Assessment Task 2:

Advanced Data Visualisation

Chanudi Neja Abeywickrama
14464693

Executive Summary

The report offers an insightful visual analysis of the Australian Open Tennis Championship through 120 years on nationality, gender, change over time of champions as well as a performance analysis of the top 7 players. The report provides recommendations for tennis professionals on potential inefficiencies and obstacles faced by players by suggesting further investigation insights. Key findings of the report include insufficient representation of champions from high-performing nationalities, possible patterns of female players scoring higher winning ratios, potential correlation between winning ratio and the number of championships won, and inconsistencies suggesting a lack of player performance by the set. Overall, the visualisations help bring out insights and patterns in the data that could potentially provide tips into improving tournament conditions for future Australian Open Tennis Championship participants.

Table of Contents

1.0 Introduction

The report presents a comprehensive, in-depth analysis of past Australian Open Tennis Championships through 120 years to identify trends and patterns and gain insights into the data. It commences with a brief exploration of the dataset, followed by a few modifications to visualise the data. The visual analysis covers information on nationalities, gender and changes over time of all champions, as well as a focused analysis of the top seven players to gain insights on their win ratio performance in the final matches. The analysis will conclude with recommendations for tennis professionals on the potential investigation factors that would lead to improved performance in future championships by overcoming current inefficiencies and difficulties faced by players.

2.0 Data Exploration

The dataset used for the visual analysis in this report contains data from the Australian Open Tennis Championships through 120 years, from 1905 to 2024. It describes men's and women's championship tournaments yearly, including the champions and the runners-up, their nationalities and countries, and the tournament results at each set.

Understanding the dataset is a crucial part of the visualisation process. In the context of this dataset, it is essential to understand how tournaments take place and how the winners are declared. A men's match consists of five sets, whereas a women's match consists of only three sets, each with a minimum of six games. The player who wins six games first, with a margin of at least two games, wins a set. The winner of the match is then determined by the best of five of a men's match and the best of three of a women's match.

The original dataset consists of 19 distinct attributes, as described in Table 1 below.

Data Attribute	Data Type	Description of Attribute
Year	Text (Categorical ordinal)	The year the championship was held
Gender	Text (Categorical nominal)	Classification of the championship by gender
Champion	Text (Categorical nominal)	Name of the champion or winner of the championship
Champion Nationality	Text (Categorical nominal)	Nationality represented by the champion
Champion Country	Text (Categorical nominal)	Country represented by the champion
Score	Text (Categorical nominal)	Score of the match by sets, as described earlier (for example, 6-4 is a set where the champion has won the set by two games). For each row of data where the gender attribute is "men's", there are five sets, and for "women's", there are three sets.
1st-won	Numeric (Quantitative ratio)	Number of games won by the champion in the first set
1st-loss	Numeric (Quantitative ratio)	Number of games lost by the champion in the first set
2nd-won	Numeric (Quantitative ratio)	Number of games won by the champion in the second set
2nd-loss	Numeric (Quantitative ratio)	Number of games lost by the champion in the second set
3rd-won	Numeric (Quantitative ratio)	Number of games won by the champion in the third set
3rd-loss	Numeric (Quantitative ratio)	Number of games lost by the champion in the third set

4th-won	Numeric (Quantitative ratio)	Number of games won by the champion in the fourth set
4th-loss	Numeric (Quantitative ratio)	Number of games lost by the champion in the fourth set
5th-won	Numeric (Quantitative ratio)	Number of games won by the champion in the fifth set
5th-loss	Numeric (Quantitative ratio)	Number of games lost by the champion in the fifth set
Runner-up	Text (Categorical nominal)	Name of the runner-up (second place) of the championship who played against the champion in the final round.
Runner-up Nationality	Text (Categorical nominal)	Nationality represented by the champion
Runner-up Country	Text (Categorical nominal)	Country represented by the champion
WinRate (Rate) (Top5+ sheet)	Percentage (Quantitative ratio)	The percentage of the overall number of games won by the champion (calculated by no of games won / (no of games won + no of games lost))
Wins (Top5+ sheet)	Numeric (Quantitative ratio)	The overall number of games won by the champion throughout all sets
Loss (Top5+ sheet)	Numeric (Quantitative ratio)	The overall number of games lost by the champion throughout all sets

Table 1: Original Dataset Attributes

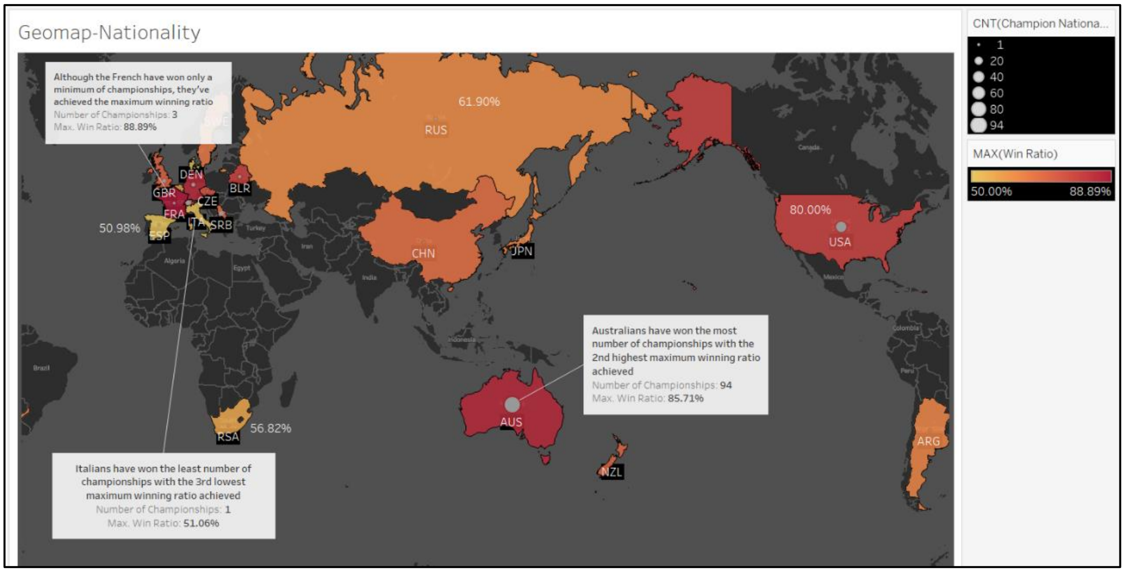
As part of the data transformation process, eight new calculated attributes were added to the original dataset to obtain winning ratio data for each set and the overall game, which are performance indicators of the champions. The calculations performed for each of these attributes are shown below in Table 2.

New Data Attribute	Calculation
Games Won	Sum of 1st-won, 2nd-won, 3rd-won, 4th-won and 5th-won
Games Lost	Sum of 1st-loss, 2nd-loss, 3rd-loss, 4th-loss, and 5th-loss
Win Ratio (percentage)	Games Won / (Games Won + Games Lost)
1st Set Win Ratio (percentage)	1st-won / (1st-won + 1st-loss)
2nd Set Win Ratio (percentage)	2nd-won / (2nd-won + 2nd-loss)
3rd Set Win Ratio (percentage)	3rd-won / (3rd-won + 3rd-loss)
4th Set Win Ratio (percentage)	4th-won / (4th-won + 4th-loss)
5th Set Win Ratio (percentage)	5th-won / (5th-won + 5th-loss)

Table 2: New Data Attributes and Calculations

3.0 Data Visualisation and Analysis

3.1 Geographic Map Analysis of Nationalities


Geomap-Nationality

Although the French have won only a minimum of championships, they've achieved the maximum winning ratio
Number of Championships: 3
Max. Win Ratio: 88.89%

Australians have won the most number of championships with the 2nd highest maximum winning ratio achieved
Number of Championships: 94
Max. Win Ratio: 85.71%

Italians have won the least number of championships with the 3rd lowest maximum winning ratio achieved
Number of Championships: 1
Max. Win Ratio: 51.06%

Figure 1: Geographic Map Analysis of Nationalities

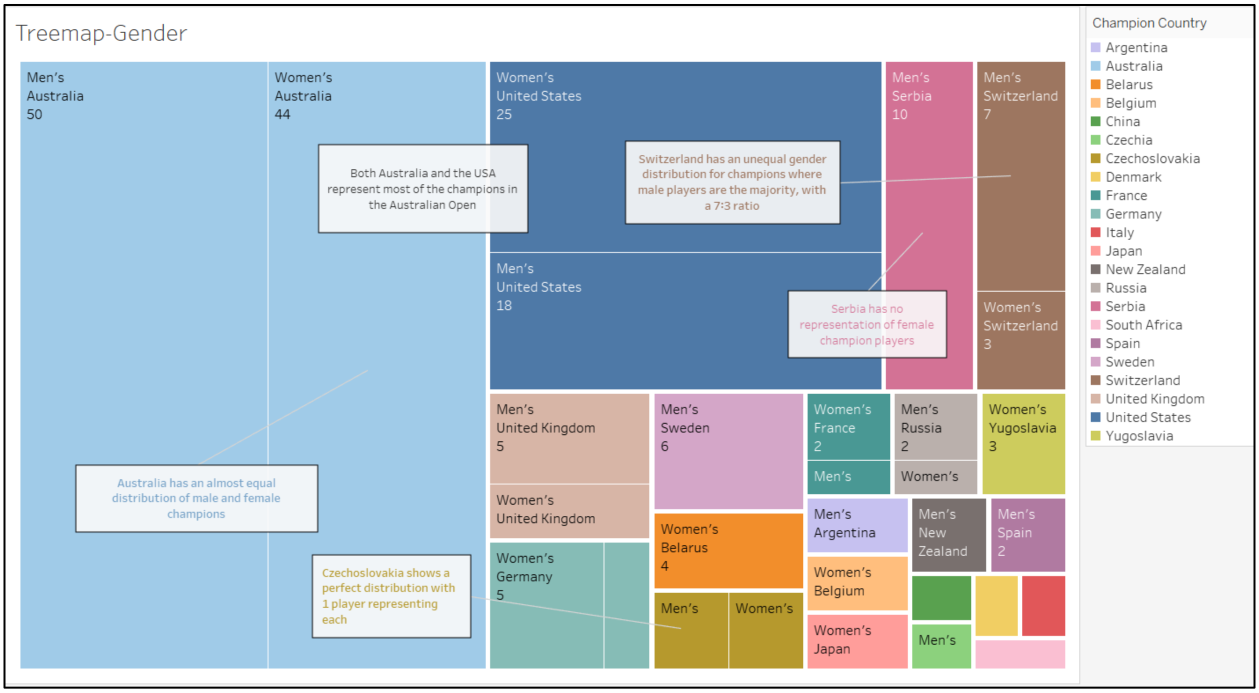
Tableau uses geo maps to visualise spatial and geographic data using country, state, or world maps, catering to the needs of the data. This visualisation can be done in two ways: a choropleth map, which colours specific regions of the map, or using symbols over particular locations. The geomap in Figure 1 utilises both these types of visualisations, layering the two maps together using dual axes for a comprehensive look at the nationalities represented at the Australian Open Tennis Championships. The choropleth map depicts the maximum winning ratio achieved by each nationality using a red-gold colour palette, where red represents higher maximum winning ratios. The circle symbols represent another attribute: the number of champions represented by each nationality, where the size of the circles increases with the number of champions. No additional dimensions were added to the symbols, such as colours and shapes, as it may overwhelm the reader with an overcrowded map attempting to communicate too much information simultaneously. The legend at the top right corner ensures that the correct information is understood.

At a glance, Australians seem to be the leading nationality, with the highest number of champions and a high maximum winning ratio recorded, with the USA leading behind them, as indicated by the large circles and the bright red shadings of the areas. However, the annotations highlight that it was the French who achieved the highest maximum winning ratio of 88.89%, although there is only a minimum number of champions representing the country. The French still lead Itay, which has the lowest number of champions represented by the championship nationalities, which is only one. Italy also has the 3rd lowest maximum winning ratio of 51.06%, indicating a low overall performance among other champion nationalities. While the shapes and colours in the maps visualised key information from the data, annotations proved helpful in quantifying and expressing this information for a more accurate analysis.

Advantages of Geographic maps:

- Visual representation of geographic data on a map allows easy understandability of country/nationality data at a glance with enhanced communication.
- The use of dual-axis maps allowed the layering of the two different types of geomaps to increase the number of dimensions shown in the visualisation and provide additional context.
- Further analysis is facilitated through interactive exploration of data by zooming in or selecting to highlight a particular region of interest.

3.2 Treemap Analysis of Gender


Treemap-Gender

Both Australia and the USA represent most of the champions in the Australian Open

Switzerland has an unequal gender distribution for champions where male players are the majority, with a 7:3 ratio

Serbia has no representation of female champion players

Australia has an almost equal distribution of male and female champions

Czechoslovakia shows a perfect distribution with 1 player representing each

Figure 2: Treemap Analysis of Gender

A treemap uses a nested rectangular structure to represent hierarchical data, using different colours and sizes and visualising data by category and values for a simple and comprehensive analysis of complex data. Figure 2 attempts to visualise the distribution of champions by gender in each country. Here, the main category is the country, represented by colour. Each country is then broken into two sub-categories (where possible), each rectangle representing the number of male and female players. The treemap was utilised in a way that represented countries in a hierarchical manner, where the countries occupied more space with more players. The size of the rectangles within each colour allows one to analyse how players are distributed by gender within a country for a more detailed analysis rather than looking at the overall distribution.

Both Australia and the United States occupy more than half of the treemap, indicating that the countries are leading in the championship. At a glance, Australia, the United States and Czechoslovakia show an almost equal distribution of champions by gender. However, this wouldn't have been identified by looking at the number of players for each gender, as the gender gaps vary broadly for each country. The visualisation communicates this by proportions, allowing for a more nuanced analysis of gender distribution by country.

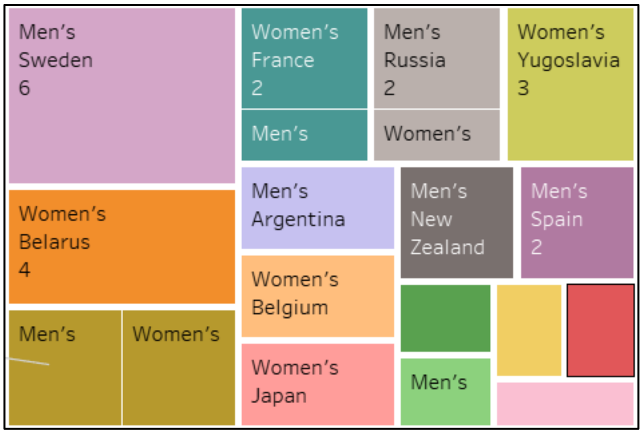


Figure 3: Treemap Analysis by Gender (Zoomed in on a specific area)

It was identified that as the number of players representing a country grew less, the gender distribution became more unequal. Figure 3 captures the countries at the lower end of the hierarchical structure, where most only represent one gender. Serbia, however, sets itself apart from this pattern, with only male players representing the country, even though the country is the third in rank for the most number of champions at the Australian Open, as highlighted in Figure 2.

Advantages of Treemaps:

- The hierarchical structure provides an instant overview of complex data by efficiently utilising space to represent large amounts of data and categories.
- The use of colours and sizes to represent categories and subcategories emphasises their relationships, adding an extra layer of information to the treemap.
- Proportion-wise comparisons of these categories help reveal interesting patterns and possible outliers at a glance.

3.3 Parallel Coordinates Analysis of Change over Time

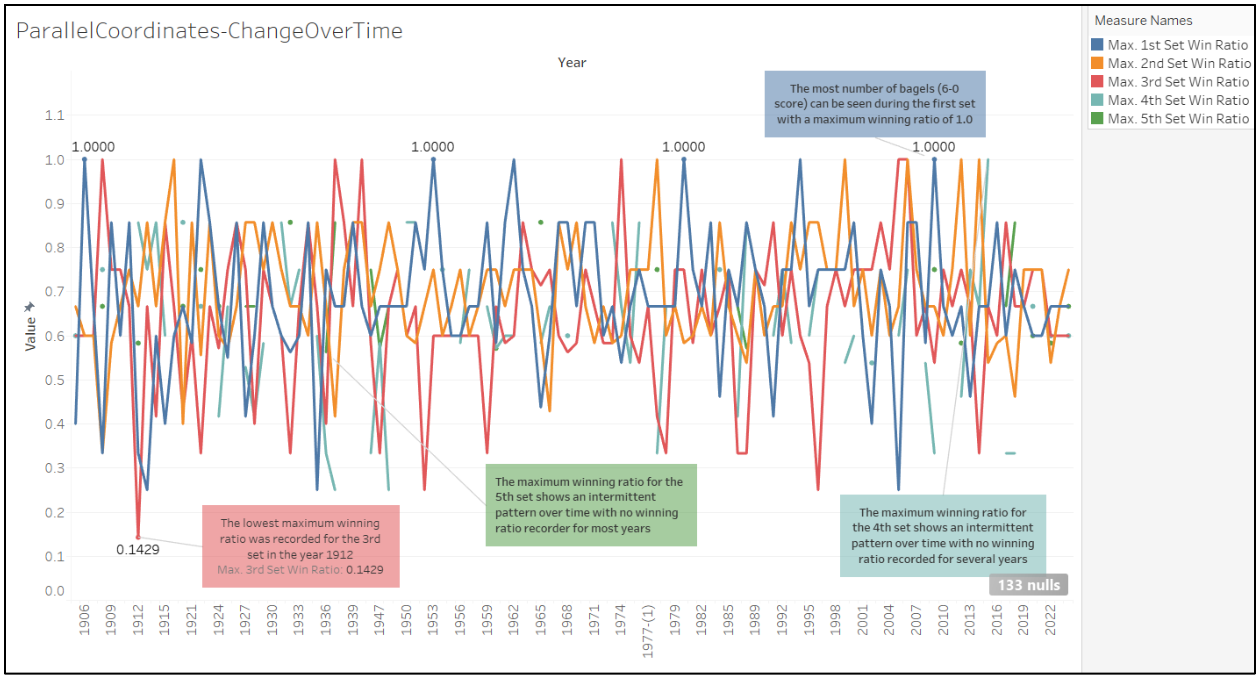


Figure 4: Parallel Coordinates Analysis of Change over Time

Parallel coordinate plots are commonly used to visualise multivariate data in a multidimensional space to analyse correlations, patterns, or trends over time. Figure 4 uses parallel coordinates to visualise the change in the maximum winning ratios achieved at each of the five sets over 23 years from 1906 to 2022. The winning ratios were aggregated by getting the maximum winning ratio achieved each year. For a more nuanced analysis, the winning ratios were categorised by the set and represented by colour to compare the fluctuations between sets as well. This categorisation also allows the viewer to analyse the trend of the maximum winning ratio for each set by filtering out the rest, as depicted in Figures 5 and 6. In comparison to other visualisations, the parallel coordinates plot is not easily readable with the multidimensional data. Therefore, annotations proved to be quite useful in highlighting any significant patterns or trends for such a graph supported by relevant data labels.

As pointed out by the red annotation, the lowest maximum winning ratio recorded over the years was 0.1429 in the year 1912. While the first and third maximum winning ratios have come close to this, the lowest maximum winning ratios don't show any consistency over the years. In contrast, the highest maximum winning ratios have reached a perfect 1.0000 several times. Analysing by colour, it is evident that the greatest number of 6-0 scores, or "bagels" in tennis slang, have been recorded in the first set.

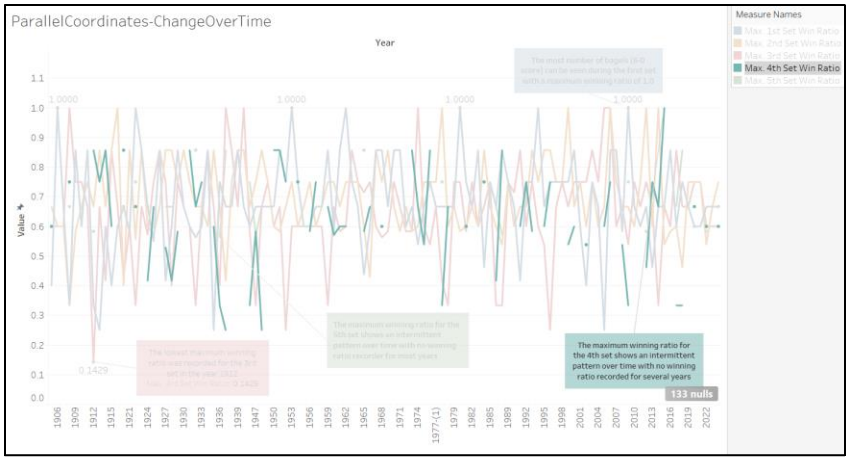


Figure 5: Parallel Coordinates Analysis – 4th Set

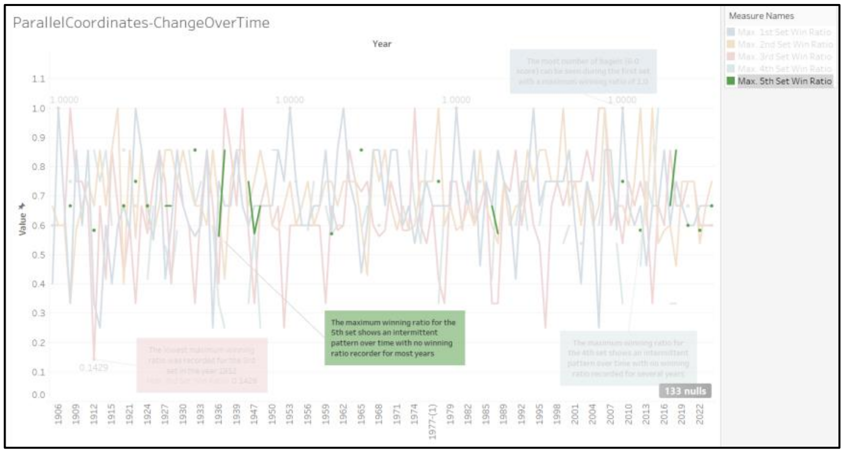


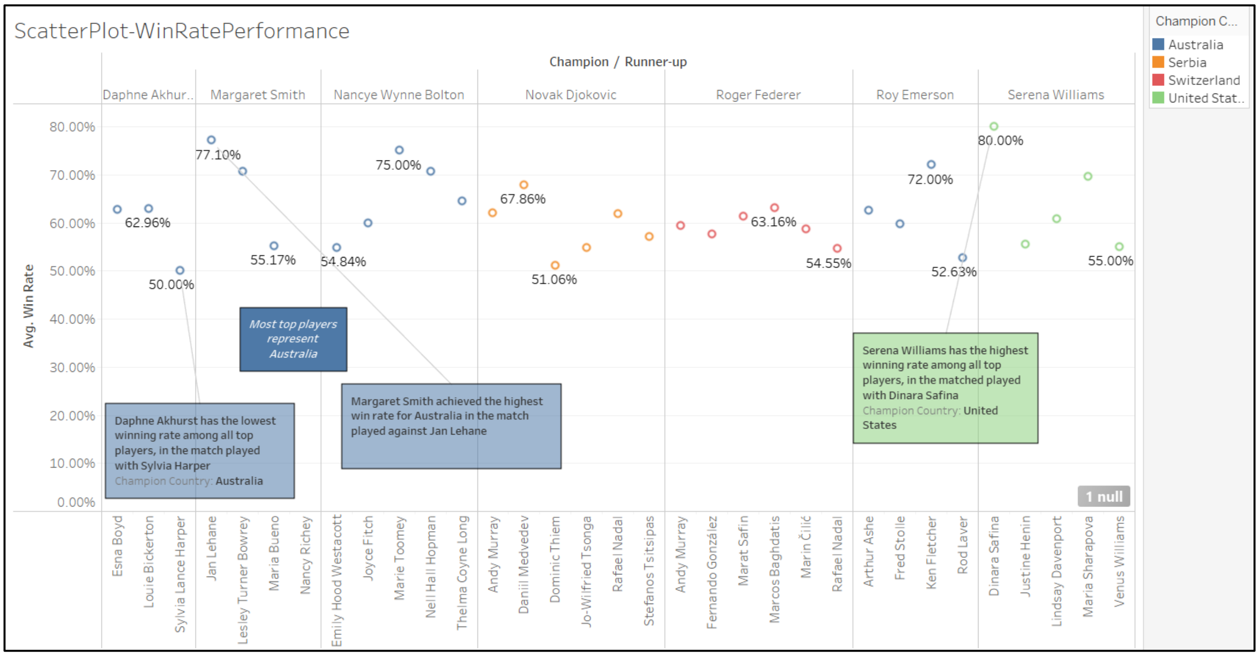
Figure 6: Parallel Coordinates Analysis – 5th Set

Looking at Figures 5 and 6, by filtering out the rest except the selected set, an intermittent pattern seems to be present for sets four and six, which wasn't noticed with the other sets. This can be explained by the fact that in the Australian Open Tennis Championship, men play five sets while women only play 3 sets. The years where no winning ratio has been recorded may suggest years in which female players have played.

Advantages of Parallel Coordinates:

- Ability to analyse data with multiple variables in a single visualisation for a comprehensive look at the data.
- Easy identification of patterns and trends with comparative analysis using tools like colour coding, highlighting, etc.
- Dynamic filtering of parallel coordinates lets a viewer focus on specific categories of data for in-depth analysis.

3.4 Scatter Plot Analysis of Win Rate Performance of Top Five Players



Most top players represent Australia

Daphne Akhurst has the lowest winning rate among all top players, in the match played with Sylvia Harper
Champion Country: Australia

Margaret Smith achieved the highest win rate for Australia in the match played against Jan Lehane

Serena Williams has the highest winning rate among all top players, in the matched played with Dinara Safina
Champion Country: United States

Figure 7: Scatter Plot Analysis of Win Rate Performance of Top Five Players

While scatter plots are generally used to plot numerical data across the x and y axes, Tableau facilitates the option to pair numerical data with categorical values as well. Figure 7 makes use of this feature to visualise the winning ratios of the top 7 players of the Australian Open Tennis Championship. For a more in-depth analysis, the winning ratios of the top players were displayed along with the runners-up by using a primary and secondary x-axis, with two attributes in the "columns" section. Further, the countries of the champions were categorised by colour to give more information about the top players.

The winning ratios for the top players in their matches against the runners-up are scattered only between 50% and 80%. Through colour analysis, at a glance, it is visible that Australia dominates the top 5+ players' board. However, the lowest winning ratio of 50% among all top players was recorded by Daphne Akhurst, representing Australia, in her match against Sylvia Harper. On the other hand, the maximum winning ratio among all top players, of 80%, was recorded by Serena Williams, representing the United States, in her match against Dinara Safina. Despite not scoring a high winning ratio among the top players, the Serbian and Switzerland champions have won the greatest number of championships among all players.

Advantages of Scatter Plots:

- Advanced features of plotting numeric data with categorical data in Tableau allow a more in-depth analysis of numeric data by category.
- Ability to visualise multidimensional data by colour, shape and size in a simple two-dimensional plot to communicate more information in one visualisation.
- Easy recognition of outliers by highlighting patterns representing relationships between variables across data points

4.0 Recommendations

In the analysis of champion nationalities in the Australian Open using Geomaps, it was identified that France, Denmark and Belarus have high maximum winning ratios, although the number of players from these countries is very low. This suggests that champions from these countries have much potential, although there's a lack of representation in the Championship. Further investigation into why this may be so should be conducted to identify the potential lack of training programs and coaches, and cultural restrictions to help rising tennis champions overcome obstacles.

Investigating the previously mentioned countries, France, Denmark and Belarus, with the gender distribution analysis using a tree map, it was identified that the majority (or only player) of players from these countries are women. This may indicate a pattern where female players tend to win by a higher winning ratio than male players. This could be an insight that can be further analysed by combining the data from the two plots to get a more nuanced understanding of the existence of such a pattern.

In the scatter plot analysis of the top players in the championship, it was identified that the Serbian and Swiss champions didn't achieve winning ratios above 70%, although they won the greatest number of championships in the tournament. This is an interesting insight found by comparing the top players in the tournament, which could benefit from further analysis of the relationship between the winning ratio and the number of championships won.

The parallel coordinates visualisation revealed an inconsistency in the pattern over time for the fourth and fifth sets that wasn't seen in other sets. While the fact that female players only play in 3 sets and male players only play in 5 sets may influence this inconsistency, it's simply an assumption based on very limited insight. This is something that should be analysed further, as this may suggest also be because of a lack of performance of players in the last sets which should be addressed if so.

5.0 Conclusion

The report commenced with a quick data exploration of the Australian Open Tennis Championship data of 120 years, followed by an in-depth visual analysis using Tableau. The visual analysis revealed key insights into the performance trends of champions by nationality, gender distribution inequalities of players by country, performance across sets in a match over the years and win ratio performance of top players. The interesting patterns revealed by these analyses have proved to be insightful in conducting deeper investigations on particular relationships that may help address potential inefficiencies or obstacles faced by players. The report ultimately attempts to suggest potential areas of improvement for future Australian Open Tennis Championships by analysing the past and current status of the championship.