Australian Open Tennis Championships Analysis

Name: Shayan Razavi
Student ID: 13532736

Exploratory Data Analysis

The dataset used in this analysis contains detailed information about the Australian Open tennis tournaments. It includes records of champions, runner-ups, match scores, and player statistics, spanning multiple years. This section summarizes the dataset's formats, values, and key characteristics, along with any observable trends or outliers.

Dataset Column Summary

Below is a table that summarizes the dataset's columns, including the data types and descriptions for each.

Column Name	Data Type	Description
Year	Ratio	The year in which the Australian Open tournament took place
Gender	Nominal	The gender category of the tournament, either Men's or Women's
Champion	Nominal	The name of the player who won the championship
Champion Nationality	Nominal	The nationality abbreviation of the champion
Champion Country	Nominal	The full name of the country the champion represents
Score	Undefined	The final match score indicating the sets won by each player
Champion Seed	Ordinal	The seed ranking of the champion in the tournament

Mins	Ratio	The duration of the final match in minutes
1st-won, 1st-loss, ..., 5th-won, 5th-loss	Ratio	The individual set scores separated into columns
Runner-up	Nominal	The name of the player who was the runner-up in the tournament.
Runner-up Nationality	Nominal	The nationality abbreviation of the runner-up
Runner-up Country	Nominal	The full name of the country the runner-up represents
Runner-up Seed	Ordinal	The seed ranking of the runner-up in the tournament

Data Types and Formats

The dataset consists of nominal, ordinal, and ratio data types. For example, categorical data like gender and country names are nominal, while seed rankings follow an ordinal structure. Match-related data such as the year and match duration are measured using ratio scales.

Missing Data

- **Mins:** has 208 out of 210 empty cells
- **Champion Seed:** has 18 out of 210 empty cells
- **1st-won:** has 1 out of 210 empty cells
- **1st-loss:** has 1 out of 210 empty cells
- **2nd-won:** has 1 out of 210 empty cells
- **2nd-loss:** has 1 out of 210 empty cells
- **3rd-won:** has 66 out of 210 empty cells
- **3rd-loss:** has 66 out of 210 empty cells
- **4th-won:** has 144 out of 210 empty cells
- **4th-loss:** has 144 out of 210 empty cells
- **5th-won:** has 185 out of 210 empty cells
- **5th-loss:** has 185 out of 210 empty cells
- **Runner Up Seed:** has 18 out of 210 empty cells

Data Distribution

The dataset consists of 210 rows, the following irregularities in the data entry was identified:

- The year column has two different champions for the year 1977 for each gender, this causes data type issues for the year column since the values were “1977-(2)” and “1977-(1)”, historically speaking the tournament was held twice that year.
- The score column has a value called “walkover”, the score column is a pretty general column anyway, but that value is an odd one out
- The champion seed and runner up seed has a value called “U” for three of the rows, it could mean unidentified
- For the score, tie breaks are usually put in brackets, but there are cases where a score like 7-6(11-9) is just 11-9

Dataset Values and trends

- **Year:** The dataset spans tournaments from 1905 to 2024.
- **Gender:** Two categories are present: Women's and Men's tournaments.
- **Champion Seed:** Seed rankings range from 1 to 17 with some blanks in between excluding not applicable and undefined.
- Runner Up Seed: Range from 1 to 31 with some blanks in between excluding not applicable and undefined
- **Most Frequent Champion:** The most frequent champion is Margaret Court.
- **Most Frequent Champion Country:** The most frequent champion country is Australia.

Outliers

- **Lower Seeds Claiming Titles**: One clear outlier is the occurrence of champions with much lower seeds, such as Seed 14 or 17, winning the tournament. Given that the majority of winners come from top seeds (Seeds 1–4), these lower-seeded victories are notable exceptions and suggest unexpected performances.
- **Unseeded Winners**: The presence of unseeded champions (marked as "U") in the dataset is a significant outlier, as the majority of champions have been seeded. This highlights moments where a player not expected to perform at a top level has managed to outperform higher-ranked opponents.
- **Extreme Seed Ranges**: While most champions come from seeds 1 to 4, the occasional victory by a seed as high as 17 or even an unseeded player deviates from the norm. This indicates that the Australian Open has witnessed significant upsets, where lower-ranked players have outperformed expectations.
- **Unusual Scores**: Some matches have unusual or uncommon score lines, such as multiple tie-breaks or large differences in game points, particularly in finals. These score lines may suggest particularly one-sided matches or highly competitive ones that went beyond the standard sets.

Data Transformations

I applied data transformations to the following columns:

- **Year:** I removed the dash and brackets with numbers in them by separating the column using this as the delimiter “-” and made year a whole number and deleted the other separated column
- **Champion Nationality:** I removed this column since i already have Champion Country
- **Mins:** There isn’t enough data, not sure where i can find data for this attribute so I’ll just delete it
- **Runner Up Nationality:** Similar to Champion Nationality Runner Up Country already exists
- **Champion Seed:** Between 1905 and 1923 the Australian open was in its early stages and didn’t have a seed for the players during that time so I replaced the null values with N/A which is suitable for visualizations.
- **Runner Up Seed:** Between 1905 and 1923 the Australian open was in its early stages and didn’t have a seed for the players during that time so I replaced the null values with N/A which is suitable for visualizations.
- **Normalized Sets** all the 1st-won, 1st loss… attributes were min max normalized into new attributes incase i want to use them in my visualizations
- **Overall win rate** is a new calculated column calculated by dividing the number of games won by the total number of games played.
- **1st Set Win Rate, … ,5th Set Win Rate** is a new calculated column that was calculated by dividing the number of games in that set (e.g. the 1st set) by the total number of games played in that set.

Findings

Total Of Champions For Each Country (Tree Map)

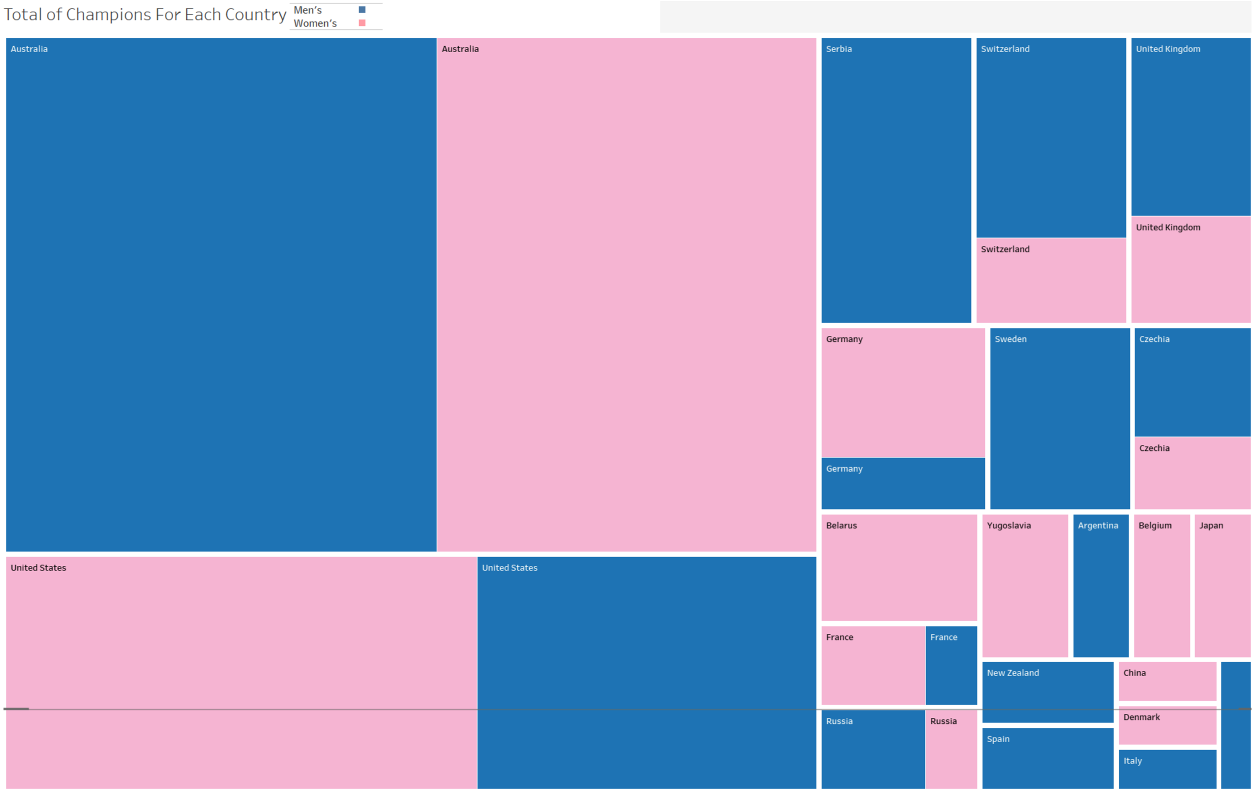
My objective was to find out which countries had the most champions by counting each occurrence of the Champion Country in the dataset.

Trends and Outliers

The treemap highlights the dominance of Australia and the United States in both men's and women's categories, with a particularly strong showing from American women. European countries such as Serbia, Switzerland, and Germany also feature prominently, especially in the men's division, reflecting more recent successes by players like Novak Djokovic and Roger Federer. The distribution of champions varies across countries, with some showing more balance (like Australia and the U.S.), while others, like Serbia, are skewed towards men's champions. The data reflects historical dominance and modern shifts, with smaller nations like Argentina, Japan, and Belarus making notable, though less frequent, contributions. The treemap reveals the increasing globalization of tennis, with champions emerging from smaller nations like Belarus and Japan. Some countries show gender balance in champions, A notable outlier is Serbia’s large representation, driven by Novak Djokovic’s influence.

Graphic Techniques

I used two colors to distinguish gender rather then having multiple distinct colors such as identifying country and gender which will over complicate the visualization, this improves readability by decreasing noise from the color, I also used a legend to point out the colors of the gender, so i don't have to add them to the label, the font size for labels are 8 and bold since i want to be able to fit the country names for each square, this way the reader can identify what country is in each square.



Win Rate For Each Set For Each Champion Seed & Gender (Parallel Coordinates)

The objective was to analyze average win rates for each set based on champion seed and gender. Most players have a consistent win rate in the first and second sets, averaging around 60-70%. Performance variance becomes more apparent as the match progresses, with higher-seeded champions showing a rise in win rates in the final sets, suggesting stronger endurance or resilience.

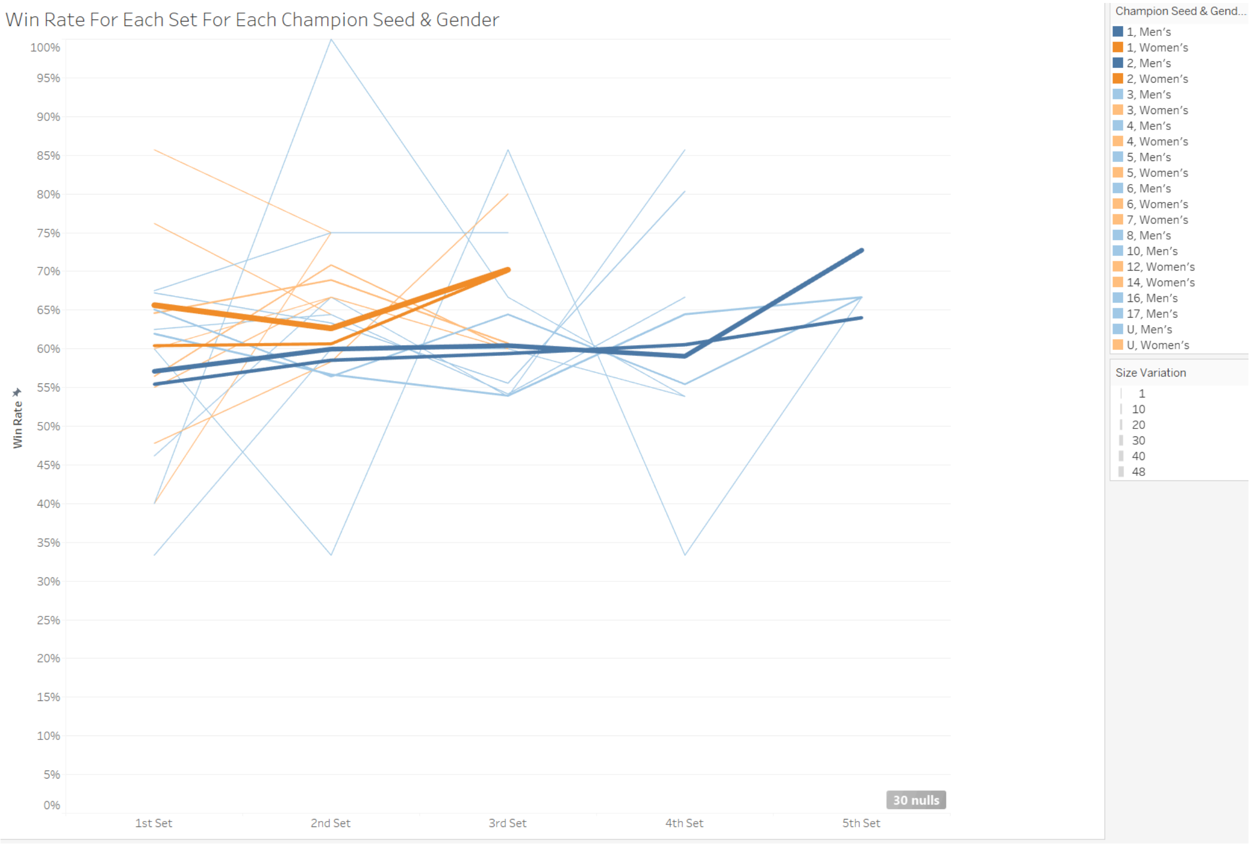
Trends and Outliers

- **Trend of Consistency in Top Seeds**: Higher-seeded champions (e.g., 1st seed) in both men's and women's categories demonstrate more consistent win rates throughout the sets. Their performance does not drop as significantly in later sets compared to lower-seeded players, indicating a strong trend of dominance by top-seeded athletes.

- **Outliers in Win Rate**: Some unseeded players or lower seeds exhibit sharp spikes or drops in win rates during certain sets (17 and 10 mens). For example, a few lines dip dramatically around the 2nd or 4th set, indicating a vulnerability for specific players or seeds in the middle of the match.
- **Rise in the 5th Set for Men's Champions:** A clear trend is visible among men's champions, where the win rate notably increases in the 5th set for some seeds, highlighting endurance and tenacity. This could be an outlier for a small number of players, as not all matches extend to a 5th set.
- **The top seeds for men and women show a similar pattern**, they both start with a neutral win rate, but when they reach the last set the win rate spikes up

Graphic Techniques

SImilarly to the previous chart i used two colors for genders to not overcomplicate the color scheme for readability purposes, i used count of champion to change the size of the lines to emphasize which lines were more important and used dark colors for seeds 1 and 2 so readers focus on those two lines since it’s the top two seeds and had more data. I made sure the y axis ranges from 0 to 100% which is the natural range. I changed both axis labels to make them more readable. I had two legends, one legend shows the different gender and seeds and the other shows the variation in sizes for the lines. Overall these changes enhance the readability of the report.



Geographic map of countries with the most champions divided into champion seeds

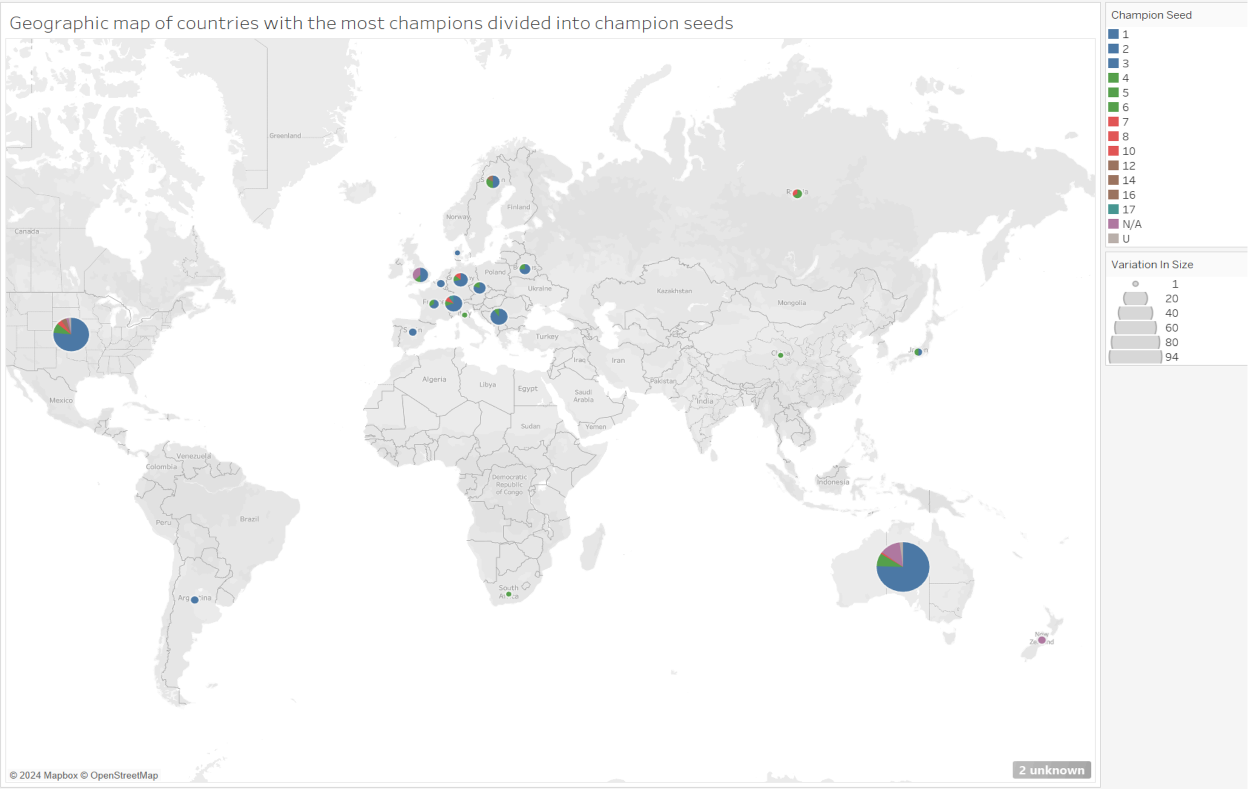
The geographic map shows countries with the most Australian Open champions divided into champion seeds..

Trends and Outliers

- **Dominance of Major Tennis Nations:** The trend shows that Australia and the U.S. dominate the chart, not only by producing the most champions but also across a wide range of binned seeds, from high to low. This suggests both top-seeded and lower-seeded champions emerge from these countries.
- **European Strength in Top Seeds:** Many European countries, like Serbia and Switzerland, show a strong presence among the top-seeded bins, indicating that their champions tend to come from higher seeds, reinforcing their position in global tennis.
- **Outliers:** Countries like Argentina, Russia, and Japan have relatively smaller circles, indicating fewer champions overall probably due to their late start. Their champions mostly come from top-seeded bins, which might suggest fewer but more dominant players emerging from these regions. The smaller size and limited seed variety make them outliers in terms of breadth, though still competitive in top-seeded categories.

Graphic Techniques

I scaled the size of the circles for each country to be bigger, so they can be more distinguishable, but not more bigger because the circles in europe would've overlapped, I made each circle a pie chart based on champion seed and then i binned the champion seed to show less variation to better help distinguish which countries had the better champion seeds.



Champions in their respective countries over the years (scatter chart)

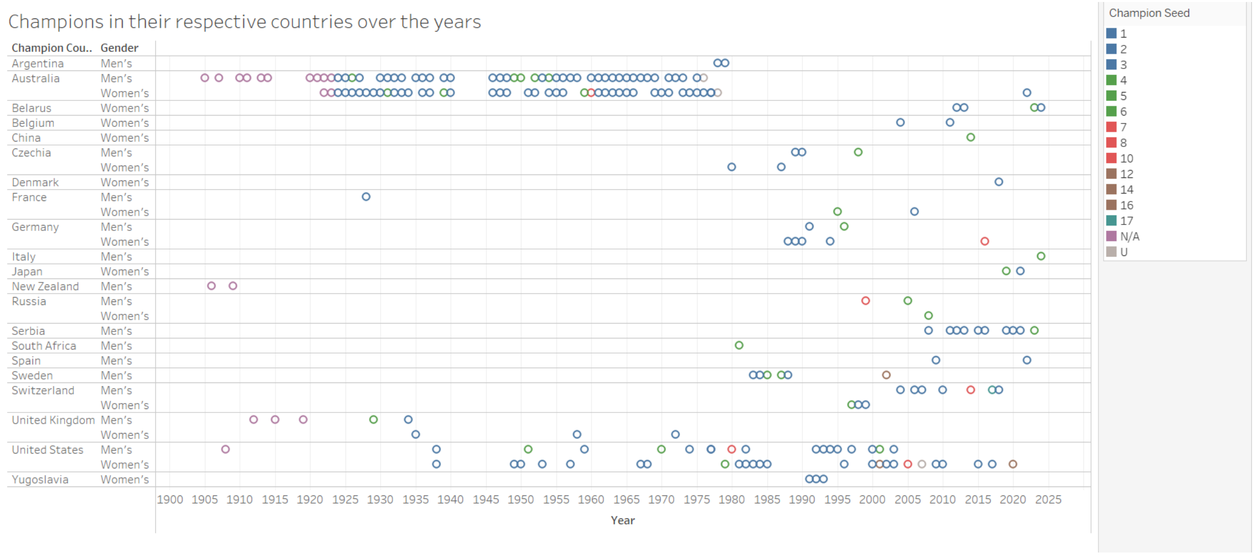
This chart presents a historical view of Australian Open tennis champions by country, gender, and seed using a scatter plot,

Trends And Outliers

- **Historical Dominance**: The chart highlights a clear trend where Australia and the United States dominated the tournament in the earlier decades, particularly from the 1920s through the 1960s, with most champions coming from high seeds (1st and 2nd) probably due to the limited people playing.
- **Recent European and Global Presence**: From the 1990s onward, more European countries and nations like Russia and Serbia start to appear, showcasing a shift in dominance toward Europe. This modern trend indicates more global participation and success.
- **Outliers**: Some outliers are noticeable in the chart. For instance, unseeded champions (grey) from countries like Switzerland and Russia emerge in recent decades, which is uncommon compared to the heavy presence of high seeds. Another outlier is the long gap in representation from countries like France, where champions appear sporadically over time.

Graphic Techniques

Champion Country and Gender were grouped together so it's easier to identify them, champion seed was binned so that there isn't so many color variations making it easier to interpret, seeds that were not applicable or undefined had their own color to identify them. A legend on the right shows where each seed matches a color for the viewer's reference improving readability.



Runner ups in their respective countries over the years (scatter plot)

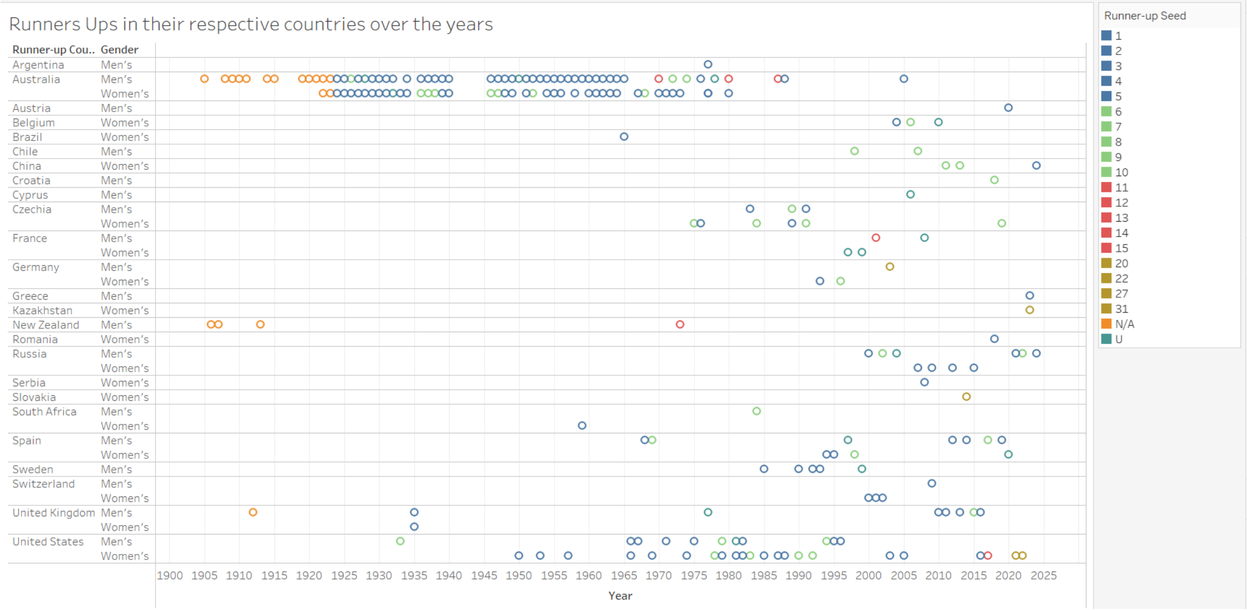
This chart visualizes the historical timeline of Australian Open runners-up by country, gender, and seed.

Trends and Outliers

- **Historical Trend of Australian Dominance:** Australia shows a significant presence among runners-up, especially in the mid-20th century, indicating that australia was closed off at the time. However, Australia's runner-up presence has decreased in recent decades, signaling a shift in global tennis competitiveness.
- **Rise of European Countries**: In the later years (since the 1990s), there is a clear trend of more European countries, including Serbia, Russia, and Spain, producing runners-up, indicating Europe’s growing dominance in the tennis scene.
- **Outliers**: Several outliers are evident, particularly unseeded players (U) from countries like Serbia and Switzerland reaching the finals, which is a rare feat. Another outlier is the relatively few runners-up from traditionally strong tennis nations like France and Germany compared to their neighboring European counterparts.

Graphic Techniques

Similar to the previous scatter chart I binned the runner up seeds so there’s less color variation for better readability, also like the other scatter chart i grouped country and gender together on the y axis so that the viewer can quickly identify the runner ups based on their country and gender, I also made sure the runner ups seeds undefined and not applicable had their own color and included a legend for runner up seeds. Overall enhancing readability as explained earlier in the other chart...



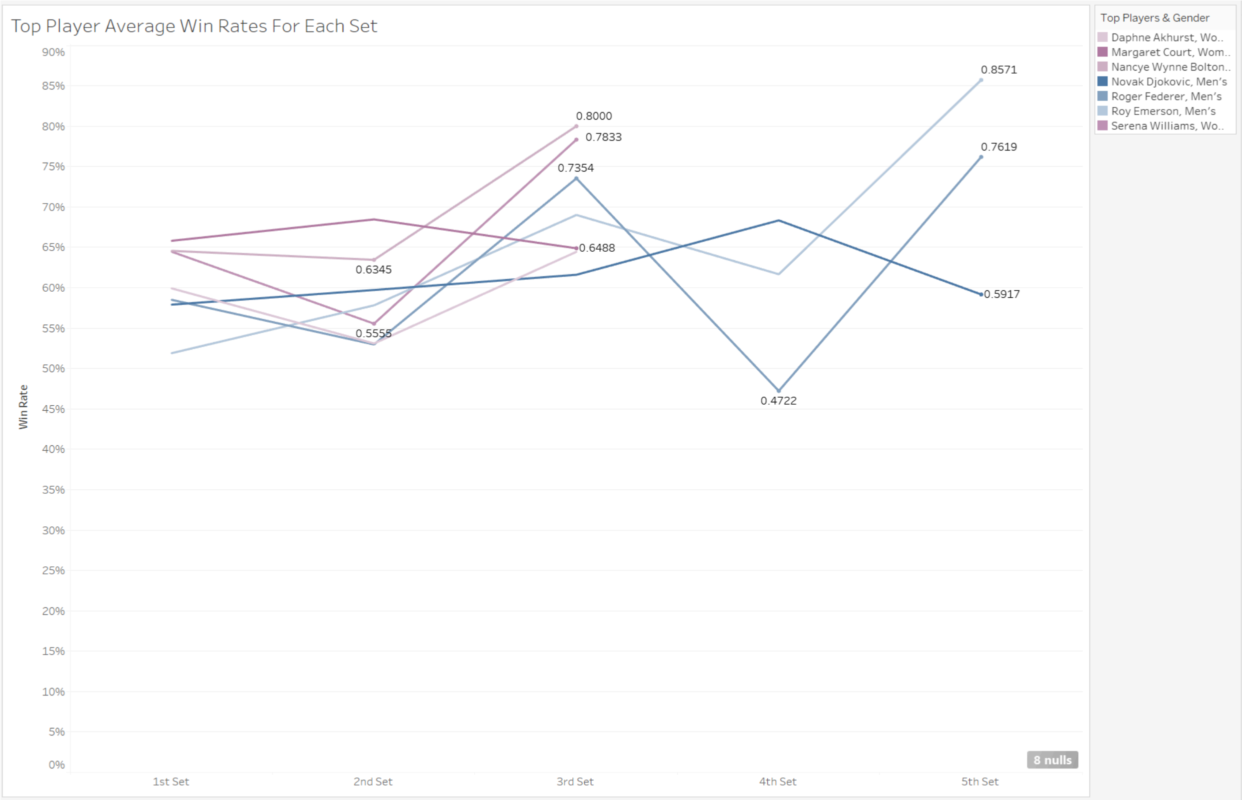
Top player average win rates for each set (parallel coordinate chart)

Trends and Outliers

- Djokovic shows consistent performance in all of his sets with a slight bump in the 4th set at 68%, but stays between 58 and 68%.
- Federer is the most volatile dipping to 47.22% average win rate in his 4th set and then jumping back up to 76.19% on the 5th set somehow regaining his performance
- Both Nancye and Serena have very good average win rates at the last set showing that they have more stamina to play the game compared to the opponent
- Roy Emerson seems to play poorly at the beginning, but starts to pick up gradually after the first set, becoming a late game monster at the last set with an average win rate of 85.71%
- Margaret Court the one with the most titles like Djokovic stays consistent in her average win rates across the sets hovering around 64 to 68%

Graphic Techniques

I wanted to keep the colors simple so i chose purple and blue, purple for women and blue for men, since I'm only working with a few data series, i just made lighter shades to distinguish each player without making things to overly complicated, this way the reader can distinguish gender and player name quite easily, the annotation in tableau weren't that great so i just used data labels for relevant sets on the chart, for the y axis i just left it at a range of 0 to 90% to make the changes in the chart more apparent.



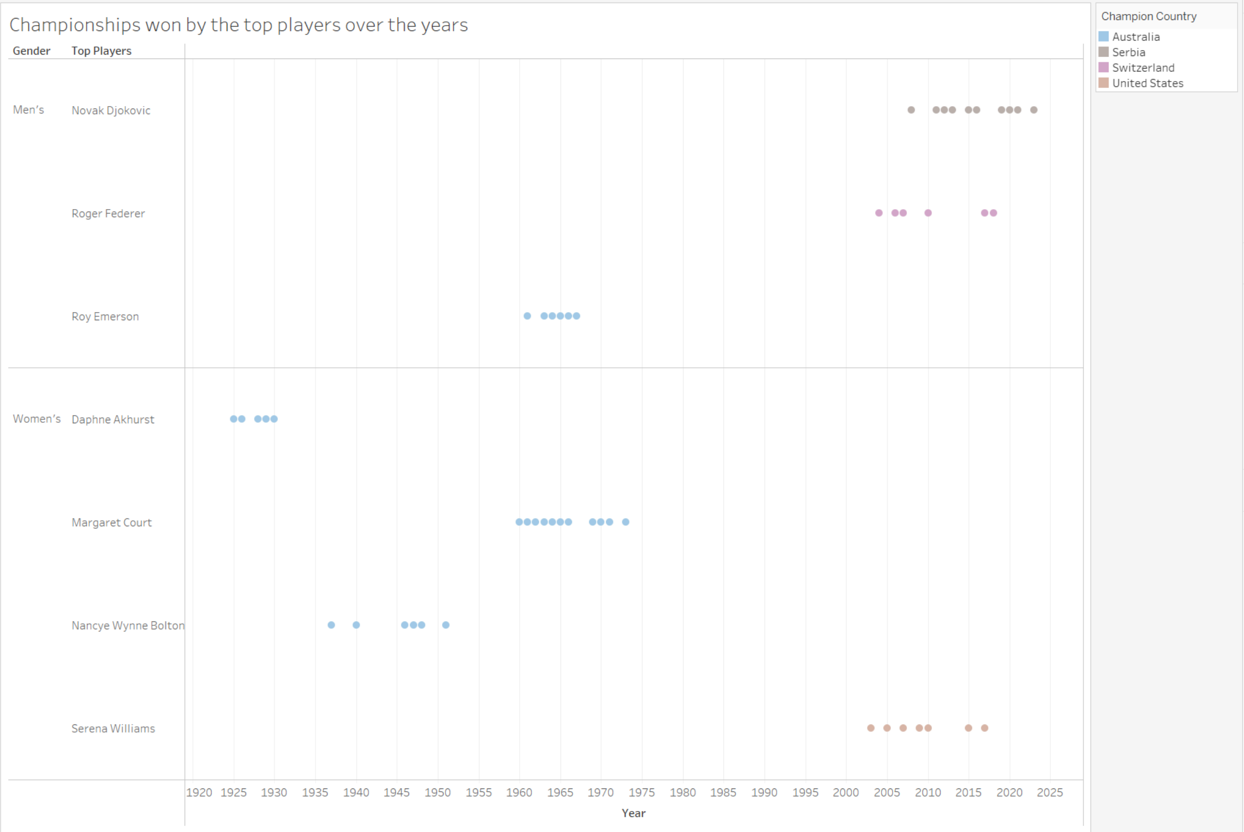
Championships won by the top players over the years (scatter plot)

Trends and Outliers

- There’s a huge gap between margaret court and serena williams (1973-2003) during this time there were no known players in the women's competition who have dominated during this time period which spans 30 years
- All these top players don't have gaps in their winning streak which makes sense since players get old and need to retire, but some notably Roger Federer, Nancye and Serena have one big gap which shows maybe they had to revise the way they played and their determination to win is strong.
- It is also clear that Australian top players were only good in the past and newer players from other countries are the top players these days.
- There’s only one top player from another country except Australia in the women's side which is Serena williams.
- Roger Federer and Novak Djokovic's streaks kinda overlap showing intense competition.
- 1967 to 2004 is also another gap where there wasn’t a dominant player in the mens dominating during that time

Graphic Techniques

I distinguished the top players by country using color and provided a legend, since there weren’t that many top players the color was less making it more readable, the champions were grouped together on the y axis based on gender, which helps focus on mens and womens top players.



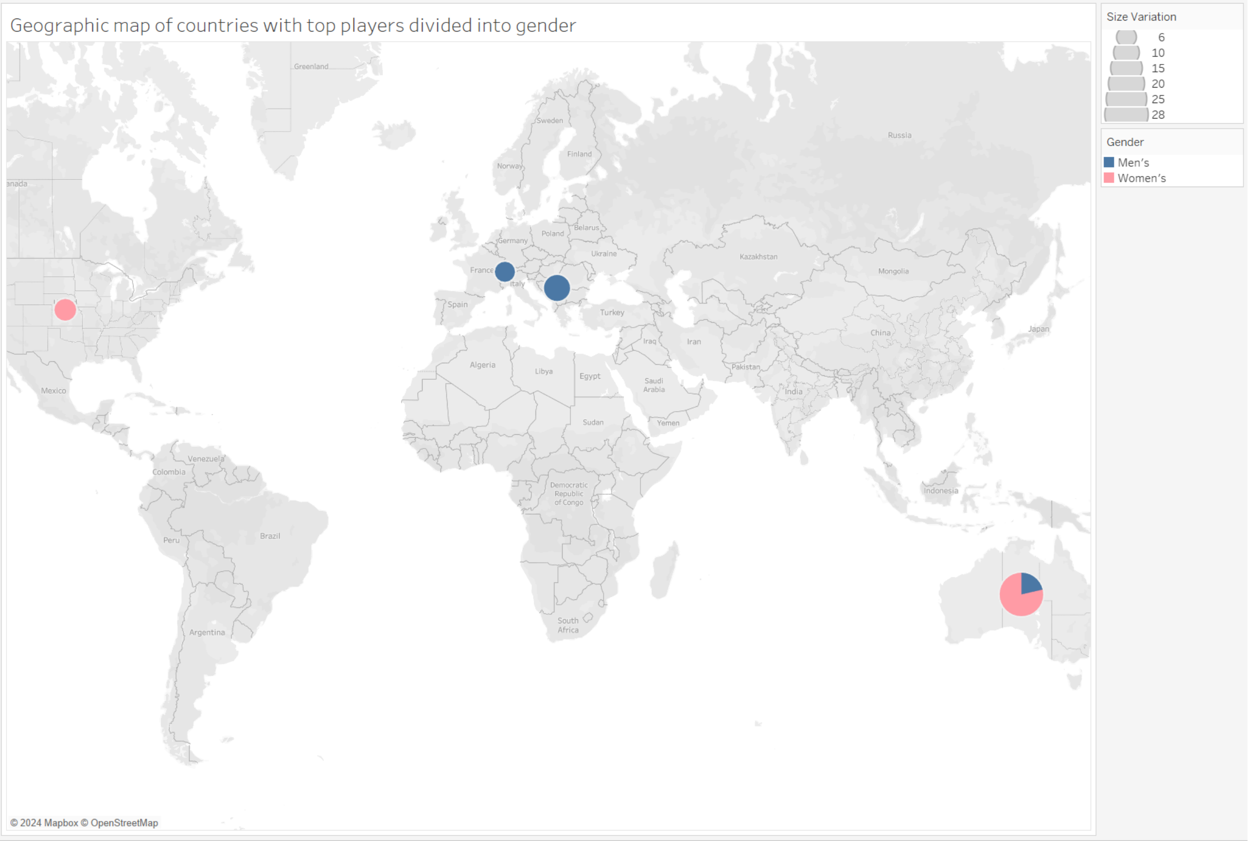
Geographic map of countries with top players divided into gender

Trends and Outliers

- United States doesn't have any male top players, but has a female top player
- Europe only has male top players
- Australia is more diverse in top players with male and female probably due to not many countries participating in the early time of the australian open

Graphic Techniques

I used pink to distinguish female and blue to distinguish male and added a legend to the right, i scaled the size of the circles making sure not to make them too big, i made the bubbles on the map into a pie chart divided into gender so readers can see the gender distribution.There's also a legend for size showing the variations in size that have been calculated.



Side by side of top players average set win rates (side by side circles)

Trends and Outliers

- The top women players seem to have more similar win rates

- The performance level of women compared to men is slightly more, this could be due to Roger Federer's average win rates skewing the performance level or also not enough data on men and also women have shorter games.
- Top men players seem to have the most outliers compared to women, maybe due to the long games

Graphic Techniques

I used pink to distinguish female and blue to distinguish male and added a legend to the right, this makes it easier for people to interpret as there are less colors, i added two annotations marking areas where there is one for womens performance level and mens making it transparent enough for the data points to be visible, i changed the y axis to range from 0 to 100% because win rates can be maximum 100% or minimum 0. I grouped the top player names by gender to help the viewer focus on either of the two. Overall all these changes help readability.

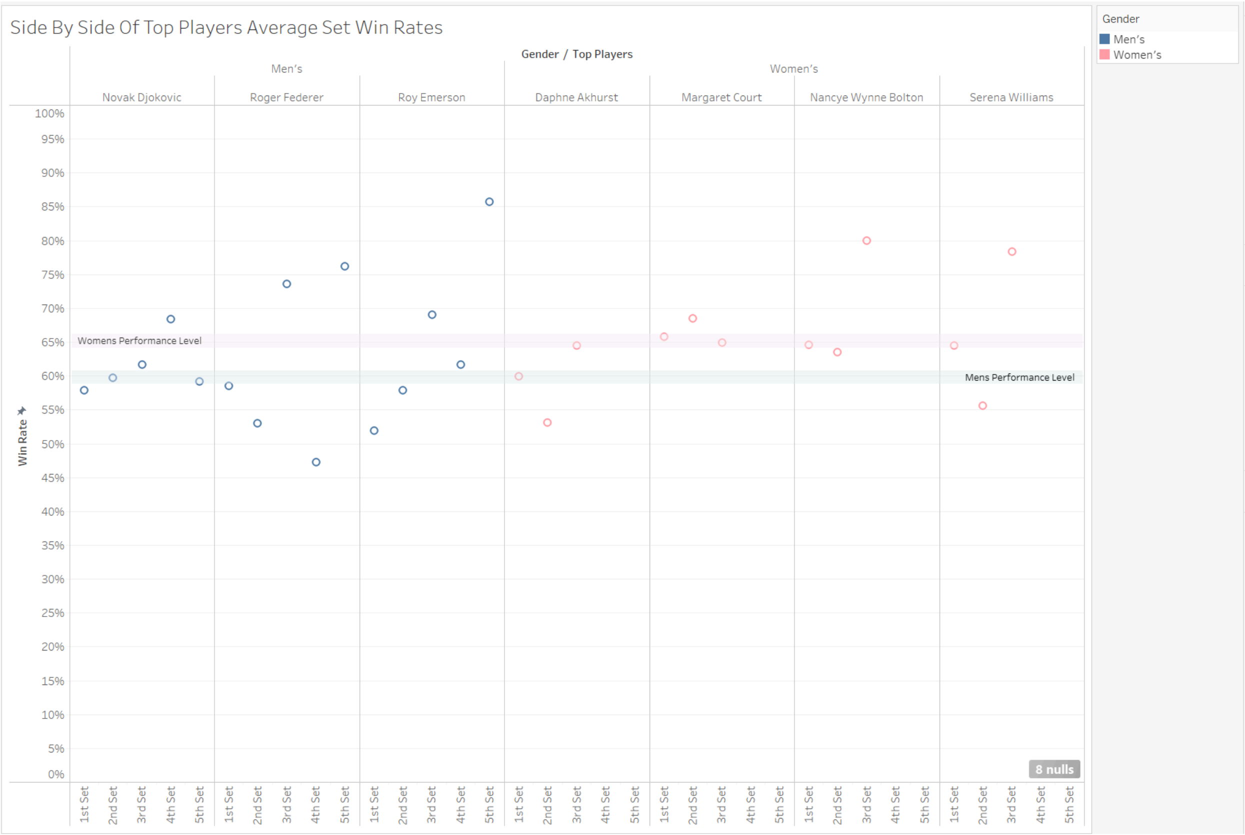


Tableau Critics

Advantages

- **User Friendly Interface:** I was quickly able to learn how to use the basics of tableau making the learning curve easy

- **Worksheet, Dashboard and Story:** I like how i could work on a worksheet and then combine everything into a dashboard and the story feature could be a good feature as well which i haven't tried out
- **The visualizations seem powerful:** I think tableau's strong suit is the visualization it can produce as it supports a broad range of chart types, coming from a PowerBI perspective.

Disadvantages

- **Limited data preprocessing capabilities:** While Tableau excels in data visualization, its data cleaning and transformation capabilities are not as robust as tools like Power BI or Alteryx, requiring users to prepare the data outside of Tableau in my case using excel.
- **Limited Formatting Options:** I had trouble changing the data labels to percentages and assigning colors was a big pain, I wish there was an easier way of doing it, otherwise it would just get complicated requiring me to make a custom calculated field.
- **Limited Customizability:** there were limited ways i can add extra features to the charts such as annotations and the ones there were available weren't that good

Conclusion

In conclusion, this analysis of the Australian Open reveals several key insights into player performance, country dominance, and evolving trends over time. Australia and the U.S. have historically led the way, but recent years have seen the rise of nations like Serbia and Switzerland, largely thanks to iconic players like Novak Djokovic and Roger Federer. Their ability to stay strong in late sets showcases their resilience and ability to handle high-pressure situations.

The data also underscores the expected dominance of top-seeded players, especially in early sets. Yet, lower-seeded or unseeded champions have shown to surprise everyone by winning the tournament, illustrating the unpredictable and competitive nature of the Australian Open.

On the women's side, Serena Williams and Nancye Wynne stand out for their impressive late-set performances, demonstrating stamina and determination as they close out matches. Meanwhile, the men’s game often sees more volatility in later sets, with players like Federer showing a dip in the middle before surging back.

Overall, the insights gained from this analysis reveal a balance between predictable patterns—like the dominance of top seeds—and the unexpected victories that make the Australian Open so compelling. As tennis has become more global, the tournament continues to feature a mix of established champions and new contenders, adding to its rich history and ongoing allure.