UNIVERSITY OF TECHNOLOGY, SYDNEY

DATA VISUALISATION AND VISUAL ANALYTICS
Assignment 2: Advanced Data Visualisation

Table of Contents

1. Introduction

This report is to serve as advanced data visualisation practice in the context of the Australian Open tennis tournament through the years from the beginning to the year 2023. To have a comprehension interpretation of the dataset, the following aspects are going to be conducted accordingly:

- Data exploration: Including identifying data formats, values, characteristics, as well as any trend or outlier in the dataset
- Additional data transformations or calculations associated with the visualisation
- Visualisations and analytics in terms of gender, nationalities and changes across time
- Spotting out and deep analytics in top-end contenders with five championships or more.
- Summarising all of the findings and analytics

Tableau along with Microsoft Excel is utilised for the project. While Microsoft Excel is mainly used for additional transformations and calculations, visualising techniques like treemap, parallel coordinates, geographic map, and scatter plotting from Tableau are applied in order to leverage the dataset readability. Finally, the report concludes with the analytics summary and tools advantages and disadvantages.

2. Dataset exploration

The Australian Open tournament, one of the four most famous grand slams, is hosted annually in Australia since 1905 (except for the period 1916-1918 and 1941-1945 due to World War I and II, respectively). With 118 years and so forth of history, the grand slam has been held 110 times in 5 Australian cities (Sydney, Brisbane, Melbourne, Perth, Adelaide) and 2 New Zealand cities (Christchurch and Hastings) (Australian Open, 2023). For the purpose of having deep insight and analytics of this tournament with respect to men's and women's singles features, a dataset containing the information of grand finals for all years has been provided. In total, it has 19 columns showing the year, champion runners-up's names, nationality and country, and the match score, along with score details for each set. For the score, the table shows in two forms, the first shows the overall result of the match, and the second one breaks the match score down into the score of the champion and the runner-up for each set. For the row number, the dataset has 208 rows, meaning 208 data points. In the original dataset, all of the columns have been formatted in general form.

For column 'Year', it has values throughout the history of the tournament, ranging from 1905 to 2023, except for the period 1916-1918 and 1941-1945 due to World Wars and the year 1986. From the present year 2023 back to 1922, the year distinct values have appeared in pairs and have appeared once for the initial era, from 1922 to 1905. When aligning to the next column, 'Gender', a pattern has been recognised. In this column, there are only two distinct values, 'Women's' and 'Men's', allocated for each distinct year value, but not from 1905 to 1921. In this period, this tournament was only held for men. Women contenders have only been participating in their own gender's champion title since 1922. For that reason, men's champion title numbers should be more than women's seventeen units. For the column showing 'Champion', the column's values are the name of the champion for the year. The column 'Champion' is followed by the column 'Champion Nationality' having country acronyms composed of 3 letters, and the column 'Champion Country' with country name values. For runners-up, we have three homogeneous columns sharing the same characteristics. With 'Score' column, via values,

we could read the number of sets of the match, and each set's score. Some matches have the scores put in parentheses, meaning there are sets decided by tie-break games. Additionally, outliers protrude in this column. There are 'retired' values in 3 matches in 1965, 1990, and 2006. We also have another outlier having value 'walkover' in 1966. Following this column are corresponding columns that have the score of the champion and the runner-up from the first set to the last set. For the column title header, it has the front part is the order of the set, and the rear part is 'won' for the champion and 'loss' for the runner-up. Finally, the column pair 'Win' and 'Loss' represents the score of the champion and runner-up in the match.

This dataset has ubiquitous data dimensions; thus, it is challenging to capture any trend or patterns underlying in gender, scores, or individuals. To facilitate these issues, extra transformations and calculations have been conducted to serve visualisation and uplifting readability.

3. Transformations and calculations

First of all, we need to transform the format of specific columns for the calculation purpose. All columns relating to the score of the champion and runner-up in each set and total are converted into numerical format. The rest have remained.

To have decent visualisation, a couple of dimensions need to be derived from inherent columns. Firstly, the 'Win Rate' column is calculated as a ratio of column 'Win' over sum of column 'Win' and 'Loss', indicating to what extend the champion overwhelms the runner-up. This column's value is set as numerical data, with 2 digits for the decimal part. Then columns 'Sets' and 'Tie-break sets' are calculated from column 'Score', showing the number of the total sets and sets decided by tie-break game in one match. These columns are formatted as numerical data, integer specifically. Next, all numerical data columns are normalised, scaling to the range from 0 to 1 for further visualisation purposes. These newly normalised values are contained in relevant columns with the title header prefixed by 'Norm'. These normalised columns are in numerical format, with 2 digits in decimal parts. For individual champions comparison, another sheet named 'Champion performance' has been generated to calculate the performance of each champion through years, by calculating average scores in, average win rate, normalised average scores, and normalised average win rate. Although most of the calculations and transformations are conducted in the Excel, in fact there are other calculations have been done directly in Tableau. There are 'Pair' as another dimension that are created to have distinct pairs of players that have had match together regardless of title 'Champion' or 'Runner up'. There also are calculated fields to calculate how many titles a champion has, a nationality has, and how many times a player participated in grand final matches as runner-up.

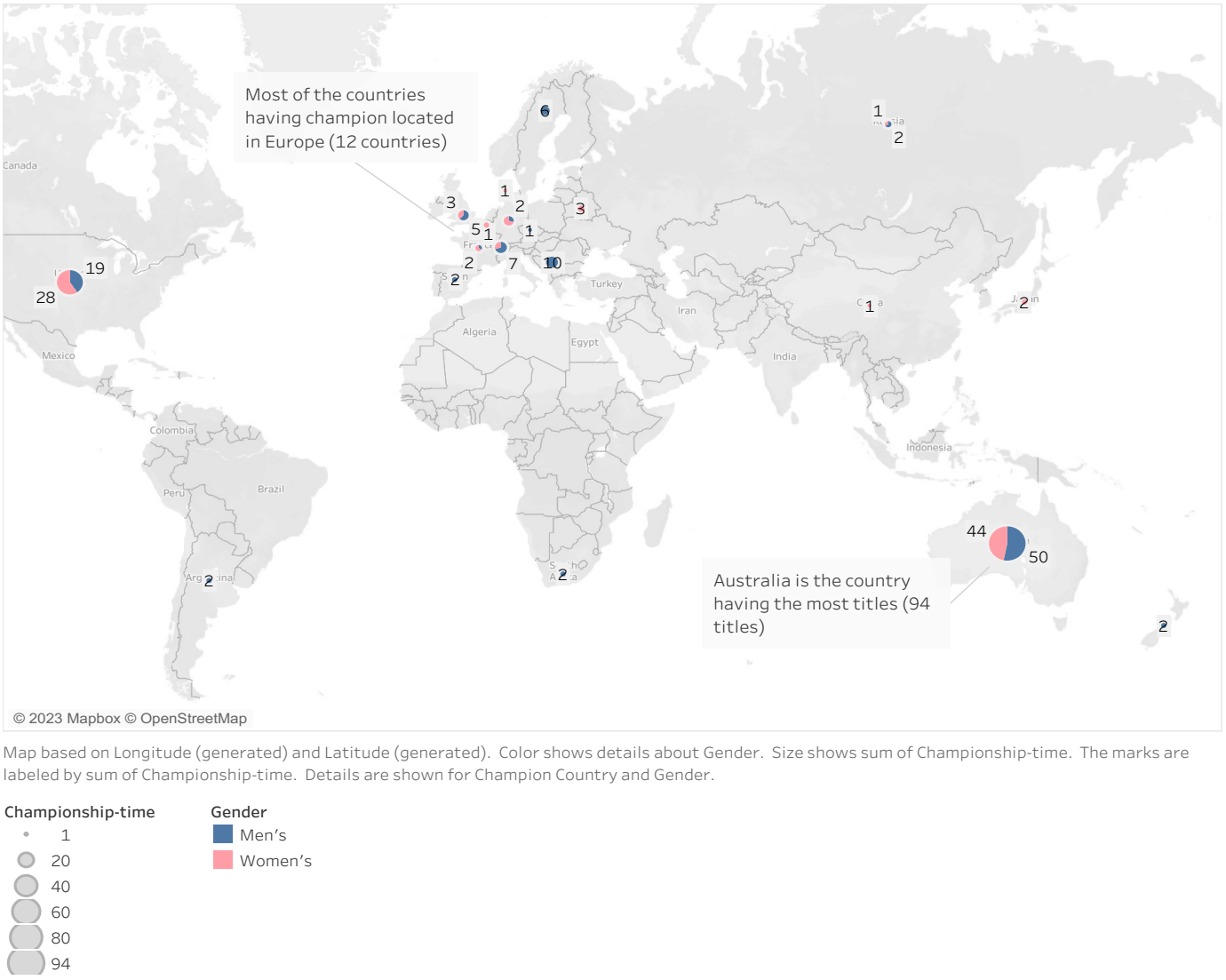
4. Visualisation and Analytics.

The first aspect that is visualised is the title number across the world. For the visualisation, the dataset with calculations and transformation has been input into Tableau for the data source. In this graph, the geographical graph technique has been utilised to plot the number of championship titles number associated with countries. To plot the graph, longitude and latitude values have been inserted as column and row, respectively, indicating the location of countries on the map. This plotting can be achieved due to Tableau has automatically interpreted values in columns 'Champion Country' and 'Nationality Country' as geographical format, For that reason, each country has been

plotted as a circle on the map. Depending on the titles number, the circle size of each country varies, getting bigger with higher number. For more detail, 'Gender' is marked with color, turning the circle into pie chart, showing the proportion of titles number in terms of men and women with blue for men and pink for women. Additionally, the numbers have also been labelled beside the circle in order to enrich the readability of the graph. Looking at the graph, Through 118 years of the Australian Open, there have been 21 countries having champions across the world, more than half of them are located in Europe, with 12 countries. For other continents, America has two countries which are the USA and Argentina; South Africa represents Africa, and Asia has Japan and China.

Last but not least, continent Australia contributes 2 countries: Australia and New Zealand. Looking at figure 1., it is obvious to recognise that Australia holds the most giant circle, equivalent to the most titles of the tournament so far, with 94 titles, to which the men's champion has contributed 50 titles, and the rest comes from women's champions. Australia is followed by the USA with 47 titles, 19 from men's and 28 from women's. Other countries hold titles ranging from 1 to 10.

Titles number across the world



Map based on Longitude (generated) and Latitude (generated). Color shows details about Gender. Size shows sum of Championship-time. The marks are labeled by sum of Championship-time. Details are shown for Champion Country and Gender.

Figure 1. Title numbers of countries across the world

To have deeper insight in how the titles are distributed through the history of the tournament, another graph has been plotted to serve this purpose.

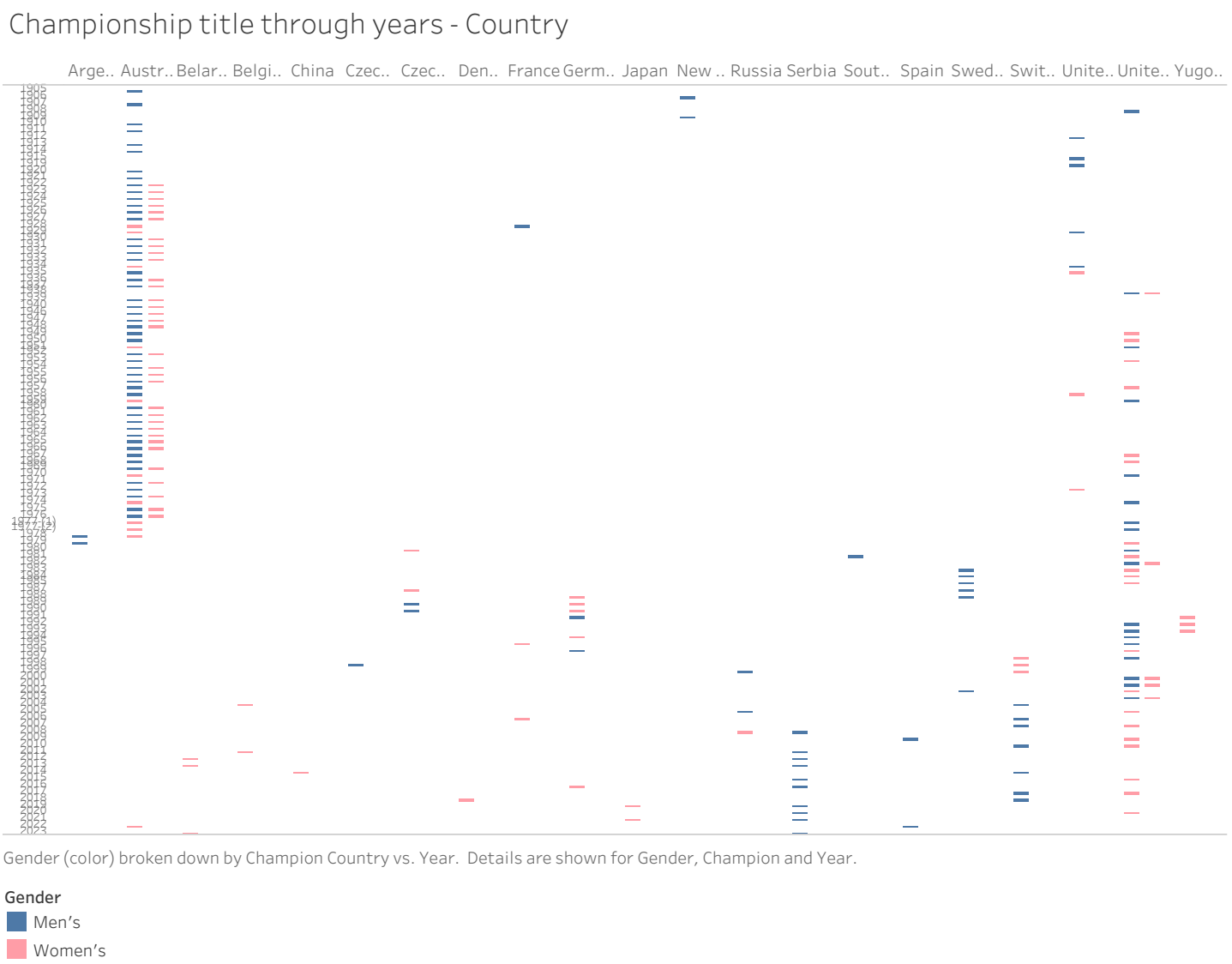


Figure 2. Title distribution through years

To have this graph, scatter plotting technique has been used, along with extra marks in genders and years. The graph figure 2 has vertical axis is year values and country names for the horizontal axis. While one dash is counted for one champion, the blue colour indicates male, and the pink indicates female. For better understanding, 'Champion' and 'Year' have been marked as auxiliary details in the label. Australia prevailed in the period from 1905 to 1922, with many years having titles for both men and women in the same year. Since then, there has been almost no title for Australia, except only one women's title in 2022 as the latest one. For the second half of the Australian Open timeline, the USA has surged as a heavy contender, challenging the stature of Australia, primarily in women's feature. In the 1980s, Sweden had 5 titles with 3 consecutive times from 1983 to 1985 and two other years in 1987 and 1988 in men's feature. From the 2000s to now, there has been a rally between Switzerland and Serbia for the men's championship, with seven times for Switzerland and ten times for Serbia. In 2023, the Belarusians won the women's title, and the men's championship belonged to Serbia for one more time.

After scheming on the country field, now it's time to examine individual achievements. In fact, there are interesting findings and insights that are spotted out and interpreted via other graphs with different plotting techniques.

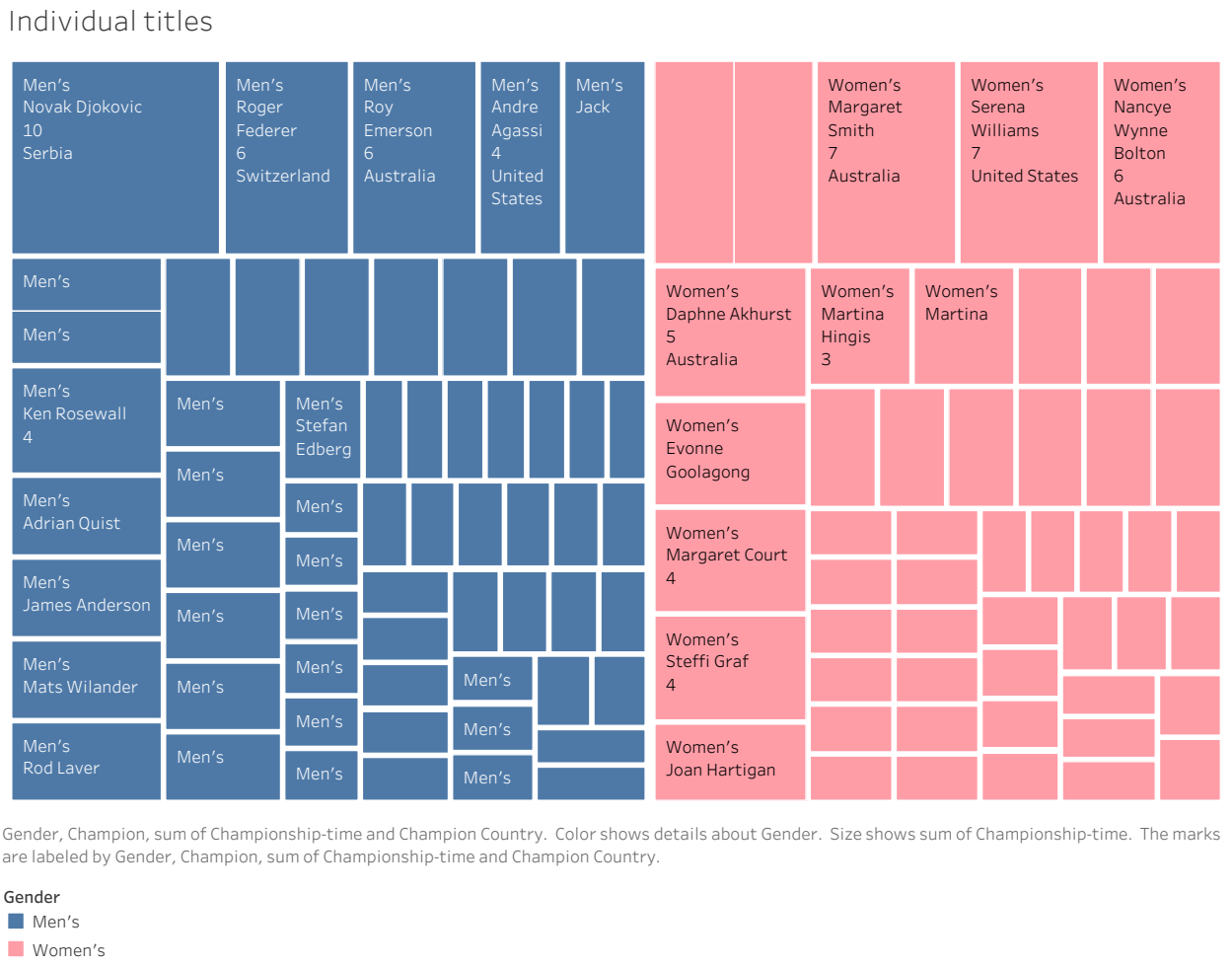


Figure 3. Individual titles

This graph used treemap plotting technique for multidimensional visualisation to find out top-end champions. With this technique, the critical dimension is the titles number of each champion from which the size of the box is measured proportionally, the higher the number is, the bigger the box is. Colour is added to express the gender, splitting the champions into two big patches in blue and pink. Furthermore, country is also added as label for the information supplement. There are only a handful of champions with astonishing profiles winning the championship 5 times or more. For men, so far the most outstanding player is Novak Djokovic from Serbia, with up to 10 Australia Open champion titles in his career. Djokovic is followed by Roger Federer from Switzerland and Roy Emerson from Australia, with 6 titles for each player. When it comes to women, we have two tennis players sharing the 1st place with 7 championships, one from Australia, Margaret Smith, and the other is Serena Williams from the USA. The second position is occupied by Nancye Wynne Bolton, another Australian tennis player. Daphne Akhurst, an Australian tennis player as well, has taken the third rank with only 5 titles. The title number seems to be an excellent metric to evaluate champions, but in order to have a comprehensive assessment, only the title number is insufficient. Hence, another graph is drawn to demonstrate player performance more thoroughly.

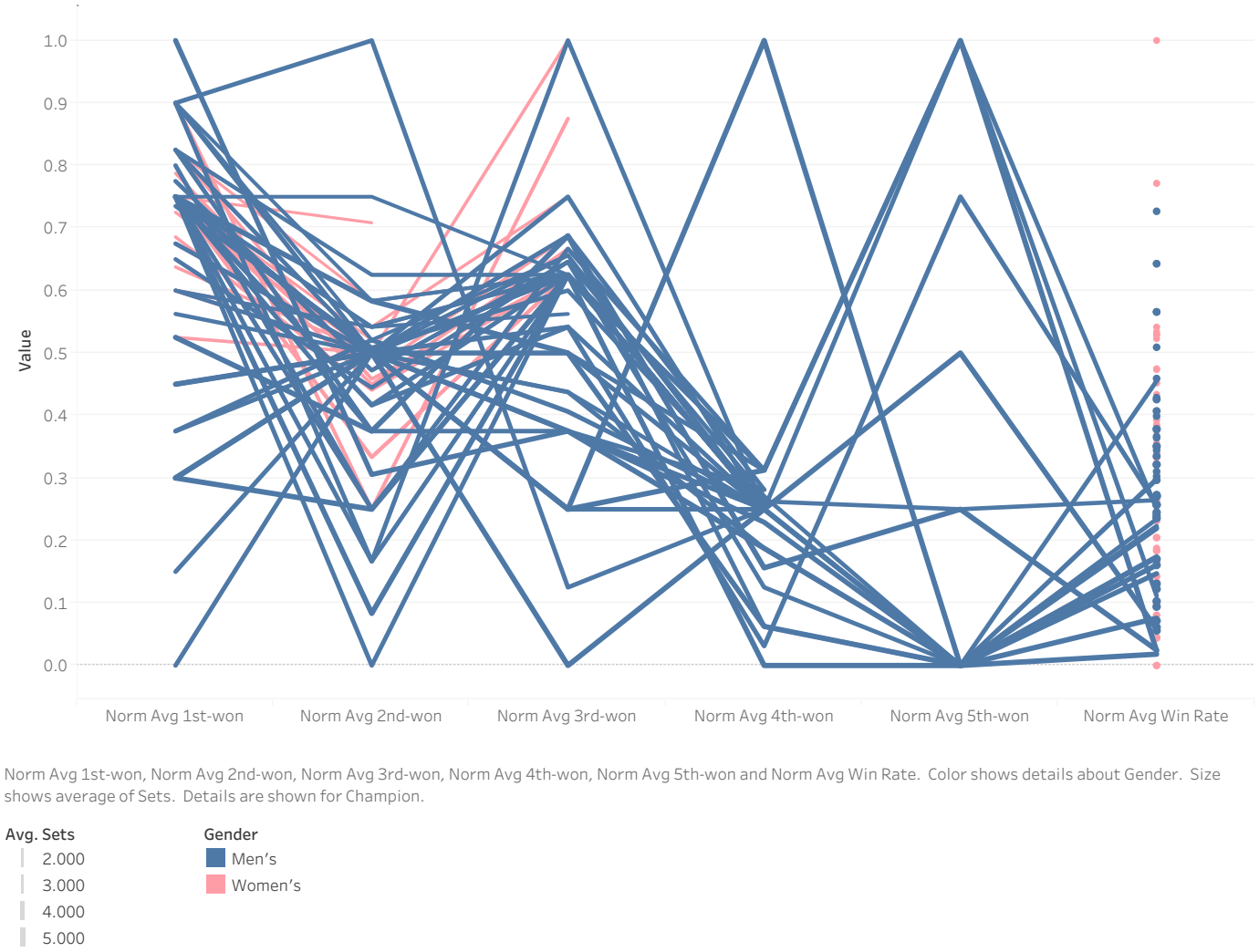


Figure 4. Champion performance

This figure shows the overall performance of champions in terms of Average score in each set and the average win rate. The technique utilised for this purpose is parallel coordinates plotting graph. Normalisation has been applied to all columns in order to scale values in each column to one identical range from 0 to 1. This helps better visualisation when the actual values of 'Avg Win Rate' column is relatively small comparing to ones of other columns, making it hard to see the distribution. Six columns, including normalised average won columns for each set with one normalised average win rate column, are evenly distributed horizontally, representing six dimensions in the form of six axes. Lines connecting values on each column indicate champions. Additionally, blue lines are men's champions, and pink lines represent women's champions. Another dimension in this graph is the thickness of the line; the thicker the line is, the more average sets the champion has for each match. One interesting finding is the maximum sets of female matches is only up to 3, while 5 for men's match. According to the tennis rules applied in this Australian Open tournament, one who reaches point 6 in the set first, by 2, will win that set, meaning the closer the champion's average point for each set to 6, the more stable the champion's performance during the whole match. For the normalised range, 6-point is 0.750, 0.500, 0.625, 0.250, and 0.000 for set 1 to set 5. Mostly values are champion's performances are condensed around these values, but there are still outliers, for Avg 1st-won, the actual average score is up to 7.667 for male champion James Anderson, For other columns, they are male champion Ken MacGregor with 12 points for

the 2nd set average point, female champion Mary Carter Reitano and male champion William Reitano that share the highest 3rd-set average point, up to 9 points. For the 4th-set, it is 18 points for male champion Gerald Patterson; really incredible!!!. For the 5th-set, they are Dinny Pails, Rod Laver, and Mats Willander with 8 points. The champion holds the highest average win rate, 0.8889, Amélie Mauresmo. The justification for this rate is that her opponent retired, and she won with the overall score: 6-2, 2-0. For the greatest of all time, Novak Djokovic, according to his specification, 5.9, 6.1, 5.8, 6.2, 6.5 for 1st-set to 5th-set average point, it seems that usually, he doesn't play well on the first and third set, and tend to beat his contender at the fourth and the fifth set. This conclusion is fortified by his average win rate, only 0.6029, which is in the moderate position in rank. He may usually win the title by using his durability.

For the women's feature, Margaret Smith and Serena Williams, who have the same highest title number, 7 times, are examined in the performance aspect. The former Margaret Smith usually won the first two sets to win the title, with 6.0 for the average 1st-set and 2nd-set point. The average win rate is also impressive, 0.7034. On the contrary, Serena Williams tends to extend the match to the third set, with the average points is 5.571, 5.286, and 6.000 for 1st-set, 2nd-set, and 3rd-set, respectively. Additionally, her average set number is 2.429, higher than 2.143 of Margaret Smith. Serena Williams's average win rate is not very impressive, only 0.6365.

Finally, there are some findings in the average win rate. While female champions have the highest average win rate up to 0.8889, the highest average win rate that male champions could reach is only 0.7826. And the average range of win rate is from 0.5172 to 0.6842.

Typically, the set is finished when a player reaches 6 or 7, by 2, but there are still exceptions in terms of points. To spot these exceptions, another parallel coordinates graph has been plotted out as below:

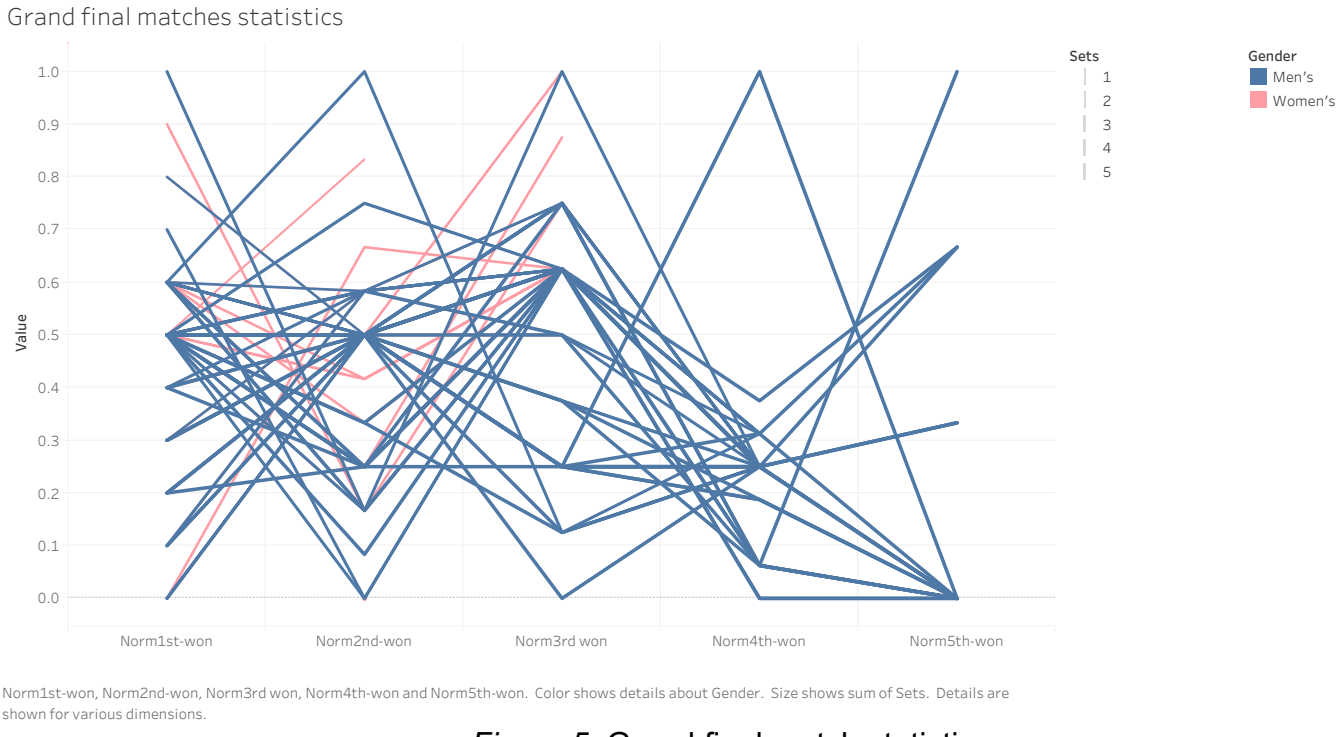


Figure 5. Grand final match statistics

In the Figure 5 graph, the same technique as Figure 4 is applied, parallel coordinates, with normalised 1st to 5th set points, and each line is one year of the Australian Open. While blue represents men's matches, pink indicates women's showdowns. Another dimension in this graph is the set number of the match, displayed via the thickness of the line. The thicker the line, the more sets the match has. Other information including 'Year', 'Champion', 'Runner-up', and 'Score' are also added to enrich the readability. For the 1st set, the record for the score belongs to the men' match in 1925 between champion James Anderson and runner-up Gerald Patterson, 11-9, and 4 sets. The 2nd set score record, 12-10, goes to the match in 1952 between champion Ken McGregor and runner-up Frank Sedgman. For the 3rd set, they are champion William Bowrey and runner-up Juan Gisbert in 1968 for men's feature; and champion Mary Carter Reitano along with runner-up Thelma Coyne Long in 1956 for women's feature. Both pairs share the score 9-7. In 1927, Gerald Patterson dominated his runner-up John Hawkes with the score 18-16 to set the score record for the 4th set. And when it comes to the final set, Andrian Quist has beaten Jack Crawford in 1936 with the score 9-7 to take the record.

For the glory of the champion, it is a mistake to ignore the runners-up. There is no champion without runner-up. To have a general statistic for the runner-up, a text could graph has been generated to show how many times a tennis player has attended the grand final match as runner-up.

Figure 6. Hall of runner-up

With this technique, all runner-up names have been plotted to form a cloud shape. The attendance number is used to adjust the size of the runner-up name, the higher the attendance number is, the bigger the name is. With this technique, it is feasible to detect

players with high attendance among the pool of runners-up. Colour is added, to mark the gender of the runner-up, with blue for men and pink for women. The runner-up attendance time was also added as an extra detail for the visualisation. Looking at figure 6, it is visible that Andy Murray and John Bromwich have the highest attendance time as runner-ups for men's feature with 5 times. When it comes to women's feature, the greatest runner-up is Esna Boyde, with 6 times. Furthermore, Esna Boyd is also the player who has the highest attendance time for both genders of all time as well.

Throughout the history of the Australian Open tournament, like any other sport, there are player pairs always expected by the audience for the grand final as their performance equilibrium fuels up the intensity of the match. To do the statistic for how many times one player has met his/her counterpart in the grand finale, a bar graph has been drawn to serve that purpose.

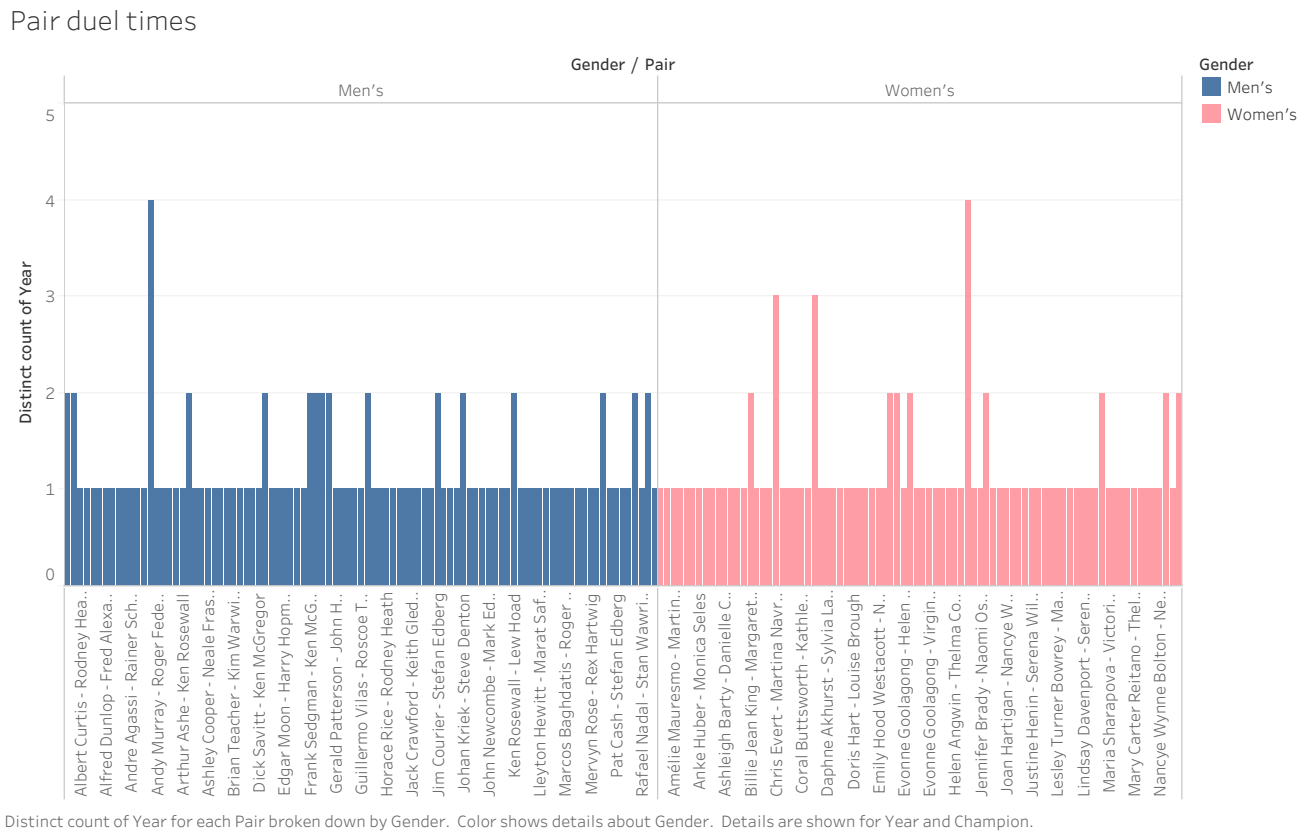


Figure 7. Pair duel times

To plot this graph, bar graph technique has been used. 'Gender' and 'Pair' have been put into the column shelf, and 'Year' in the row shelf with distinct counting. The bar graph, one of the most popular graphs, is used due to its simplicity and straightforwardness in the context of low dimensional space. The height of the bar helps to distinguish one pair from other pairs transparently. Colour is added to categorize pairs into genders accordingly, with blue for men's pairs and pink for women's pairs, elevating the data readability. 'Year' and 'Champion' are also added for supplementary information. Looking at Figure 7, it is obvious to notice that the pair 'Novak Djokovic-Andy Murray' has the highest duel times with 4 times in 2011, 2013, 2015, and 2016. All of the duels happened in the 2010s, showing the era of these two players. Looking back at the individual achievements, with 10 titles for Novak Djokovic, and no title for Andy Murray, this only confirms the greatness of Novak Djokovic for one more time. Andy Murray is still

one of the most outstanding tennis players, but maybe it was bad luck for him when his talent blossomed at the same time with the vivider flower Novak Djokovic.

For women champions, we have a similar pair Jan Lehane and Margaret Smith, with 4 duel times in a flush of 1960,1961, 1962, and 1963. Their talents have reigned for the first half part of the 1960s. Jan Lehane seems to be Andy Murray's female version, when she got no victory for all duels, and her attendance seems to be for polishing the stature of Margaret Smith.

Finally, in the last part of the visualisation and analytics, the intensity of the match is considered thoroughly with proper visualisation technique.

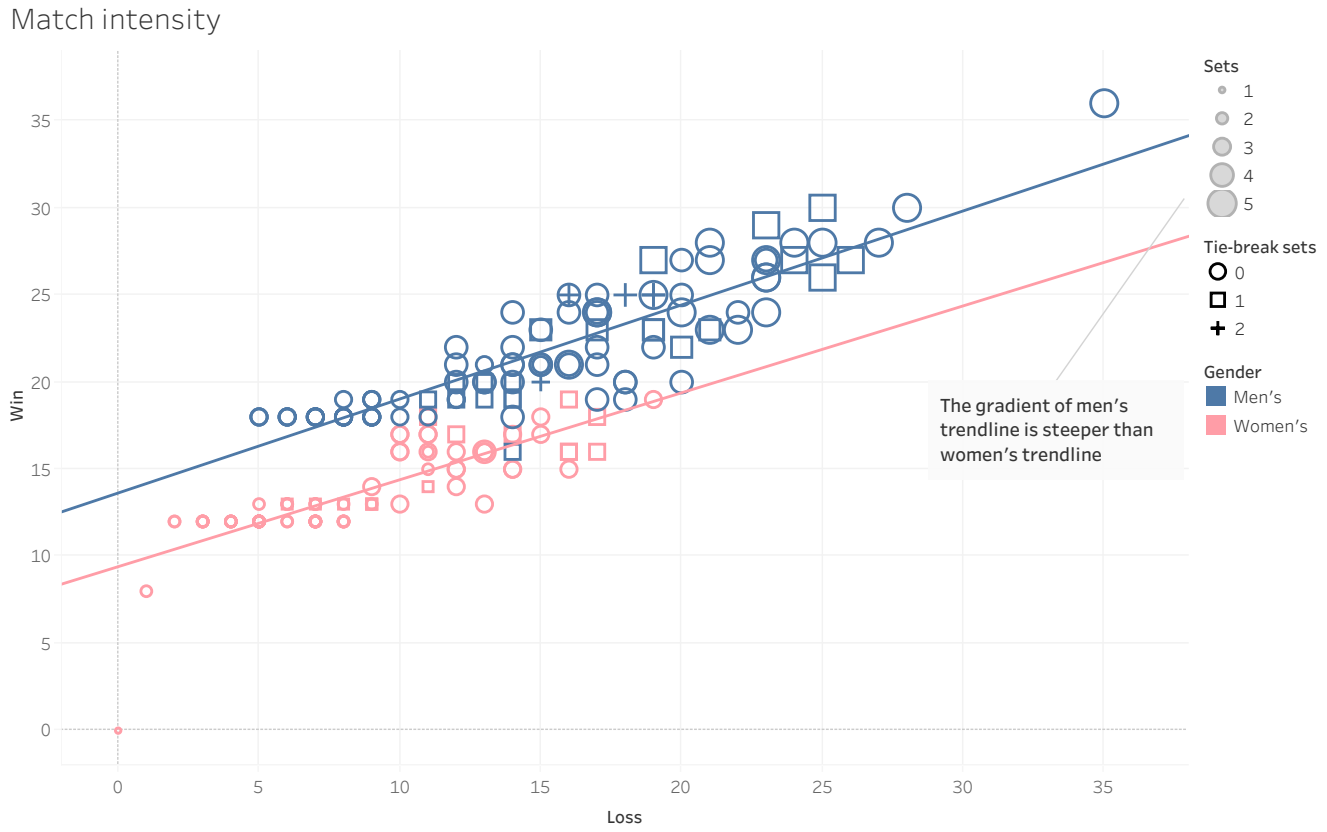


Figure 8. Match intensity

The technique is used is scatter plotting. There is a supposition in data field that if there are sufficient data, it is possible to plot out a pattern. Hence, 'Win' and 'Loss' are put into row and column shelves, respectively, due to their considerable value varieties. Each symbol is a match, with color blue and pink to distinguish men's matches and women's matches. Other dimensions are applied, including tie-break sets as different symbols, and total sets proportional to the size of the symbol. Furthermore, 'Year', 'Champion', 'Runner-up', and 'Score' are also added to the information label. Looking at the graph, there are three outliers. The first outlier, located at the right top of the graph, is one men's match held in 1927 of champion Gerald Patterson and runner-up John Hawkes with merely one point for 'Loss'-'Win' discrepancy. The next outlier is the women's match in 2006 between Champion Amélie Mauresmo and runner-up Justine Henin. The reason for its loctation at the low left corner of the graph is the runner-up

retired, leading to a humble total score 6-1,2-0. The final outlier is an extreme point, right at the origin of the graph, meaning zero for both 'Loss' and 'Win'. This point indicates a women's match in 1966 of champion Margaret Smith and runner-up Nancy Richey. The runner-up didn't show up, so Margaret Smith had a walkover victory. There is a tendency in the plotting, the loss and the winning score increase proportionally when moving upward and moving left.

Similarly, the size of the symbols increases with respect to these movements. This tendency is obtained by the nature of the tennis player. For the rest of the points on the graph, all of the women's matches lie in a region below the horizontal line intersecting the y-axis at value 20, and on the left side of the vertical line intersecting the x-axis at value 20. This could be explained by the limit in the total sets of women's feature, 3 sets for maximum. For men's feature, primarily men's matches are plotted on the upper part of the graph. While there are only 15 women's matches with one tie-break set and no women's match having 2 tie-break sets, men's feature has up to 21 matches having one tie-break set and 5 matches having two tie-break sets. Interestingly, no men's 5-set match has 2 tie-break sets, only 3 men's 4-set matches and 1 men's 3-set match have 2 tie-break sets. Another finding is 3 of 5 men's matches with 2 tie-break sets have Novak Djokovic as the champion, and the pair Novak Djokovic-Andy Murray appeared in 2 of those 3 matches (2015 and 2013). Despite the number of men matches having 2 tie-break sets is more than one of women matches having 2 tie-break sets, another metric that should be considered is the trendline for all men's matches and women's matches. The trendline of men's matches has the gradient of 0.539686, and 0.4989 for women's matches trendline. The higher the gradient is, the bigger the discrepancy between 'Win' and 'Loss' is, and vice versa. The lower gradient shows higher intensity in the match, where the champion has to race against the fierce runner-up for every single point.

5. Conclusion

The Australian Open tournament, with more than 118 years of history, is indeed one of the most attractive grand slams in the world. And this statement has been proved with this report. Starting only with a plain dataset composed of numbers and strings, statistics and analytics in terms of country and individual performance have been gradually conducted properly thanks to valuable tools, including Microsoft Excel for calculations and transformation; and Tableau for visualisation. While Microsoft Excel shows robustness in the computing field, its visualisation built-tin functions pose inferiority when compared to Tableau. Tableau provides more comprehensive solution in visualisation, especially in the context of multidimensional data. By using different techniques, multidimensions have been integrated into one graph, enhancing the data readability and connectivity, making the storytelling flow of the report smoother and more coherent, and avoiding the fragmentary condition due to using excessively separate graphs and charts. Having said that, although snippet is provided in Tableau to create calculated fields, it is more convenient to let Excel do the data pre-processing in advance. A clean, well-processed and organised dataset imposes the likelihood of a decent analytical report.

6. Reference

Australian Open. (7 October 2023). In *Wikipedia.* https://en.wikipedia.org/wiki/Australian_Open.