32146 Data Visualisation and Visual Analytics
Autumn Semester 2025
Assessment 2: Report

Australian Open Tennis Championships: A Data-Driven Performance Analysis

Student Name： Zheng Wang
Student ID：14403000

Executive Summary

This report explores the Australian Open tennis championships data since 1905, highlighting long-term trends and top players' performance through data visualisation. Using Tableau, I constructed a variety of interactive visuals including Treemaps, Parallel Coordinates, Geographic Maps, and Scatter Charts. These visual tools allowed us to uncover national and gender-based patterns and compare the playing styles and success patterns of top athletes (players with five or more titles). I conclude with a critical reflection on Tableau's strengths and limitations for tennis data storytelling.

1. Data structure and characteristics

The Australian Open dataset includes tournament results from 1905 to 2025, encompassing over a century of matches. The dataset contains:

Structured Fields: Year (numeric), Gender (categorical), Champion & Runner-up (categorical), Nationality, Match Score (numeric), Champion Seed (ordinal), Country, and Set Scores (numeric)

Data Format: Provided in Excel format; consistent with long-form table structure. The original dataset comprises 22 columns and 213 rows, with each row recording one match.

Data Content: Includes gender, nationality, Champion Nationality, seed ranking, match score, match duration, and win/loss details for each of up to five sets.

Data type:

Name	Description	Data type
Year	Year of the open	Quantitative (interval)
Gender	Gender of the athletes	Categorical (nominal)
Champion/Runner-up	Name of the Winner/Runner-up	Categorical (nominal)
Champion/Runner Nationality	Winner and Runner-up's nationality (3-letter code)	Categorical (nominal)
Champion/Runner Country	Winner and Runner-up's Country (full name)	Categorical (nominal)
Champion/Runner_up Seed	Seed of the champion	Categorical (ordinal)
Score	Match score	Quantitative (ratio)
Mins	Duration of the match in minutes	Quantitative (ratio)
1st won -5th won	Games won by the champion in each set	Quantitative (ratio)
1st loss _5th loss	Games lost by the champion in each set	Quantitative (ratio)

Trends & Outliers Observed:

a. Only 4 records contain valid entries in the Mins column, indicating significant missing data in the column.
b. Many matches concluded in just 3 sets, so fields representing the 4th and 5th sets often contain no values (missing data by design).
c. In 1977, the Australian Open was held twice—once in January and once in December—resulting in four total records for that year that required year disambiguation.
d. In 1965 and 1966, the final matches were decided by "retired" and "walkover" respectively. These conditions were reflected in the Score column and also impacted win/loss calculations.

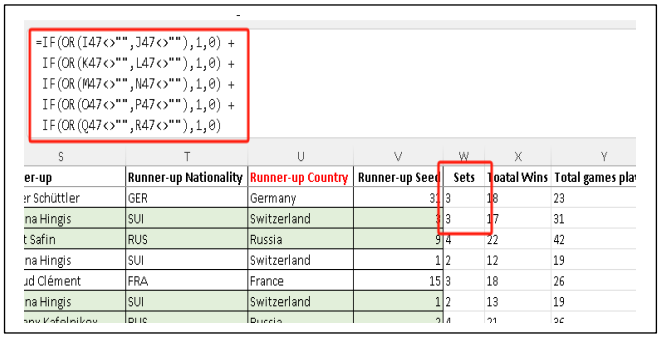
e. Women's competitions were first recorded in 1922. All matches before that year featured only men's.
f. The Australian Open uses different match formats: men's finals are played as best-of-five sets, while women's finals are played as best-of-three sets. This results in women's matches do not contain data for the 4th and 5th sets.
g. Seed rankings for players began in 1924. Matches before that year do not contain seed information.

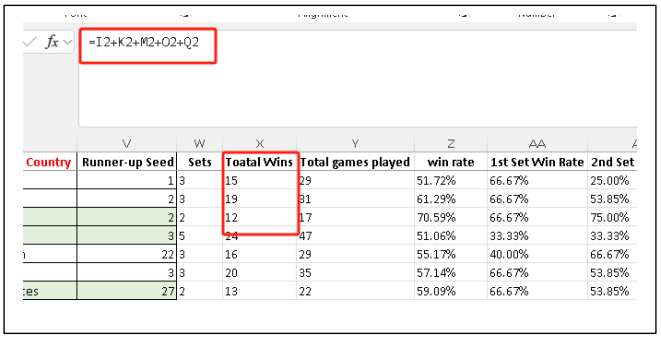
2. Data Transformations & Calculations

Using Excel, we calculated the following performance metrics for each match:

Sets: Total sets played. Total sets played, used to measure match length and player endurance.

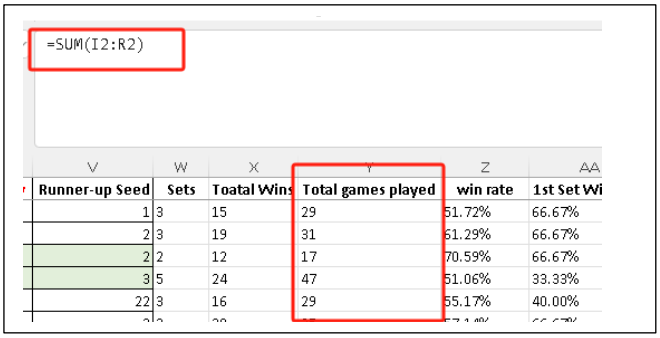
Total Wins: Number of sets won by the champion. The number of sets won by the champion reflects dominance in the match.

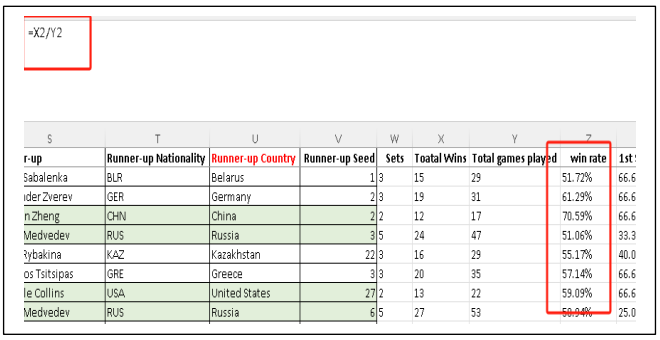




Total Games Played: Sum of games played in the match

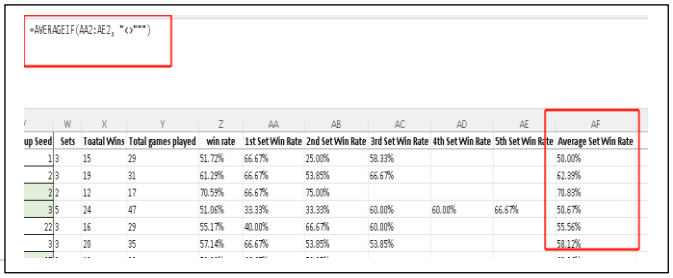
Win Rate: Proportion of sets or games won, indicating overall match intensity.





Set-by-Set Win Rate: Win rate calculated separately for each of the 1st to 5th sets. Used to compare consistency.

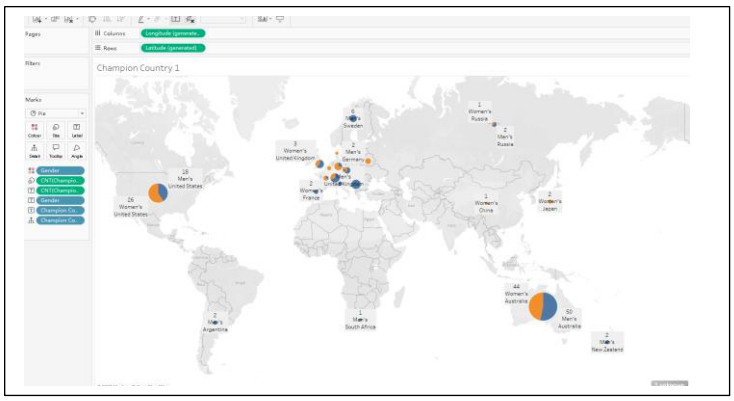
Average Win Rate: Average win rate across all sets played. Provides a indicator of performance efficiency.

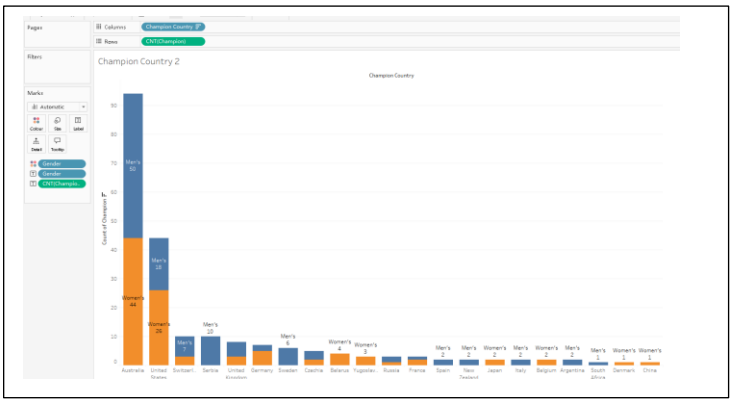


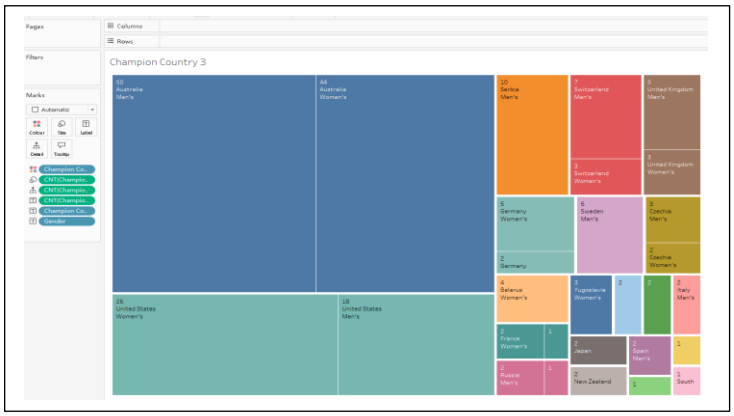
An attempt was made to convert the Year column from Text to Date format in Excel. However, rows 4 to 7 could not be successfully transformed. Fortunately, the Text format did not affect functionality or analysis within Tableau.

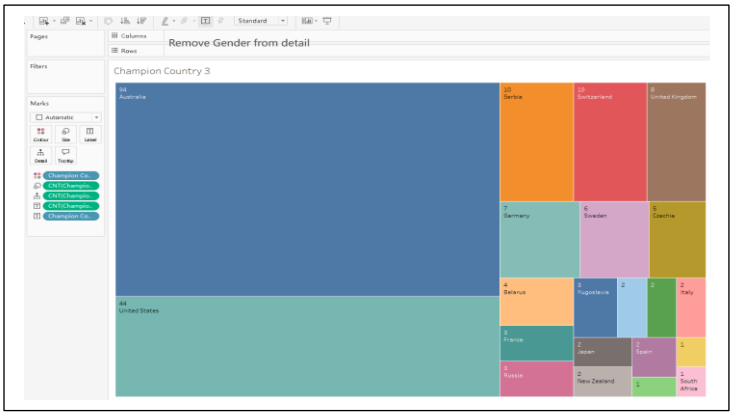
3. Visualisation Analysis

Champion Country Analysis
Three complementary visualisations were used to analyse the distribution of champions across countries and genders:







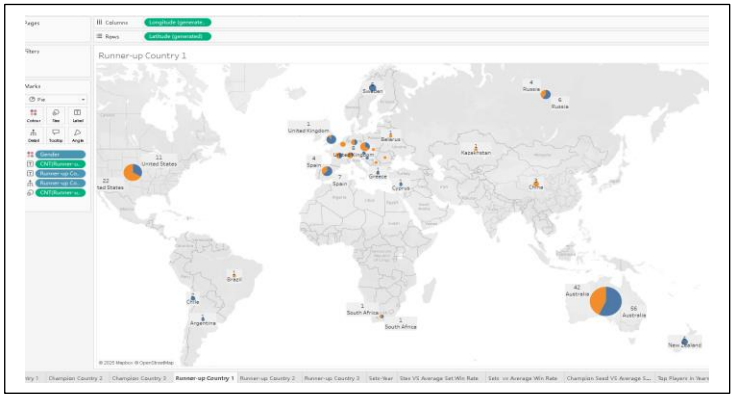


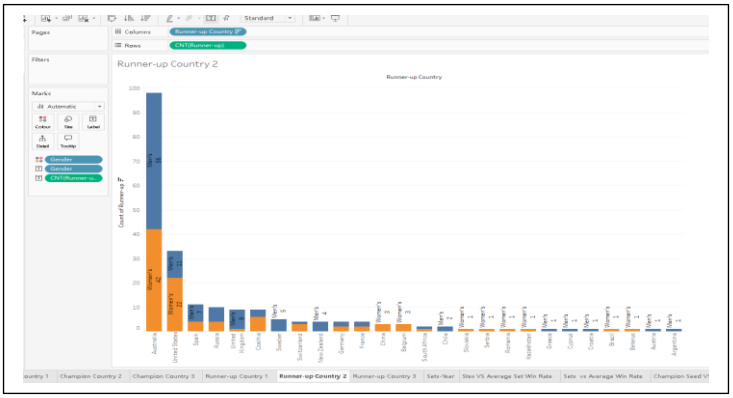
Geographic Map: This chart visualised both the number and gender distribution of champions by country. It revealed that Australia and the United States have been historically dominant in both men's and women's competitions. The map also illustrated a growing European presence with Serbia, Switzerland, and Spain emerging as strong tennis nations.

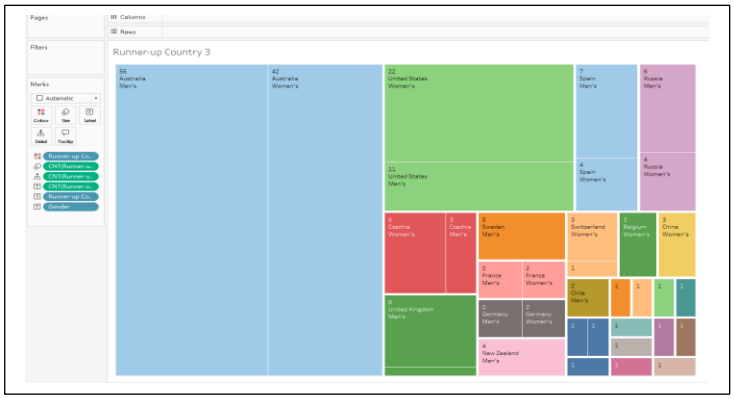
Bar Chart: The stacked bar chart ranked countries by total titles, broken down by gender. It reinforced Australia's lead and highlighted Western Europe's rise. The chart also helped identify gender skews in some nations—for instance, the United States had more women's champions, while Serbia had only men's champions.

Treemap: By combining country and gender dimensions, the treemap offered a proportional comparison of titles. Australia held the largest area for both men (50 titles) and women (44 titles), followed by the US (18 men's, and 26 women's titles). Notably, Serbia and Switzerland had a concentrated contribution from single standout players like Djokovic and Federer.

Runner-up Country Analysis
I used a geographic Map, stacked bar chart, and treemap to analyse the distribution of runner-up finalists by country and gender. Together, they reveal a complementary perspective to champion data, showing national competitiveness.







Geographic Map: Showcases the global spread of runner-up appearances. Australia, the United States, and Spain appear frequently, indicating strong. Pie charts by gender highlight a relatively balanced presence of men's and women's finalists in countries like the U.S. and Australia.

Stacked Bar Chart: This visual gives a direct count of runner-up appearances by country and gender. Australia leads with 98 total runner-up appearances (56 men, 42 women), followed by the United States (33 combined). It demonstrates how dominant tennis nations frequently appear in finals.

Stacked Bar Chart: This visual gives a direct count of runner-up appearances by country and gender. Australia leads with 98 total runner-up appearances (56 men, 42 women), followed by the United States (33 combined)

Treemap: Reinforces the bar chart insights, offering an intuitive area-based view of runner-up frequency. It emphasizes Australia's and the U.S.'s lead.

Key Finding:
Australia leads both in champions (50 men, 44 women) and runner-ups (56 men, 42 women), confirming home court strength.

The United States showed higher consistency in women's titles but also significant runner-up presence.

Emerging countries like Serbia and Switzerland demonstrated recent male dominance.

Match Sets Comparison Over Time
To understand match dynamics, I analysed the number of sets played each year, split by gender.

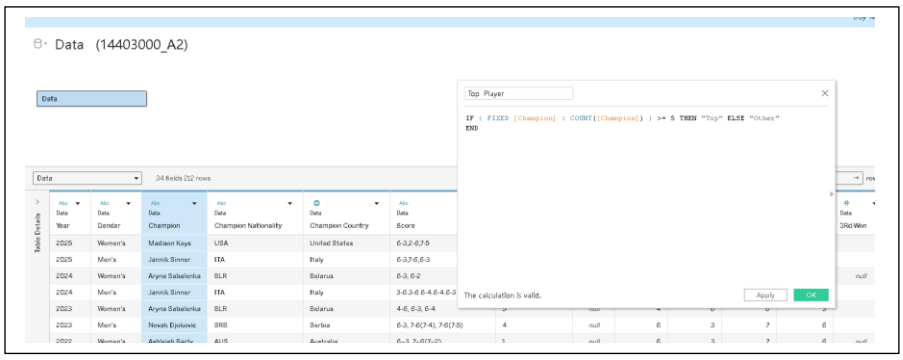
Key Finding
Men's Matches: Typically played as best-of-five sets, with most matches ending in 3, 4, or 5 sets.

Women's Matches: Consistently played as best-of-three sets. Most finals ended in 2 sets, with a smaller portion extending to 3.

Outliers: The 1966 matches appear 0, consistent with special cases of "walkover".

Despite the differing match formats, men's matches demonstrated greater competitiveness, with most finals being decided in the fourth or fifth set.

Sets vs Average Set Win Rate
First, we used Tableau to calculate the number of championships won by each player. Those with five or more titles were classified as "Top" players, while all others were labelled as "Other" in a newly created calculated field**.**



This scatter chart plots total sets played against the average set win rate for each top player, offering insight into both match volume and efficiency. Gender is represented through colour and top players are represented through shape encoding.

Key Finding
Margaret Court and Novak Djokovic stand out with the highest average set win rates. Djokovic leads in both set volume and consistent high performance, indicating endurance and skill over a large number of matches.

Roger Federer and Roy Emerson show high total sets but slightly lower efficiency. On the women's side, Serena Williams and Evonne Goolagong also demonstrate strong performance.

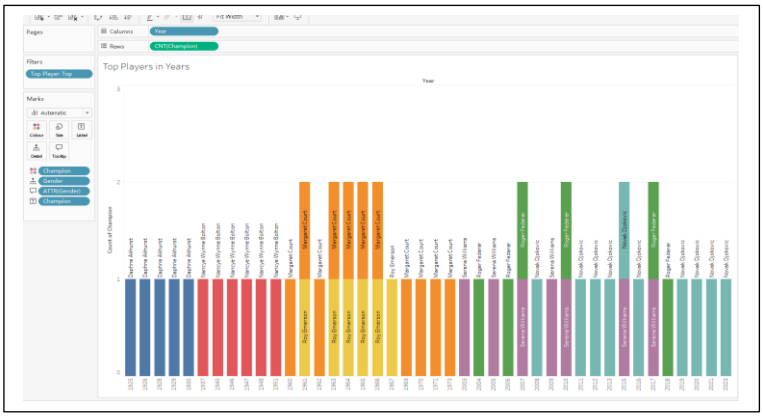
Seed Ranking vs Average Set Win Rate
This visual investigates whether higher-seeded players perform better in terms of set win rate.

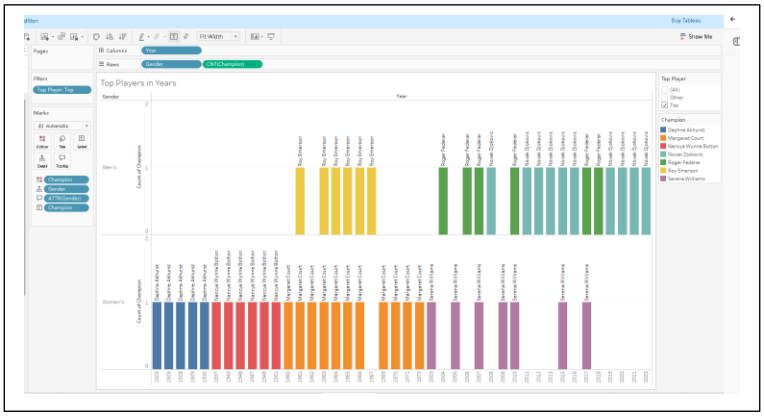
Key Finding:
High-performing players like Djokovic and Court maintained excellent set win rates regardless of seed.

Players with fewer wins (small bubble size) generally cluster near the bottom left—indicating low seed and low win rates.

Top Players Title Timeline

Bar charts segmented by year and gender show title distribution across the top players (⩾5 wins):

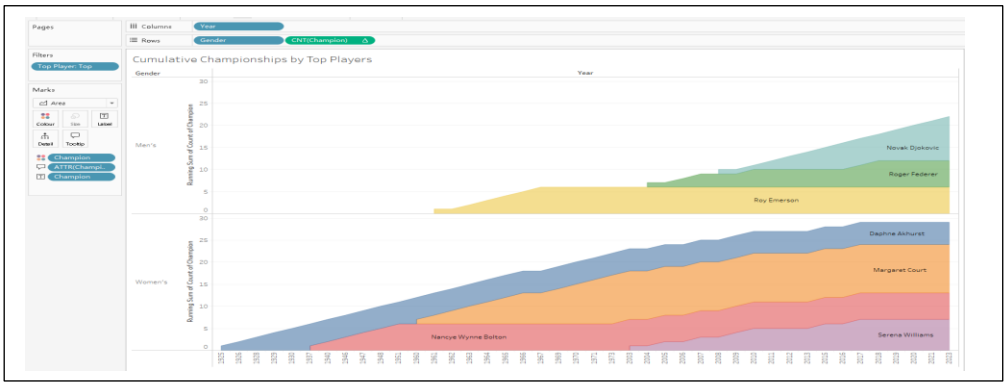




Key Finding:
Men:
Roy Emerson dominated the 1960s.
Roger Federer and Novak Djokovic alternated dominance from 2004 to 2023.

Women:
Margaret Court won consistently across two decades (1960s–70s).
Serena Williams exhibited long-term impact with scattered but repeated wins (2003–2017).
Daphne Akhurst and Nancy Wynne Bolton led during the 1920s–50s.

Cumulative Championships by Top Players
The cumulative area chart provides a clear visualisation of the long-term dominance of top players in the Australian Open history.

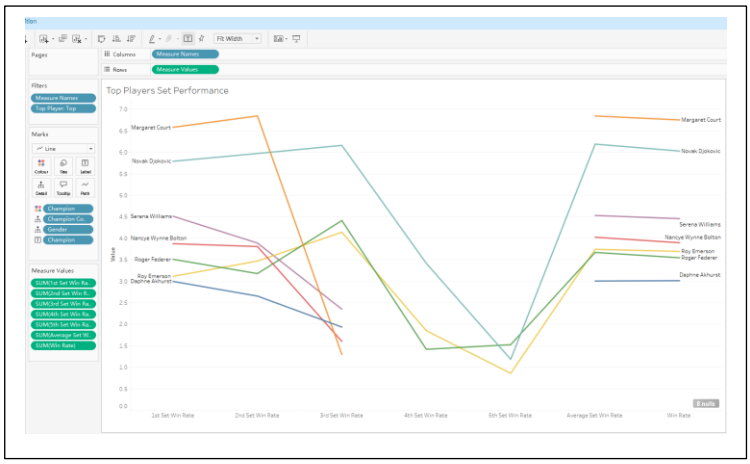


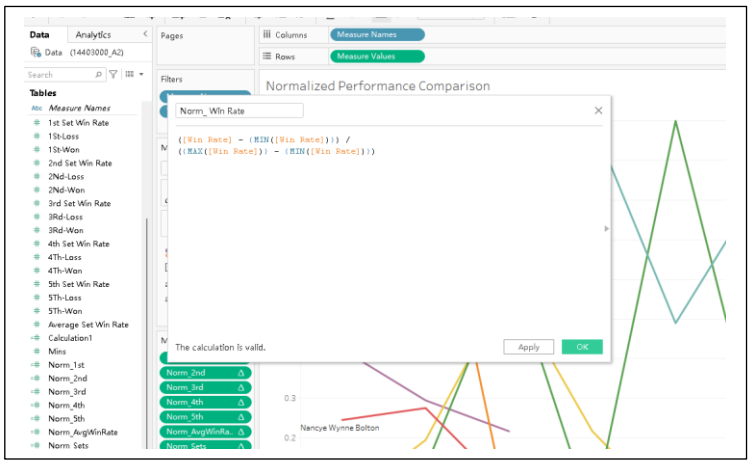
Men:
Roy Emerson was dominant in the 1960s, holding the cumulative lead for many years.
Roger Federer surpassed Emerson from the mid-2000s to 2018.
Novak Djokovic overtakes all others by 2023 with the highest total number.

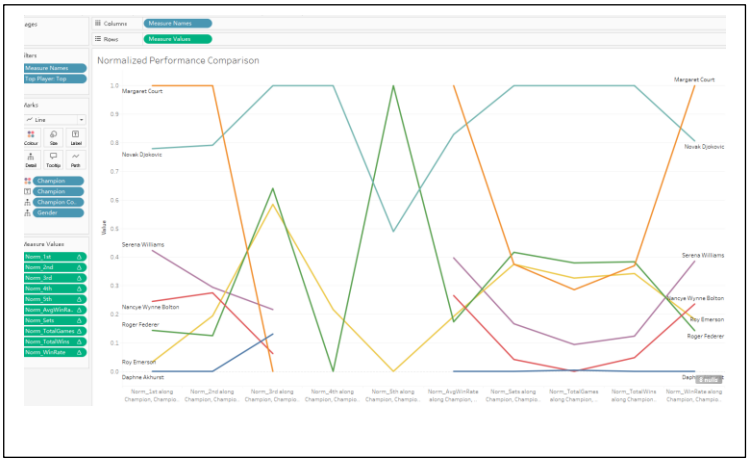
Women:
Daphne Akhurst and Nancye Wynne Bolton contributed heavily during the early and mid-20th century.
Margaret Court showed dominance through the 1960s and 70s, with the highest cumulative total among women.

Serena Williams maintained a strong performance, marking her as a top female player.

Top Players Performance
We first analyzed the data directly from the source and then normalized the relevant fields to re-examine the analysis.







Key Finding:
Margaret Court showed overwhelming dominance in the first two sets, leading to a win without requiring a third set in most matches—reflected by her 0 win rate in the third set, indicating that she typically secured victory early.

Novak Djokovic had a consistently high overall win rate but showed a drop in the fifth set.

Roger Federer performed particularly well in the third and fifth sets, especially the fifth. This may reflect his exceptional endurance and physical resilience in long matches.

For female players, except Daphne Akhurst, most showed declining win rates as matches progressed.

In contrast, male players tended to show peak performance in the third set.
On the Average Set Win Rate scale, Margaret Court and Novak Djokovic stand out as the top-performing players in the women's and men's categories, respectively.

Novak Djokovic also secured the highest total number of match wins among all top players analyzed.

Word Cloud Analysis of Champions and Runner-up

The word cloud visualises all players who reached the finals of the Australian Open, including both champions and runner-ups. The size of each name reflects how frequently a player appeared in the final match.

Larger names such as Margaret Court, Novak Djokovic, Serena Williams, and Andy Murray indicate frequent appearances in finals, with or without victory. This helps uncover not only dominant champions but also consistent high performers who reached multiple finals over the years.

Conclusion

This project successfully leveraged Tableau to analyze and visualize a century of Australian Open data. We explored tournament trends through treemaps and maps, revealing Australia and the United States as dominant countries in championship appearances. The scatter plots comparing sets played and average win rate showed Novak Djokovic and Margaret Court as performance outliers with consistently high win rates. The gender-based breakdown revealed that men's matches often extended into fourth or fifth sets, while women's matches typically ended by the second or third, reflecting their different formats.

Further, seed-based scatter plots showed that not all top seeds win consistently—indicating performance is not always aligned with ranking. The cumulative area chart effectively traced the championship growth of top players over time, with Djokovic and Court emerging as the most dominant figures. The parallel coordinate and normalized comparison charts enabled cross-set performance comparison, emphasizing Djokovic's late-match endurance and Margaret Court's early-set dominance.

Finally, the word cloud highlighted frequent finalists by name size, capturing not only champions but also consistent contenders. Overall, Tableau enabled a clear, interactive way to connect raw historical data with storytelling insights, despite its limitations in data preprocessing and statistical depth. Through thoughtful visual

strategy, this analysis uncovered both expected legends and surprising patterns in the history of Australian Open tennis.

Personal Reflection on Visualisation Tools

In Assessment 1, I used Power BI, and I've also applied Python (Jupyter) and R in previous data analysis coursework. Compared to Python and R—which require considerable coding to achieve similar visualisation outcomes—Tableau and Power BI make it easier to build interactive visual storytelling. Tableau in particular allows quick access to advanced chart types like treemaps, maps, and parallel coordinates.

Tableau's interface is mostly drag-and-drop, enabling users to build visuals efficiently—especially helpful for beginners. It performs strongly in special-purpose visuals, especially geographic charts and multivariate comparisons.

Power BI, on the other hand, is more powerful in terms of data modelling supported by DAX and Power Query. Its built-in AI features also provide surprisingly accurate analytical suggestions. However, for highly customised or advanced statistical plots, Python and R remain superior due to their flexibility.