Assignment 2

Executive Summary

The AUS Open Tennis dataset was provided for analyzing the tennis results throughout the 120 years of the tournament. The insights will be used for people interested in tennis. High-dimensional visualization methods were used to produce graphs that are friendly to the general audience or statistically heavy. Treemap, line chart, symbol map, parallel coordinates, bar chart, scatter chart and word cloud were used for visualization. It was found that Australia displayed a strong win rate particularly in men's category, a few fluctuations on overall win rate was found throughout the years, and only male players have won from 4th round. The two most significant figures were found to be the Australian woman player Margaret Court and Serbian man player Novak Djokovic. Thus, the visualizations communicate the key insights from the dataset, making the patterns easier to detect.

Introduction

The AUS Open Tennis dataset is used for visualization and interpreting the insights. First the dataset was analyzed, and the summary of each column and their data format are shown below.

Attribute	Description	Format	Data Type
Year	The year that the match was held	Date	Categorical (Ordinal)
Gender	Gender of the champion and runner-up	String	Categorical (Nominal)
Champion	Name of the champion	String	Categorical (Nominal)
Champion Nationality	Nationality of the champion	String	Categorical (Nominal)
Champion Country	Country that the champion represents	String	Categorical (Nominal)
Score	Scores throughout the match. E.g. 6-3,7-6(13-11) means the champion won by 6-3 in the first round and 7-6 in the second round, with (13-11) tiebreaker on the 6-6 mark.	String	Categorical (Nominal)
Champion Seed	Indicates the player's ranking at the start of the match. Based on	Integer (whole)	Categorical (Ordinal)

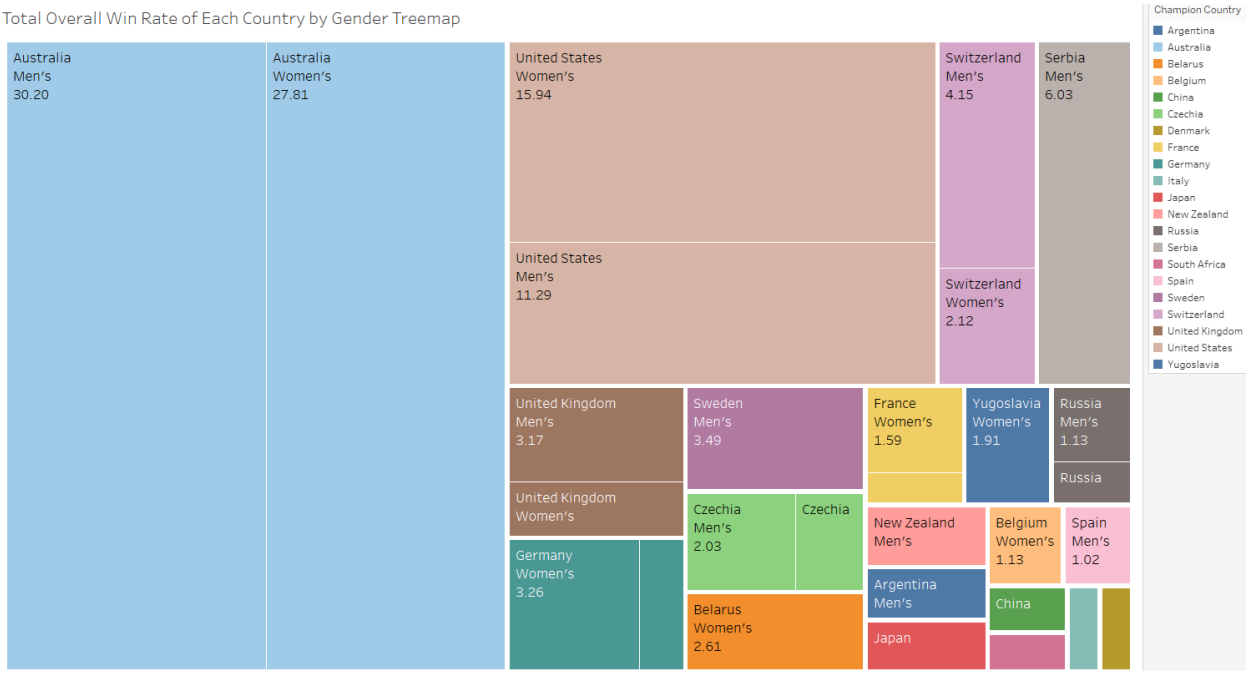
	overall prior performance and used to avoid matching up two top players.		
Mins	Time taken for each match	Integer (whole)	Quantitative (Interval)
1st won ~ 5th won	Number of points won for each set	Integer (whole)	Quantitative (Ratio)
1st loss ~ 5th loss	Number of points lost for each set	Integer (whole)	Quantitative (Ratio)
Runner-up	Name of the runner-up	String	Categorical (Nominal)
Runner-up Nationality	Nationality of the runner-up	String	Categorical (Nominal)
Runner-up Country	Country that the runner-up represents	String	Categorical (Nominal)
Runner-up Seed	Similar to champion seed (See champion seed above).	Integer (whole)	Categorical (Ordinal)

New columns were also created for further analysis. New columns 1st Set Win Rate ~ 5th set win rate were created by dividing the number of games won in that set by the number of games played in that set. And the column Overall Win Rate was created by dividing the number of games won in that match by the number of games played in that match.

The following visualization methods are used – treemap, line chart, symbol map, parallel coordinates, bar chart, scatter chart and word cloud. The null values in the dataset are left as is since it is not suited to fill them with zero or remove those rows. The Tableau software automatically omits cells with null values as well. And since the column Mins only has two cells with values, it is not used during visualization.

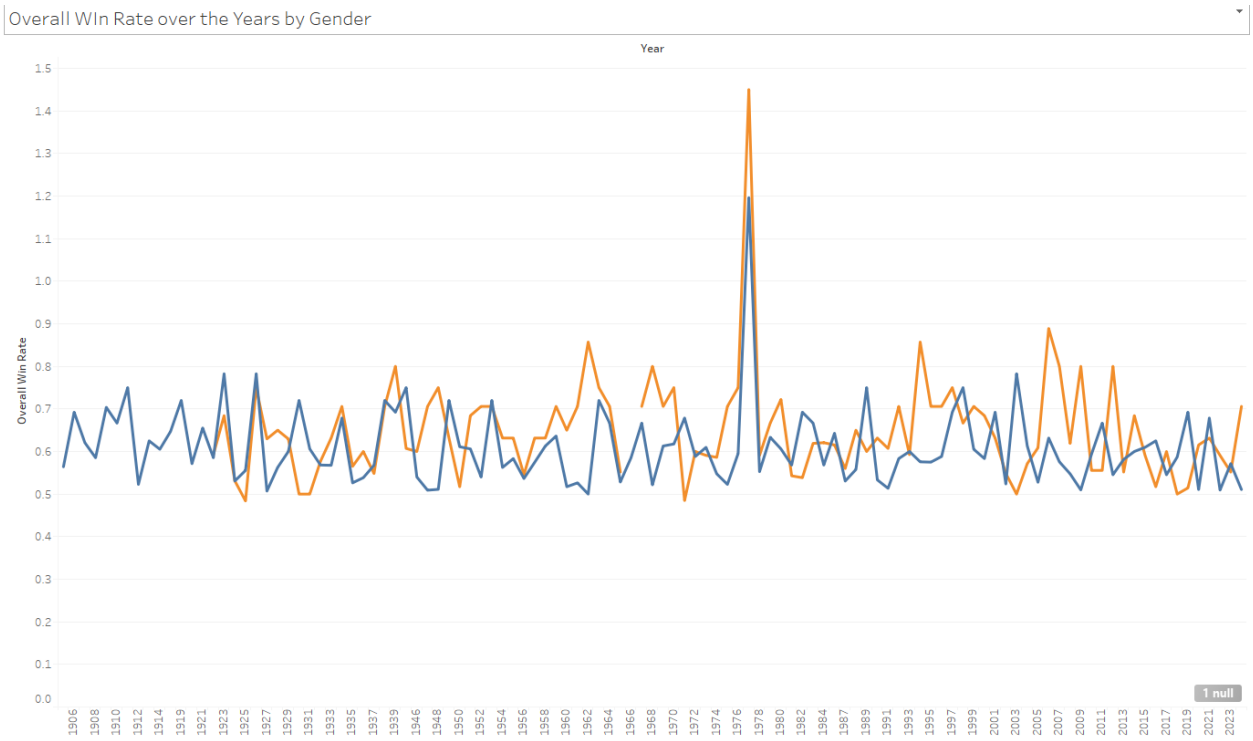
Data Visualization Charts and Explanation

Treemap for total overall win rate of each country by gender



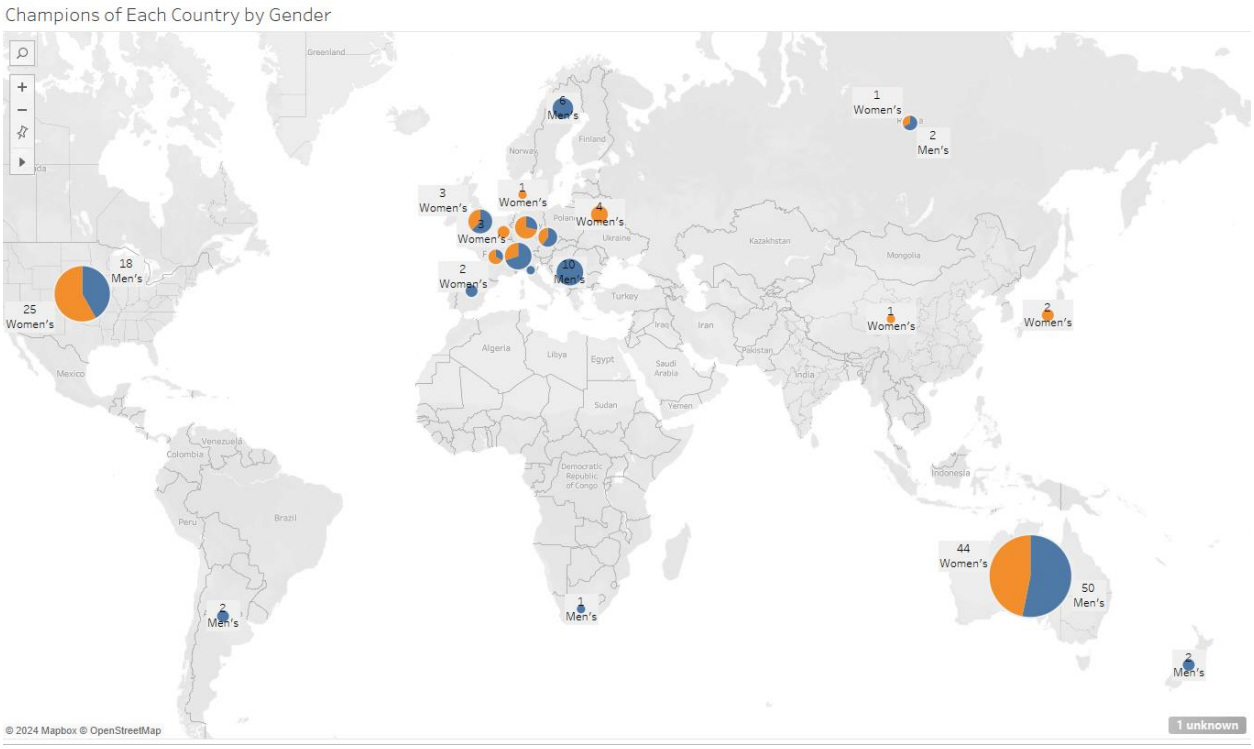
The graph above shows the total overall win rate by each country, divided by gender. It is apparent that Australia has the largest total overall win rate, with value for men higher than that for women. The second largest value is achieved by United States, followed by Switzerland and Serbia. The lowest value is achieved by Denmark which only had women champions and achieved 0.5, followed by Italy with score of 0.51. It is also discovered that there are seven countries with mens-only champions and six countries with women-only champions.

Line chart for overall win rate over the years by gender



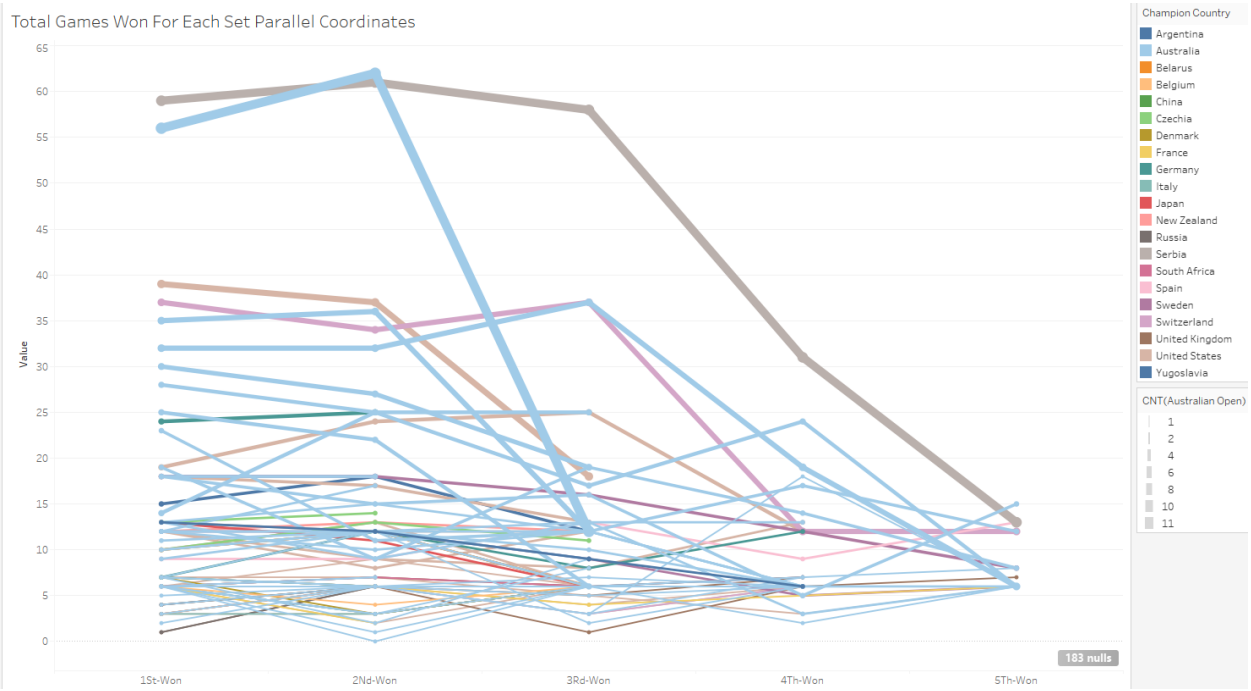
The figure above shows the overall win rates achieved by each man and woman champion over the years. It is evident that the women's category did not start until 1922, and one female contestant walked over in 1966, thus automatically making the opponent champion. Both men and women had highest overall win rate in 1977, achieving scores of 1.1961 for men and 1.45 for women.

Symbol map for winning count of each country by gender



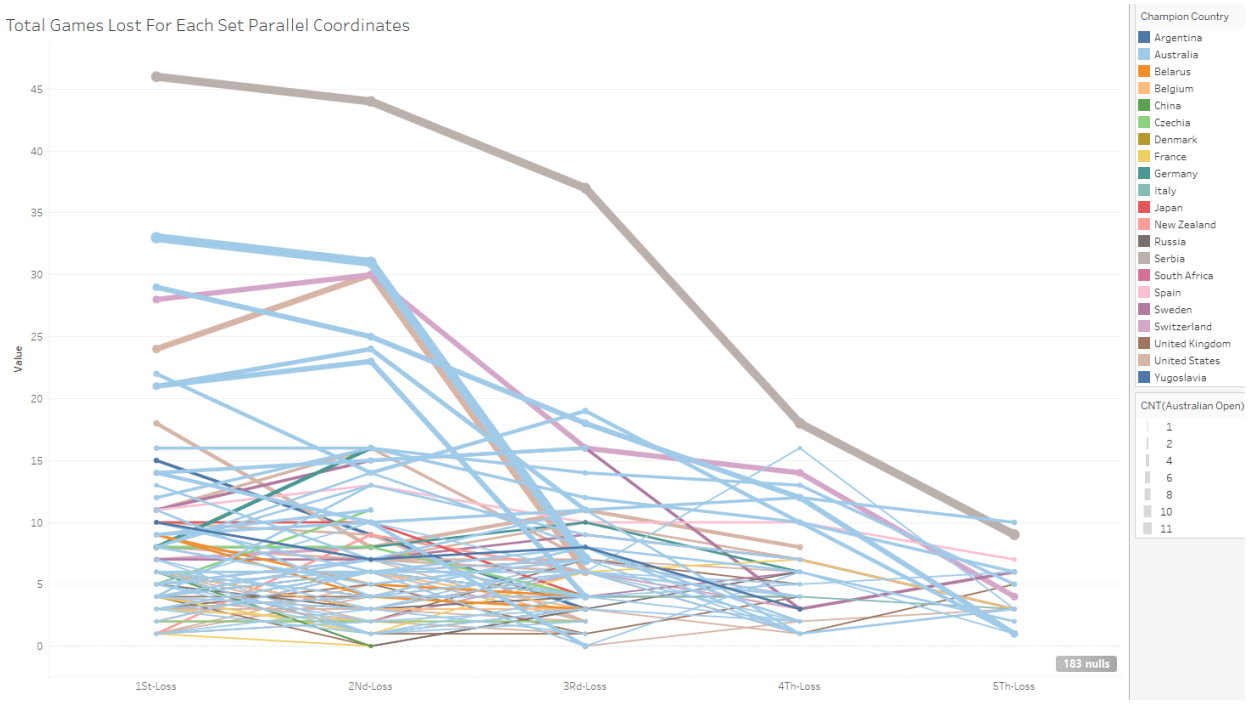
The figure above shows the winning count of each country, categorized by gender. It Australia has the largest number of winnings, with 50 men and 44 women. United States comes in second with 18 men and 25 women. Like treemap shown above, Switzerland and Serbia come in 3rd and 4th places respectively.

Parallel coordinates for total games won for each set



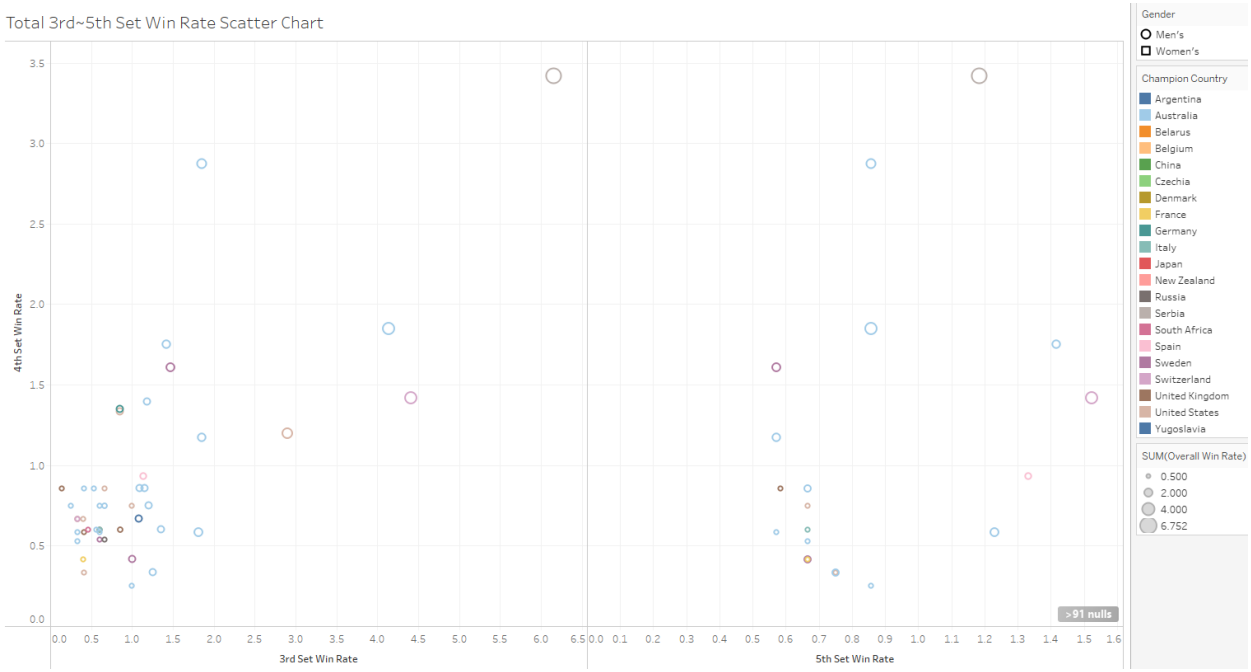
The figure above shows the total games won in each set, colored by champions' countries and line size defining the player's winning times count. The female champion Margaret Court from Australia was the most dominant in the second round with 62 points, and does not have any 4th and 5th rounds, indicating that she won early in her matches. The highest points for 1st and 3rd round were achieved by the male Serbian Novak Djokovic. It was also noticed that Czechia players had a steady count of games won and ending in the 3rd round, which might indicate their country's playstyle.

Parallel coordinates for total games lost for each set



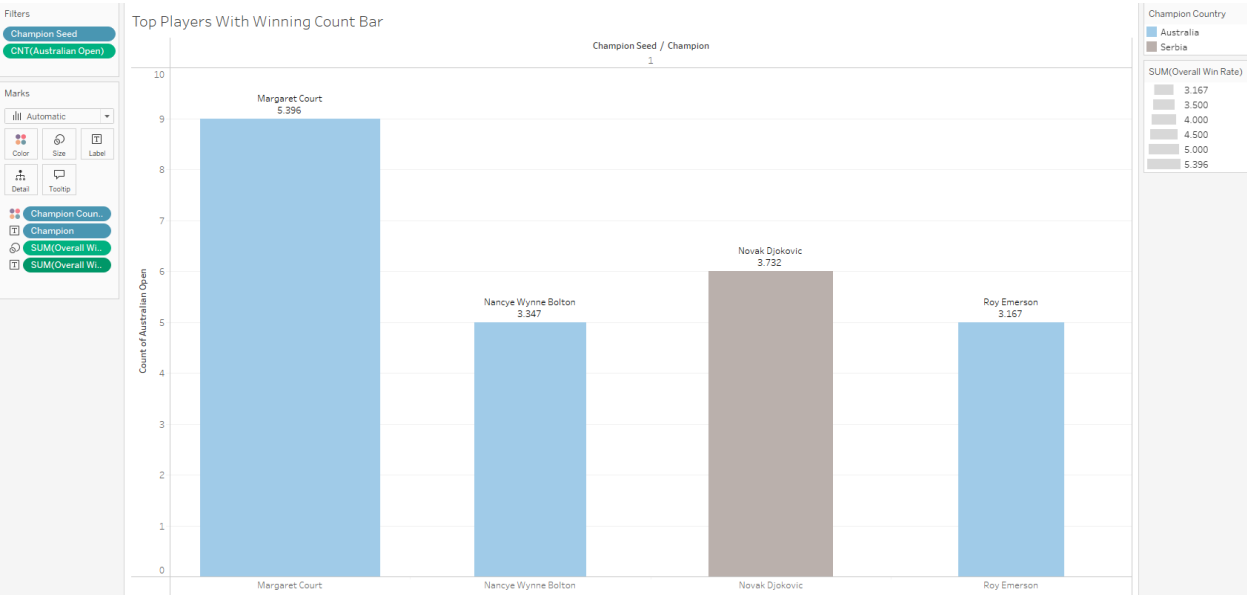
The figure above shows the total games lost in each set, colored by champions' countries and line size defining the player's winning times count. Although most of the lines are similar to the winnings graph counterpart mentioned above, some patterns are also found such as how Yugoslavia player had a steady decline of points lost throughout all his games.

Scatter chart for total 3rd~5th set win rate



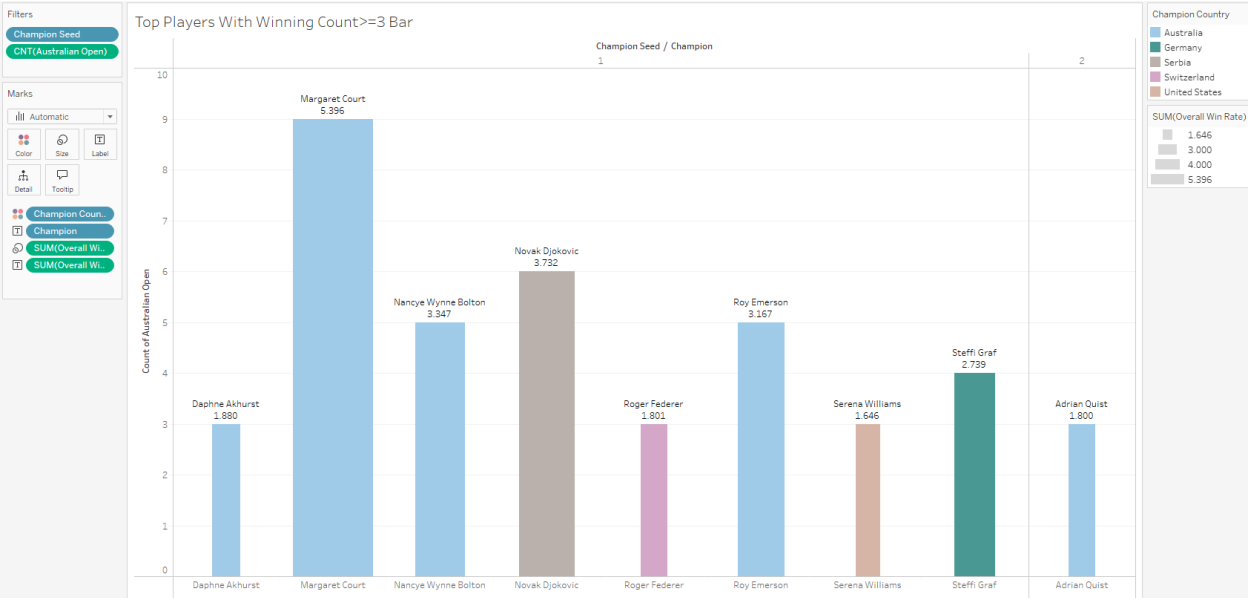
The figure above shows the total win rate for 3rd ~ 5th set, with countries as colour codes, total overall win rate for size and gender for shapes. Based on this figure, it is evident that only male players make it till 4th set since there are no datapoints for women. The Serbian champion Novak Djokovic has the highest win rate for all three metrics.

Bar chart for top players with winnings count

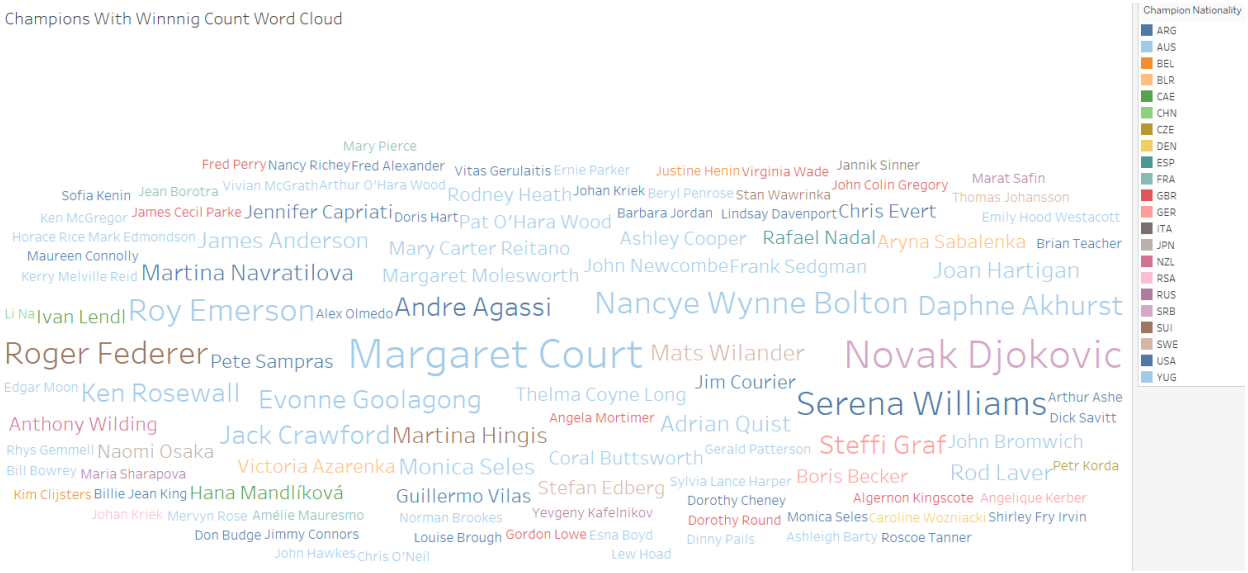


The figure above shows the top players (players with champion seed of 1) who have won five or more times, with total overall win rate as size. The four top players are Margaret Court with 9 times win, Novak with 6 times win, Nancye Wynne Bolton and Roy Emerson with 5 times win. Only Novak is from Serbia and the other three players are from Australia.

It was also noticed that there are no other players who have won 5 times with champion seed below 4. The graph was also viewed with champion seed < 3 and winning count >=3. It was found that 5 out of 9 champions are all Australian as shown below.



Word cloud for champion with winning count



The figure above describes the word cloud of champions, with their nationality as color code and size as winning count. The most noticable names are Margaret Court, Novak Djokovic, Roger Federer, Nancye Wynne Bolton and Serena Williams. Similar to the treemap, there are a lot of blue color text which are Australian players.

Word cloud for runner-ups with winning count

Similar to the champion counterpart, the figure above describes the word cloud for runner-ups, with their nationality as color code and size as winning count. Some of the most noticeable names are Esna Boyd, Chris Evert, Andy Murray, John Bromwich and Maria Sharapova. One noticeable fact is that there are a considerable number of runner-up players from the United States.

Conclusion

New columns were created for further calculation and different visualization techniques were used to effectively convey the patterns and stories. The timeline chart showed that there was a breakthrough point in 1977, and that women have overall win rate in most years despite men having more win counts throughout the years. Using several parallel coordinate graphs, it was proved that some countries could have a certain playstyle that keeps their games won or lost path similar. Also, the fact that there are only male champions from 4th round could either mean that female players do not have enough stamina, or they finished the match early. Treemaps were used to show the total overall win rate of each country by gender. This combined with the insights from symbol map showed that Australia was dominant in these tennis matches, followed by the US. This fact was further reinforced by the results from the bar chart that shows the top players, and the word clouds. The Australian players also took the top spots of matches throughout the years, which was proved by the bar charts. The two players – Margaret Court and Novak Djokovic were significant in most of the graphs visualized.

Each graph has its own advantages and disadvantages, which are mentioned below.

- Treemaps: These can effectively show proportions and relative sizes of each category while handling a large number of them. But labelling can become difficult with small categories and can be hard to understand if the heirarchy becomes complex.
- Symbol maps: These are great for visualizing geographical data and can be easily interpreted. But not many information types can be displayed and are not suited for comparing categories with similar values.
- Line charts: These are best for visualizing trends over time, and they can display continuous data changes well. But they are not good at comparing data points in a single time point and can have overlapping data.
- Parallel coordinates: These are great at showing multiple variable relationships and can effectively show patterns and outliers, but they can be difficult to understand when there are too many data points, and not good at showing trends over time.
- Scatter charts: These are good at displaying relationships between two variables on top of other variables and can display patterns. But they are limited to two variables only and can be overwhelming if there is too much data.
- Bar charts: These are the best at showing different categories at a single time point. But like other graphs, too much data can become overwhelming.
- Word clouds: These are great at displaying prominent terms or word frequency, but information that can be shown is limited and bad at analyzing the context.

To conclude, documenting the Australian Open dataset allowed the professional tennis community to gain insightful results which were obtained through higher-dimensional visualizations. The tennis players can observe these results to create new strategies for upcoming games while businesses could also leverage these data to improve fan engagements and training programs.