32146 Data Visualisation and Visual Analytics - Autumn 2024

Assessment Task 2

By Aishwarya Panhale- 24587976

Table of Contents

Table of Figures

Executive Summary

This report provides in-depth exploration of advance visualisation of data from the Australian Open tennis championships. The dataset was used to understand the high-dimensional championship match using Excel and visualisation tool such as Tableau. The dataset consists of champions from the year 1905 to 2024, including men's and women's singles with scores and Runner up. With 120 years of record, it contains 19 attributes and 211 rows with represent 211 matches played over the years. It is essential to have a thorough understanding of the source of the dataset, which includes a variety of attributes with numeric/string formats. The columns in the dataset represented by an attribute hold a significant data about the sets and records of the match. Data cleaning was carried out to find any outlier or any data cleaning protocols were required. Data transformation was utilised to forecast the outcomes of these players.

With the help of analysis in Excel, there are Top players which includes 4 women and 3 men. These Top players have won 5 or more titles in the Australian Open championship over the years. To narrate the data story effectively, visualisation techniques are used in tableau such as geographic maps, treemaps, and parallel coordinate plots. These visualizations provided insights into player performance across genders and different nationalities.

The dataset requires a factor to calculate the percentage of win rate for these champion by frequently securing a set scores of 6-any number less than 3 or 6-0. Champion Amelie Mauresmo from France secured highest win rate of 89% with set score as 6–1, 2–0 retired. Champions with a higher win rate percentage typically won more matches, played fewer sets, and defeated their opponents 6-0.Regardless of gender, the win rate of each player is heavily influenced by the total number of wins or losses. The visualisation techniques used has maintained the consistency in result patterns.

This report includes evaluation of the advantages and disadvantages of each visualization technique and offering insights of understanding their efficacy in narrating the data story. It highlights the importance of selecting appropriate visualization methods and improve understanding to generate insights. With the help of conclusion, the summarises of the key findings are mentioned through the in-depth analysis.

Data exploration

Data exploration is the initial phase in the process of data analysis and visualization which involved thorough examination of the dataset to find meaningful insights and make interpretation. Analysing the dataset, unique patterns and trends can be found which contributes to a deeper understanding of the data. This exploration process involves summarizing the format and description of the attributes, and interpretation of the data.

Following are the data format of the attributes described in the dataset:

Attribute	Type	Description
Year	Numeric- year format (YYYY)	Tournament year
Gender	String	Gender of the champion
Champion	String	Winner of Australian open Championship name
Champion Nationality	String	Nationality of the champion

Champion Country	String	Country of the champion
Score	String- separated by commas	Score of the champion and runner up of the championship.
1st-won- 5th won	Numeric	Game won in 1st-5th set
1st -loss- 5th loss	Numeric	Game lost in 1st-5th set
Runner-up	String	Name of the Runner up of the tournament
Runner-up Nationality	String	Nationality of the Runner up of the tournament
Runner-up Country	String	Country of the Runner up of the tournament

Key findings through data exploration on dataset of the Australian Open Championship:

1. Since 1905 to 2024 Champion Novak Djokovic from Serbia has secured the maximum number of championships (10) in Men's category.
2. Champion Serena Williams(United states) and Margaret Smith(Australia) has secured second highest maximum number of championships (7).
3. Since 1905 to 2024, in the Australian Open Championship, Australia secured the highest number of championships.
4. Serbia has only one Champion with maximum number of championship title.

Data Preparation

Data preparation is essential to achieve desired output, it consists of data cleaning and transforming. Calculations and additional variables are added into the dataset to enhance the visualization. Few missing values were identified in Champion Seed, Runner-Up Seed, and few wins and loss columns. Removing the missing values in wins and loss would have removed data, which was essential for visualization, this missing values indicates the instances in which the champion secured the title before completing all games, prompting their exclusion and the missing values were kept in the dataset. Columns such as Champion seed, Runner-up Speed and Mins have been removed because these columns had no values. New attributes are also added then into the dataset.

Following are the attributes added into the dataset:

Attributes	Description
1st- 5th Win Rate	Win rate of the game based on the score secured in each game
1st-5th Loss rate	Loss rate of the game based on the score secured in each game
Total Win	It is summation of all the wins for the particular tournament year
Total Loss	It is summation of all the loss for the particular tournament year
OverallMatch-WinRate	It is total percentage value of win and loss rate of all the game played in that year to get total win Rate of that game.
Champion Count	Number of times the player has secured the Champion title

RunnerUp Count	Number of times the player has secured Runnerup title.

Visualization Techniques

Treemap

Treemaps, also referred as Slice and Dice, has an innovative way to display data. The different size and colour rectangles convey information based on the filters applied in tableau. This visualization technique is effective as it aids in navigating complex datasets by visually depicting various values and categories. Rectangle sizes in treemap corresponds to a values with are depicted as higher values get larger rectangles and vice versa. This hierarchical structure of the data effectively communicates by offering a structured and understandable data relationships.

Interpretation and analysis of the Treemap

In Figure1, In tableau treemap visualization is created where the win rates of champions categorized by country. The map includes Detail labelling of the champion name on rectangles and the scaling techniques of the visualization is between the champion's maximum Win rate for the tournament. the rectangles representing champions with higher win rates are represented by larger rectangles, while those with lower win rates appear smaller. From Australia, John Hawkes in Men's Category has highest win rate of 0.7826 whereas Margaret Smith in women's category has highest win rate of 0.8571.
From the Figure 1, showcase that Australia which is shown using light blue colour has maximum number of champion with win rate. The United States is at second number for maximum number of champion. Billie Jean King and Serena Williams in women's category have same Win rate as 0.8000 and Andre Agassi in men's category has 0.7826-win rate. Countries like Serbia, Yugoslavia, China, New Zealand, Argentina, South Africa, Japan, Spain, Italy, and Denmark have only one champion over the years. Among these countries Yugoslavia has highest Win rate of 0.7059 and Denmark has lowest win rate of 0.5000 in women's category

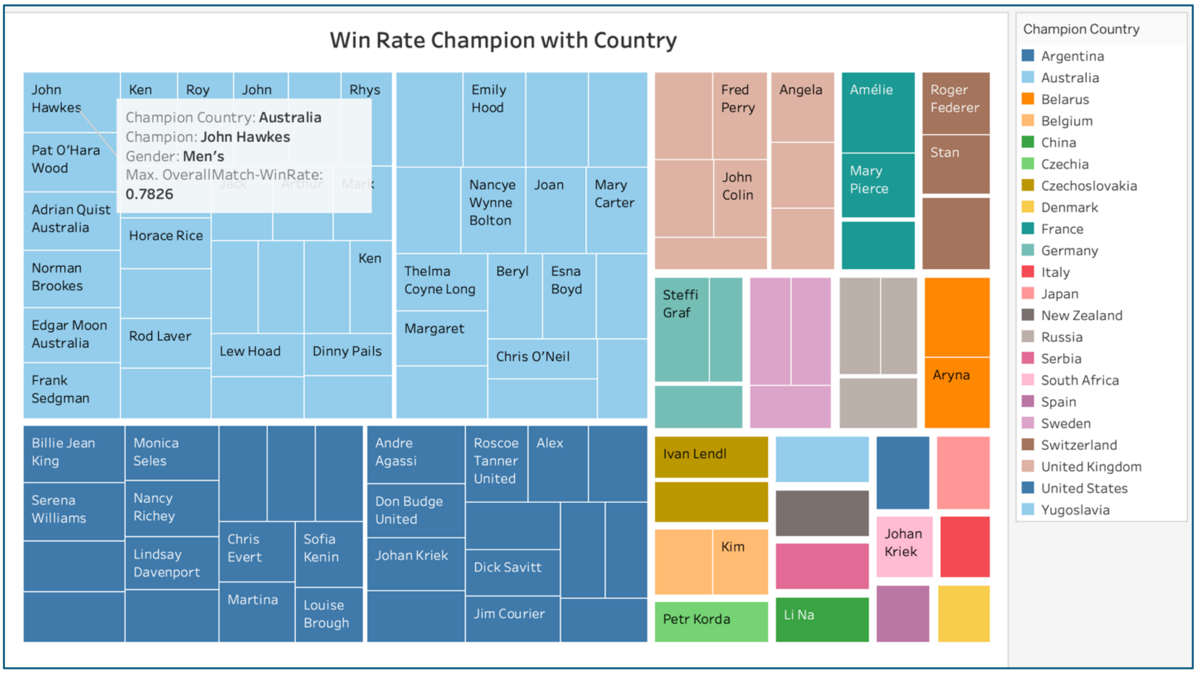

Figure 1 Treemap- Win Rate Champion with Country

Advantages and Disadvantages of Tree Maps

Advantages:

1. Tree maps provide a hierarchical representation of data and allows to visualize the data structure and understand patterns of smaller attributes nesting into larger categories.
2. This Feature enables the understanding of data and its relationships and provides an easy grasp of the hierarchy present in the dataset. It also highlights the identification of trends and outliers.
3. With the help of colours and size of the rectangle, the visualization provides quick analysis by differences in pattern. The use of colour coding enhances the visualization which represent different categories. Category such as country in the above figure1 and distinguishes between various data segments, enhancing comprehension and interpretation.
4. Moreover, tree maps are known for their user-friendly nature, providing quick insights of the concepts due to which it makes them accessible to users of varying levels of expertise, providing efficient data exploration and interpretation. This visualization technique is a valuable tools for visualizing hierarchical data, offering a combination of hierarchical representation.

Disadvantages:

1. Treemaps at times get cluttered and complex specially with large datasets. It can challenge the visualization interpretation due to numerous rectangles which can overwhelm the ability to extract the insights from the visualization.
2. Being easy in nature, treemap are not universally suitable for all data types. With string and discrete numeric format, this visualization is not suitable. This inherent limitation restricts the use of tree maps in certain data visualization scenarios.
3. It can be challenging to read labels within tree maps due to the numerous rectangles. This difficulty in label readability can lead to misinterpretation in understanding the data represented by each rectangle, hampering the effectiveness of the visualization.
4. Treemaps excel in illustrating differences, there are limited capabilities for comparing the relationships between multiple attributes. This affects the ability to explore complex and large datasets and limits the depth of analysis and insights which can be derived from the visualization.

Parallel Coordinate Plots

Parallel coordinate plots are reliable techniques for visualizing data that exists in multiple dimensions. They use of several axes representing a quantitative variable with its unique scale and measurement unit. The visualizations ensure the representation of data provided thorough comparative analysis. The arrangement of axes is important to understand the trends and patterns within the dataset. The Parallel lines between adjacent variables signify a positive correlation, whereas the intersecting lines signify a negative inverse relationship. This visualization technique allows users to explore relationships and aggregation patterns in the complex, multidimensional and large datasets, with axis showing the quantitative value.

Interpretation and Pattern Analysis of Parallel Coordinate Plot

Using Tableau, The parallel Coordinates plot is created based on 1st -5th win rate score along with Overall win rate. By integrating the top 10 Champion with colour coded lines and separate plots for Men and Women Championship allows to get detailed insights into player performance. In Figure 2, Plots are divided into two category where Margaret Smith has highest Win Rate of 0.8571 followed by Serena Williams with a win rate of 0.8000 in

Women's category. In men's category Andre Agassi is at the top with win rate of 0.7826 followed by Ken Rosewall with 0.7200 win rate.

As seen in the figure 1 below, Margaret Smith being the highest win rate champion, there is a negative correlation as the average of 1st win rate 1.0000 dropped to 0.7500 in 2nd average win rate. Similarly, Second highest women's category champion with high overall win rate- Serena Williams, showcase a negative correlation as the average of 1st win rate 1.0000 dropped to 0.6667 in 2nd average win rate. Daphne akhurst having low Overall Win rate compared to Serena and Margaret. Her 1st to 3rd wins rate is showing positive correlation from 0.1429 to 0.5714 and a slight increase to 0.6000 in 3rd average win rate. Daphne has shown 319.67% growth based on her average game win rate.

In Men's category, Novak Djokovic shows both positive and negative correlation, Negative correlation is seen over the 1st, 2nd and 3rd win rate where the average value 0.600 dropped to 0.400 and further dropped to 0.2500. The positive correlation was seen in 4th average win rate(0.6667) the increase of 166.7% was seen.

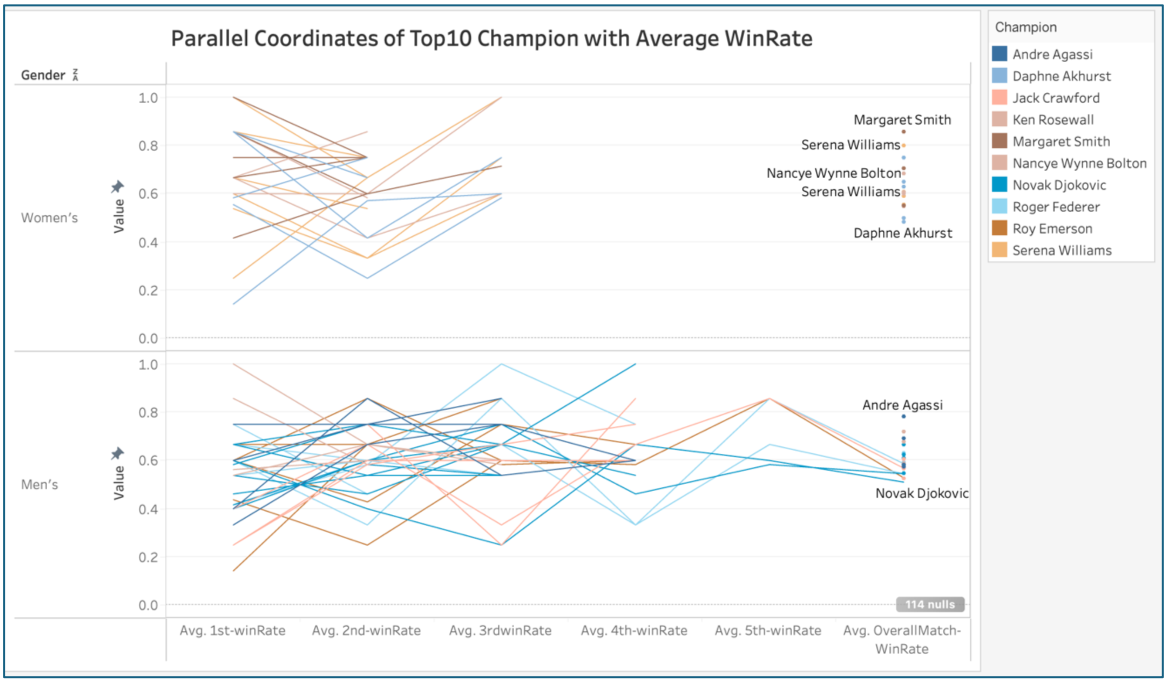

Figure 2 Parallel Coordinates- Top10 Champion with WinRate

Advantages and Disadvantages of Parallel Coordinate plot

Advantages:

1. Parallel coordinate plots allow multiple variables on different axes which simplifying the analysis of complex and large datasets.
2. To identify patterns and relationship between variables, with positive or negative correlations parallel coordinates visualization technique helps to provide meaningful insight.
3. parallel coordinate plots also help in the detecting different data points, ease to locate these data points and identify the outliers within the dataset.
4. Comparative analysis of different data subsets can be carried out through parallel coordinate plot and examining the paths of lines across multiple axes, provides deeper insights into dataset variations.

Disadvantages:

1. Parallel coordinate allows multiple axis, the plots may become complex and cluttered, because of large datasets which can at times be difficult to perform analysis and interpret the insight. .
2. Data scaling issue can provide distorted visualization which affects the accuracy of analysis.
3. Labels and annotations can make the visualization too cluttered and make it difficult to read due to overcrowding.

Geographic Map

Geographic map termed as "geo-maps," is a visualization technique to represent the data over the geographical map. The data points on the map helps to understand the information and facilitates data analysis. In Tableau, the data points are represented using circle with different sizes indicating the values of measures like champion count as shown in the Figure 3.

Interpretation and Pattern Analysis of Geographic Map

In figure 3, to create geographic map in tableau, the rows and column are latitude and longitude which are generated by adding champion country in filters. Champion country along with champion count which is calculated for every player who has been champion over the years from 1905 to 2024. The circle over Australia is bigger as it has maximum Champion count with 94 champions over the years followed by second is united states with 43 champion count and third is Serbia with 10 champion count. Countries like Denmark, Italy and China have only one champion count.

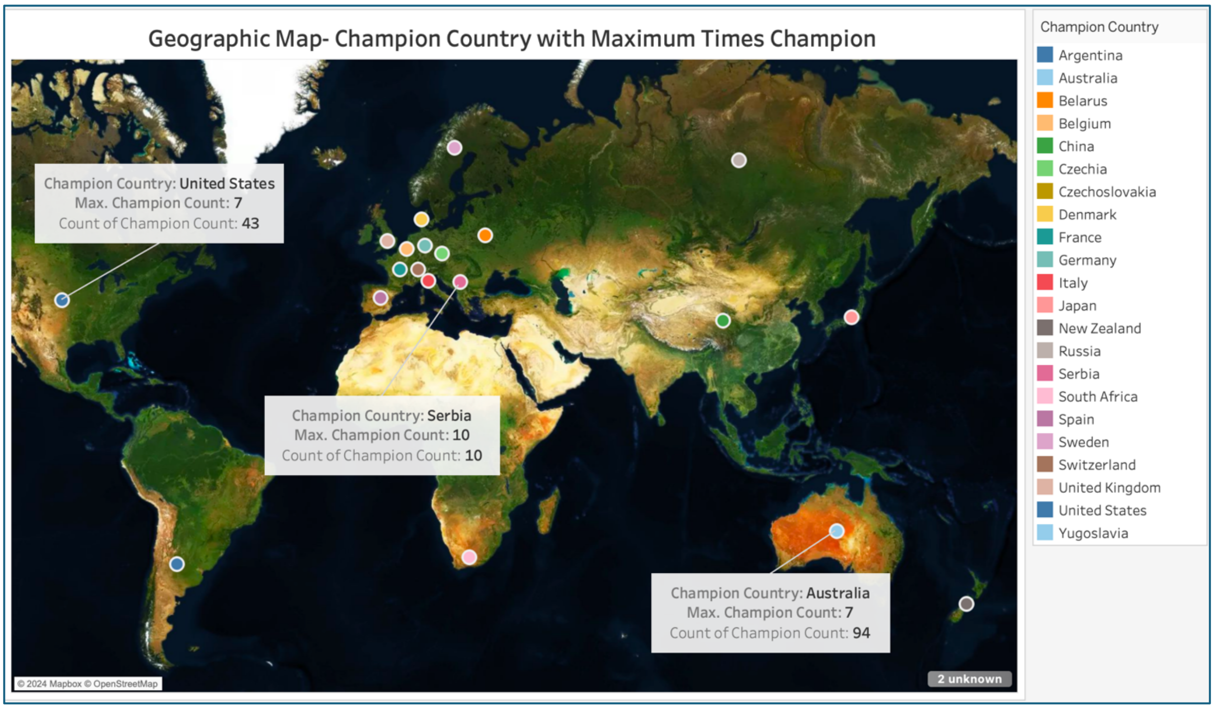

Figure 3 Geographic Map- Countries with Most Champion Count

Advantages and Disadvantages of Geographic Map

Advantages:

1. Geographic maps provide a geographical context that enhances understanding of data over the location-dependent data such as Champion's Nationality, helps for pattern recognition over the geographical map.

2. Geographic map simply data to get diverse perspective by offering visual clarity.
3. Multiple attributes are allowed to provide enhanced visualization and data analysis.
4. With the help of Spatial Analysis, the countries in relation to Champions of the tournament over the year are showcased with data point circle size help in identifying patterns, outliers, and trends.

Disadvantages:

1. Usually with large data points and complex data sets multiple attributes plotted on the map can make the visualization cluttered, making it difficult to interpret the data.
2. Lack of legends can make it difficult to understand the data and due to misinterpretation insights will be incorrect.
3. Some countries at times are not recognized in tableau which hinders the analysis.
4. Identifying trends over the time can be difficult to interpret by using geographic map.

Scatter Plot

Scatter plot is a data visualization technique where data is graphically presented as individual data points usually depicted as dots on a two-dimensional graph. The dot on the plot correlates to a specific value which is represented on the x-axis and y-axis. Scatter plots are used to analyse the relationship between the two variables (in the Figure 4 it is using champion's total loss and wins). Using this technique patterns trends, and outliers within the dataset are identified which provides a insight for data interpretation and analysis. The main goal of using scatter plot is to access the data points in relation to variables and also identifying their direction and impact.

Interpretation and Pattern Analysis of Scatter Plot

In Tableau, the scatter plots the values required are column and rows. To differentiate the data points, gender column acts as a differentiator and with help of different colour and circle sizes identing the data points is easier for the two different category.

This scatter plot contains champion's average win rate and also their total win and loss value which helps in identifying the player's performance. Overall, Novak Djokovic has highest win count of 222 along with 154 losses. As shown in the figure 4, Novak Djokovic, Roger Federer in Men's category has the second highest value with 132 wins over 92 loss which is similar to 59% win rate. The cluttered part near 0 axis shows that there are many players that have Total win less than 60 and Total loss less than 50. Identifying the outliers is easier, in the below image the outlier is used to identify the player's performance who either performed well or not well. Trendlines are sued to identify the trend and make it easier to identifying the outlier and data points that are in a group with similar data variables of total loss and total win.

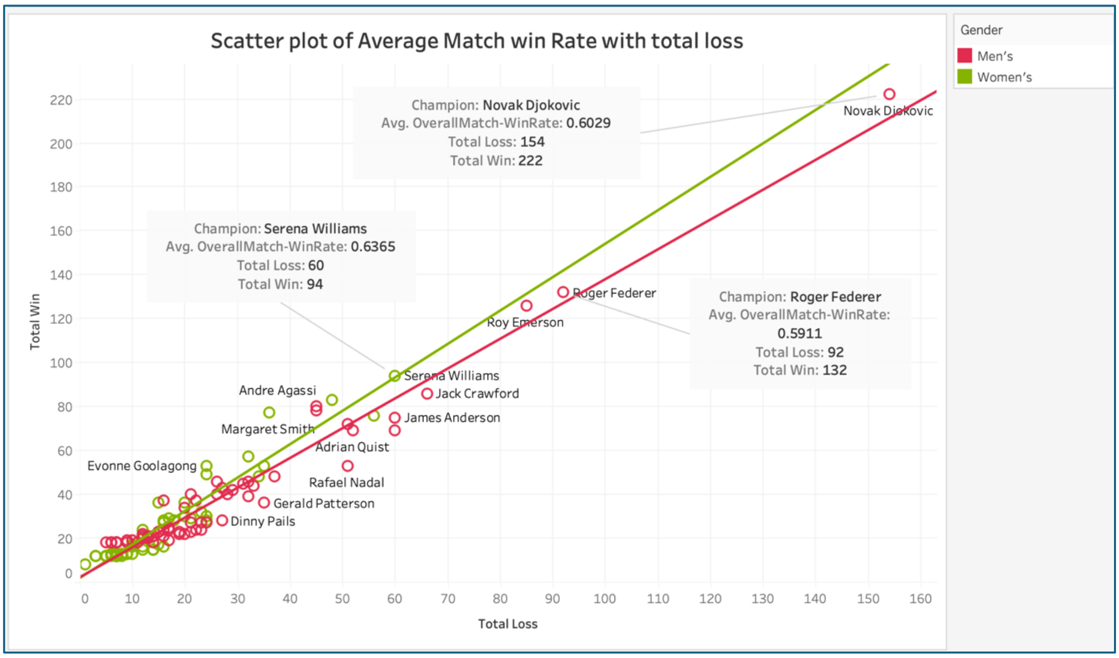

Figure 4 Scatter plot- Average Match win of champion

Advantages and Disadvantages of Scatter Plot

Advantages:

1. Scatter plots provide a relationship between two variables and helps in identify the patterns, trends, and correlations between data.
2. This visualization technique makes it easy to identify the outliers that is deviate from the general data points which enables to identify pattern and get insight of the data trend.
3. It also allows to assess the data distribution which is plotted across the axis which helps in providing insights of data reach.
4. This visualization is simple which makes it easy to interpret with limited statistical knowledge and provides effective insights of findings.
5. Scatter plots are flexible as it accommodates various types of data and are suitable for both numeric and categorical variable.

Disadvantages:

1. Scatter plot visualization has limitation when providing the visualizing is to have relationship between only two variables, with complex multi-dimensional dataset it does not provide visualization.
2. With large datasets , datapoints can overplotting which makes it difficult to distinguish and understand insight resulting in misinterpretation
3. At times, the visual representation which might suggest relationships between data points, but they are actually statistically incorrect.

Top Player Analysis

To get performance analysis of the top players, In tableau the multi-axis is used to plot two different variables which are essential to analyse the players performance. Top 10 champion are selected based on their number of titles won along with their runner-up these top players are differentiated based on their country and gender. Based on the Runner-up, the top champion performance can be analysed.

The bar graph is colour coded based on the country which helps in differentiating the country that has maximum championship. Australia has maximum champions followed by united states, Serbia having only champion standout in performance of the player due to highest won rate and maximum championship title.

In Men's category champion Novak Djokovic has highest win rate and a record 10 titles which makes this champion’s performance standout. The scatterplot of champion count also highlights when he received his championship with the runner-up details as well. In women's category, Margaret Smith from Australia has the Highest win rate and a record of 7 titles which makes her the performance standout. The scatterplot also includes the champion count with the runner-up details as well which helps for performance analysis to identify the players best performance among these multiple titles.

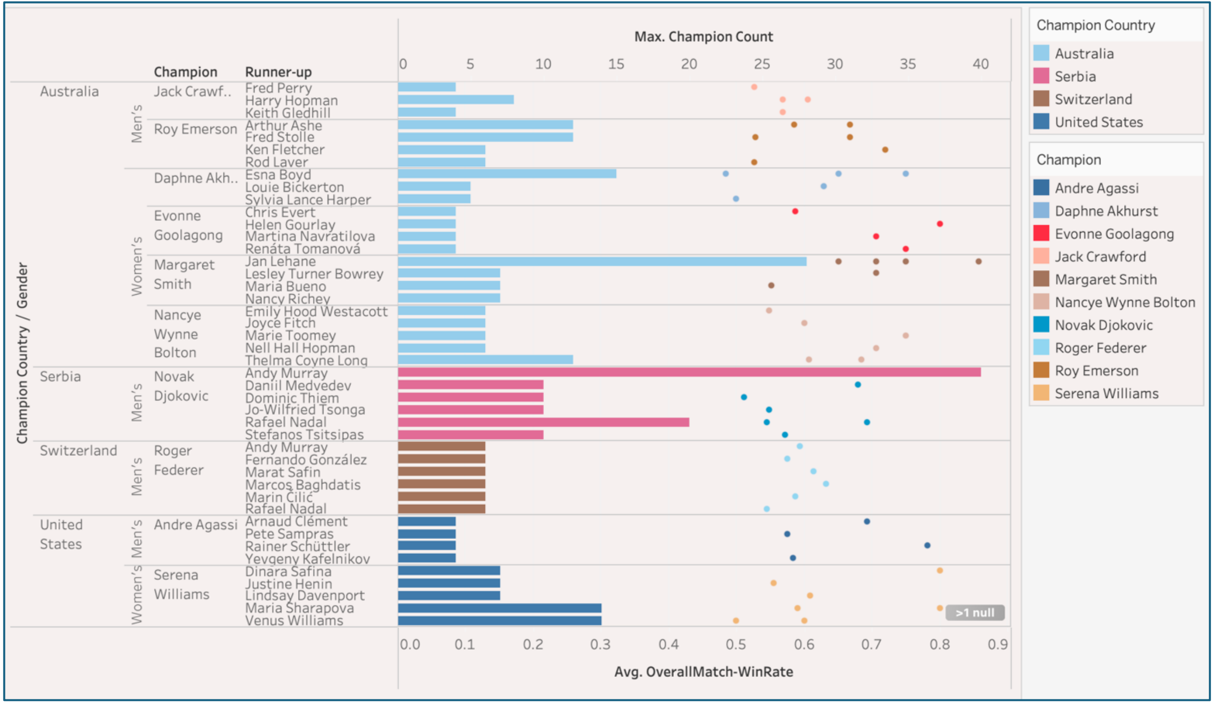

Figure 5 Top Players Performance Analysis

Conclusion

With this comprehensive exploration of dataset and visualization techniques which includes tree maps, parallel coordinates, scatter graphs, and geographic maps. These visualization techniques provide a valuable insights and statistics that have uncovered the Australian Open Championship dataset for the year 1905 to 2024. Essential Key findings includes the dominance of Australia and its players , particularly in women's category. Margaret Smith(Women’s category from Australia) leads the women’s category with 7 championship title. In Men’s category, Novak Djokovic from Serbia standout due to his performer in men's category and securing 10 championship titles while being the only one player from Serbia along with Andy Murray following closely to Djokovic. This analysis reveals a correlation between players winning pattern which is an interesting because of loss in second set after winning the first one which showcase the difficult competition among players and changing their overall win rate to perform well in the tournament.

Further exploring deeper into the analysis, women have higher average win rate compared to men, suggesting a greater margin of victory for female players. Novak Djokovic and Roger Federer stand out due to their excellent performance , while Margaret Smith and Serena

Williams overpower the women’s category. In Countries, Australia is the leading country for the total championship titles won over the years, along with USA and Serbia . Tableau and Excel helped for understanding the data patterns through different visualization techniques and enhances effectiveness of data analysis and provides ease of access for a beginner.

In conclusion, the analysis provides the insights of the Australian Open Championship along with its competitive dynamics among players to achieve the championship title. With the help of data visualization techniques these patterns and trends were discovered providing valuable insights within the dataset.