Assessment Task 2: Advanced Data Visualisation

John-Paul Martin 14508648

Table of Contents

Summary of Dataset Formats, Values, Characteristics, Trends, and Outliers

The following dataset features data pertaining to the men's and women's singles winners of the Australian Open from the most recent winner, down to the first ever winner in 1905. This tournament is the first of four annual Grand Slam Tennis Tournaments, held in Melbourne in January each year.

The dataset is structured in a tabular format, with rows and columns being displayed within an Excel Spreadsheet. Each row represents the tournaments winner for that year, and for each year we have two rows that represent the female and male winners respectively. For each player sample, we can see various columns holding information such as the players nationality, gender, score in the winner's final game, minutes played, their performance in the finals match, as well as data regarding their opponent in the final. The dataset has 22 columns with 213 rows.

Attribute Name	Type	Description
Year	Integer (interval)	Year of the tournament
Gender	String (nominal)	Category of the competition
Champion	String (nominal)	Name of winning player
Champion Nationality	String (nominal)	3 letter nationality code of champion
Champion Country	String (nominal)	Full country name of champion
Score	String (nominal)	Final match score
Champion Seed	Integer (ordinal)	Champion's tournament seed
Mins	Integer (ratio)	Duration of match in minutes
Set (1st, 2nd, etc) won	Integer (ratio	Games won by champion in each set
Set (1st, 2nd, etc) lost	Integer (ratio)	Games lost by champion in each set
Runner -up	String (nominal)	Name of losing finalist
Runner-up nationality	String (nominal)	3 letter nationality code of runner up
Runner-up country	String (nominal)	Full country name of runner up
Runner-up seed	Integer (ordinal)	Tournament seed of runner up

The dataset overall displays consistent information and formatting, however only the first five values for the minute's column are filled, requiring a need to research match durations to populate this column. Trends observed include a historical dominance by Australian winners in the early decades of the tournament, shifting to more cultural diversity in recent years, a rarity of men's singles finals extending to five sets, women's matches consistently limited to three sets due to rule constraints, and a frequent occurrence of runners up being within the top 5 seeds.

Outliers include exceptionally long sets, such as Gerald Pattersons 18-16 set in 1927, and one-sided victories like Victoria Azarenka's 6-3, 6-0 win in 2012, alongside rare occurrences of unseeded or lower-seeded winers such as Madison Keys at seed 19 this year and Rodger Federer at seed 17 in 2017.

Summary of Data Transformations and Calculations Performed

In preparing the dataset for analysis, several steps were taken to enhance its usability and interpretability.

Data Completion: Minutes Column

The Mins column initially contained values for only the first five rows. To ensure consistency and completeness across the dataset, estimated values were generated and populated for the remaining records based on the scores of each match.

Win Ratio Per Set

Win ratios were calculated for each set to provide a performance metric. The formula applied was:

Win ratio = Set Won / (Set Won + Set Lost)

An example for the first set: *1st Set Win Ratio = 1st Won / (1st Won + 1st Lost)*

Total Games Won and Lost

To capture overall performance in the finals, total games won and lost were computed by summing the respective wins and losses across all sets.

Elite Performers (5+ wins)

A new worksheet was creating to focus on the generational players. To do this a frequency column was added using the Excel formula:

COUNTIF(C: C2)

This tracks the number of times a player appears in each row, essentially tracking the number of Australian Open titles won by each champion. A filter was then applied to retain only players with more than five wins, allowing targeted analysis of dominant champions.

Parallel Coordinate Chart

Explanation of chart type

This is a chart type that displays data frequencies, relationships, and aggregation patterns for multidimensional data. The chart represents each dimension of data as a parallel axis and draws polylines between each axis at appropriate values. They are ideal for comparing many variables and observing the relationships between them. Parallel coordinate charts usually come in two forms, one where all the axes are normalised to keep all the scales uniform, and one where each variable works off different units of measurements.

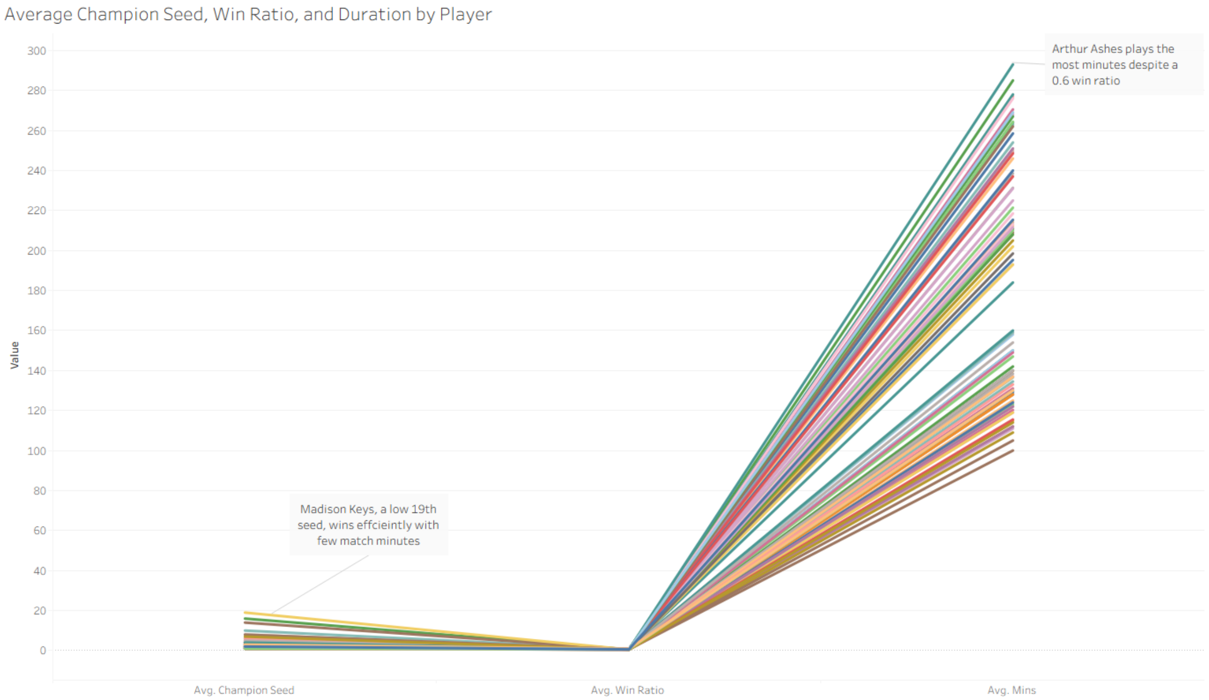


Figure 1. Parallel Coordinate Chart for Player Seed, Ratio and Duration

Analysis: This chart has not been normalised, as indicated by the varying directions of the lines. Champion seed values show limited variability, highlighting a strong relationship between higher seeds and their likelihood of winning a Grand Slam. The lowest-seeded champion recorded was Madison Keys, who secured her title as the 19th seed in a three-set final against top seed Arnya Sabalenka earlier this year.

Win ratios for most players cluster between 0.5 and 0.6, though there is considerable variation in average minutes played. Sofia Kenin recorded the shortest final at 100 minutes, while Arthur Ashe played the longest at 293 minutes in 1970, defeating Dick Crealy (6-4, 9-7, 6-2).

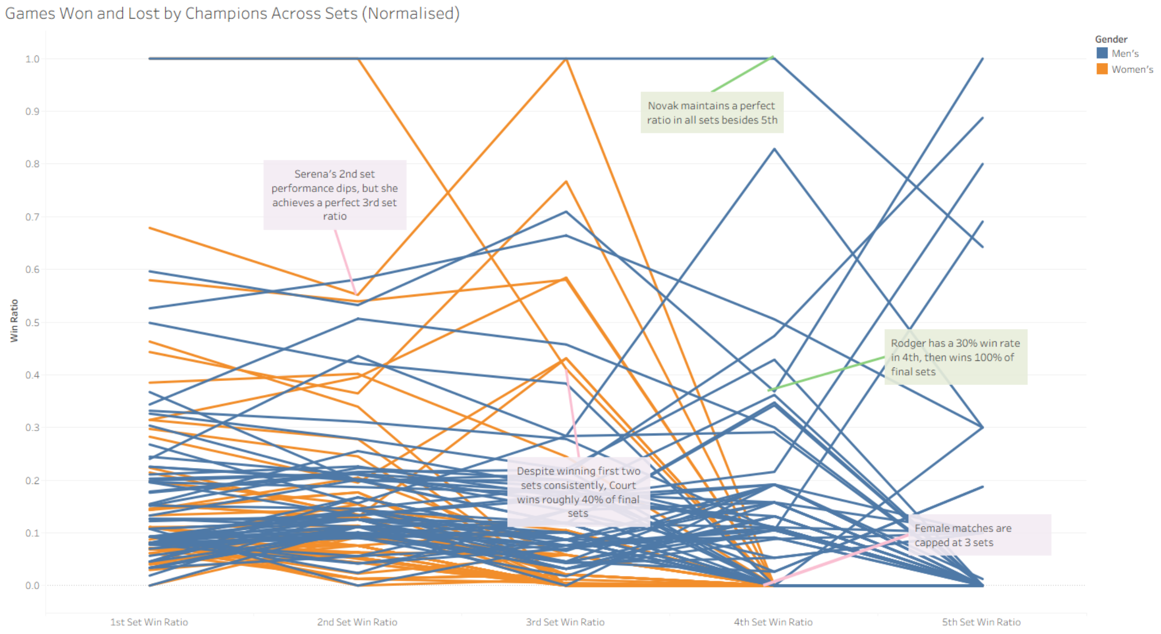


Figure 2. Parallel Coordinate Chart for Set Performance by Gender

Analysis: The chart values were normalised, evident from the spread of polylines across the axes. Female players converge at the fourth set due to the best-of-three format, with only Serena Williams and Margaret Court achieving perfect win ratios. Serena rises from 0.7 in the first two sets to 1.000 in the third, showing strong performance under pressure, while Court maintains perfection through two sets before falling to 0.411 in the final. Novak Djokovic shows dominance with a perfect ratio until the fifth set, dipping slightly to 0.643. Most other players remain in the 0.0–0.2 range, particularly in extended matches, reflecting less consistent dominance.

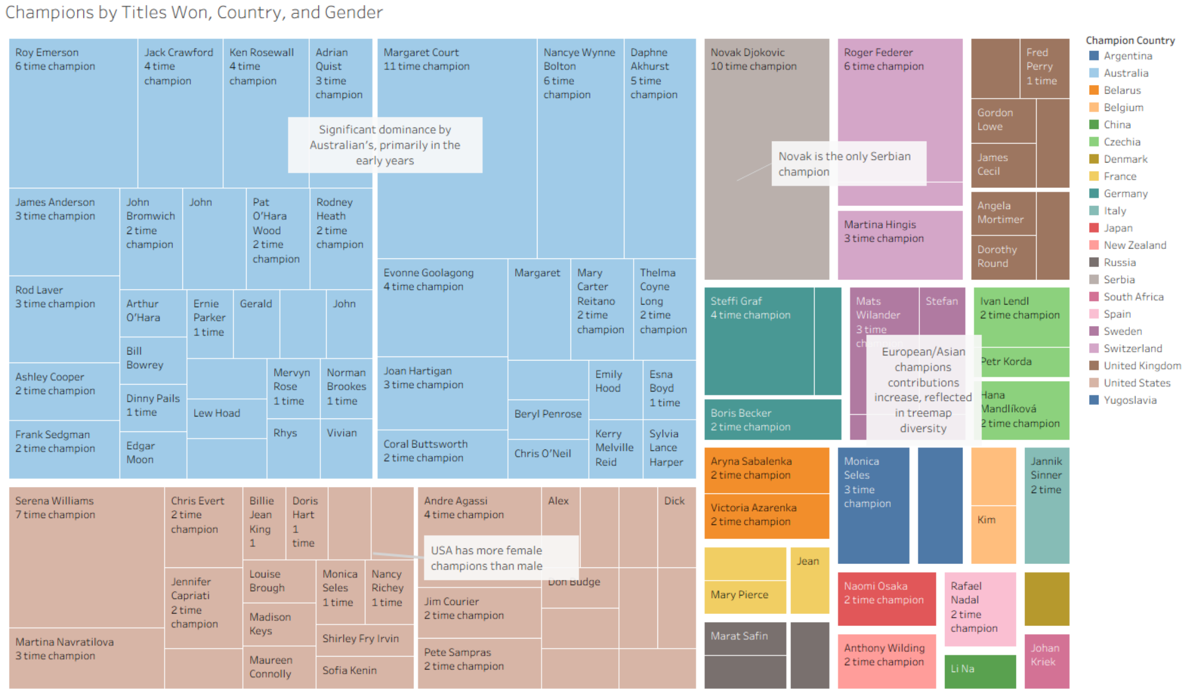
How the charts were made: Both charts were constructed using Measure Names in columns and Measure Values in rows, detailed by Champion and extended to Entire View. Figure 1, which was not normalised, differentiates players by Champion colour, with selected measures of Champion Seed, Win Ratio, and Minutes (all aggregated by average to account for players with multiple titles). In contrast, Figure 2 uses normalised win ratios across all sets and colours by Gender to distinguish between men's and women's champions.

Advantages and Disadvantages: The parallel coordinate chart has several advantages such as its ability to effectively display high dimensional data. Traditional 2D charts struggle to show the relationships between more than two variables, but through multiple parallel axes, these relationships can be effectively displayed. Similarly to a line chart, parallel coordinate charts are effective at revealing correlations and relationships between variables making it efficient for spotting trends and outliers.

That said, it also has its limitations such as the clutter displayed in Figure 2. This can reduce readability and hide important patterns. Additionally, it may be difficulty to interpret charts with different scales, and normalising values can even confuse readers. This chart is also only limited to numerical data make it unusable for categorical.

Tree Map

Explanation of chart type: This chart type consists of nested tiles that visualise hierarchical data, with each tile sized proportionally to its data value. The tiles represent specific categories within a chosen dimension and are organised according to hierarchy. Tree maps are effective for highlighting patterns, proportions, and relationships between variables.



Significant dominance by Australian's, primarily in the early years

Novak is the only Serbian champion

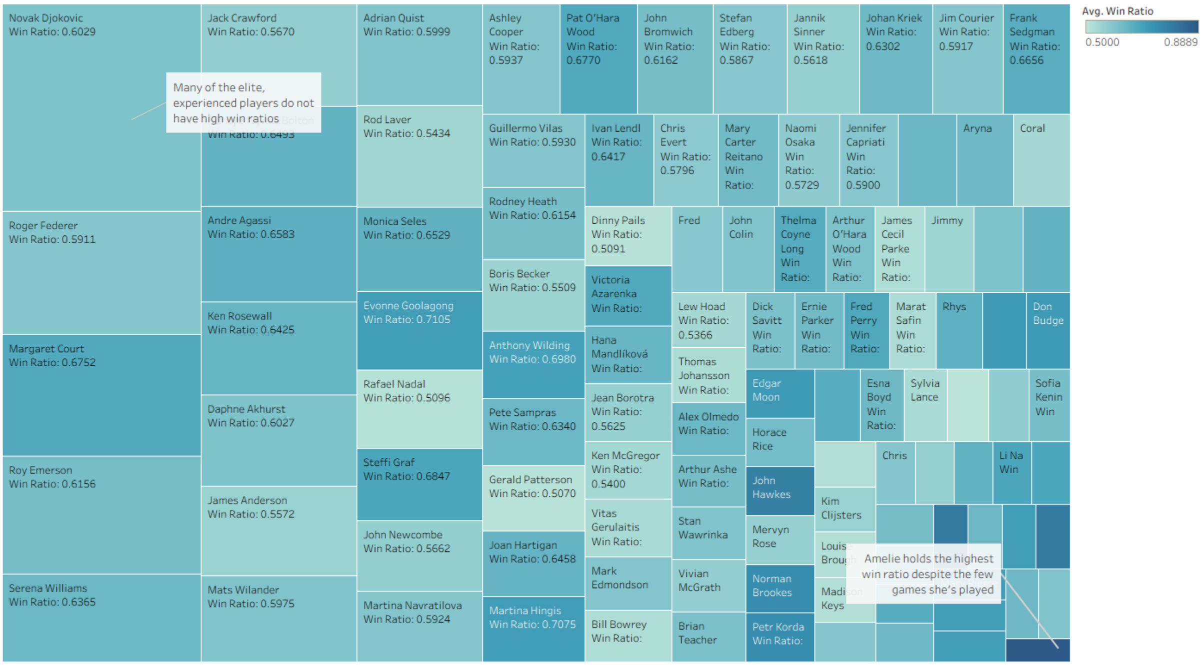
European/Asian champions contributions increase, reflected in treemap diversity

USA has more female champions than male

Figure 3. Tree map displaying champions by titles, country and gender

Analysis: Australian players occupy around one-third of the tree map, highlighting their dominance in this Grand Slam. Much of Australia's success is attributed to Roy Emerson and Margaret Court. Comparable individual achievements are evident with Serbia's Novak Djokovic, Switzerland's Roger Federer, Japan's Naomi Osaka, and Spain's Rafael Nadal. While one half of the chart is concentrated in just two countries, the other half reflects a broad mix of nations, pointing to a growing diversity of global talent.

Champions by Win Ratio and Games Lost



Many of the elite, experienced players do not have high win ratios

Amelie holds the highest win ratio despite the few games she's played

Figure 4. Tree map displaying champions by win ratio and games lost

Analysis: This chart highlights both skill and experience, with colour representing a player's average win ratio and tile size showing the total number of games lost. These pre-attentive attributes allow us to quickly identify two key players. In the top right is Novak Djokovic, who has the most games lost of all players, more than 60 above Federer. This suggests that his matches often extend to a fifth set, reflecting both endurance and high-level competition. In the bottom right, a dark blue tile represents Amelie Mauresmo. She holds the highest win ratio at 0.889, with eight games won. Her title was claimed in 2006 against Justine Henin-Dardenne, where she went up 6-1 in the first set and 2-0 in the second before her opponent retired due to stomach cramps caused by a misuse of medication.

How the charts were made: Tree maps in Tableau require two measures for size and colour, and one dimension for text. In this case, Champion was used for text, CNT(Australian Open) (number of titles won) for size, and Champion Country for colour. Gender was added to detail to highlight the contribution of men and women within each country. Finally, CNT was placed on the label to quickly show how many times each player has won the tournament.

Advantages and Disadvantages: Tree maps have several inherent advantages. They are particularly effective at differentiating between categories, which improves readability, and their use of size and colour enables quick pattern recognition. This allows readers to immediately grasp key insights, such as which categories dominate or how values compare relative to one another. Tree maps are also useful for visualising hierarchical relationships, giving a sense of structure across multiple levels of data.

However, tree maps also have notable limitations. When a large portion of the data has similar values, it can be difficult to distinguish between individual tiles, reducing interpretability. The visual complexity created by many tiles of varying sizes and colours can also be overwhelming, particularly in larger datasets, making it harder for readers to focus on specific trends or insights. Compared with simpler chart types, such as bar or line charts, tree maps require more effort to interpret, especially for audiences unfamiliar with this visualisation format.

Geographic Map

Explanation of chart type: Geographic maps are a chart type that display data on a map using information relating to locations whether that be country, state, city, etc. As opposed to presenting information in a typical chart format, geographic maps as suggested by the name, display them on a map, making it easier to interpret regional trends. In the context of this report, geographic maps can visualise where champions come from, or even the number of titles a country has won.

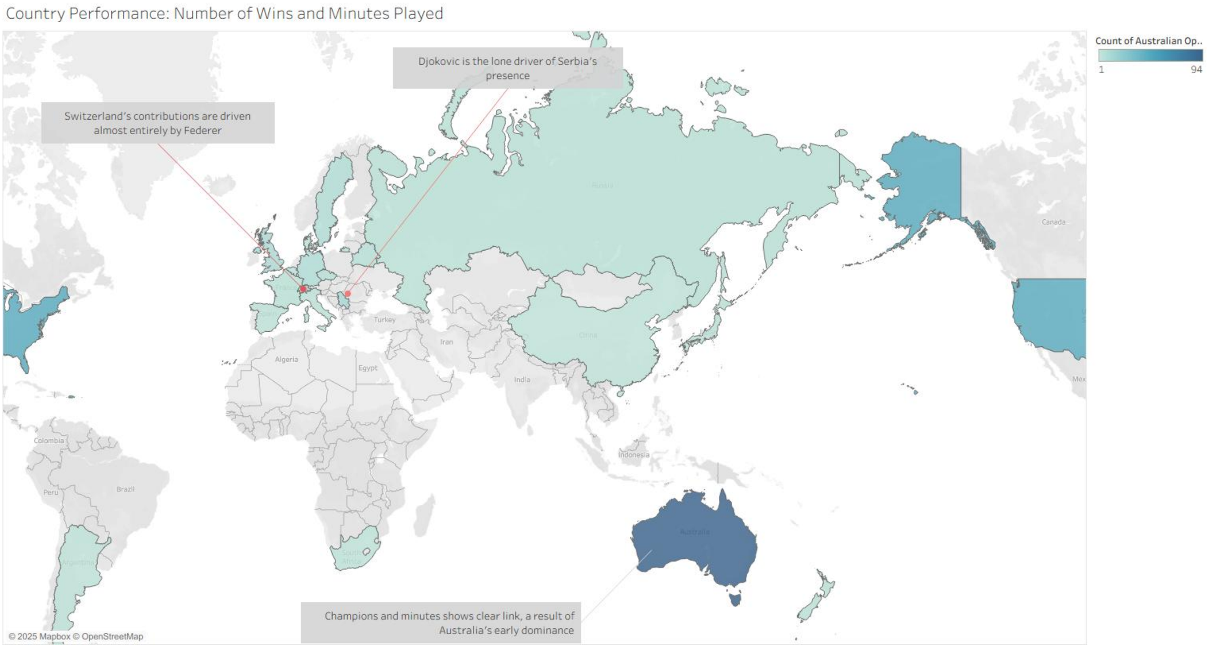
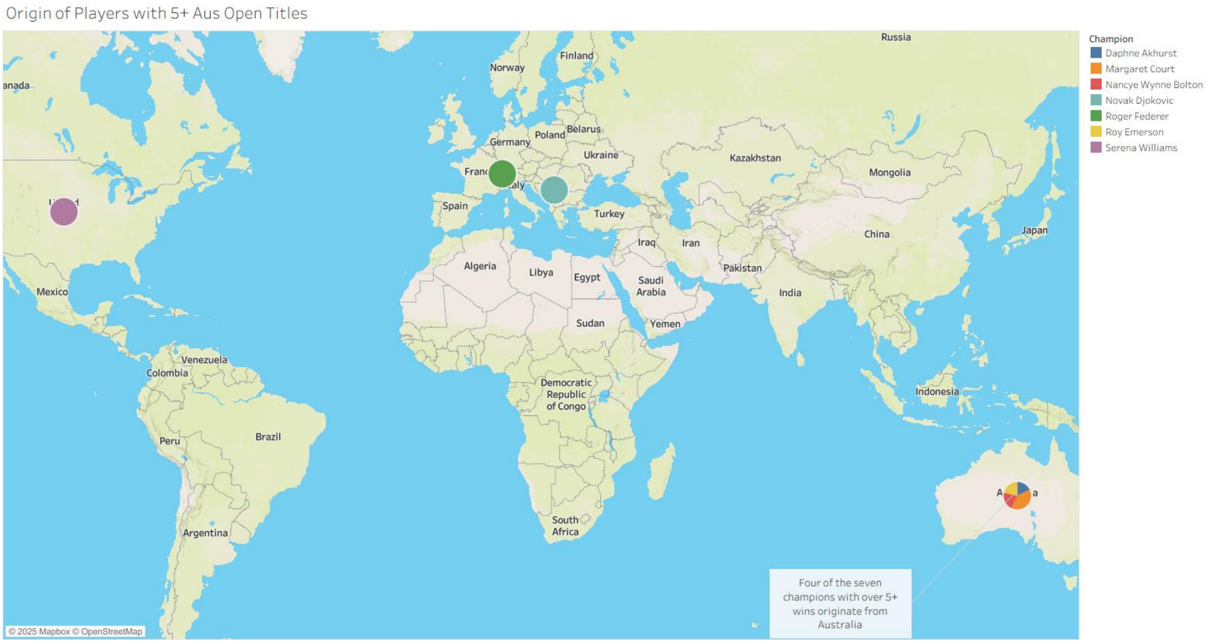

Country Performance: Number of Wins and Minutes Played
Count of Australian Op..
1
94
Djokovic is the lone driver of Serbia's presence
Switzerland's contributions are driven almost entirely by Federer
Champions and minutes shows clear link, a result of Australia's early dominance
Russia
Canada
Turkey
Iran
China
Algeria
Egypt
India
Indonesia
Colombia
Peru
Brazil
Australia
Mex
© 2025 Mapbox © OpenStreetMap

Figure 5. Geographic map showing country performance

Analysis: This map illustrates the overall performance of each country based on the number of Australian Open titles won and total minutes played. Darker shades represent countries with more Grand Slam victories, with Australia emerging as the clear leader, well ahead of second place United States. Australia has secured 94 titles, more than 50 ahead of the United States, and significantly more than all other countries combined. The total minutes played further reinforces Australia's dominance, exceeding 16000 minutes across its champions.



Four of the seven champions with over 5+ wins originate from Australia

Figure 7. Geographic map showing elite player's origins

Analysis: This map displays the countries of players who won more than five times. Again, Australia not only dominates, but does so at the elite level, with four of the seven players being Australian. As demonstrated by the pie chart, Margaret court contributes to over a third of Australia's total wins. All seven players have combined for a total of 51 Australian Open Grand Slams, with Australia accounting for 28 of these.

How the charts were made: To make these charts I first selected Champion Country, ensuring it was assigned a Geographic Role. Once selected Longitude was placed in the Columns and Latitude in the Rows shelves. I then added Champion Country to detail and Champion to colour to distinguish between each, before changing the Marks type to pie to account for countries with multiple winners and compare performances. Finally, I added CNT (counts number of wins) to the Angle for better readability and added Champion, Country, and CNT to the Tooltip for a more detailed view.

Advantages and Disadvantages: A benefit of geographic maps is that they are very digestible, in that they are easy for readers to interpret. This allows for quick comparison, and to extract actionable insights efficiently. Additionally, this chart type can reveal trends in geography and regions that most charts cannot.

Some limitations of geo maps include the tendency to oversimplify complex data. The information presented is often one-dimensional, typically displaying only a single variable at a time, and it lacks the depth and detail that other chart types can provide. Additionally, small regions such as Serbia can be difficult to see clearly, making their shading or colour challenging to interpret accurately during analysis.

Scatter Plot

Explanation of chart type: A scatter plot represents data points using dots, with each dot positioned according to two different numeric variables. The horizontal and vertical axes indicate the values for each data point. Scatter plots are typically used to identify patterns, trends, and relationships between variables.

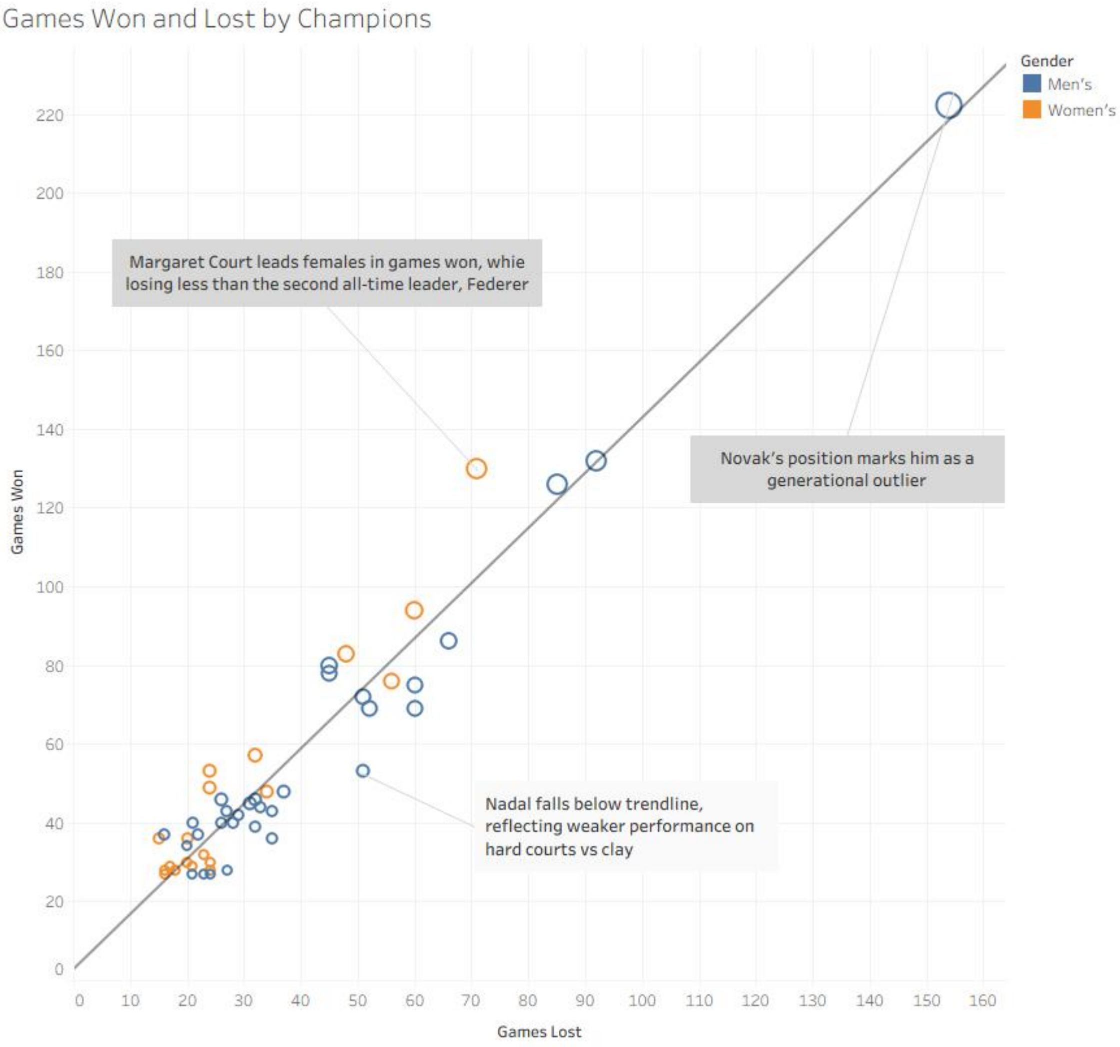


Figure 7. Scatter plot for players games won and lost

Analysis: The chart compares the performance of the top 50 players by looking at the number of games won against the number of games lost. The trend line shows a balanced split and players above tend to win more than they lose, while those below are less dominant. Rafael Nadal, despite being known as "The King of Clay", and one of the greatest players ever, falls below the line significantly. This suggests that while he thrives on clay courts, he is less dominant on hard courts. More recent champions such as Naomi Osaka (30 wins, 24 losses), and Janik Sinner (43 wins, 35 losses) also appear below the line. On the other hand, players like Margaret Court and Novak Djokovic are positioned well above the line, with Novak standing out as the most dominant.

How the chart was made: Games Lost for Columns, Games Won for Rows, and Gender for Colour, allowed me to create this chart. To make it more insightful, Champion was added to detail, where it was then filtered to display the top 50 players by games won to reduce clutter. Finally, the size was adjusted using SUM(Games Won), so that players with less wins had smaller dots, and a trend line was added using a linear model.

Advantages and Disadvantages: The scatter plot is one of the most effective chart types at displaying the relationship between two numeric variables, allowing for trends to be easily spotted and detecting outliers such as Novak. They are also straightforward to create, and simple to understand.

However, its simplicity can be seen as a weakness since it is only limited to two variables, limiting its capabilities particularly when dealing with complex datasets. Additionally, clusters and overplotting is extremely common, particularly in this instance as I had to limit entries to the top 50 only.

Conclusion

The analysis of Australian Open champions reveals patterns that extend beyond individual performance, providing insights into what drives success in tennis. Novak Djokovic and Margaret Court have reached the pinnacle of combining skill and consistency to dominate over multiple years, establishing themselves as benchmarks for male and female tennis players. Their performances highlight how athletes are not only physically capable, but also strategic and adaptable to match conditions and dynamics.

The relationship between seed and performance shows that while higher seeds generally have an advantage, tennis results are unpredictable. Players such as Madison Keys demonstrate that other factors can overcome expected outcomes, suggesting that elite performance involves external variables.

At a national level, Australia's historical dominance warrants further investigation, with their success potentially being due to early development, or being used to the conditions. However, the increasing diversity of champions signals a globalisation of tennis talent, where emerging nations are producing world class players that can compete at the highest level, indicating shifts in the competitive landscape.

Surface conditions and match duration further emphasis the importance of versatility. Champions who maintain high win ratios across different court types such as Roger Federer (typically known for his performances on grass courts) or endure extended matches (such as Novak) demonstrate not only stamina, but the ability to adapt strategies to different contexts, highlighting adaptability as a critical component of success.

In summary, success in tennis also comes in skills that cannot be measured such as mental fortitude and adaptability. While past champions have shaped the sport, the evolving competitive landscape highlights the influence of global talent and changing dynamics. These trends help us understand what makes a top player and can act as a framework for future players aiming for success at the elite level. As the sport continues to globalise and grow, exploring these variables further may reveal even more insights into the pathways to success.