32146 Data Visualisation and Data Analytics

Assessment Task 2: Advanced Data Visualisation

Jun Hyun Lim
25175391
Issue date:
13 October 2024

Table of Contents

1 Understanding Dataset

This dataset provides match records of Australian Open tennis championships over 120 years. The dataset captures important measures for the performances of male and female players and, therefore, enables analysis of trends, outliers, and patterns over time. The structured nature of the dataset allows match scores, seed rankings, nationalities, etc., to enable a range of visualization techniques by effectively allowing categorical and numerical comparisons. The next section gives an overview of how this dataset is structured and describes the importance of the handling and visualisation of null values, which often play a key role in transparent analysis.

1.1 Data Structure and Key Variables

This dataset comprises variables about players and matches. The key variables in this dataset are the following:

Year	The year in which the tournament was conducted.
Gender	The category of the match, whether men's or women's.
Champion	The player who emerged as the champion.
Champion Nationality	The nationality of the champion as a code and full country name.
Champion Seed	Seeding rank is given to the champion at the time of the commencement of the tournament's
Score	Score of the match – number of sets taken by the winner.
Mins	Length of the match in minutes.
Games Won/Lost	Number of games won/lost by the winner.
Win Ratio	Calculated measure to give the proportion of games won by the winner from the total games played.
Set Win Ratios	Games won per set ratios.
Runner-up Seed	Ranking of the runner-up player.
Runner-up Country	Nationality of the runner-up.

1.2 Importance of Handling Null Values

Handling null values-especially those missing or incomplete-plays a prime role in any kind of data analysis, which can decide the overall quality of the analysis. This is the juncture where proper detection and understanding of where null values occur in this dataset become important.

Importance: Missing values may occur in key fields like Match Duration, Win Ratios, Set Win Ratio, Games Won/Lost, and Champion Seed which may cause incomplete analysis if not treated.

Visualisation of the null values: In visualising the null values, the trends of the missing data are brought into view. Maybe some certain years of players have more information lacking, and identification of such breaks assures broader knowledge of the data set.

Bonus Point Consideration: The presence of null value analysis in the assignment is an added step for data transparency. It will describe what part of the dataset is missing. Also, there are times when such gaps become useful in a report. The solutions also call for visualisations that show the position of the null values in the dataset and these are easily available in visualisation tools like Tableau.

Across all visualisations, null values were handled by creating custom fields and labels using the **IFNULL()** function, ensuring complete data representation.

1.3 Purpose of the Dataset

This dataset gives a detailed overview of the performance of players over the years. Comparing various metrics for the duration of matches, win ratios, and ranking of players is thus possible. In this regard, the dataset covers data for both male and female players, hence enabling gender comparison. Added to this, null value analysis is enabled for completeness in the quality of the data. All this would present a complete picture of the performance of the players om the Australian Open, by pointing out the key trends and investigating the gaps in the data.

2 Data Exploration

The Australian Open dataset consists of variety of variables that differ in terms of their measurement scales. Proper classification of these variables into appropriate categories is essential for performing meaningful analysis and generating accurate visualisations. In following sections, the dataset's variables are classified as nominal, ordinal, interval, ratio, and derived data, providing a structured foundation for further analysis.

2.1 Nominal Data

Nominal data refers to categorical variables, where the categories do not have a natural order. In this dataset, nominal variables include:

- **Gender**: Classifies matches as either male or female categories.
- **Champion's Nationality**: Represents the nationality of the tournament champion, recorded as country names or abbreviations.
- **Runner-up's Nationality**: Denotes the country of the player who finishes as the runner-up.

These nominal variables allow for the grouping of data, which facilitates comparative analysis of trends based on nationality and gender.

2.2 Ordinal Data

Ordinal data consists of variables that indicate a meaningful order, but the differences between consecutive valued are not necessarily consistent. The ordinal variables in this dataset are:

- **Champion Seed**: The initial seed ranking of the tournament winner.
- **Runner-up Seed**: The initial seed ranking of the runner-up player.

While these rankings provide a relative ordering of performance, they do not imply equal intervals between ranks. Therefore, ordinal data requires specific analytical techniques that respect the inherent order without assuming uniform differences between values.

2.3 Interval Data

Interval data are comprising variables with meaningful intervals between values, but without a true zero point. The most obvious interval variable in this data set is:

- **Year**: Denotes the year in which each tournament took place. As an interval variable, it enables trend analysis over time; however, it cannot be used for ratio-based comparisons due to the absence of an absolute zero.

2.4 Ratio Data

Ratio data includes variables that have both meaningful intervals and a true zero point, allowing for the calculation of differences and ratios. The ratio variables in this dataset are:

- **Match Duration (Mins)**: Represents the total duration of each match, measured in minutes, allowing for direct comparison of match lengths.
- **Games Won/Lost**: Records the number of games won and lost by both the champion and runner-up in each match.
- **Win Ratio**: The proportion of games won by the champion out of the total games played.

The presence of a true zero point for these variables permits full quantitative analysis, such as the calculation of averages, proportions, and performance comparisons over time and across categories.

2.5 Derived Data

Derived data refers to variables that are computed or transformed based on the original dataset. Although certain statistics, such as set win rates, are not explicitly provided, they can be derived through calculated fields. These derived metrics enhance the depth of analysis, offering more nuanced insights into player performance and trends. By incorporating derived data, the analysis can uncover patterns that may not be immediately apparent in the raw data.

3 Visualisation Techniques

3.1 Treemap

The **Treemap** is a method of hierarchical visualisation that efficiently presents categorical data. In this report, the researchers have employed treemaps while comparing champions according to nationality, number of wins, and seed positions that capture the trend of dominance across countries and seeding ranks. A treemap is an effective technique for hierarchical data, such as nationality and individual wins, as it could visually compare countries and players.

- **Dimensions**: Champion's Nationality, Champion's Name, Champion Seed (Null Correction), Gender
- **Measures**: Sum of Games Won, Win Ratio

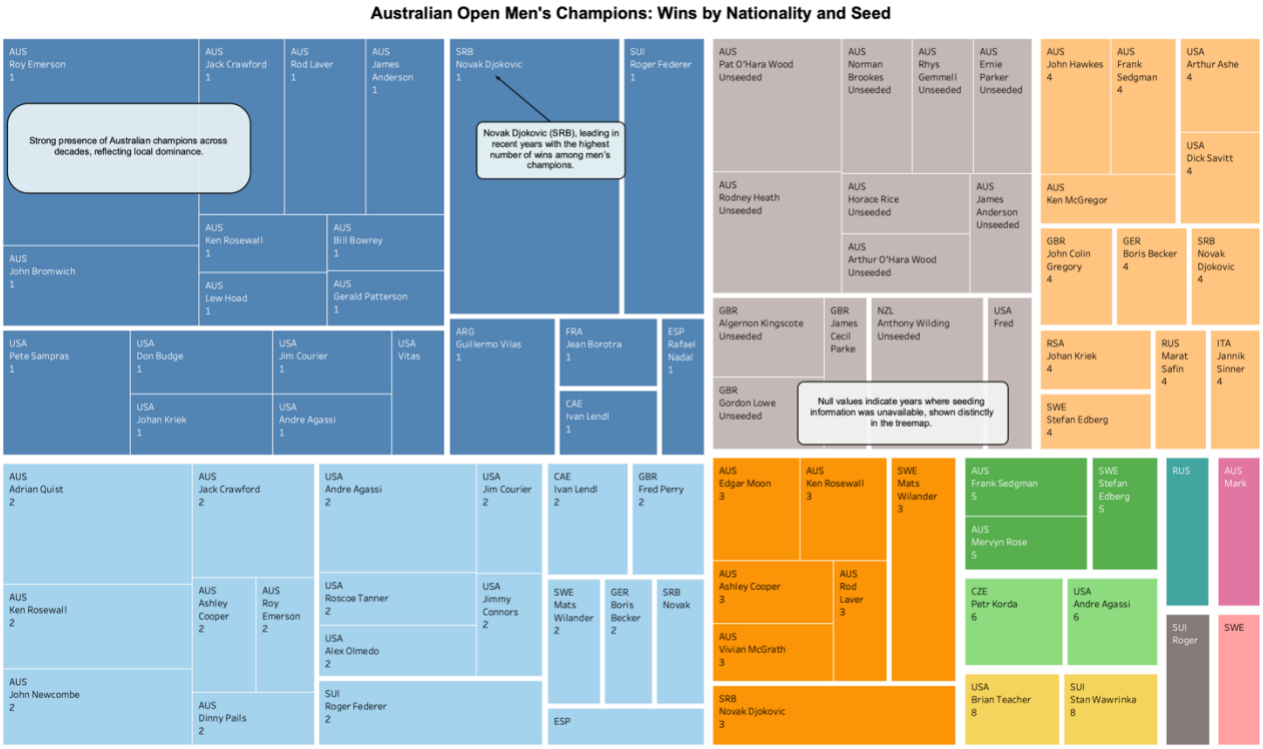

Figure 1. Men's Australian Open Treemap by Nationality and Seed

This treemap (Figure 1) efficiently visualises the distribution of Australian Open men's champions by their nationality and seeding positions. The size of each block in this treemap represents the number of wins by champions, with colour intensity showing the seeding position of those champions. Lower seed numbers are coloured in darker shades for higher-ranked players, while lighter shades indicate lower seeds or unseeded champions. This visualisation highlights dominance by nationality, particularly Australia's, and provides insights into performance patterns related to seed ranking. The compact treemap structure allows for easy comparison of champions by nationality and seeding to get an overview of player dominance over decades.

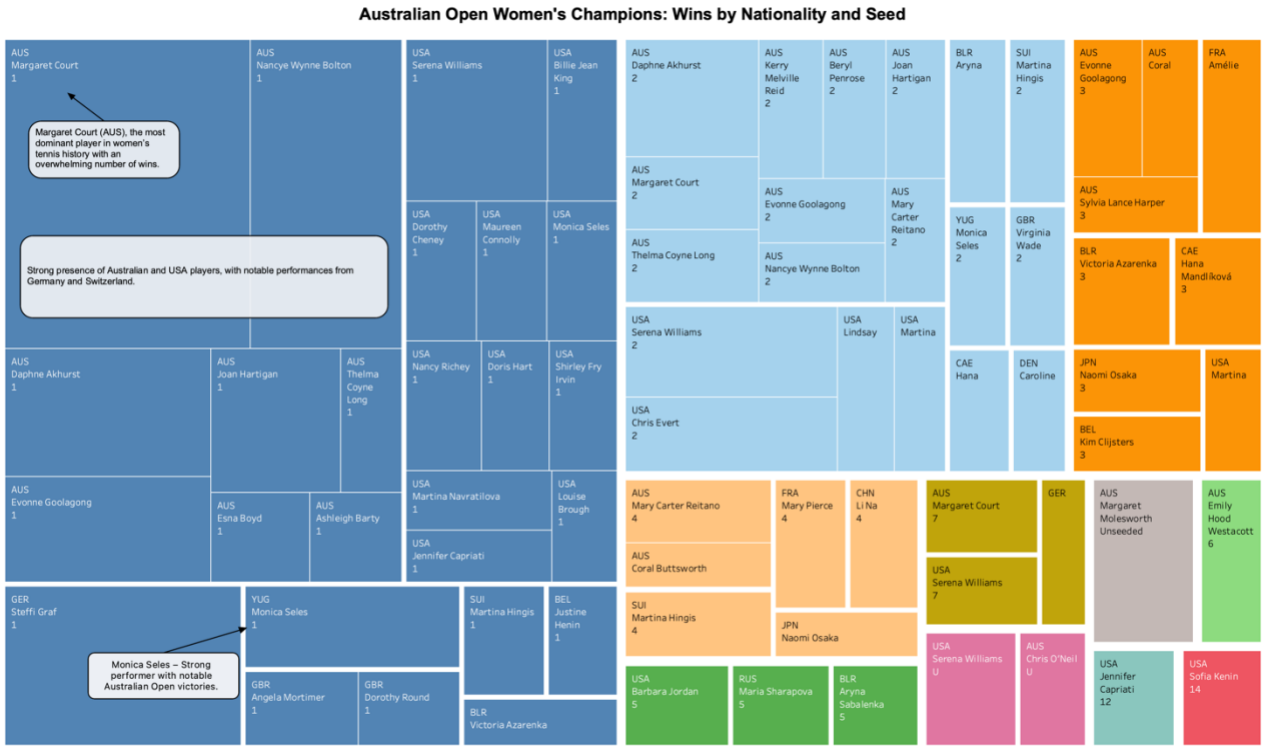

Figure 2. Women's Australian Open Treemap by Nationality and Seed

The women's treemap (Figure 2) offers a hierarchical comparison of Australian Open champions, segmented by nationality and seed. Similar to the men's treemap, this visualisation utilises block size to represent the number of games won and shades of colour to represent the position of the seed. This technique is useful for exploring patterns of dominance amongst top women players – including Margaret Court – alongside the representation of unseeded champions. While the treemap focuses on size and colour, it allows analysis of national performance and individual seed success simultaneously, bringing some important trends to the fore.

Advantages of the Treemap

- **Space-efficient**: The treemap layout effectively represents a large volume of data within a compact space, providing a holistic view of winners across gender categories.
- **Clear Hierarchy Representation**: The hierarchy of champions by country, combined with seed and wins, makes it easy to detect dominance patterns briefly.
- **Dual Metrics**: The use of size and colour to encode data adds an additional layer of information, making it easier to grasp both the frequency of wins and player success relative to their seed.

Disadvantages of the Treemap

- **Comparison Challenges**: It may be difficult to compare categories with small differences in size, particularly when the number of wins is close between players.
- **Scalability Issues**: When the dataset grows large, the blocks may become too small for effective visual comparison.
- **Complexity for Beginners**: For viewers unfamiliar with treemaps, interpreting both colour and size dimensions can be complex without guidance.

Handling of Null Values

In the treemap, a calculated field was created to handle missing Champion Seed data. Use **IFNULL()** to replace the null values in the Champion Seed column with a custom value called "Unseeded." This assures complete visualization since unseeded champions are being captured and visualized in the treemap. The colour coding then will correctly depict the performance of all the champions, as will the size of the blocks. This approach maintains data integrity while allowing a clear representation of missing values, scoring extra points for completeness.

3.2 Parallel Coordinates

A **Parallel Coordinates** chart was used to illustrate the comparisons between champions of the Australian Open by several performance metrics. These charts enable the visualisation of relationships among multiple variables such as Games Won and Lost, Win Ratio, and Set Win Ratios that permit the drawing of detailed comparisons of performances across several dimensions. In this report, parallel coordinates have been applied to the champions, segmented by gender. This chart effectively highlights the general trend of Games Won and Lost, Win Ratios, and Set Win Ratios for Sets 1 through 5, underling the gender divide in performances between male and female champions.

- **Dimensions**: Gender, Champion's Name, Champion Nationality
- **Measures**: Normalised (Win Ratios, Set Win Ratios)

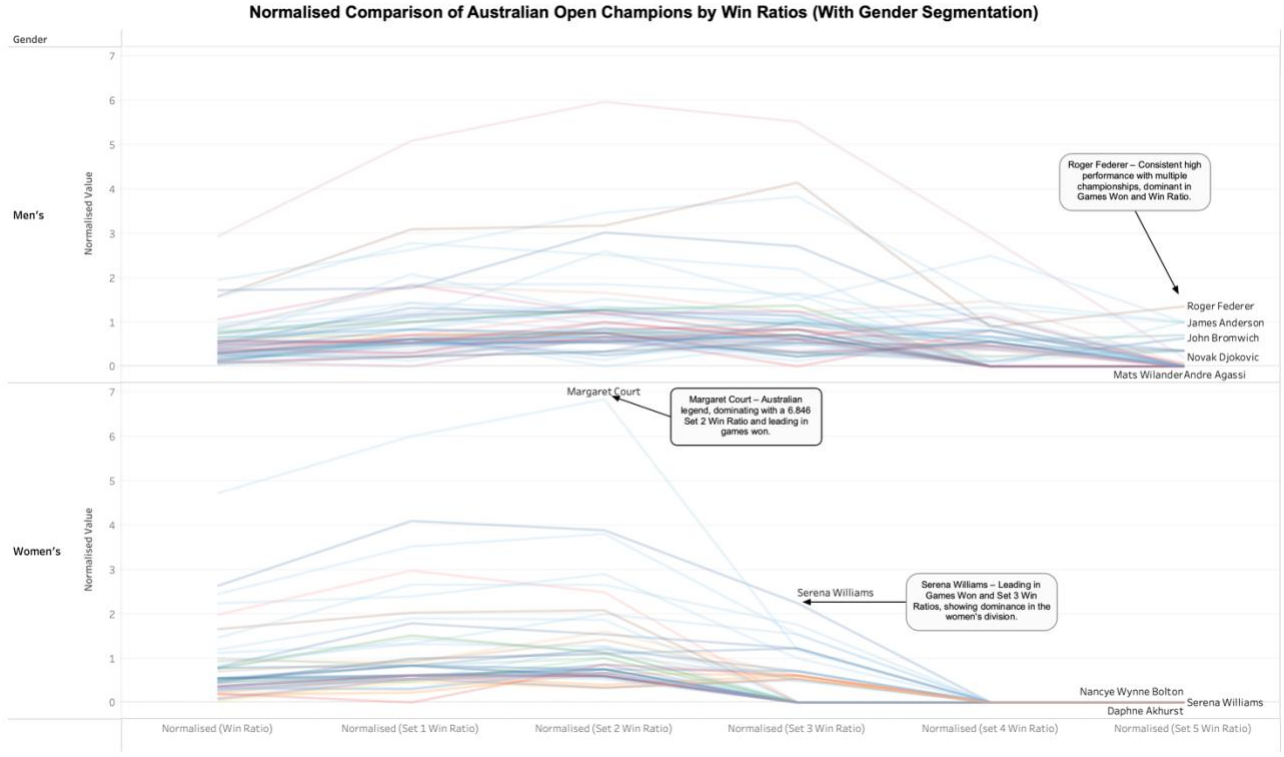

Figure 3. Men's and Women's Australian Open Parallel Coordinates by Games Won, Lost, and Win Ratios (With Gender Segmentation)

This parallel coordinate chart (Figure 3) visualises the performance of male and female champions across multiple metrics, including games won, games lost, and win ratios. The chart normalises these values against the same scale bias created by different metric scales that have been removed.

It is a line graph depicting the players individually, divided into segments by gender, thus showing a direct comparison in the men's and women's performances. Key champions, such as Roger Federer, Serena Williams, and etc…, are highlighted to emphasise their consistent dominance. This approach effective in observing a performance trend, including changes sets and provides an overview of the success of player within the tournament.

Advantages of the Parallel Coordinates

- **Multi-Variable Analysis**: These parallel coordinates setting enables multiple variables to be assessed together in such a way it is easy to notice which champions have performed across multiple metrics.
- **Gender Segmentation**: Having different visualisations for men's and women's categories, in turn, has allowed the chart to convey more clearly gender-specific performance trends.
- **Highlighting Key Players**: The ability to annotate and highlight certain champions, such as Federer and Williams, brings into immediate view which players are dominant in which categories.

Disadvantages of the Parallel Coordinates

- **Overlapping Lines**: When dealing with a large dataset, the lines of the parallel coordinates chart intersect, making it hard to discriminate between performers. It can be tuned by adjusting opacity, but it can still be a problem for reading.
- **Scalability**: The graph carries a lot of information, but over-populating the same thing makes it chaotic. It may be useful to limit the number of metrics included on the plot or modify line thickness based on that.

Normalisation Calculation

Normalization is the first step in the work, where all the performance metrics are brought to comparable scales. In this regard, all performance variables were normalized using the formula:

Normalized Value = (Original Value - MIN(Value)) / (MAX(Value) - MIN(Value))

This formula normalizes each value to 0 and 1; in this way, all metrics can be consistently compared. Normalizing Win Ratio and Set Win Ratios (Set 1 through Set 5) The parallel coordinates chart displays all dimensions on a common scale. Identifying performance trends and anomalies becomes easier, regardless of the original measurement scale.

Normalization ensures that all key indicators, such as Win Ratios, are meaningfully comparable and differences in the performance of players across metrics are visually apparent.

Handling of Null Values

In the parallel coordinates chart, nulls appeared in the Set 3, Set 4, and Set 5 Win Ratios columns. These nulls were created by the matches that did not extend to the later sets. Using the **IFNULL()** function set these null values equal to zero to indicate that the match ended before then. Such an approach is efficient in retaining the visual continuity of the chart for visualization of gaps in the dataset. That also underlines the matches that ended before the later sets, which shows

the performance trend based on the length of the match. The use of a calculated field will guarantee the accuracy of visualization and the inclusion of all applicable data points.

3.3 Geographic Map

The **Geographic Map** was used to represent the champions of the Australian Open regarding their nationality and total games won across the world. This visualization shows shifts in dominance by country over time, which enables regional trends in tennis successes to be more easily communicable. Consequently, mapping the nationality of champions shows a clear view of where the top players have come from, enhancing the reader's further understanding of global dominance in the sport.

- **Dimensions**: Champion Country, Champion Nationality
- **Measures**: Games Won

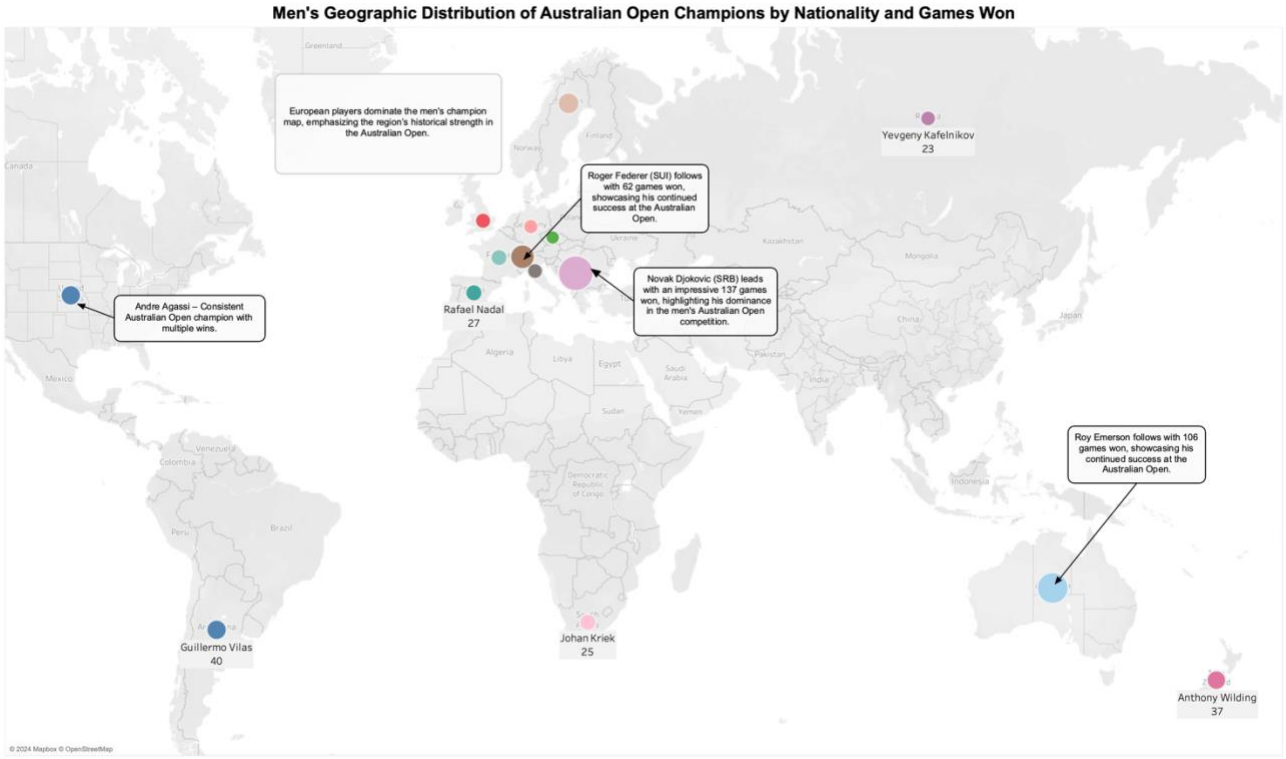

Figure 4. Men's Australian Open Geographic Map by Nationality and Games Won

The geographic map (Figure 4) visualises the global distribution of Australian Open men's champions by nationality and the total number of games won. Each champion's country is represented by a circle whose size is proportional to the number of games won. This map showcases regional dominance, highlighting the stronghold of European and Australian players. Key champions, such as Novak Djokovic, Roger Federer, and Roy Emerson, are highlighted with annotations that supply a spatial understanding of player success. This is an intuitive glimpse into the global nature of the tournament, with a focus on the dominance of players from certain regions

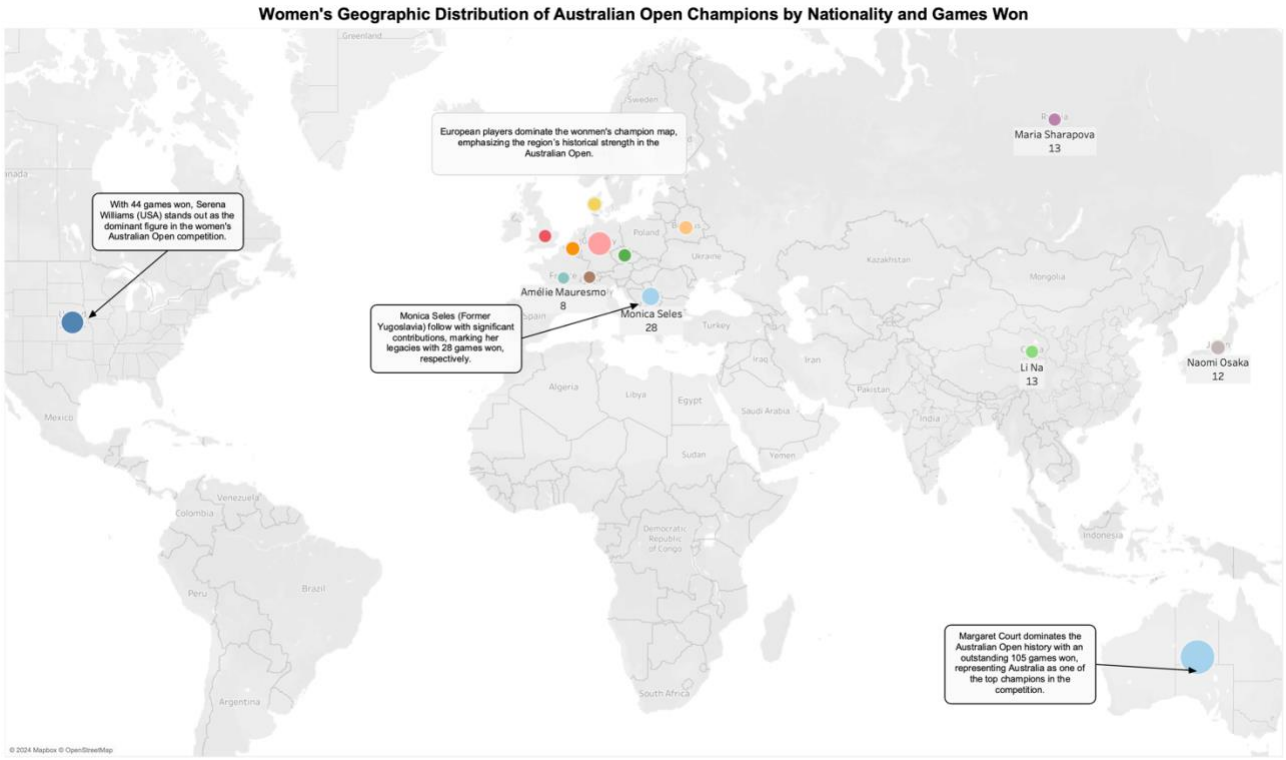

Figure 5. Women's Australian Open Geographic Map by Nationality and Games Won

Similar to the men's map this visualisation (Figure 5) plots the global distribution of Australian Open women's champions, with each circle representing a champion's country and the number of games won. Indeed, dominant players in this map include Serena Williams and Margaret Court. Also, one can compare not only performances but also nationality. This allows users to perceive trends in the geographic representation that takes place and the global diversity of the tournament. The size of circles immediately allows comparison in the visual sense of player success; it lets the user see which regions have produced the most champions over time.

Advantages of the Geographic Map

- **Global Context**: The map is an overview of the geographic representation of tennis champions, which brings clarity into which countries have been able to produce top-performing players over time.
- **Intuitive Visualisation**: The geographic map serves as a quick way for viewers to understand; therefore, it is an easy method of comparison for performance by country via spatial location and size variation in the visual elements, for example, circles.
- **Highlighting Key Players**: Annotation were used to emphasise top players, such as Novak Djokovic and Serena Williams, drawing attention to the standout athletes in each category.

Disadvantages of the Geographic Map

- **Limited Detail**: This map provides an effective overview of the distribution of champions by country. It is, however, not detailed enough for any deeper analysis other than nationality and the total games won.
- **Overlapping Regions**: If multiple champions came from the same country, the circles would overlap each other, not clearly distinguishing between players and capturing their contributions.

Handling of Unrecognised Geographic Data

During the creation of the Women's Geographic Distribution map of Australian Open champions by nationality, an issue arose with unrecognised geographic data for Yugoslavia. The country of Yugoslavia no longer exists but is supplanted by several successor states, so the entry could not be recognised. Considering this, the entry was recorded manually in Serbia to show the information accurately (Central Intelligence Agency, n.d.). This correction makes the geographic distribution complete and accurate and provides an accurate representation of champions from the former Yugoslavia region. Since this has been tackled, all nationalities are reflected clearly in this visualisation, retaining the integrity of the data analysis.

3.4 Scatter Chart

The **Scatter Chart** is a valuable tool for identifying relationships and distributions between two key variables. This report assesses the relationship between champions' games won and the games lost by categorizing them by the champion seed and gender. By plotting the games won against one axis and those lost against the other, viewers can easily spot how champions perform relative to their seeding. Highly seeded champions, like Novak Djokovic and Roger Federer, often come with a high number of games won to show dominance in the tournament.

- **Dimensions**: Champion's Seed (Null Correction), Champion's Nationality, Gender
- **Measures**: Games Won, Games Lost

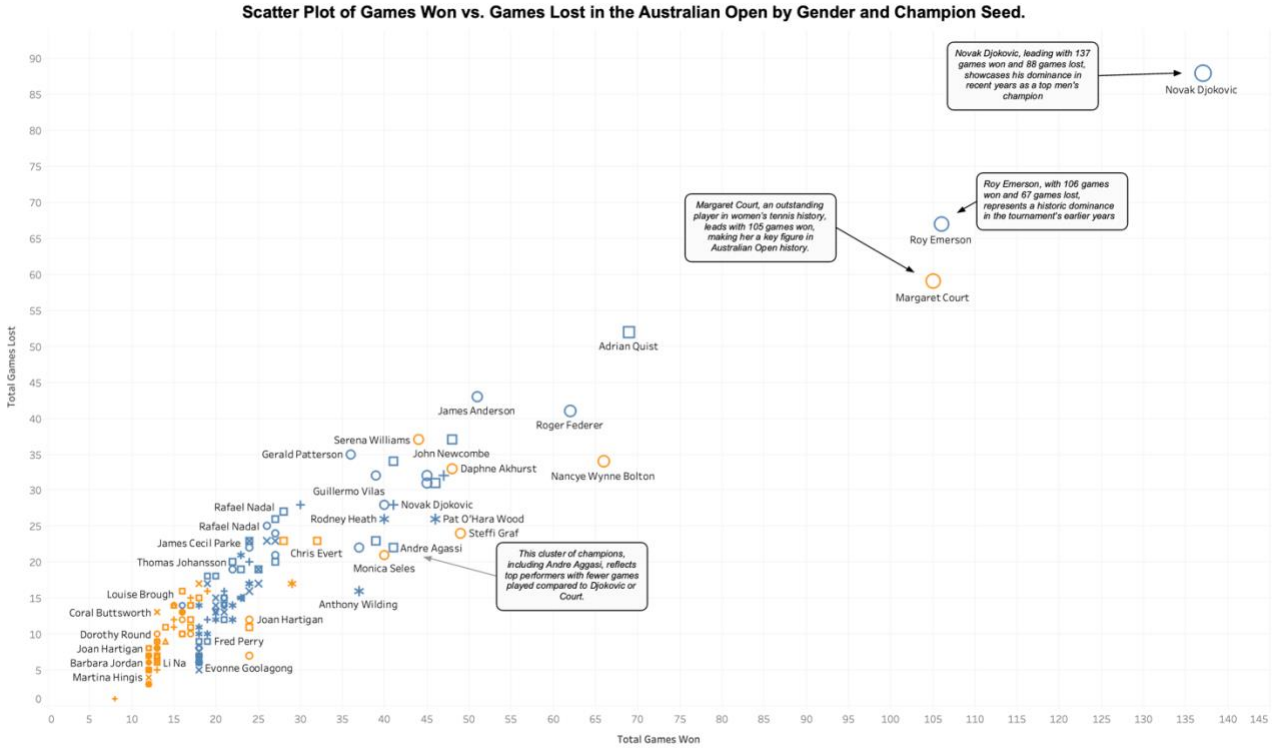

Figure 6. Men's and Women's Australian Open Scatter Chart by Gender and Champion Seed

This scatter chart (Figure 6) visualises the relationship between games won and lost by Australian Open champions, segmented by gender and seed position. Each player is represented with a point in this scatter chart, where the size of the point shows the number of games won by each player, while

the shape of the point depicts the seed position. Colour doing separates men from women; blue for men and orange for women, allowing easy visual comparison of performances across genders. This works particularly effectively to highlighted outliers; Novak Djokovic won about twice as many games as anyone else the most, and Margaret Court also won far more games than anyone else. In the scatter chart, it is clear how high seeding is related to good performance, but also how variable the unseeded champions are.

Advantages of the Scatter Chart

- **Correlation Identification**: It will be helpful in establishing the relationship between the seed, games won and lost.
- **Outlier Detection**: It identifies outliers such as Novak Djokovic, who has been winning many games while losing a few.
- **Gender Comparison**: The colour used for gender will make it easy to compare the trend of the performance of gender in general.

Disadvantages of the Scatter Chart

- **Clutter**: With many data points close together, some points may overlap, making it difficult to differentiate between players with similar records.
- **Limited Scope**: The scatter chart only compares two variables at a time, making it less ideal for multi-dimensional comparisons.

Handling of Null Values and Outliers

In the scatter chart, null values for Champion Seed were treated by creating a calculated field. Using the **IFNULL()** function, null values in the Champion Seed column were assigned the label 'Unseeded,' ensuring that unseeded champions were included in the visualization. This approach guarantees that every champion, seeded or unseeded, is represented in the chart, allowing for a complete analysis of player performance. By accounting for missing data, the scatter chart maintains data integrity and provides a full view of the distribution of games won and lost.

3.5 Explanation of Champion Seed on Games Won

The Champion Seed field plays a critical role in understanding how Games Won are distributed in the tournament, as it reflects a player's ranking and likelihood of performing well. Higher-seeded players are typically expected to win more games, while unseeded or lower-seeded players are generally less likely to progress in the tournament. Including this dimension in the analysis uncovers key insights about performance trends, offering a more nuanced view of the competition. For example, high-seeded players like Novak Djokovic and Serena Williams consistently win more games, which aligns with expectations. In contrast, lower-seeded and unseeded champions such as Andre Agassi and Monica Seles emerge as important outliers, performing beyond what their seeding might predict. This contrast highlights the importance of including seed data to fully understand the dynamics of player success in the Australian Open. By incorporating Champion Seed, the analysis ensures that all player performances - whether expected or surprising - are represented accurately, providing a more comprehensive understanding of how players meet or exceed expectations based on their initial seeding.

4 Top Players Performance Analysis

Analysing the performance of the top players in the Australian Open, several key trends stand out, revealing the dominance of both male and female champions.

Men's Division

In the men's division, Novak Djokovic leads with the most games won, a trend evident across multiple visualisations. In the treemap, he has the biggest block due to his high number of wins. This is further demonstrated in his high seeding and very few losses of games, making him an outlier in the scatter chart for great consistency. The parallel coordinates chart enforces the view of top-tier performance across different metrics but mainly on his high win ratios.

Roger Federer and Roy Emerson also performed strongly. Federer's win/loss ratio is balanced, as shown in the scatter chart, where his high seeding consistently aligns with his results. Emerson is an outstanding player from the earlier years who also reflects good performances in the treemap and scatter chart, in which his performances are visually highlighted.

Women's Division

Margaret Court and Serena Williams lead in the women's division. Court is the most successful player, with a record 105 games won, as reflected in the treemap and scatter chart. As far as the parallel coordinates chart goes, it further shows her efficiency continuously wins with fewer and fewer losses.

Serena Williams' modern-day dominance is reflected in her 44 games won represented by her outlier position in the scatter chart. Her contribution to the sport is mapped on a geographic map, showing the achievements of Serena Williams, depicted by an unusually large representation of her wins.

Outliers and Trends

While high seeding generally correlates with greater success, as seen in players like Djokovic, Emerson, Court, and Williams, there are notable outliers who defy this trend. Andre Agassi and Monica Seles are prime examples, as they performed exceptionally well despite lower seeding. In the scatter chart, these players are positioned in areas that reflect unexpected performances, where they won substantial numbers of games despite not being top seeds.

These visualisations not only present the relationship between seeding and performance but also give an understanding of the most important outliers and trends, with Djokovic and Court always within the top echelons in most of these metrics. Performance analysis of top players therefore gives a total view of their lasting impact on the Australian Open, propelled into greatness by their remarkable consistency, high win ratios, and strategic plays.

5 Executive Summary

This report presents a comprehensive analysis of the performances of tennis champions during the Australian Open over the past 120 years using advanced data visualization techniques. Key variables such as player nationalities, champion seeds, games won, and win ratios were analysed to enable visual comparisons across gender, nationality, and performance metrics.
Visualisation techniques such as treemaps, parallel coordinates, geographic maps, and scatter charts were employed to uncover trends, patterns, and outliers in player performances.

Key observations include the consistent dominance of Novak Djokovic and Margaret Court, as their superior performances were highlighted through multiple visualisations. Treemaps showed national dominance by highlighting which countries produced the most champions. Parallel coordinates enabled a multi-metric comparison, revealing how key players like Djokovic and Court maintained consistent win ratios. Geographic maps helped illustrate regional trends of tennis success, showing the historical dominance of Australian and European players. Scatter charts showcased outliers such as Andre Agassi and Monica Seles, whose lower seed positions contradicted their outstanding performance.

Each of these methods in visualization helped to tell a continued story about the history of the tournament: from national dominance and top player performance to the players performing well with low seeding that introduced some level of unpredictability. This report highlights how the correlation between seeding and success generally holds, yet exceptions are evident through specific outliers. The comparison of top players' performance, including Djokovic, Court, and Williams, is well-concluded and provides insight into their sustained dominance. Overall, the visualizations used in this report offered deep insights into patterns and outliers, making it an effective tool for sports analytics.

6 Conclusion

This report provided a comprehensive analysis of players' performances at the Australian Open, using multiple visualisation techniques such as treemaps, parallel coordinates, scatter charts, and geographic maps. These visualisations effectively uncovered major trends, patterns, and outliers throughout the tournament's history, making it clear how different metrics—such as games won, seed positions, and nationalities—correlate with success.

Breakthrough points included the domination of Novak Djokovic in the men's division and Margaret Court in the women's division. Djokovic's consistency in winning a high number of games while losing very few seems crystal clear in both the scatter chart and parallel coordinates, while the dominance of Court is underlined in both treemap and geographic maps. Serena Williams' record of 44 games won further cemented her as one of the all-time greats, clearly shown in both treemap and scatter chart visualisations.

However, striking outliers like Andre Agassi and Monica Seles broke that expectation when they won several games as lower-seeded players. Their deviations are most apparent on the scatter chart because their performances run contrary to the typical correlation between seeding and success. These outliers are an interesting counter-narrative to the general trend of seed-based performance.

Each of these techniques of visualisation told part of the story in terms of the history of this tournament and player performances. Treemaps showed the dominance hierarchically by nationality and seed, and the parallel coordinates chart compares multi-variables across key performance metrics. A geographic map intuitively provided regional dominance, showing which countries have constantly churned out strong champions. The scatter plot broke down the relation between won and lost games and pointed out some extreme outliers, Agassi and Seles.

Another strong aspect of the analysis was the labelling and annotating involved in walking the audience through the analysis. Labels and annotations on the scatter plot made pointing at outliers easier; color-coded treemaps and parallel coordinates visualisations made the comparison of the dominance by nationality and seed more accessible.

While the visualisation techniques were largely successful in conveying key trends, limitations such as overlapping records in scatter charts (due to similar player performances) were noted. The combination of all the above techniques formed a comprehensive toolkit for data analysis in complex datasets such as this Australian Open.

Overall, seeding has been a reliable predictor of success at the Australian Open, though notable exceptions, such as Agassi and Seles, provide compelling counterpoints. These findings have been represented clearly and in an easily understandable format through treemaps, parallel coordinate scatter charts, and geographic maps, reinforcing high-level data visualization in sports analytics. Overall, these were quite complete visualisations for the understanding of player success at the Australian Open for the fulfillment of this report's objectives.

7 Reference

Central Intelligence Agency. (n.d.). *Serbia – Country Factsheet*. https://www.cia.gov/the-world-factbook/countries/serbia/factsheets/