# Assessment Task 2: Advanced Data Visualisation

Hasim Rahman

Student ID :-24674715

Masters in Business Analytics (Ext)

Australian Tennis Championships from 1905 - 2023

# Summary of the Dataset

This report focuses on exploring key trends of the Australian open dataset provided. The Australian Open is an annual tennis tournament which is held at the Melbourne Park in Melbourne, Victoria, Australia. This competition is one of the four most prestigious competitions in tennis. The Australian open dataset contains data from 1905 to 2023 containing 208 matches in total, the original dataset contains 19 columns with data about the champions and the runner ups and details about the finals.

After using the count of titles on Tableau we could conclude that there are 7 players in the history of the competition to win more than 5 championships.

- 4 women - competition started in 1922
- 3 Men - competition started in 1905

## Data Exploration

### Key Conclusions from the analysis of the Australian open dataset

Through analysis and interpreting the datasets, these highlights/ patterns or trends can be noticed

1. Novak Djokovic has won the championship 10 times which is the highest number of titles won by a single player, he has never been a runner up so we can see he has won 10/10 of the finals he has played. He also has the highest combined win rate of any player in the history of the competition.
2. Serena Williams and Margaret Smith  have shown dominance in the female game, winning the championship 7 times each in completely different eras.

3. Australian Esna Boyd has been the person who has lost the competition most times in the finals by coming short 6 times.
4. Australia has had a dominant show in the competition's history, although most of their victories came in the earlier years of the competition.

Next we can see all the Variables which were given in the original dataset.

| Number | Name of the variable | Type | Description |
|---|---|---|---|
| 1 | Year | YYYY Format, Quantitative (interval) | Year the game was played |
| 2 | Gender | Binary, Category Nominal | The gender of the competition |
| 3 | Champion | String format, Nominal Category | Name of the Champion |
| 4 | Champion Nationality | 3 String format, Nominal Category | A code for the nation used for the winner |
| 5 | Champion Country | String format, Nominal Category | Geographical data to use for the nation for the winner |
| 6 | Score | General, Pair of integers | Scores of finals played between the champion and the runner up |
| 7 | 1st-won | Quantitative Scoring between 1-6 | Scores of the winner in the first set |
| 8 | 1st-loss | Quantitative Scoring between 1-6 | Scores of the loser in the first set |
| 9 | 2nd-won | Quantitative Scoring between 1-6 | Scores of the winner in the second set |
| 10 | 2nd-loss | Quantitative Scoring between 1-6 | Scores of the loser in the second set |
| 11 | 3rd-won | Quantitative Scoring between 1-6 | Scores of the winner in the third set |
| 12 | 3rd-loss | Quantitative Scoring between 1-6 | Scores of the loser in the third set |
| 13 | 4th-won | Quantitative Scoring between 1-6 | Scores of the winner in the fourth set |
| 14 | 4th-loss | Quantitative Scoring between 1-6 | Scores of the loser in the fourth set |
| 15 | 5th-won | Quantitative Scoring between 1-6 | Scores of the loser in the fifth set |
| 16 | 5th-loss | Quantitative Scoring between 1-6 | Scores of the winner in the fifth set |
| 17 | Runner-up | String format, Nominal Category | Name of the Runner up |
| 18 | Runner-up Nationality | 3 String format, Nominal Category | A code for the nation used for the runner up |
| 19 | Runner-up Country | String format, Nominal Category | Geographical data to use for the nation for the runner up |

## Interesting Findings

- Women's Competition started in 1922 while the men's competition started in 1905.
- There were multiple tournaments in the year 1977 as we have multiple divisions for that year.
- During the WW2 periods, no Australian open tournaments were played.
- There are three blank columns for Champion Seed, Mins, and Runner-up Seed. Those three values hadn't been appropriately recorded before 1945, so we removed them from the dataset.

## Data Format

Using the already given data we can create more variables to help us in further visualisations. As we are looking to work with higher dimensional data for our visualisation as we need 4+ dimensions for all of our data. Under this we will discuss the new variables and dimensions which we created for the visualisations.

1. Winner and loser points: the sum of all the points scored by the winners and the runner ups.
2. Winrate%: winner points divided by the sum of winner and loser points then its format changed to percentage.
3. 1st to 5th win rate: the points of the winner for the specific set divided by the sum of the points of the winner and the loser of that set.
4. Normalised win rates for 1st to 5th win rate: using a formula and calculated fields on tableau we figured out the win rates in a normalised format so we can use them for parallel coordinate charts.
5. Number of games: using and if statements we can count the number of games which were played in every match.
6. Eras: the time periods were grouped in order to do time series analysis and compare how the game changed over the years.
7. Continents and runner up continents: the countries were further divided into their continents to compare their individual successes.

# Tree Map

A treemap displays structured hierarchical data as a set of nested rectangles with varying sizes and colours. A tree map is a rectangle with more rectangles inside of ot that are quantitatively proportional to the whole tree map, the largest rectangle is in the top left corner and the smallest in the bottom right corner. Colour coding can also be a useful tool to further differentiate between its rectangles by either using intensity of the colour or just different shapes.

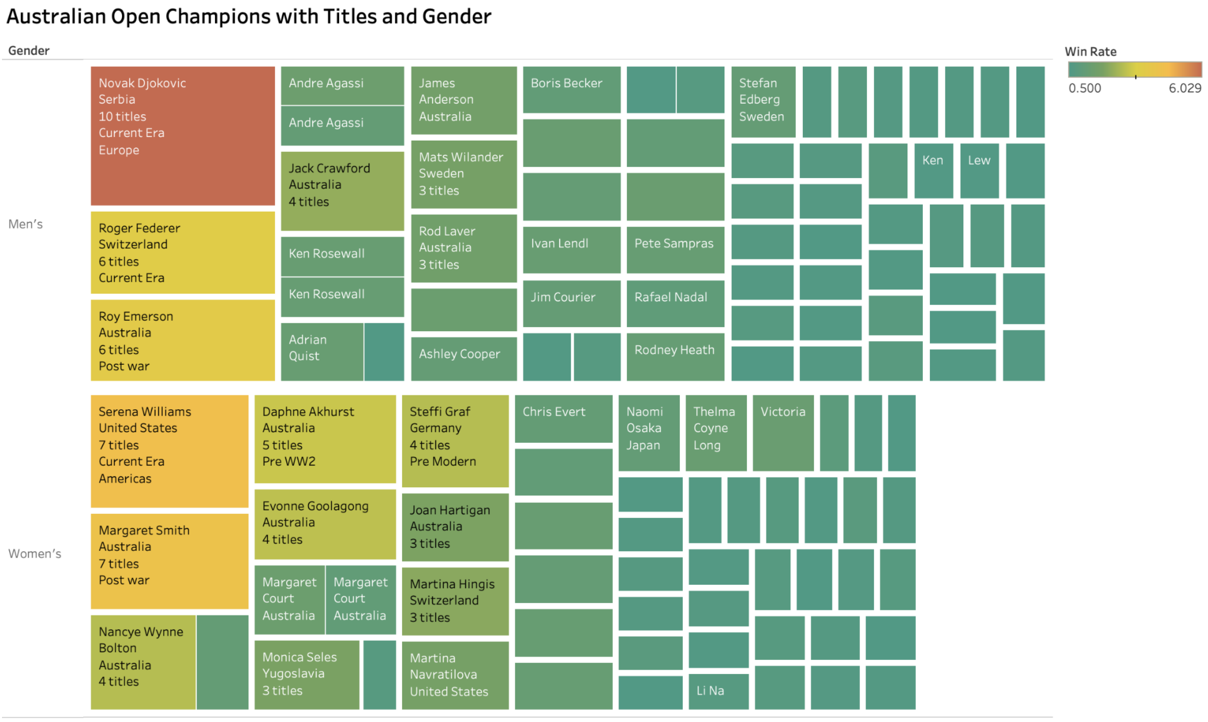


*Figure 1*

*Figure - 1*:- In this visualisation, we are using highest win rate to have the highest intensity in the rectangles to show the most dominant players in their finals, the players who have players would have a higher win rate as their sum is used.

The size of the rectangles show the number of titles a person has won, and also other details about the players are also given. We have 2 levels of hierarchy and 2 quantitative measures.

*Figure - 2* :-In this Visualisation, the colours show the continents and the sizes show the number of times these players have played the finals and became runner up, through this graph we can see the players continents division. Here we have an additional level of hierarchy which is the continent.

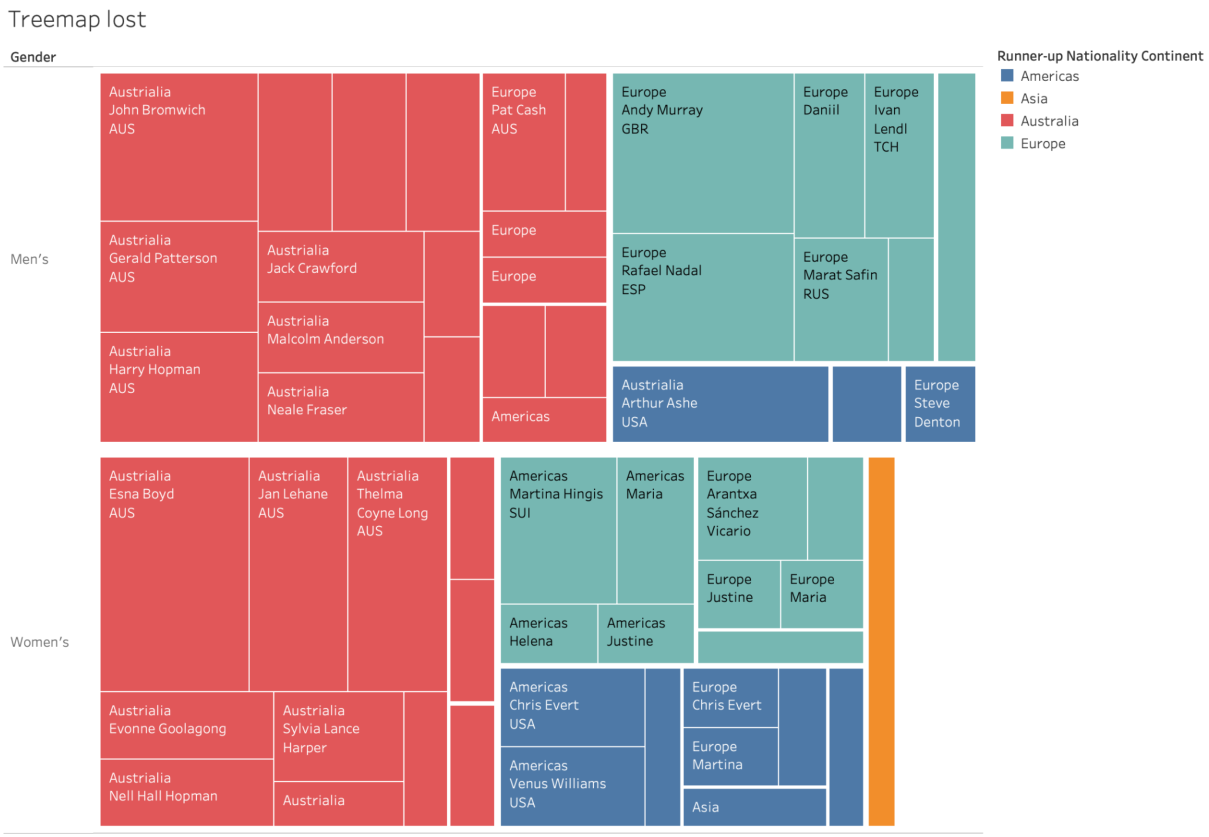


*Figure 2*

**Advantages of using Treemaps**

1. Hierarchical Representation: Treemaps excel at showing hierarchical data structures, making it easy to visualise the relationships between broader and narrower categories or groups.
2. Comparative Analysis: Treemaps allow for easy comparison of the sizes of different categories within a hierarchy.
3. Colour Coding: Treemaps can use colour coding to convey additional information.
4. Interactivity: Interactive treemaps can provide detailed information upon hovering or clicking on specific rectangles, allowing users to explore data in depth.

**Disadvantages of using Treemaps**

1. Labelling Challenges: It's hard to label smaller rectangles without cluttering the text.
2. Limited Usage: It might not be suitable for some data types and is very limited to hierarchical data.It might be unusable for time series or temporal data.

# Parallel Coordinate Maps

Parallel coordinate is used for multidimensional data using a set of polylines linked between axes at appropriate values between different dimensions.The data between axes shows the data relationship and the aggregation pattern.

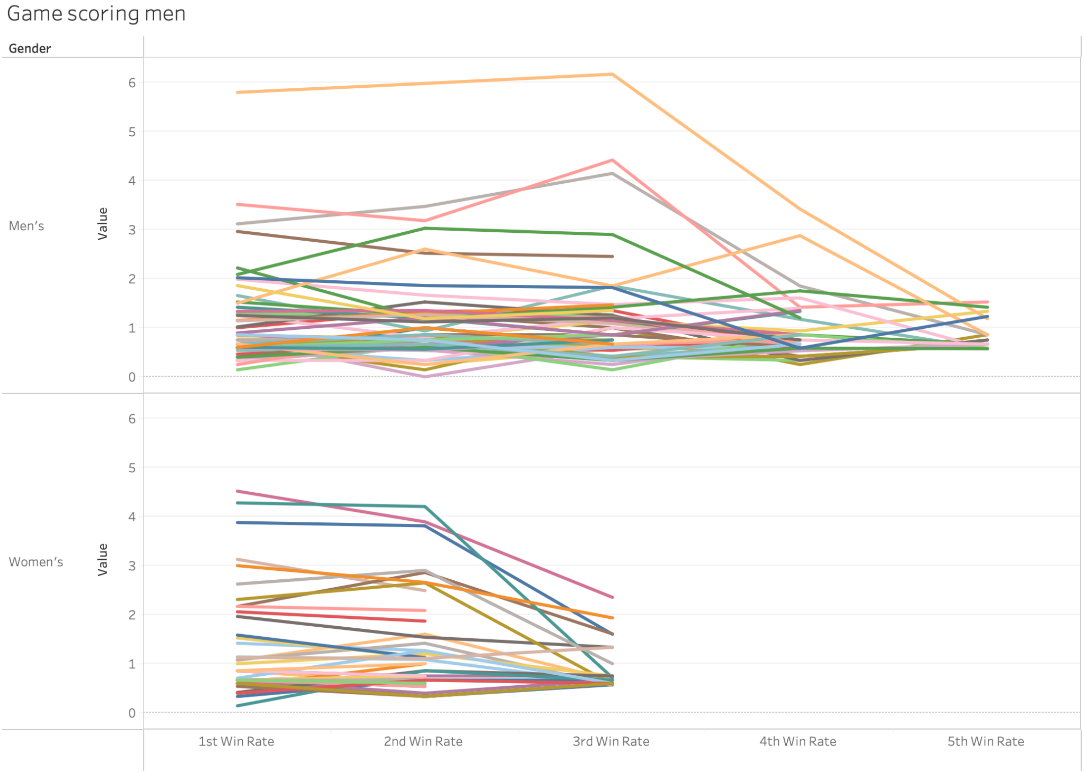


*Figure 3*

In this figure we are using non normalised data, we are comparing the win rates for every player in all the sets to compare the trajectory of the map throughout their game. Using gender on the row we can get separate divisions for male and female players as they play with different rules, this is a 5D and 3D parallel coordinates respectively.

**<u>Normalising Data</u>**

Data normalisation organises and transforms data to eliminate redundancies and improve data integrity within a single dataset or database.Normalized data is easier to sort, filter, and analyse, leading to better data exploration. With less number of columns and improved organisation, users can enhance visualisation, understanding and pattern recognition.In this dataset we normalised all the sets' win rates data to create a more comprehensible parallel coordinate map regarding the players performance in every set.

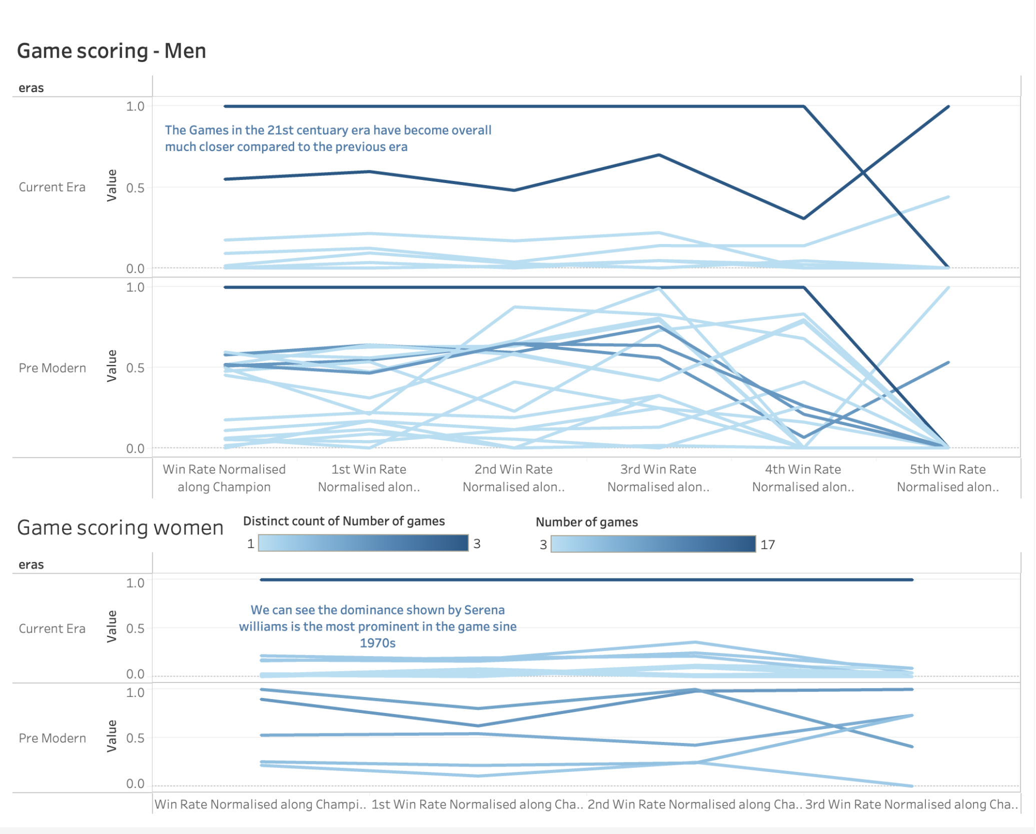


*Figure 4*

In this figure we use a 6D axes graph all the while comparing the win rates for every set and the over all winrate, we have divided it into 2 separate graphs for men and women as they have separate rules and women need to pay less sets, furthermore to make our chart more understandable we have tried to compare the 2 latest eras that were created in order to gauge more information from the chart.

## **Advantages of using Parallel Coordinate Maps**

1. Multivariate Analysis: Parallel coordinate charts are most prominently used for visualising and analysing datasets with multiple variables.
2. Pattern Recognition: PCPs make it easier to identify patterns, trends, and correlations within the data.
3. Highly Interactive: They can be used very well for data exploration.

## **Disadvantages of using Parallel Coordinate Maps**

1. Very tough to read for a person without prior context and hard to infer information.
2. It is very limited to numerical data, they are not as useful with categorical or textual data.
3. Limited for Time-Series Data: While PCPs can display time-series data, they may not be the most intuitive choice for visualising temporal trends over time.

## Geographic Maps

Geographic maps are visual representations of the Earth's surface, using cartographic projections to display countries of the world. They include elements like scales and legends. Topographic features, cities, and transportation networks are commonly featured, with interactive maps for enhanced accessibility. Maps may use GIS technology for data integration. Geographic maps serve various purposes, including navigation, urban planning, environmental analysis, and research, with each type of map tailored to specific informational needs.

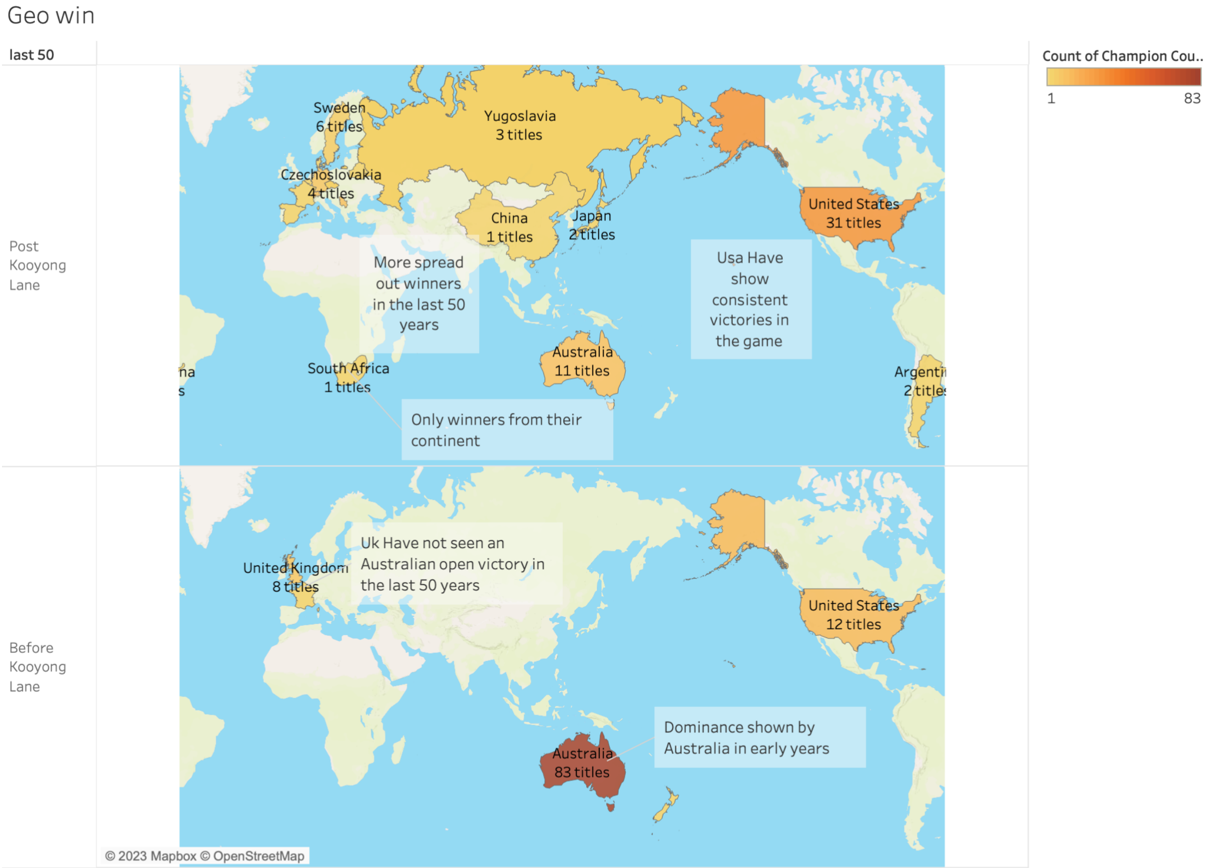
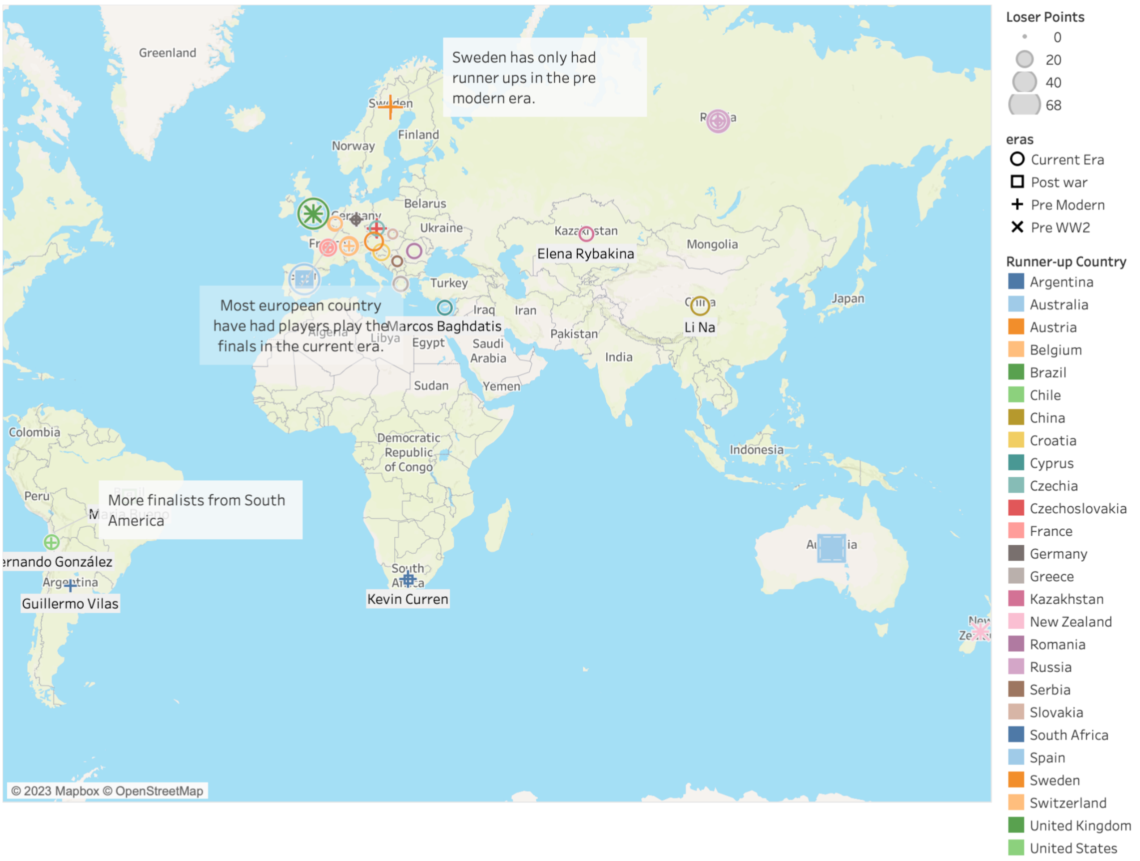


*Figure 5*

This figure illustrates the winners from 2 different eras - before and after the Australian open was played in Kooyong lane, to show how different the winners distribution has been. This map shows which country has won the title and the shading donates the amount of times a country has won.  The map denotes the countries geographically.

Geo lose



Sweden has only had runner ups in the pre modern era.

Most european country have had players play the finals in the current era.

More finalists from South America

*Figure 6*

This figure illustrates the runner ups for every tournament and the colours donate the countries, the shapes denote the eras the matches were played in and the size of the shapes denotes the number of loser points scored by a player.

## Advantages of geographical maps

1. Visual Clarity: Geographic maps offer a clear visual representation of spatial data, aiding understanding and communication.
2. Data Analysis: Maps facilitate the analysis of spatial data, revealing patterns and relationships.
3. Spatial Insight: Maps help illustrate and comprehend the relationships between geographic features and regions.

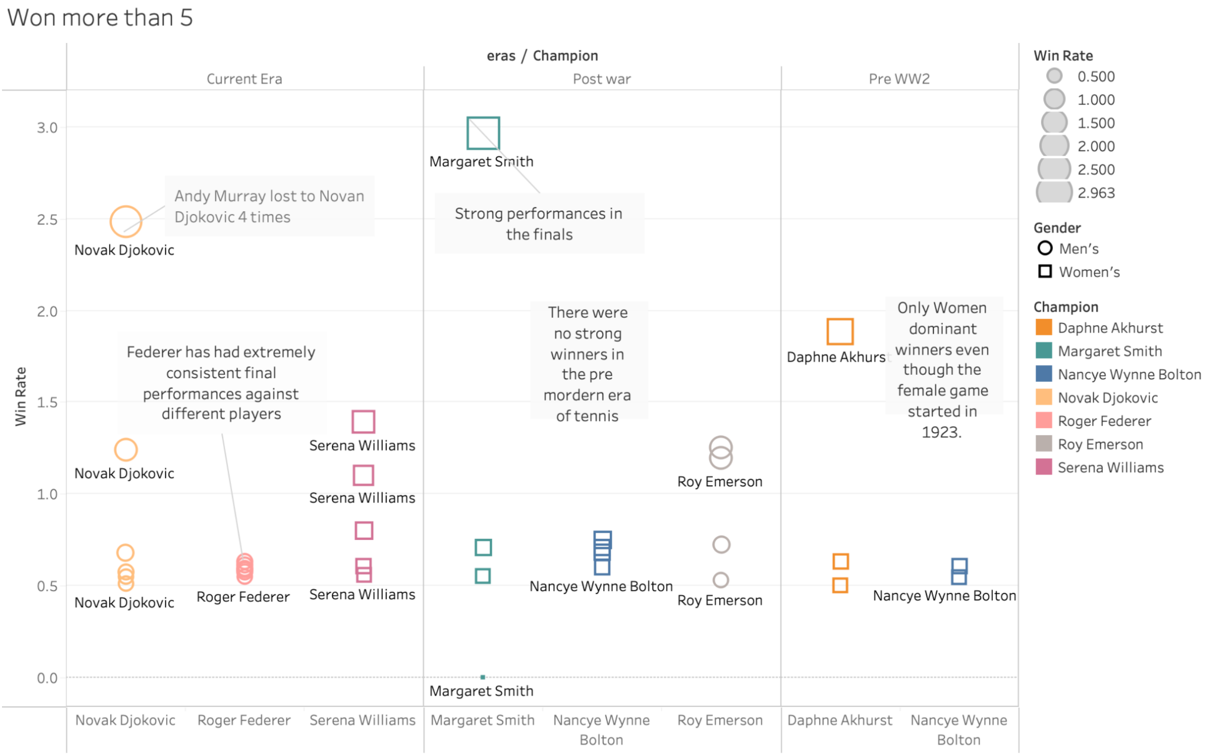
## Disadvantages of geographical maps

1. Visual Clutter: Overloading maps with excessive detail can lead to visual clutter, obscuring essential information.
2. Geographic Literacy: Effective map interpretation often necessitates a basic level of geographic literacy.

# Top Players performance

## Scatter Chart

7 winners have won the australian open more or equal to 5 times, this chart analyses the winners performances. These are the top players of the competition so we need a closer look into them.


Won more than 5

Andy Murray lost to Novan Djokovic 4 times

Strong performances in the finals

Federer has had extremely consistent final performances against different players

There were no strong winners in the pre mordern era of tennis

Only Women dominant winners even though the female game started in 1923.

*Figure 7*

This chart illustrates the top winners of the game and the colours donate the winner, the shape shows the gender of the champion and the winrate shows the size of the shape against their particular opponent. With this chart we are able to compare the win rates of all the top players and how they performed in their specific bouts against their respective opponents in their games.

We are using this scatter graph in order to get out our top contenders and have them compete with each other and analyse patterns and trends from the visualisation.

## Lollipop Chart

To further illustrate the top performers we will use another chart for the top performers.

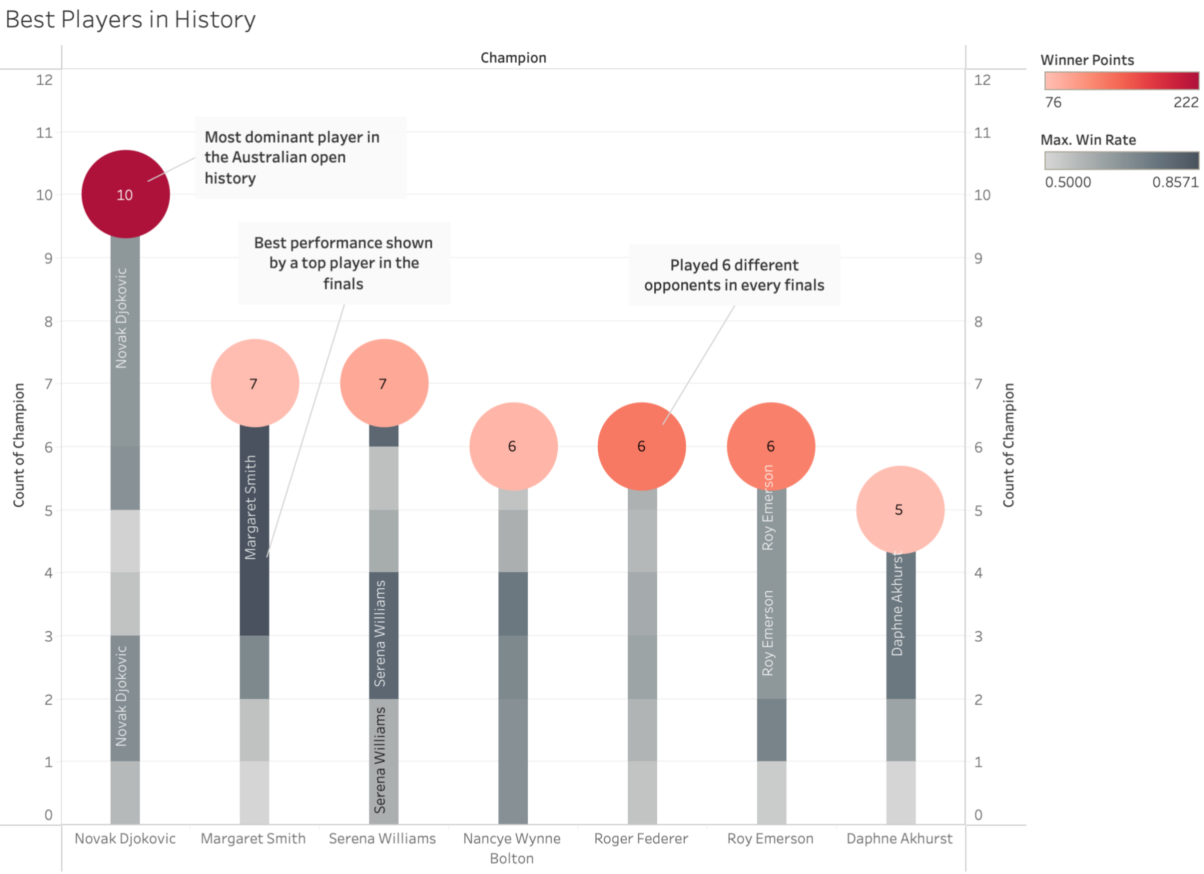


*Figure 8*

In this illustration, we use the circles to indicate the sum of points earned in the finals by all of the players, the division of the bars indicate the number of title matches they played against specific opponents, the sum of those matches are the height of the bars which indicate the title count for the players, the colours of the bar show the maximum win rate by the champions against their specific opponents.

This chart summarises the high dimensional data in a very readable and an interactive way, a lot of information is being conveyed through it but its very visibly clear and easy to explain with the correct legend.

## Word Cloud

A visualisation method that displays how frequently words appear in a given body of text, by making the size of each word proportional to its frequency. All the words are then arranged in a cluster or cloud of words.

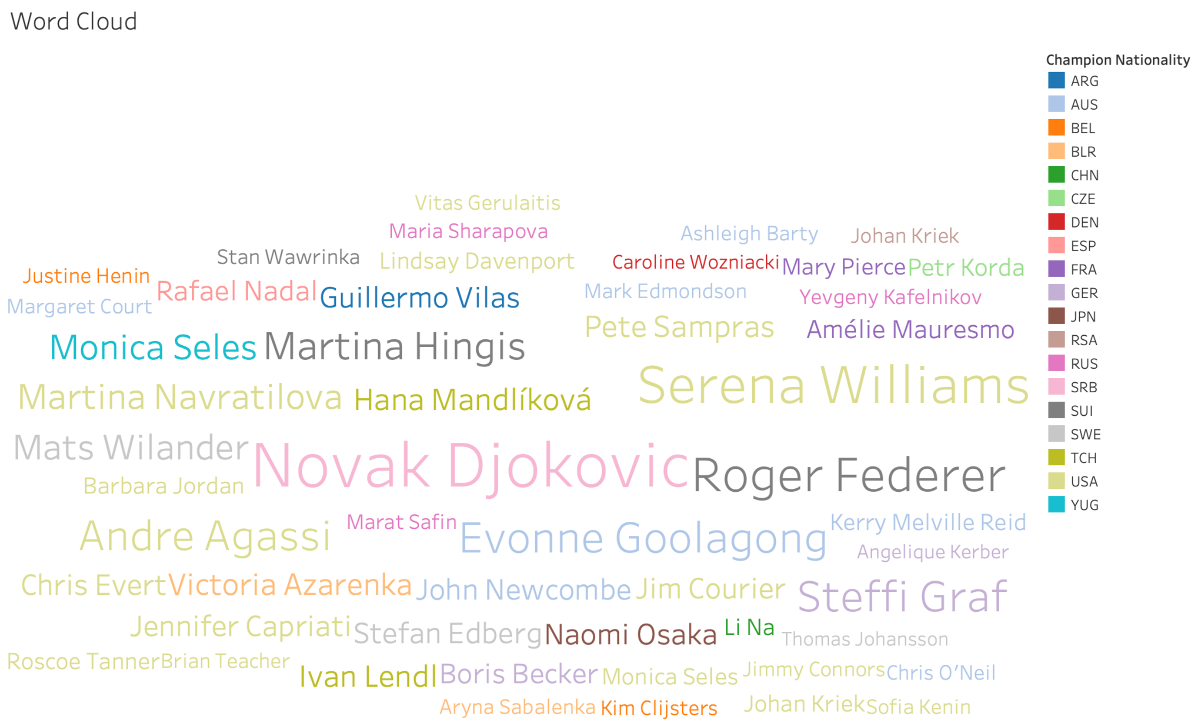


*Figure 9*

This figure illustrates the size of the names as their sum of the win rate and the colours as the nationality, the data is further filtered to only include players from the last 50 years.

From this figure we can clearly define the dominance of the player and their nation as the colours give a good representation for the number of players who have won from the country and the text shows the players individual performance.

## **Executive Summary**

This assignment focuses on visualising and analysing high-dimensional data from the Australian Open tennis championships, from 1905. The primary goal is to identify key trends, changes over time, and top winners in the competition and to provide valuable insights for tennis enthusiasts. While using several visualisation tools such as treemaps, parallel coordinates, geographic maps, and scatter charts, they were employed to enhance data understanding.

**Data divisions:** Through visualisations, we observed the distribution of champions' nationalities and gender over time. This highlighted changes and patterns in the competition. We further on created other groups like eras to divide times and continents to divide nations into separate groups and gain further understanding.

**Top Players:** We identified and visualised top players who have won the Australian Open five or more times. These legends of the sport showcased remarkable dominance in the tournament.

**Visualisations:**

- A treemap provided a hierarchical view of champions and their performances in the finals and then further dividing their genders
- Parallel coordinates revealed relationships between various data attributes, enabling us to explore patterns and trends in the competition.
- Geographic maps illustrated geographical diversity among champions and runners-up.
- Scatter charts allowed us to delve into specific match-related statistics for individual players and their games.
- Lollipop Chart to delve more into the top players statistics. It also helped us in comparing all of them.
- Word Cloud - We used a word cloud to get information on the champions and their countries by their colours and sizes.

## Advantages of Using Tableau:

- Data Visualization: Tableau is highly effective in creating visually appealing and interactive data visualisations, making it suitable for presenting complex data to a broad audience.It offers an intuitive, user-friendly interface that allows users to create high level visualisations.
- Interactivity: Tableau provides interactivity features, enabling users to explore data by filtering, sorting, and drilling down into details, which enhances data analysis.
- Data Transformation: The platform can handle data transformation and cleaning tasks, helping to prepare the dataset for analysis.Easy to create new variables on Tableau using calculated fields or creating groups using the group function.

**Disadvantages of Using Tableau:**

- Resource-Intensive: Working with large datasets or creating intricate visualisations may require substantial computational resources, potentially leading to slower performance.
- Learning Curve: For users new to Tableau, there is a learning curve, and mastering advanced features may take time.

# Conclusion

In conclusion, the visualisations have given us these key insights, patterns and findings from the Australian open Dataset after the analysis and visualising the various charts to gain a deeper understanding of one of tennis's biggest tournaments.

- The Games in the 21st century era have become overall much closer compared to the previous era. This can be seen from the parallel coordinate chart in which we see the average win rate for the players is lower so they are less likely to beat their opponents with ease.
- Serena Williams has been an extremely dominant player in her games in the finals as she has a noticeably higher win rate to her contemporaries.
- Novak Djokovic has had the strongest performance in the third set compared to any other player in any other set.
- Esna Boyd has lost the competition for a record 6 times while reaching the finals.
- Martina Hingis has reached the finals 6 times and has converted the finals 50% of the times she has played.
- Nancy Wayne Bolton has played the finals 8 times, the highest for the women's game, winning 6 of those finals and losing only 2.
- Australia won 83 titles before 1972 and since then they have only won 11 titles.
- The United Kingdom have not won a single title since the ground was changed to Koonyang lane in Melbourne, Victoria.
- Johan Kreik from South Africa is the only player from Africa to win the Australian open in the history of the competition.

- Guillermo Vilas from Argentina has won the title two times and he is the only player from South America to win the title in the competition’s history.
- Monica Seles from Yugoslavia has won the competition 3 times and is the only player from a country which does not exist anymore  to win the title in the history of the competition.
- Serbia has won 10 titles, one of the highest in the competition and all of them were won by a single player, Novak Djokovic.
- Andy Murray and Jan Lehane have lost the finals 4 times to the same opponent - Novak Djokovic and Margaret Smith Respectively.
- Roger Federer and Nancy Wayne Bolton had extremely consistent performances in their finals, they are also the top performers who have lost finals in their Australian open career.
- Margaret Smith has given us the most dominant performance between all of the top performances in a final with a win rate of 0.85.
- Roger Federer played 6 different finalists in each of the finals he played in the Australian open.
- There were 0 top performers in the pre modern era of chess(1970-2000).
- Daniel Medvadev is the best runner up in the sports history, he got 26 points in a finals and still went on to lose the finals.
- Novak Djokovic is the king of the Australian open with the highest titles won and the highest titles won in a row by any player.

## References

- Wrobel, M. (2023, June 28). Data Normalization, Explained: What is it, Why it’s Important, And How to do it. Invgate. Retrieved from https://blog.invgate.com/data-normalization#:~:text=Normalized%20data%20is%20easier%20to,visualization%2C%20understanding%20and%20pattern%20recognition.

- Tennis Australia. (2019). The Happy Slam: A History of the Australian Open. Google Arts & Culture. Retrieved from https://artsandculture.google.com/story/the-happy-slam-a-history-of-the-australian-open-tennis-australia/twWhuFJOkd3ELA?hl=en.