

Australia Open Visual Analytics

Report by – Aryan Bansal

Contents

Understanding the dataset:	1
Data Exploration:	2
Tree Map:	2
Advantages of Tree Map:	3
Disadvantages of Tree Map:	3
Geo Map:	3
Advantages of Geo Maps:	4
Disadvantages of Geo Maps:	5
Parallel Coordinates:	5
Advantages of Parallel Coordinate Graph:	6
Top Players Analysis.....	6
Scatter Plot	7
Executive Summary	8
Conclusion	8

Understanding the dataset:

The dataset given is of Australian Open Championship for Tennis and is provided with 120 Years of data, which consists of the champion players for the particular tournament, Runner-up, their nationalities and the scores for the sets they played. The data provided does not consists of the Year 2020 and 2021 because of Covid-19 pandemic, due to which the tournament was not played for those 2 years.

Data Exploration:

The key point to start an analysis and the 2nd stage of performing an analysis is data exploration to have a better understanding of how the data is to perform various statistical and visual analysis.

The data consists of different types of attributes available in the data, such as: -

- Nominal Data – We are provided with Champion, Nationality, Country, Runner-up, etc.
- Ordinal Data – We are provided with Year in the dataset
- Categorical Data – We have been given Gender as categorical data
- Text Data – Scores can come in Text data
- Ratio Data – Win ratios for different sets and total sets can be categorized as ratio data.

Tree Map:

A tree map is a kind of visualization which consists of graduated sizes of rectangles of a similar kind formed in a hierarchical structure, or tree. These rectangles represent certain variables and categorical data which is created using one or more dimensional data and one – two measure data. It can be created in different formats to be visualized, such as varying in size and varying in colour to represent the intensity of the data within a category.

It is best used to represent a data when created to show a part-to-whole relationship.

Tree Map



This tree map created for the dataset is to visualize the win ratio of different players which is sorted using different attributes like, Gender, Champions, Champions Country, Win Ratio.

The map is divided amongst Countries which is further divided into gender. Different colour shows the different countries, and the size of rectangles shows the win ratio. Bigger the size is higher the win ratio and vice-versa. Win Ratio is calculated by dividing Games won with Total number of games played by a player.



This map shows the details of a player while hovering upon the player in the map. The details show the country name of the champion, the name of the champion player, player name who lost that match and the win ratio of the player won.

Advantages of Tree Map:

- Data is fragmented and is largely concentrated into rectangles which are sorted in size according to the data input.
- Extremely efficient when used to create data which is hierarchical and can showcase large amount of data on a single map.

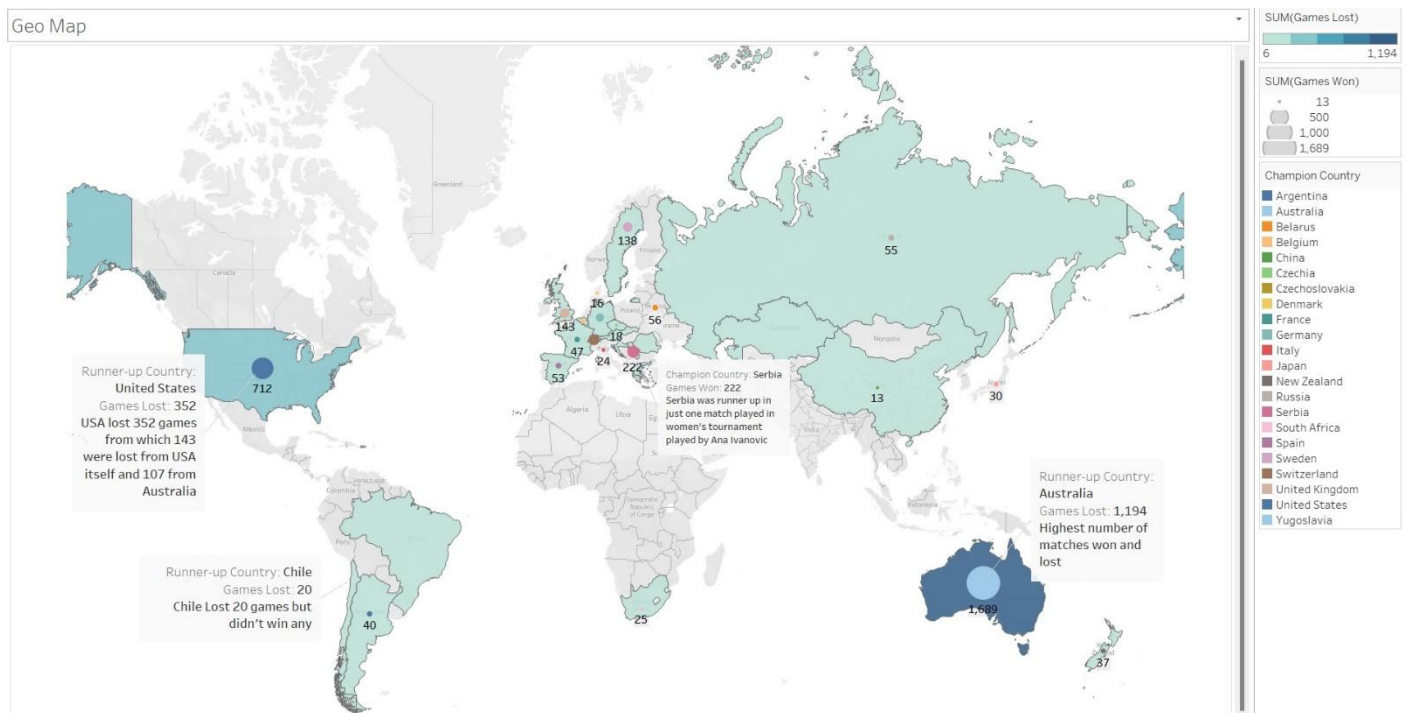
Disadvantages of Tree Map:

- When the rectangles are too small, it hides the data labels which does not gives proper visualization.
- It is difficult to make precise comparison between different attributes as the sizes of rectangles seems very similar.
- Only zero or positive values can be added to the map which is a limitation in visualization.

Geo Map:

Geo Map is a type of visualization which is created to connect and analyse data into geographical data instead of showing names of countries. It can be used in two distinct ways, one of which is a symbol map and other one is geographical map.

Both the maps are a bit similar to use, although symbol maps can have upto two measure values and geographical maps can have upto one measure value.



In the above map, the attributes used are champion country, games won, games lost and runner up country. I have connected two different kind of maps discussed above by adding dual axis and showing symbol maps as Games Won and champion country and geo map as Games lost and runner-up country.

For Symbol maps, champion countries have been filtered with colours and the size of the symbol shows the number of wins by that country.

For Geo map, Games lost by that country is filtered with colour which shows blue as lowest games lost to Teal as highest number of games lost.

From the geo map created, it can be seen that Tennis is a sport played all over the world with Europe being the biggest participant of the competition as almost all of the countries from that continent has participated which is missing from Africa.

Advantages of Geo Maps:

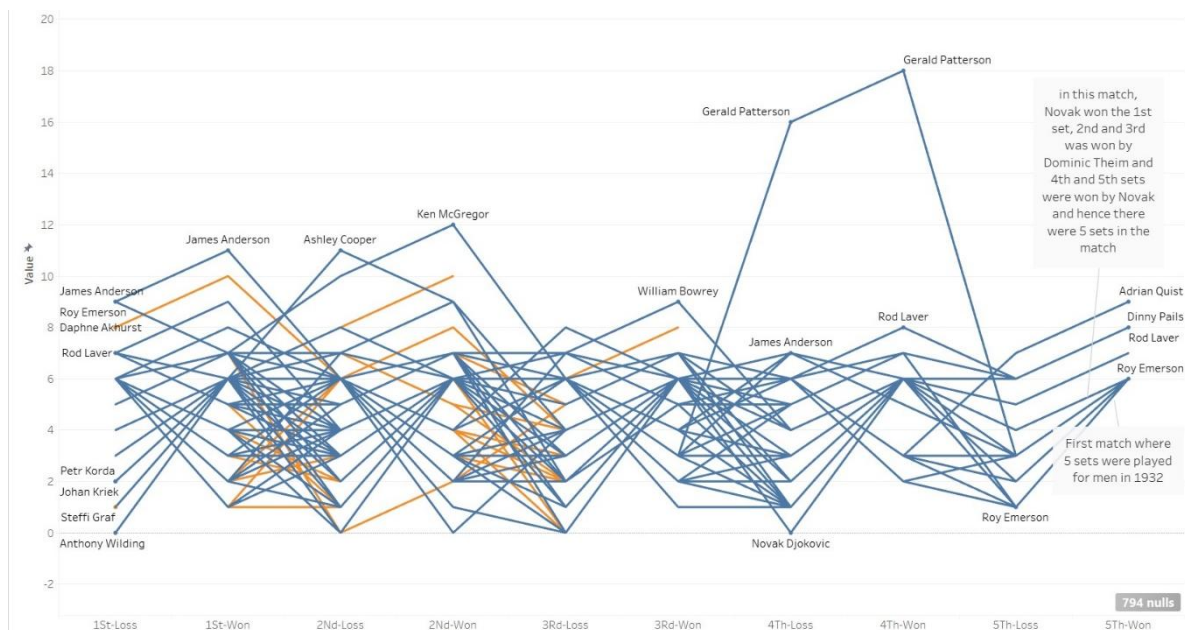
- Market segmentation, trend analysis based on regions can be analyzed through geo map giving a creative and organized way to visualize data.
- It avoids overlapping and multiple dimensions to prevent confusion.
- It has a variety of options to visualize data in different customization options like shading, symbols, sizing, etc.

Disadvantages of Geo Maps:

- Overuse of different shapes and colours can increase the complexity of visualizing the data

Parallel Coordinates:

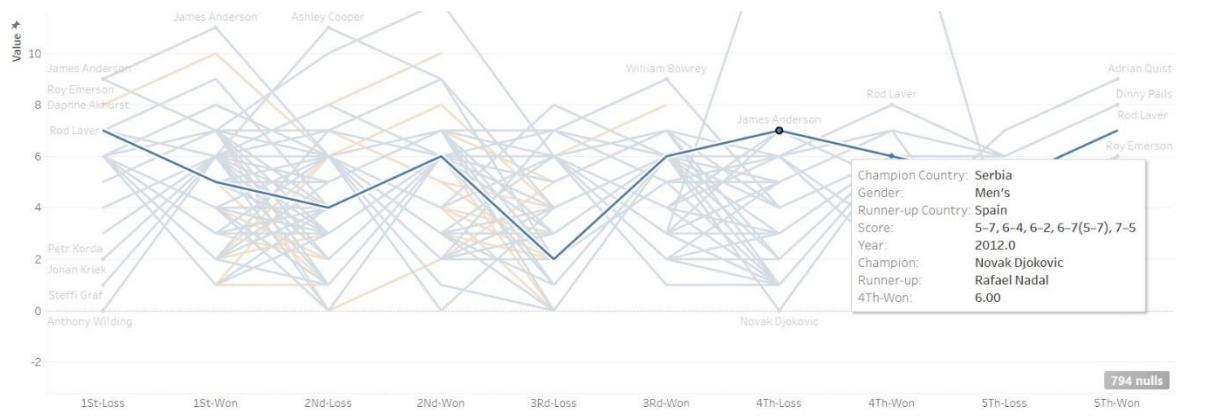
This visualization helps in creating individual axis for all the variables and putting them parallel to each other. Each data point is depicted as line traversing to parallel coordinates which corresponds to a specific variable or dimension.



The above parallel coordinate graph shows the variables such as Gender, Champion, Score, Year, Runner-up, Champion Country, and Runner-up Country.

The coordinates are bifurcated into two colours, in which blue is for men and orange is for women. It shows the number of games won and lost by a player since the player started to play in this tournament, which can also be seen as players evolution.

It can be seen that in 1932, first match of 5 sets were played in mens tournament which was played between Jack Crawford and Harry Hopman which was won by Jack.



In the above visual, it can be seen that Novak played 5 sets which not always happens as usually the match gets over on two or three sets. Here the match was against Rafael Nadal in which 1st set was won by Rafael, 2nd and 3rd by Novak and 4th set was again won by Rafael and hence the 5th match which was won by Novak and hence the champion for the tournament.

Advantages of Parallel Coordinate Graph:

- High scalability while visualizing large dataset as various filtration options are available and axes can be arranged to accommodate that data.
- Here, the colours of the line can help us depict the relation between number of games champions played and the match scores.

Top Players Analysis

Top Players Analysis has been performed for those players who have won the championship for 5 or more times. To perform this analysis, new win ratio has been calculated and has been visualized as scatter plot.

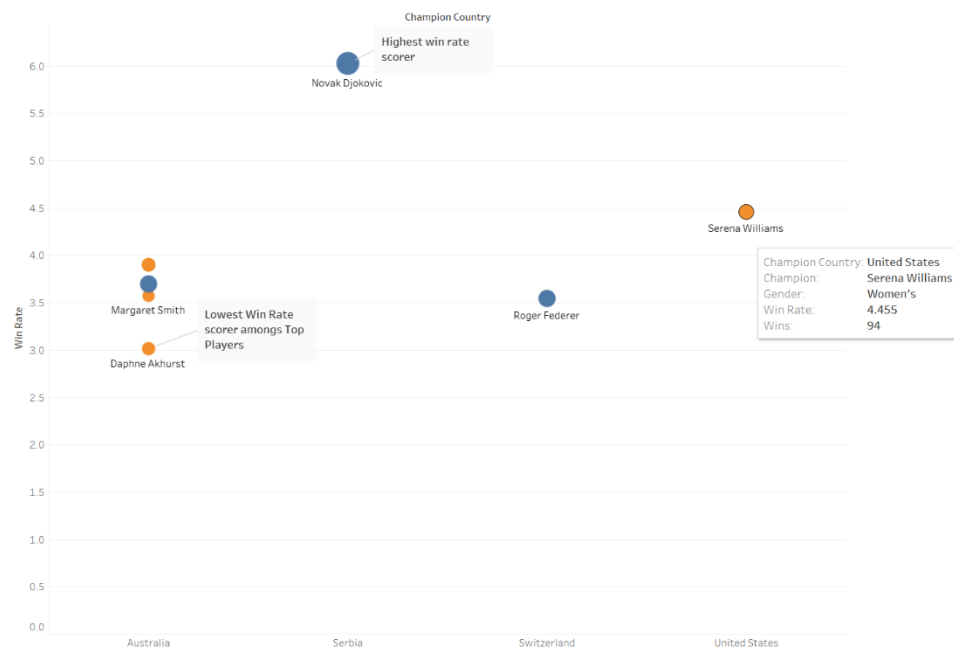
The performing players are:

- Daphne Akhurst
- Margaret Smith
- Nancye Wynne Bolton
- Novak Djokovic
- Roger Federer
- Roy Emerson
- Serena Williams

Players with 5(+) wins

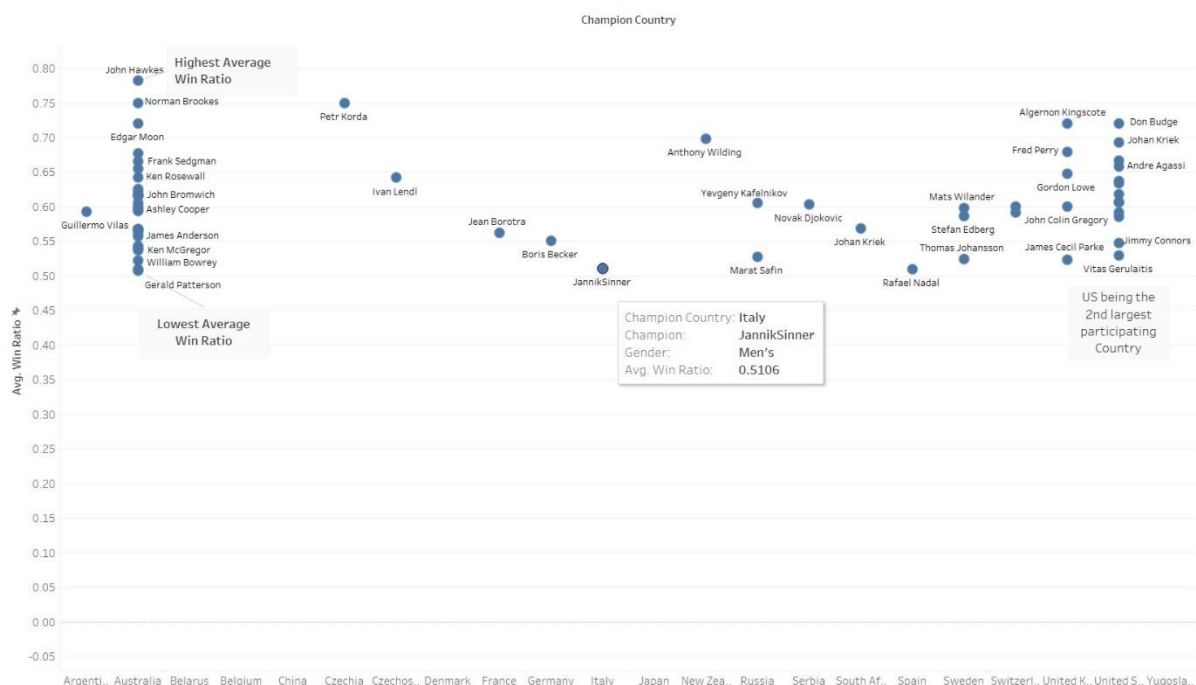
Player	Number of Wins
Daphne Akhurst	5
Margaret Smith	7
Nancye Wynne Bolton	6
Novak Djokovic	10
Roger Federer	6
Roy Emerson	6
Serena Williams	7
Grand Total	47

Scatter Plot



This graph shows the top players win ratio on the Y-axis and countries on the X-axis. The genders have been bifurcated in colours, blue depicts men and orange depicts women. The graph shows the total games won by the player with the win ratio.

Scatter plot - Men's



The above graph represents the average win ration for all the men played in the tournament having the variables such as, Win ratio, Champion Country, Champion, through which it can be visualized the countries participation, highest and lowest average win ratio and comparison

between players. Novak being the highest winner of the matches, he is not at the top of average win ratio.

Executive Summary

The dataset covers 120 years of the Australian Open Championship for Tennis, omitting 2020 and 2021 because to the COVID-19 pandemic. It contains information on champion players, runners-up, nationalities, and set scores. Data research showed a variety of qualities, including nominal, ordinal, category, textual, and ratio data. Analysis was carried out using visualization approaches such as tree maps, geo maps, and parallel coordinate graphs. Tree maps effectively displayed different players' victory rates, whereas geo maps demonstrated global involvement in the competition. Parallel coordinate graphs represented player evolution and match information across time. Top players with five or more wins were studied, with Novak Djokovic emerging as the most successful. A scatter plot was used to visually represent victory rates and nation involvement, showing important patterns and similarities among players. Overall, the research gives useful information about players' past performance in the Australian Open Championships.

Conclusion

Finally, the analysis of the information for the Australian Open Championship for Tennis provides significant insights into the tournament's rich history and top players' performances. We got a thorough grasp of numerous facets of the championship, including player progression, geographical involvement, and performance measurements, using data exploration and visualization tools such as tree maps, geo maps, parallel coordinate graphs, and scatter plots. Despite the limitations given by the COVID-19 epidemic, which resulted in the omission of data for 2020 and 2021, the research gave a comprehensive overview of the tournament's 120-year history. By identifying top-performing players and evaluating their win ratios, we discovered significant trends and patterns, with Novak Djokovic standing out as the player with the most wins. The study emphasizes the necessity of using data analytics and visualization approaches to unearth significant insights and guide decision-making in sports and beyond. Overall, this study helps to a better understanding of the Australian Open Championship's legacy and the accomplishments of its most famous players throughout history.