32146 Data Visualisation and Visual Analytics - Autumn 2024

Assessment Task 2: Advanced Data Visualisation

Report: Australian Open tennis championships

Student Name: NIPUNN KHURANA

ID: 25009200

Date: 24/03/2023

Table of Contents

INTRODUCTION

The Australian Open is one of the four Grand Slam tennis tournaments, along with the French Open, Wimbledon, and the US Open. It takes place every January at the Melbourne Park complex in Melbourne, Australia. The event was first organized in 1905 and was dubbed the Australasian Championships before changing the name to the Australian Championships in 1927 and the Australian Open in 1969. The Australian Open offers men's and women's singles, men's, women's, and mixed doubles, and junior and wheelchair tournaments. The event's hard-court surface is renowned for being blue Plexicushion since 2008 and previously green Rebound Ace from 1988 to 2007. The Australian Open has a place in tennis history for several reasons. It was the first Grand Slam to organize indoor games on three of its central courts: Rod Laver Arena, John Cain Arena, and Margaret Court Arena, all of which have retractable roofs. With these, Australian Open continues to lead the way in innovation and player health, ever setting the stride for other Grand Slams to follow. Historically, the Australian Open is the least popular Grand Slam because of its distance and passing right about the holidays, Christmas, and New Year. Still, since it was changed to mid-January and advanced travel options have been made, the event has improved in stature every year. Some notable Australian Open title record-holders include Novak Djokovic, who has won several men's singles titles, making him the greatest male player of the sport. Margaret Court made an unrivaled 11 women's single title wins, including in the sport's amateur days. It is known as a "slam" festival that predates other southern hemisphere sports.

DATA DESCRIPTION

Year	**SHEET**	**The year of the tournament.**	**INTEGER**
Gender	Australian Open & Top 5+	The gender category of the tournament (e.g., Men's, Women's).	CATEGORICAL
Champion	Australian Open & Top 5+	Name of the champion.	STRING
Champion Nationality	Australian Open & Top 5+	Nationality of the champion. (e.g., CHN for China).	STRING
Champion Country	Australian Open & Top 5+	Country represented by the champion	STRING

Score	Australian Open & Top 5+	The final score of the championship match.	STRING
1st won	Australian Open & Top 5+	Number of games won in the 1st set	INTEGER
1st lost	Australian Open & Top 5+	Number of games lost in the 1st set	INTEGER
2nd won	Australian Open & Top 5+	Number of games won in the 2nd set	INTEGER
2nd lost	Australian Open & Top 5+	Number of games lost in the 2nd set	INTEGER
3rd won	Australian Open & Top 5+	Number of games won in the 3rd set	INTEGER
3rd lost	Australian Open & Top 5+	Number of games lost in the 3rd set	INTEGER
4th won	Australian Open & Top 5+	Number of games won in the 4th set	INTEGER
4th lost	Australian Open & Top 5+	Number of games lost in the 4th set	INTEGER
5th won	Australian Open & Top 5+	Number of games won in the 5th set	INTEGER
5th lost	Australian Open & Top 5+	Number of games lost in the 5th set	INTEGER
Runner Up	Australian Open & Top 5+	Name of the runner-up in the championship.	STRING
Runner-up Nationality	Australian Open & Top 5+	Nationality of the runner-up.	STRING
Runner-up Country	Australian Open & Top 5+	Country represented by the runner-up	STRING
Wins	Top 5+	Total games won in all sets	INTEGER
Loss	Top 5+	Total games Lost in all sets	INTEGER
Win Rate	Top 5+	Total Games/Total Games played	INTEGER

DATA PREPERATION

In this analysis we have introduced several new variables and performance metrics, the dataset now offers a more detailed view of player performance and match dynamics throughout the tournament.

Summary of Performance Variables Added

- Games Won: A new variable representing the total number of games a player won across all sets in a match was calculated using the formula:
- Games Won = 1st won + 2nd won + 3rd won + 4th won + 5th won
- This metric provides a comprehensive view of a player's dominance within a match.
- Games Lost=1st lost+2nd lost+3rd lost+4th lost+5th lost. The total number of games lost by a player was aggregated across all sets to understand resilience and the challenges faced during a match.
- Win Ratio: This key performance indicator, the ratio of games won to total games played, was introduced to measure player efficiency and overall performance in a match. The formula used is Games Won/(Games Won + Games Lost)
- Set-Specific Performance Ratios Developed
- Set Win Ratio: For each set within a match, a win ratio was calculated to indicate the proportion of games won, providing insights into how players perform during different phases of the match. The formula used is Set Won/(Set Won + Set Lost) and similarly for subsequent sets.
- Set Loss Ratio: This metric mirrors the set win ratio but focuses on the games lost, highlighting potential vulnerabilities. It is calculated for each set as Set Lost/(Set Won + Set Lost) and similarly for other sets.

The advantages of creating such variables are:

- Descriptive Analytics: It enables a more detailed evaluation of player performances across matches, examining set-by-set effectiveness.
- Strategic Insights: The dataset assists coaches and analysts in identifying player strengths and weaknesses, facilitating targeted training and match preparation.

Limitations

There are instances where a men’s champion player whitewashes the runner up by 3-0 (3 sets to zero), in that scenario, the last two games are not meant to be played since the winner has been declared, so we get an error value for the set win and loss ratio since the divisor is ‘zero’. Thus, it depends on how many spares games are remaining in the match.

DATA VISUALISATION TECHNIQUES

1. TREEMAP

A treemap is a visualization of hierarchical data, which shows a set of nested blocks. Each branch of the hierarchy is represented by a rectangle, which is then filled with rectangles of smaller size – they show either the sub-branches of the hierarchy or the data themselves. The area of the tree's rectangle varies according to the evaluated dimension, typically a numerical parameter. This makes it easy to figure out where the data is proportional to what we need, which helps identify the distribution of different items. Features of the treemaps are as follows:

Hierarchical structure – a treemap shows how levels of the hierarchy are organized. The treemap hierarchy is represented as a layer of rectangles within a bigger block.

Size encoding – the area of the element is proportional to some kind of metric, such as the number of games won, or other factors implying quantitative measures.

Color coding – the colors are used both to distinguish categories and show the variations within the category. For example, darker colors may signify higher values, while lighter one's lower values.

Space efficiency – a treemap is good for general data analysis because of its compact format. Even though there is a lot of information on it, the effective use of space and no need for additional axes help not to overload the chart.

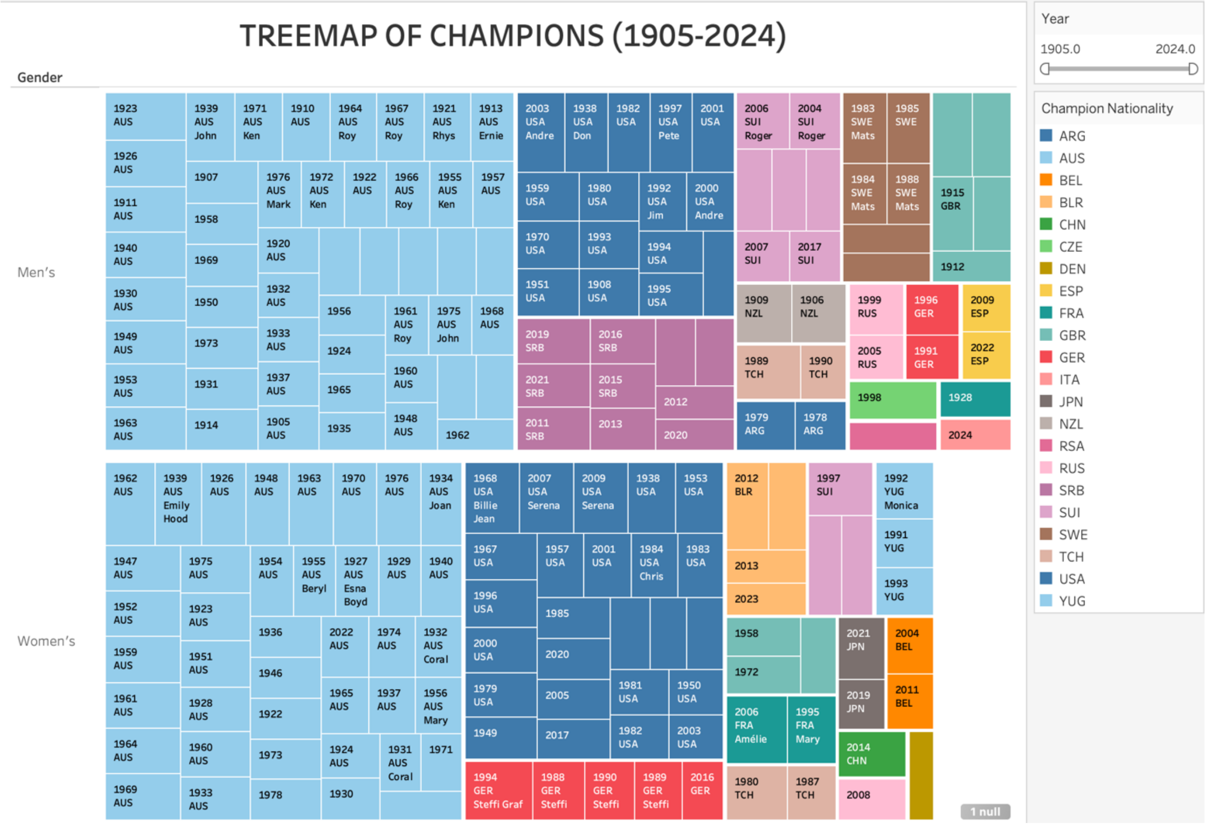
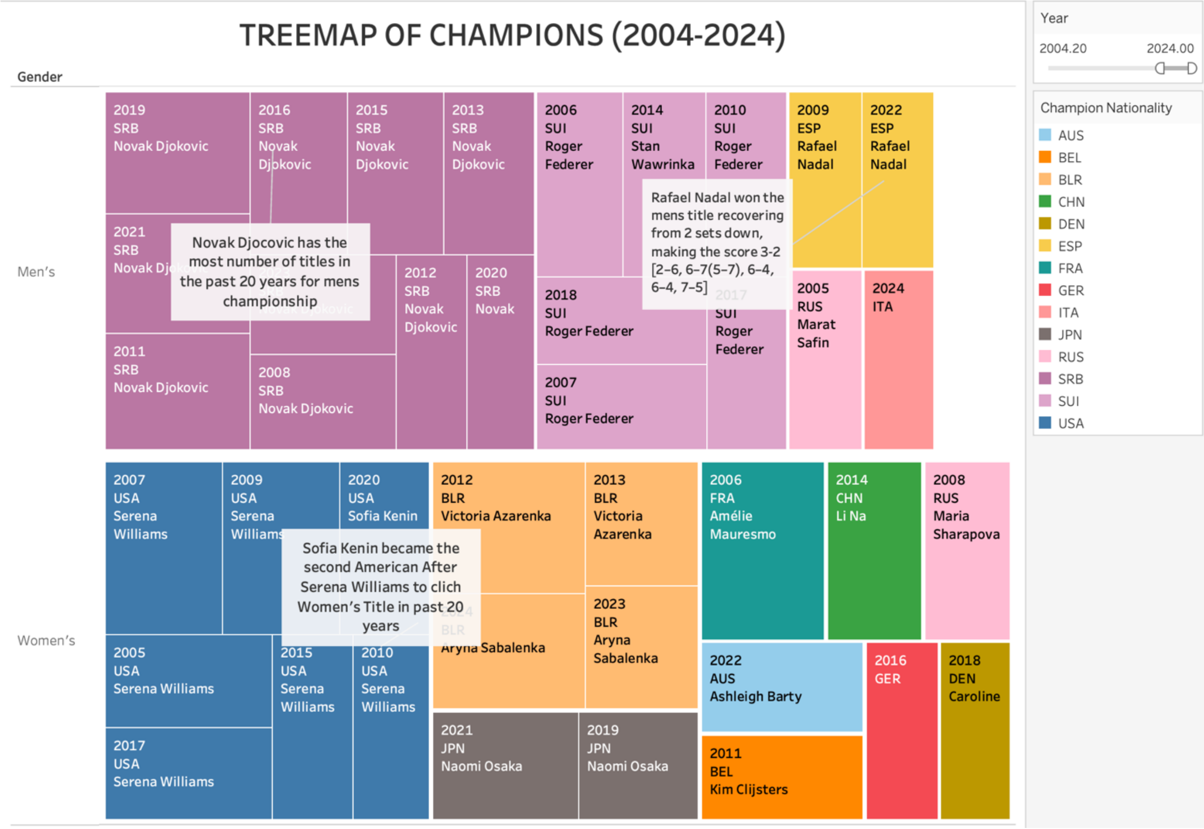


FIG 1.1

The treemap (Fig 1.1) is segmented into two major categories representing Men's and Women's championships. Each champion's tile is sized perhaps according to the significance of their win ratio, and is color-coded by the player's nationality, enhancing the visual differentiation between different countries. Over the years, Australia and USA has the greatest number of titles in the history of Australian Open.



Novak Djocovic has the most number of titles in the past 20 years for mens championship

Rafael Nadal won the mens title recovering from 2 sets down, making the score 3-2 [2–6, 6–7(5–7), 6–4, 6–4, 7–5]

Sofia Kenin became the second American After Serena Williams to clich Women's Title in past 20 years

FIG 1.2 Treemap of Champions (2004-2020)

The treemap (Fig 1.1) is segmented into two major categories representing Men's and Women's championships. Notably, Novak Djokovic and Serena Williams appear as dominant figures in their respective categories, with Djokovic's successes highlighted for the Men's championships and Williams for the Women's, emphasizing their significant impact and dominance in the sport during this period. The visualization also includes specific noteworthy achievements, such as Rafael Nadal's comeback in 2022 and Sofia Kenin's notable win as the second American after Serena Williams to clinch a title in the past 20 years. This treemap not only highlights the frequency of wins by various champions but also effectively showcases trends and shifts in dominance over the years at the Australian Open.

METHODOLOGY

For this treemap visualization (FIG1.1), we focused on analyzing the distribution and performance of Australian Open champions over selected periods. The primary metric for analysis was the Win Ratio, which provides insights into the relative performance of each champion during their respective tournaments.

Data Selection and Preparation

- Variables Used: The key variables selected for the visualization included the Win Ratio, Champion Nationality, Year, and Match Score.
- Calculation and Metrics: The average Win Ratio (AVG(Win Ratio)) was utilized as the primary measure to determine the size of each rectangle within the treemap, establishing a hierarchical structure based on performance.
- Conversion and Adjustment: The Year variable, initially a quantitative measure, was transformed into a continuous dimension. This conversion facilitated dynamic temporal analysis through interactive elements.

Visualization Design and Structure

- Hierarchical Layout: The treemap was structured hierarchically with Champion Nationality as the top layer, influenced by the average Win Ratio, which dictated the size of each rectangle.
- Color Coding: The rectangles were color-coded based on Champion Nationality to visually distinguish the dominance of different countries across the tournament's history.
- Labels and Tooltips: Each rectangle was labeled with the Year and the Champion's name to provide immediate context. Additional details such as the exact match score were included in the tooltips, appearing upon hovering over each segment, to enrich the user's interaction with data points.

Interactivity and User Engagement

- Filtering Capability: To enhance user interaction and allow for targeted analysis, we implemented filters based on Gender and Year. These filters enable users to refine the visualization to display data for specific genders or time periods.

- Dynamic Time Selection: A slider was introduced to filter the data dynamically by year. This feature allows users to select any specific period for detailed examination, particularly useful for observing trends and shifts in champion performance over time.
- Dual Treemap Layout: Two treemaps were designed to compare the performances segmented by gender, facilitating an in-depth comparative analysis across male and female champions.

Analytical Goals and Output

- Purpose: The primary aim of this visualization is to provide a clear and interactive representation of dominance by nationality and individual performance in the Australian Open, with an emphasis on trends over the last 20 years.
- Insights: Each rectangular tile in the treemap represents a champion's nationality, name, and the year they were crowned, offering a quick visual interpretation of which countries have been most successful and which champions stood out in specific years.
- Utility: The color coding and size variations help in quickly identifying patterns of dominance and performance, making it easier for sports analysts, enthusiasts, and other stakeholders to draw conclusions about the effectiveness of players and strategies over time.
- This methodological approach ensures that the visualization is not only informative and relevant but also engaging and accessible for users, allowing for easy exploration of historical data and extraction of actionable insights.

2. PARRALLEL COORDINATES

Parallel coordinates are a common type of visualization used to display high-dimensional data by plotting each data point as a polyline that intersects a set of parallel lines (axes), each representing one dimension of the data. This method is particularly useful for exploring relationships and patterns across multiple dimensions within a dataset.

Creation of Parallel Coordinates Plot

Data Preparation:

We began by selecting all the variables related to Set Win Ratios and Set Loss Ratios from the dataset.

To consolidate these variables into a manageable format, we right clicked on one of the selected variables and chose the 'pivot' option from the transform menu. This transformation resulted in the creation of 'pivot field names' and 'pivot field values', effectively restructuring the data for multi-dimensional analysis.

Visualization Setup:

- We then moved 'pivot field names' into the column shelf and 'pivot field values' into the row shelf. This arrangement set up the basic structure of the parallel coordinates plot.
- We adjusted the measure for 'pivot field values' to 'Average', ensuring that the plot reflects the average values across all sets, providing a clearer view of overall performance trends.

Converting to Line Graph:

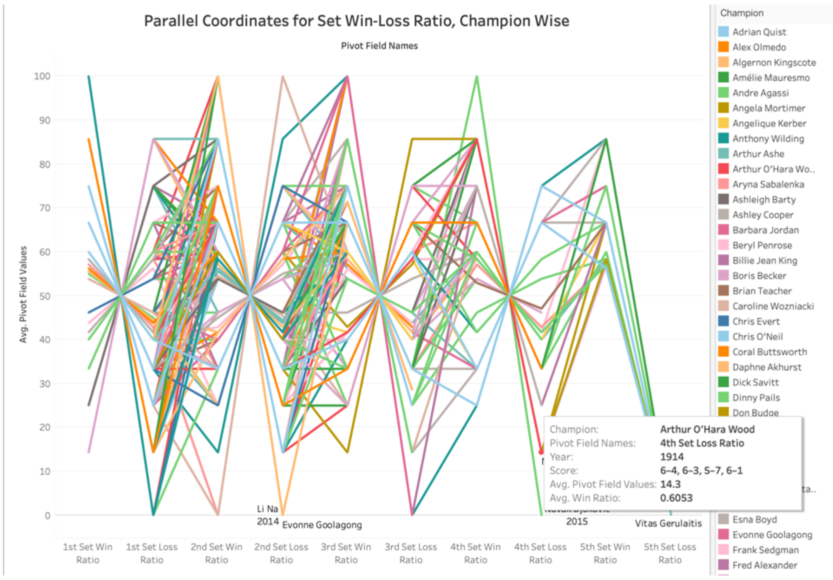
- Initially, the default visualization appeared as a bar chart. We converted this into a line graph to align with the typical layout of parallel coordinates, where each line represents a data point across multiple dimensions.

Enhancing the Visualization:

- Color Coding: We used the 'Champion' variable to color-code the lines, which helps in distinguishing the trajectories of different champions across the sets.
- Labels: For easier identification, we labeled the lines with the 'Champion' name, coupled with the 'Year' to provide temporal context.
- Details: We added 'AVG(Win Ratio)' to the detail shelf to include average win ratios in the visualization, enriching the data context.
- Tooltips: We utilized 'ATTR(Score)' for the tooltips, allowing additional match-specific information to be accessible when hovering over any part of the line.

Interactive Filtering:

- To facilitate comparative analysis and to allow for focusing on specific champions, we applied a filter on the 'Champion' variable. This feature enables viewers to selectively view and compare performances of different champions over the years.



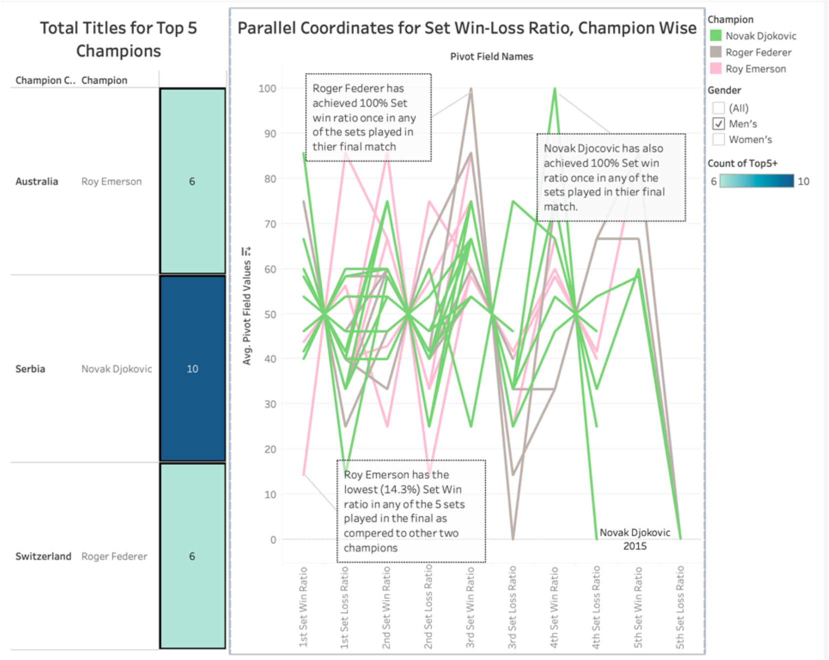
Champion: Arthur O'Hara Wood
Pivot Field Names: 4th Set Loss Ratio
Year: 1914
Score: 6–4, 6–3, 5–7, 6–1
Avg. Pivot Field Values: 14.3
Avg. Win Ratio: 0.6053

Fig 2.1 Parallel Coordinates for Set Win-Loss Ratio, Champion Wise

Creation of Championship Titles Won Table (TOP 5)

In creating a new visualization on a sheet named (Total Titles for Top 5 Champions), we leveraged data from the 'top5' sheet focusing on "Champion" and "Champion Country," which we placed in the rows section to organize the display. We used the "Count of top 5" measure for both color coding and labeling to effectively highlight differences and trends. To tailor the analysis specifically for men's and women's championships, we included "Gender" as a filter. This setup allows for distinct and comparative analysis across genders, enhancing insights into top performances by champions from different countries.

We combined both the sheets(“Parallel Coordinates for Set Win-Loss Ratio, Champion Wise” & “Total Titles for Top 5 Champions”) to create a dashboard and used “Total Titles for Top 5 Champions” sheet as a filter to generate parallel coordinate graph accordingly.



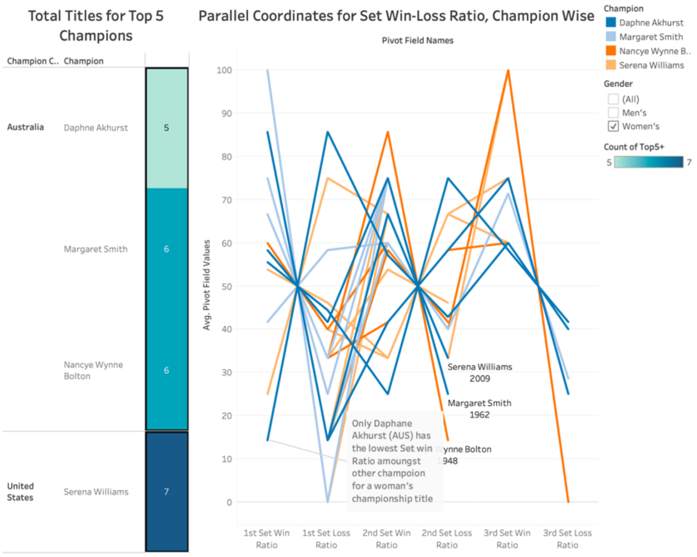
Roger Federer has achieved 100% Set win ratio once in any of the sets played in thier final match

Novak Djocovic has also achieved 100% Set win ratio once in any of the sets played in thier final match.

Roy Emerson has the lowest (14.3%) Set Win ratio in any of the 5 sets played in the final as compered to other two champions

Fig 2.2 Men's Comparative Analysis

The analysis (FIG2.2) reveals that Novak Djokovic leads with 10 major titles, followed by Roy Emerson and Roger Federer, each with 6 titles, as depicted in the bar chart. The parallel coordinates plot highlights that both Roger Federer and Novak Djokovic have achieved a 100% set win ratio in some of their final match sets, contrasting with Roy Emerson's lower performance, who has the lowest set win ratio (14.3%) among them.



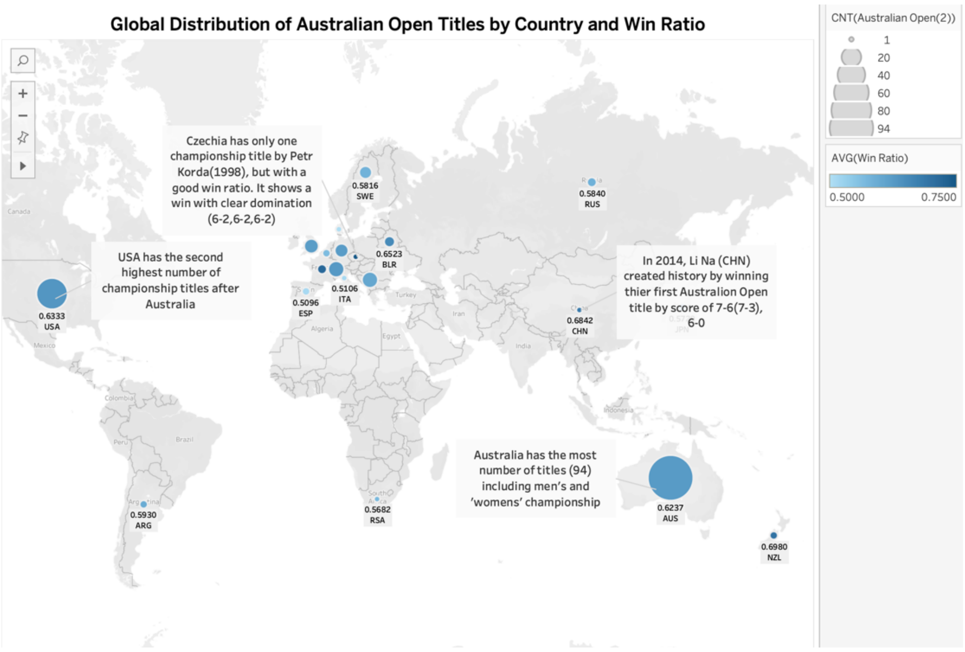
Only Daphane Akhurst (AUS) has the lowest Set win Ratio amoungst other champoion for a woman's championship title

FIG 2.3 Women's Comparative analysis

Fig 2.3 highlights Serena Williams as the leader with 7 major titles among the top women champions, followed closely by Margaret Smith and Nancye Wynne Bolton with 6 titles each. In the parallel coordinates, the variation in set win-loss ratios across different matches underscores Daphne Akhurst's notable performance, marking her as having the lowest set win ratio among these champions, which contrasts with the more variable performance patterns of Serena Williams and other top competitors.

3. GEO MAP

A Geomap, often referred to as a geographic map or just a map, is a data visualization tool that graphically represents a physical area to show information pertaining to that location. This might involve a basic map highlighting different regions or countries, or more complex interactive maps featuring multiple layers of data. Typically, data is represented through symbols like circles, squares, or specialized icons, with variations in size, color, or other characteristics depending on the values of the data they depict.



Czechia has only one championship title by Petr Korda(1998), but with a good win ratio. It shows a win with clear domination (6-2,6-2,6-2)

USA has the second highest number of championship titles after Australia

In 2014, Li Na (CHN) created history by winning thier first Australion Open title by score of 7-6(7-3), 6-0

Australia has the most number of titles (94) including men's and 'womens' championship

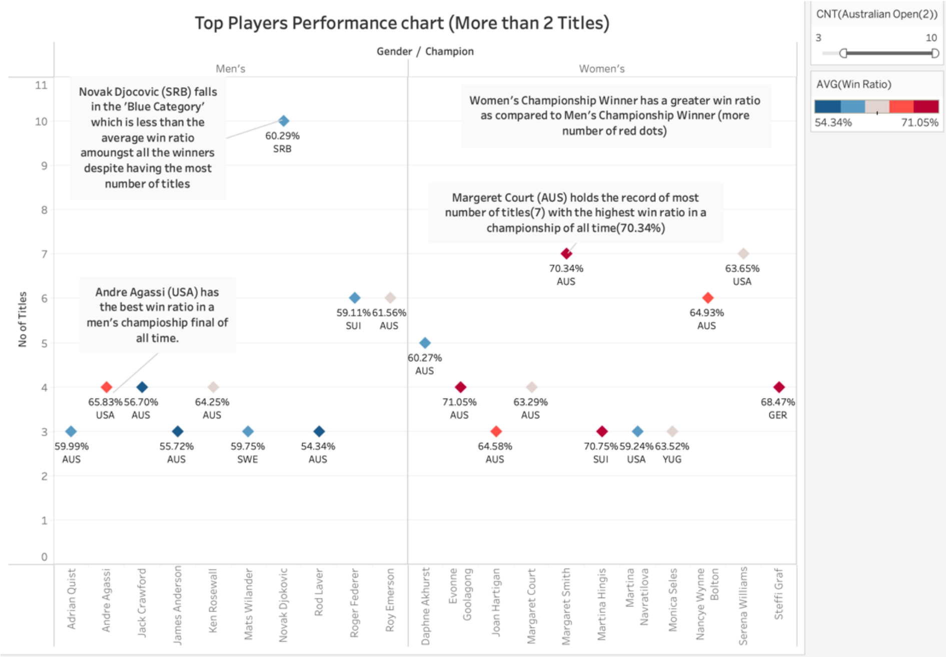
Fig 3.1 Global Distribution of Australian Open Titles by Country and Win Ratio

To create the geomap, we began by assigning the "Geographic Role" to the 'Champion Country' measure, setting it as 'country/region'. Next, we dragged 'Champion Country' onto our sheets pane, which automatically generated a map adorned with circles representing each champion's nation. To visually depict the number of championship titles, we adjusted the size of each circle by adding 'COUNT([Australian Open(2)])' (No of titles Won) to the size section of the marks card. For color coding, we used 'AVG(Win Ratio)', where darker circles indicate a higher win ratio of the country. To enhance the descriptive quality of our geomap, we included 'AVG(Win Ratio)' and 'Champion Nationality' in the detail section of the marks card and incorporated 'Score' into the tooltip section, providing additional contextual information when hovering over each circle.

The geomap visualization (FIG 3.1) illustrates that Australia leads with the highest number of Australian Open titles (94), followed by the USA, highlighting dominant tennis performances. Notable insights include Czechia's high win ratio from a single title and Li Na's historic win for China in 2014, showcasing significant global contributions to the tournament.

4. SCATTER CHART

A scatter plot is a graphical tool that utilizes Cartesian coordinates to illustrate the values of typically two variables within a dataset. Points on the plot represent the data, with each point's horizontal position determined by one variable and its vertical position determined by the other variable. Scatter plots are crucial because they offer a visual insight into the relationship and correlation between two numerical variables. This visualization aids in recognizing patterns, trends, and possible anomalies in the data. Consequently, scatter plots are essential for exploratory data analysis as they help uncover correlations, causations, and form hypotheses for more detailed statistical analysis.



Top Players Performance chart (More than 2 Titles)

Novak Djocovic (SRB) falls in the 'Blue Category' which is less than the average win ratio amoungst all the winners despite having the most number of titles

Andre Agassi (USA) has the best win ratio in a men's champioship final of all time.

Women's Championship Winner has a greater win ratio as compared to Men's Championship Winner (more number of red dots)

Margeret Court (AUS) holds the record of most number of titles(7) with the highest win ratio in a championship of all time(70.34%)

FIG 4.1 Top Players Performance Chart

In Fig 4.1, we aimed to create a 'Top Players Performance Chart' to analyze champions who have won more than two titles. Initially, we selected "Champion" and "Gender" and dragged them to the columns section to bifurcate our analysis by gender. To capture performance metrics, we combined 'Number of Titles Won' and 'Average Win Ratio' by dragging 'COUNT([Australian Open(2)])' to the Rows section and 'AVG(Win Ratio)' to the Marks section of our scatter chart. For visual differentiation, we used color coding: red to denote the maximum win ratio, blue for the minimum, and grey for mid-range win ratios among top players.

In the tooltip section, we included the 'Year' using the 'maximum' metric to display the most recent title won by each champion. For labeling, we chose 'Champion Nationality' and 'AVG. Win Ratio' to enhance clarity on our scatter plot. We also used 'COUNT([Australian Open(2)])' in our filter section, displayed the filter on the visualization, and set it to include only champions who have won between 3 to 10 titles. This approach helped us focus on the significant achievers, making our analysis both targeted and comprehensive.

This visualization(FIG4.1) presents a scatter plot titled "Top Players Performance Chart (More than 2 Titles)" which compares the number of titles won to the average win ratio of top tennis players, segmented by gender. For the men, Novak Djokovic stands out with a lower average win ratio despite having the most titles. In contrast, the women's chart shows Margaret Court leading with the highest number of titles and an impressive win ratio. Each point on the plot is colored to represent different win ratios, highlighting significant outliers and the overall distribution of performance metrics across top champions. The plot efficiently encapsulates performance trends and excellence in tennis, distinguishing between genders and indicating the consistency versus peak performances of champions.

EXECUTIVE SUMMARY

Our report provides a comprehensive analysis of the performance trends and achievements of tennis players at the Australian Open, utilizing advanced data visualization techniques facilitated by Tableau. Our analysis focuses on various dimensions of the data, including gender-specific performance, geographical distribution of champions, and historical trends in player success and match outcomes.

Key Findings:

- Gender-Specific Performance Trends:
    - Men’s Category: Novak Djokovic emerges as a standout, with the most titles won over the period analyzed. His performance, alongside other male champions, is visualized through parallel coordinates(FIG 2.2) that highlight his unparalleled win ratios in final match sets.

    - Women’s Category: Margaret Court dominates with the highest number of titles. The scatter plot (FIG 4.1) analysis provides insights into her consistent high performances and contrasts them with other leading female players, emphasizing her historical significance in the tournament.

- Geographical Insights:
    - The geomap visualization (Fig 3.1) reveals Australia as the dominant nation in terms of title wins, with significant contributions also observed from the USA and emerging champions from countries like China, as exemplified by Li Na's historic victory in 2014.

- Performance Metrics and Visualization Advantages:
    - Using treemaps, the distribution of titles and performance metrics such as win ratios are effectively communicated, showing how different players and countries have performed over the years. These visualizations not only depict data but also tell compelling stories of player dominance and emerging trends.

The parallel coordinates and scatter plots highlight the comparative performance of top players, providing a clear view of how individual tactics and consistency vary across different matches and conditions.

CONCLUSION

Our study has illuminated several key aspects of tennis excellence and strategy:

- Cultural and Strategic Impact on Tennis: The data underscores how the cultural backdrop and strategic decisions within different countries influence their success in the tournament. For instance, the evolution in training techniques, dietary changes, and technological advancements in sports equipment could be correlated with the increasing competitiveness and changing dynamics observed in the tournament outcomes.

- Predictive Insights for Future Tournaments: By analyzing historical data and current trends, we can offer predictive insights that could influence player scouting, training focus, and game strategies. For example, understanding the conditions under which certain players excel can help coaches tailor training programs that maximize player strengths and minimize weaknesses.

- Role of Analytics in Sports: This report also highlights the transformative role of analytics in sports. By leveraging data visualization tools like Tableau, stakeholders can achieve a more nuanced understanding of complex data sets. This not only enhances viewer engagement through dynamic and interactive visual stories but also empowers decision-makers with data-driven insights.

In summary, this report not only charts the historical achievements and patterns within the Australian Open but also acts as a blueprint demonstrating how data-driven approaches can revolutionize understanding and strategies in sports. As we move forward, the integration of more granular data, including player biometrics and in-match analytics, will likely offer even more sophisticated insights, further enhancing the strategic depth of competitive tennis.