, distinguishing a clear front-runner in terms of win rate becomes challenging, indicative of closely contested and fiercely competitive matches among the male elite.

Advantages:
1. Visualization of Relationships: Scatter plots excel in revealing the relationship between two variables, helping identify patterns such as linear, exponential, or clusters.
2. Identification of Outliers: Outliers can be easily spotted as points that deviate from the general pattern of the rest of the data.
3. Distribution Insights: Scatter plots provide a clear view of the distribution and concentration of data points in a dataset.
4. Flexibility: Can be used for a broad range of data types, whether continuous or categorical.
5. Ease of Interpretation: With the right scale and axis labeling, scatter plots are straightforward and intuitive, making them easily interpretable even by non-technical audiences.

Disadvantages:
1. Limited to Two Variables: Traditional scatter plots are limited to showcasing the relationship between two variables at a time.
2. Overplotting: For large datasets, points can overlap, making it difficult to discern the density or number of points at a particular location.
3. Less Effective for Categorical Data*: While scatter plots can represent categorical data, they are most effective for continuous data. Other visualizations may be more suitable for purely categorical datasets.
4. Requires Meaningful Axes: The chosen axes must have a meaningful relationship for the scatter plot to provide valuable insights. If they don't, the plot might be misleading or nonsensical.
5.*No Information on Time: If the dataset has a temporal dimension, scatter plots don't inherently convey time progression, unless time-series scatter plots or animations are used.

In conclusion, while scatter plots offer a powerful visualization technique for examining relationships between two variables, it's essential to consider the nature of the data and the specific insights desired to ensure they are used effectively.

## Top Player Visual Analysis

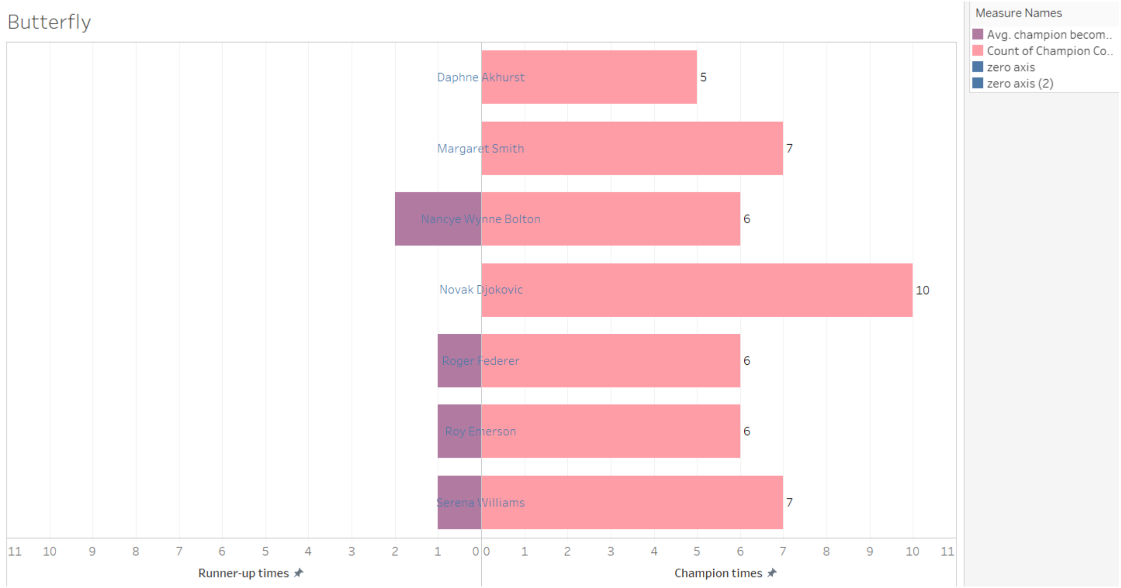


Figure 9: The butterfly bar chart shows the runner-up times and champion times of top players

In the butterfly chart, an effective tool for comparison, the left axis delineates the number of times players have finished as runners-up, while the right axis quantifies their championship wins. Standing prominently on this chart is Novak Djokovic, whose remarkable record boasts of 10 championship victories without a single loss in a final match, epitomizing dominance. Daphne Akhurst and Margaret Smith mirror this trend, both displaying flawless records by clinching titles each time they reached the finals. Another notable presence on the chart is Nancye Wynne Bolton, who secured the champion's title six times, but also faced the bittersweet reality of being the runner-up twice, showcasing her consistency and tenacity in reaching the final stages multiple times.

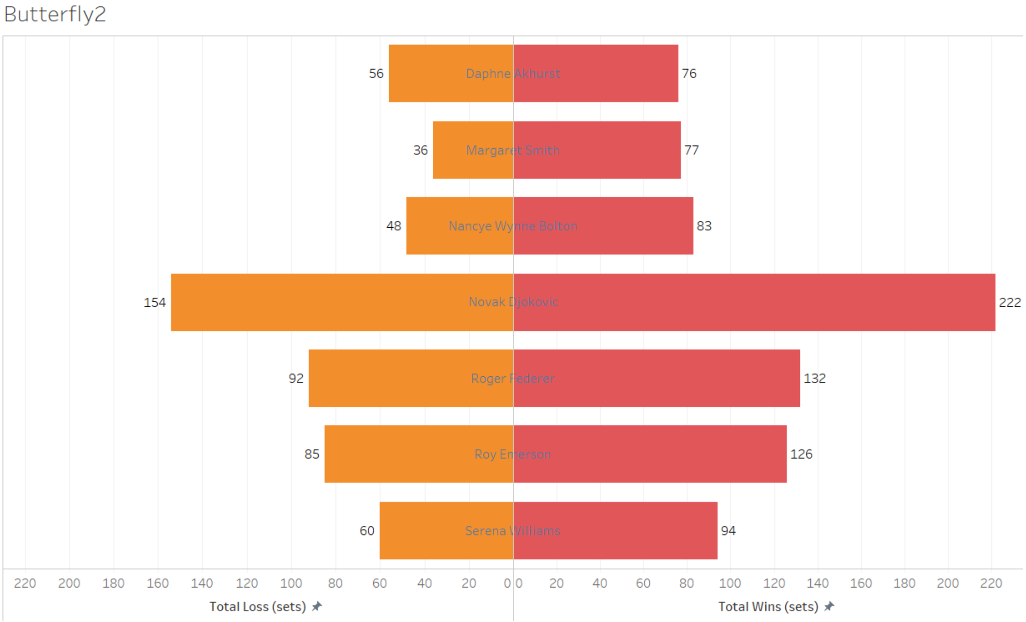


Figure 10: The butterfly bar chart shows the total losses and wins of the top champion

In this chart, the left axis highlights the total sets players lost, while the right axis represents the total sets they've won. Novak Djokovic's statistics stand out prominently; as a player with the most tournament victories, it's unsurprising that his set wins significantly outnumber his losses, illustrating his consistent performance and dominance. On the women's side, both Margaret Smith and Nancye Wynne Bolton exhibit a pattern where their total losses amount to only half of their total wins. This suggests that not only did they consistently progress through the tournament stages, but they also demonstrated remarkable skill and strategy in most matches they played.

Conclusion:

- Women consistently display a higher win rate compared to men, with certain female players demonstrating pronounced dominance.
- Jan Lehance and Andy Murray have both faced the disappointment of finishing as runner-up on four occasions, highlighting the challenges and unpredictability of top-tier tennis.
- Margaret Smith's win rate is unparalleled, placing her at the pinnacle of the sport. Serena Williams closely follows her in terms of achievements.
- The men's matches appear to be tightly contested, with no clear dominant figure in terms of win rate, reflecting the intense competitiveness of the male elite.
- In the butterfly chart analysis, Novak Djokovic stands out with a significant number of set wins compared to losses, underscoring his dominance, especially given his record number of tournament victories.
- Margaret Smith and Nancye Wynne Bolton both exhibit patterns where their set losses are substantially lower than their wins, testifying to their consistent and dominant performances throughout the tournaments they participated in.