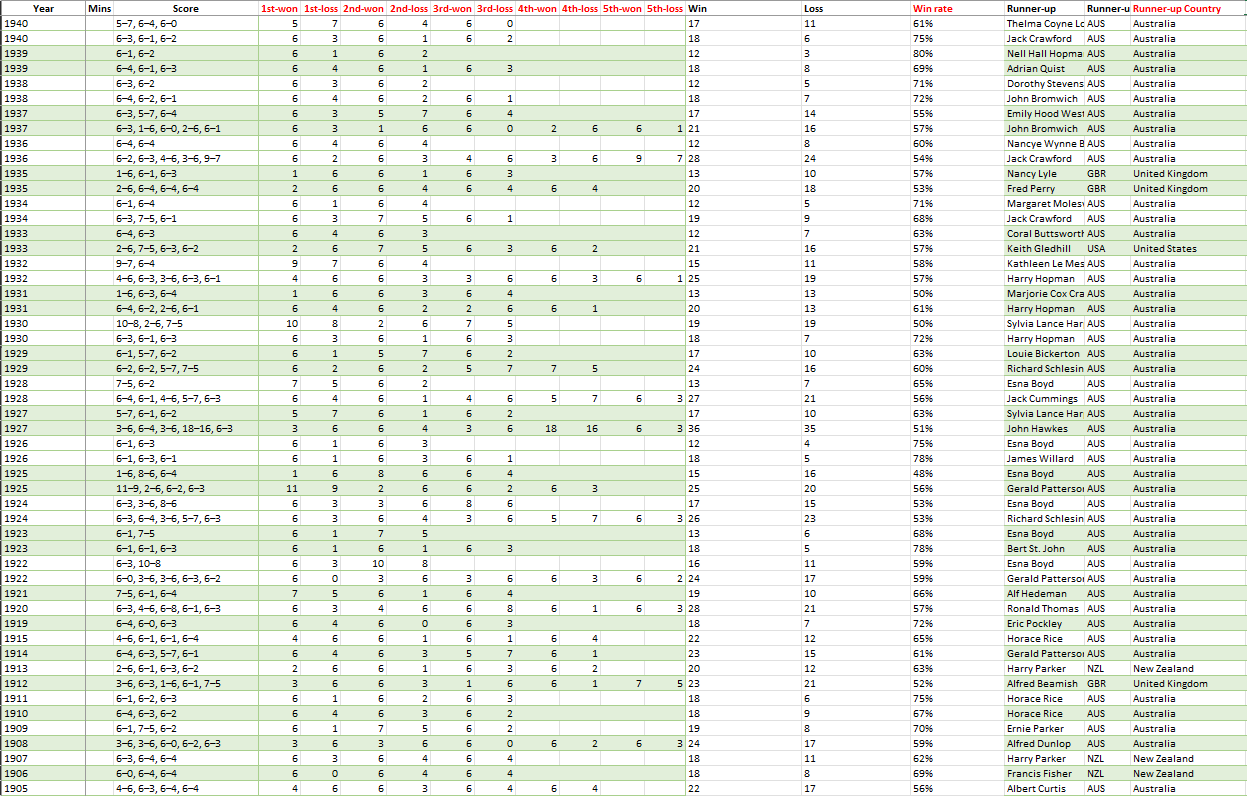
**320146 Assessment Task 2:**

**Advanced Data Visualisation**

Wilcent Wijaya - 14378917

Sample Dataset:



Figure

The dataset above consists of data based on the Australia Open Championship that have been collected throughout 118 years of matches for both men and women between the year 1905 to the year 2023. The dataset is made out of various data types or categories this includes year, gender, name, runner up, nationality, runner up country, score, and runner up’s nationality. The dataset will be visualized and analysed by creating each respective illustrations by using Tableau in order to be compared. This report will also take a look at each category and analyse the relation between other categories in the dataset by using the method available in Tableau.

|  |  |  |
| --- | --- | --- |
| **Category/Attribute name** | **Data Type** | **Description** |
| Year | Quantitative | Year of the specific match |
| Gender | Categorical | Championships is divided into 2 genders Men and Women |
| Champion | Categorical | The winner’s name |
| Champion Nationality | Categorical | Nationality of the winner |
| Champion Country | Categorical | The country name of which the Winner represents |
| Champion Seed | Quantitative | The ranking of the winner for the purpose of when the match ends in a draw |
| Mins | Quantitative | Total Minutes of the match |
| Score | Quantitative | Scores of each set of games |
| Runner up | Categorical | Name of the runner up in the Championship |
| Runner up Nationality | Categorical | Nationality of the runner up |
| Runner up Country | Categorical | The country which the runner up is representing |
| Runner up Seed | Quantitative | The ranking of the runner up for the purpose of when the match ends in a draw |

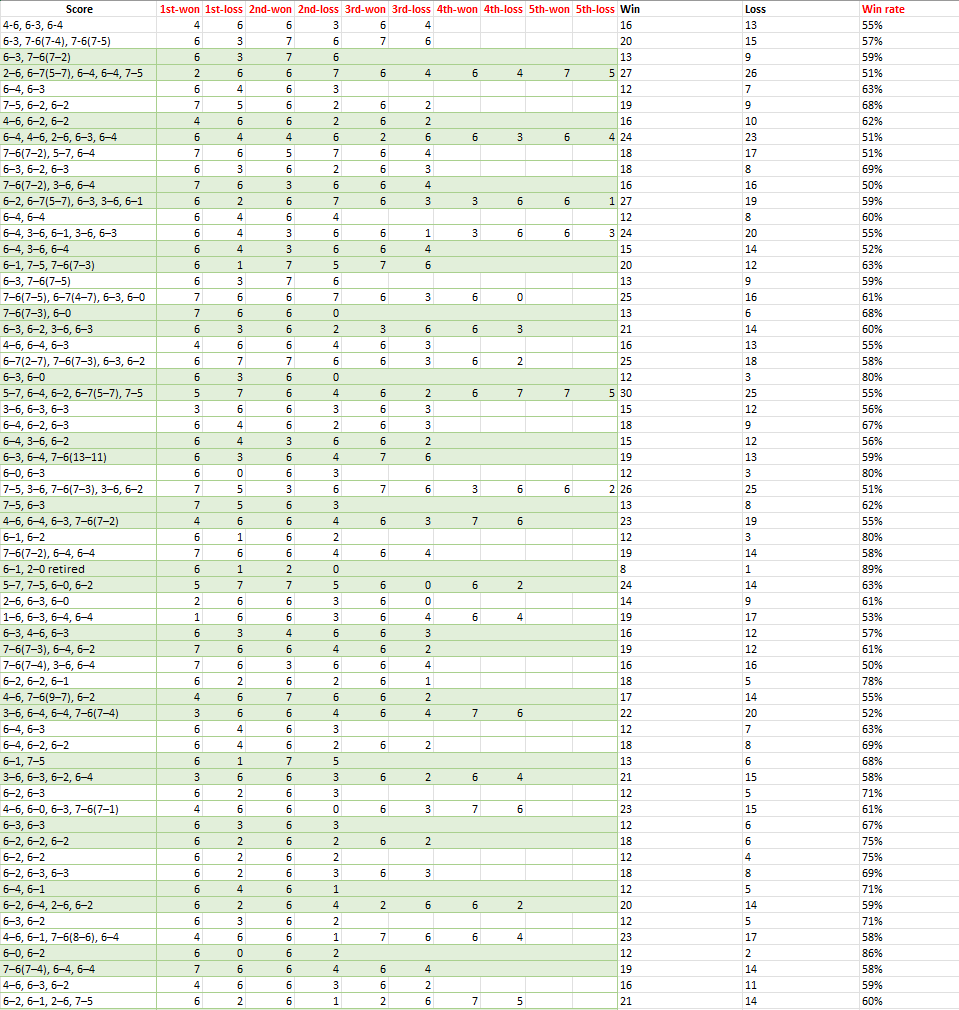
Data Preparation:

Missing Values

|  |  |
| --- | --- |
| **Attribute Name** | **Number of Lines** |
| Champion Seed | 208 |
| Mins | 208 |
| Runner up Seed | 208 |

There are always missing values when it comes to huge dataset, so it is not a surprise when there are missing data in the dataset above. There are a total of 3 categories with no data values in the dataset, those includes Champion seed, Mins, and Runner up seed. It is not an optimal move to replace the values with 0s since it could negate or interfere with the analysation methods that will be used on this report therefore it is better to ignore those missing values

Winning-Rate

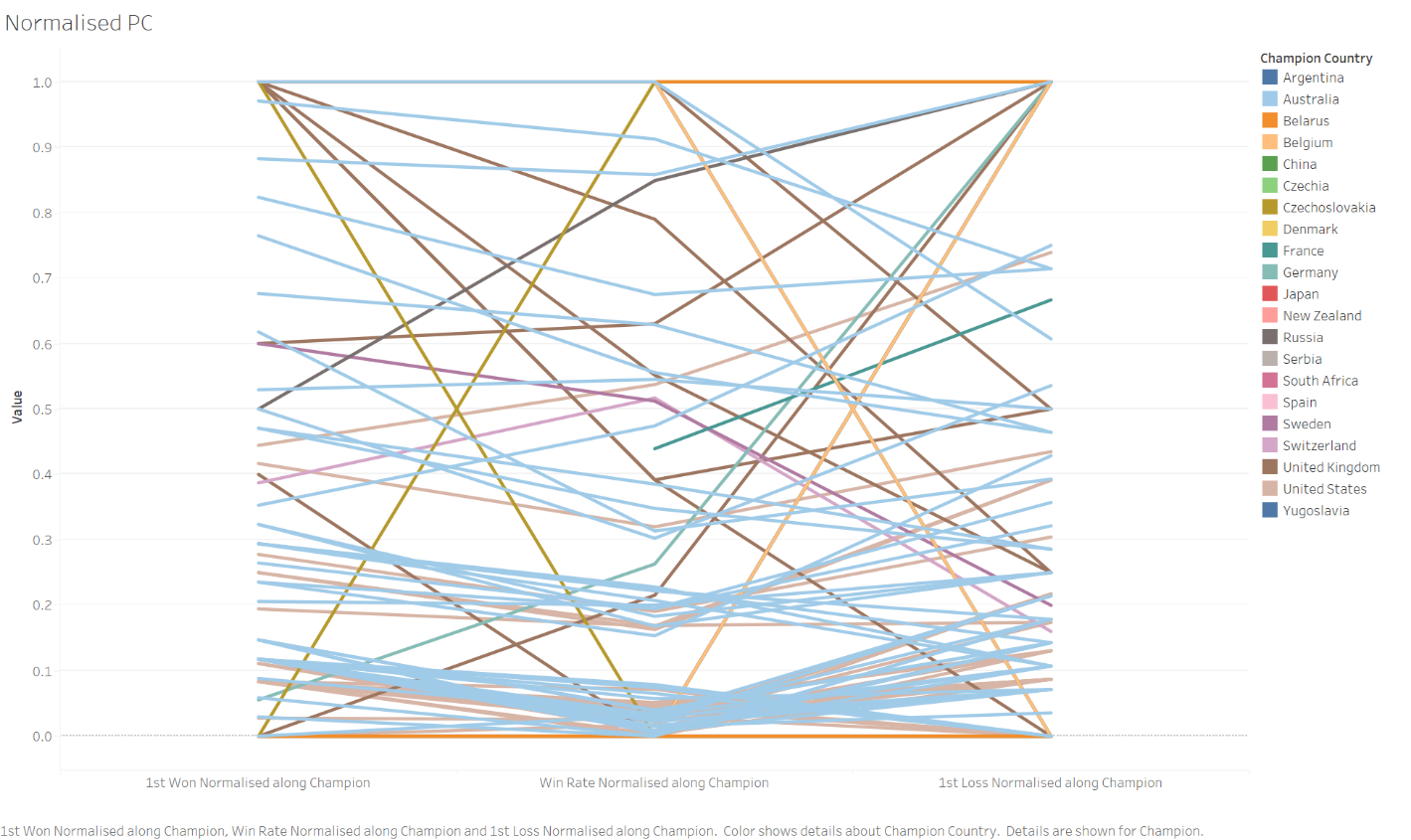


Figure

There is an additional of 3 categories called Win, Loss, and Win rate. The Win category is based on the total or sum values of 1st won to the 5th loss values, similarly the Loss category is also comprised of the total values of 1st won to the 5th loss values. After calculating the total Win and Loss values, the addition of the Win rate category is added which made up of the total win divided by the total of wins and losses.

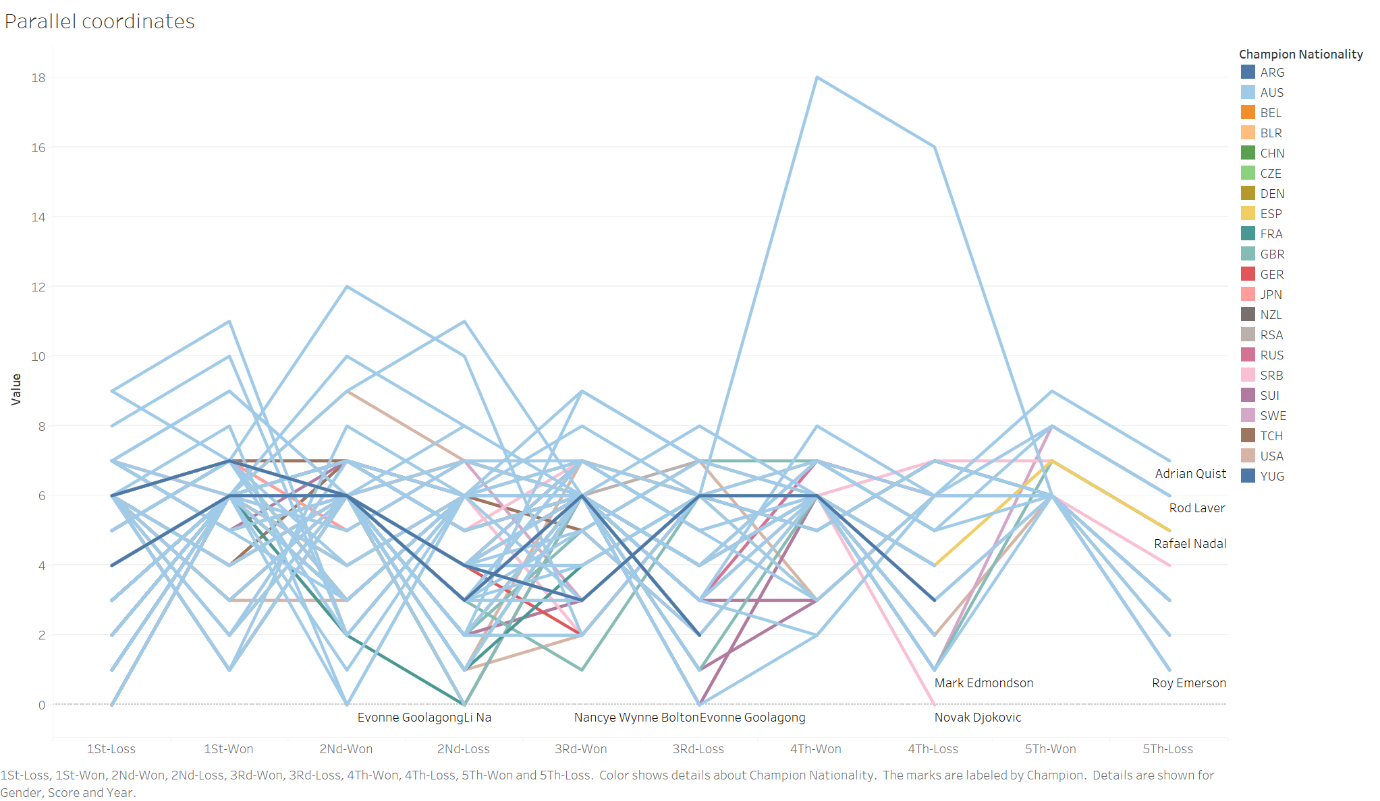
Visualisation Method 1. Parallel Coordinate:

Parallel coordinate is a great tool to display the relationship between variables needed to be measured. Furthermore, Parallel coordinate helps to illustrate the patterns formed from the various correlated variables. Parallel coordinate is the optimal choice when it comes to display high dimensional data since it can display a row of data by utilizing different axis to help illustrate the different factors in those variables.

****

Figure

From figure 3 graph above we can see the correlation between the 1ST won that have been normalised with each respective champion, the normalised win rate of each respective champion, and the normalised 1st loss of each respective champion. The figure above shows the normalised win rate by taking in the 1st won of each champion and the 1st loss of each champion.



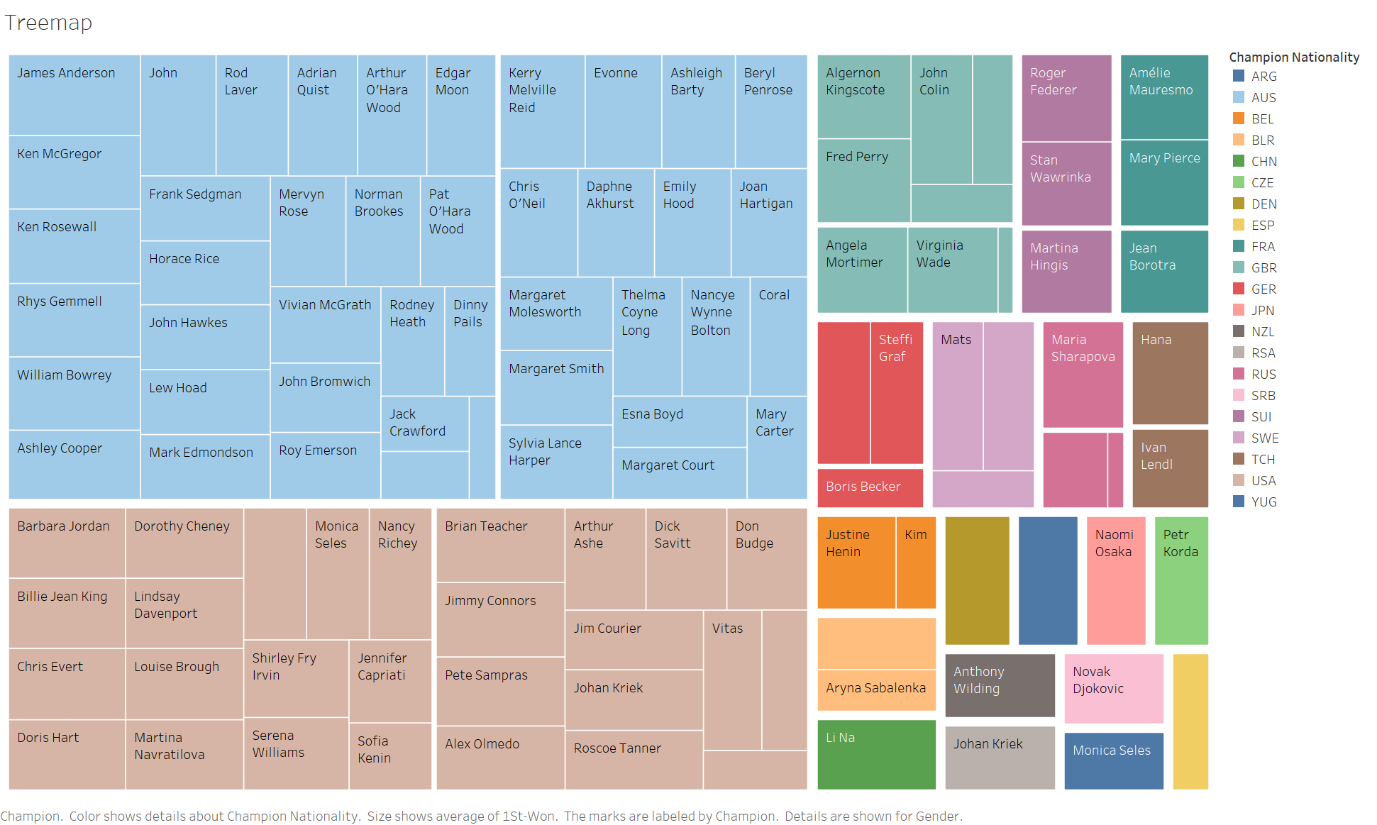
Figure

Figure 4 above shows the scores of each player and we can see the relationship between win and losses of the players. The lines are coloured by each respective countries based on the champions, those line are also labelled by the champions names followed by the addition of detail with the variable of gender. From the graph above we can see that player Gerald Paterson scored a high of 18 on the 4th won variable.

The advantages of parallel coordinates are that it allows for the visualization of multivariate data and complex format data. Furthermore, it also allows the comparison of relationships between assigned variables. On the other hand, the disadvantages of a Parallel coordinates would be the clutter it shows on the graph, when there is a lot of data represented in the graph, it can be hard to read and identify the patterns showed in the graph. Furthermore, Parallel coordinate can be hard to be identified and understand by those who are not familiar with the graph since it shows high amounts of data in a complex manner.

Visualisation Method 2. Treemaps:

Tree maps is one of the best tools for data visualizations, tree maps illustrate the data in a hierarchical format and each data is represented in a rectangle. Each rectangle will be formatted based on the values it’s been assigned. Furthermore, the rectangles will be coloured based on the assigned values for improved readability.



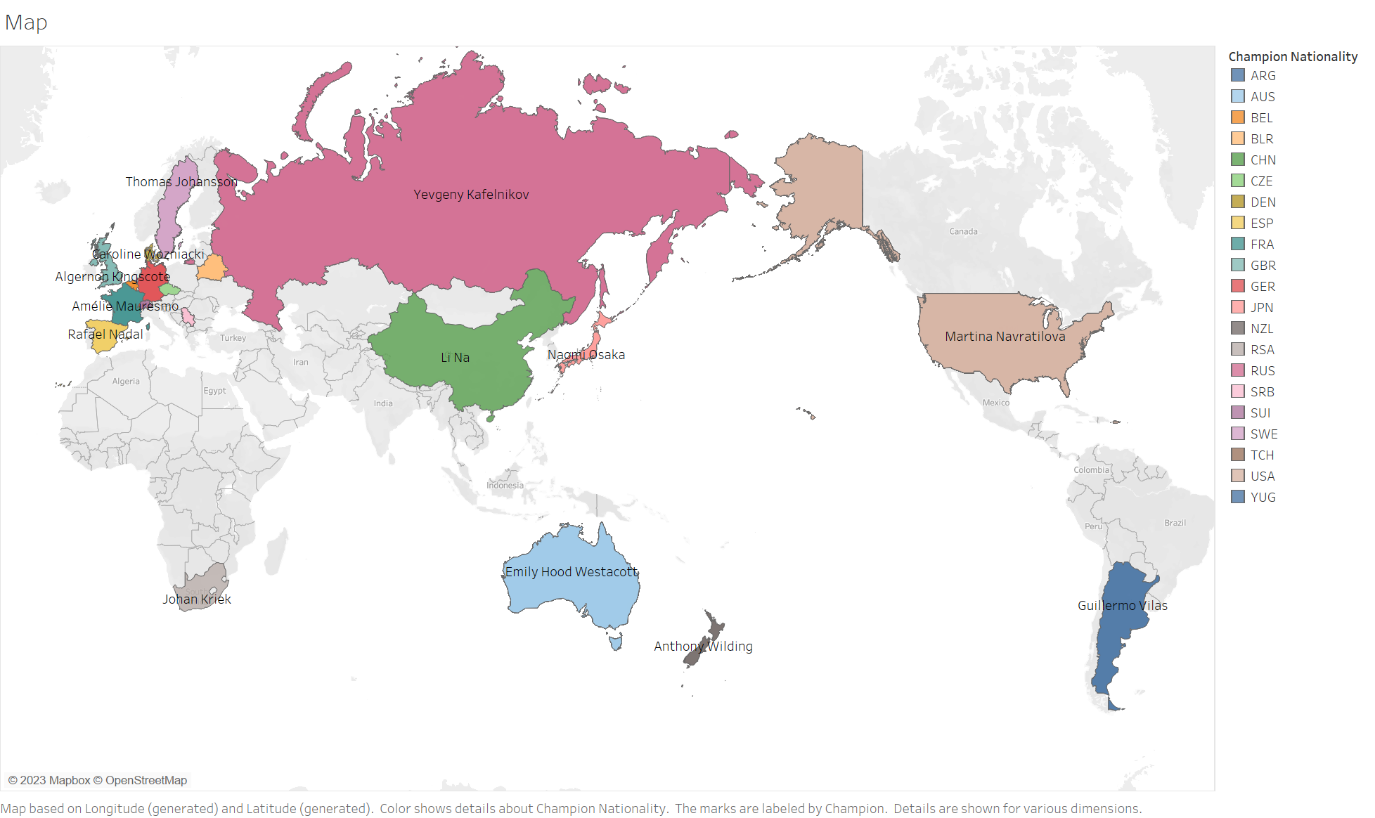
Figure

From figure 5 we can identify that Australia which is represented by blue rectangles takes up most of the space in the graph, this means that players from Australia win the most Wimbledon championships unlike the other countries which is represented by different colours. Similarly, the second country that takes most of space in the graph is the USA represented by the colour of beige. Both the USA and Australia has a significant different with the other countries in the chart since both countries won championships by a significant different when compared to other countries. The third country that has the most championship wins would be the United Kingdom represented by the colour light turquoise.

The advantages of tree maps would be the efficient use of space since its rectangles showed in the graph is proportional to its values. Furthermore, tree maps allow the representation of hierarchical data in the graph. On the other hand, the disadvantages of tree maps would be the limitation to hierarchical data since tree maps are best suited only for hierarchical data. Furthermore, tree maps also are best suited to only displaying static data.

Visualisation Method 3. Geographic Map:

Geographic map can be a great tool to display the origin of each champion since it illustrates a geographic map of the world and given each respective colour to identify each country from which the champion is from. Geographic map is a great tool to identify the location of data of variables assigned.



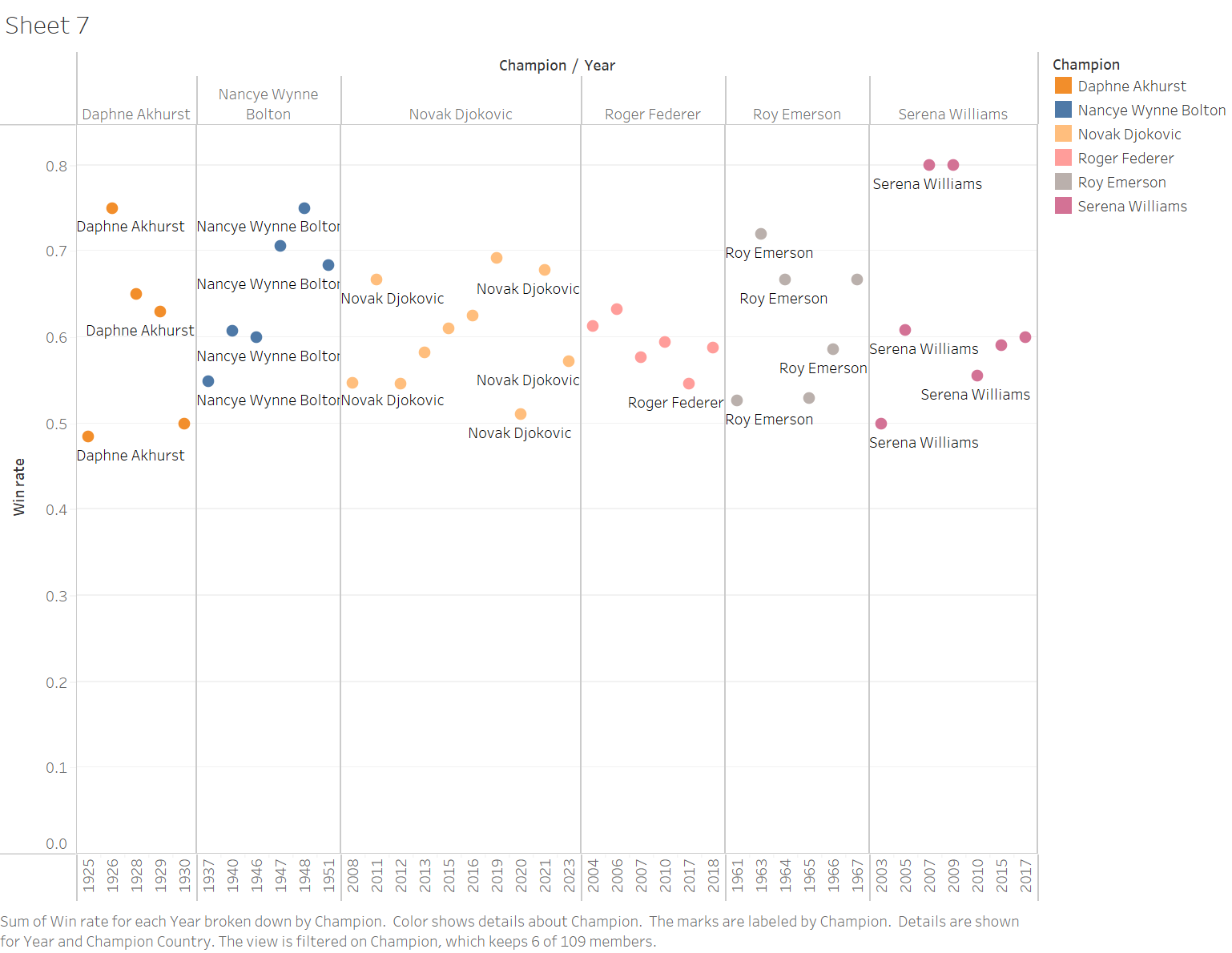
Figure

From figure 6 we can see all the countries that have participated or won the championship. Each country has been represented by each colour. Furthermore, each country has been labelled by the name of the champion that has won the championship, followed by the addition of stats of each player such as the score, gender, year of the championship, and nationality of the champion.

The advantages of Geographic map would be the easy readability for readers. This is because the champion of each country is represented by the country on the geo map. However, geographic map is only optimal for geographic data. Furthermore, if the data is too complex or the data is too cluttered, it would be difficult to interpret the data accurately. Even though geographic data is only limited to geographic data it can be optimal for visual appeal and only used for specific parts of the assigned data.

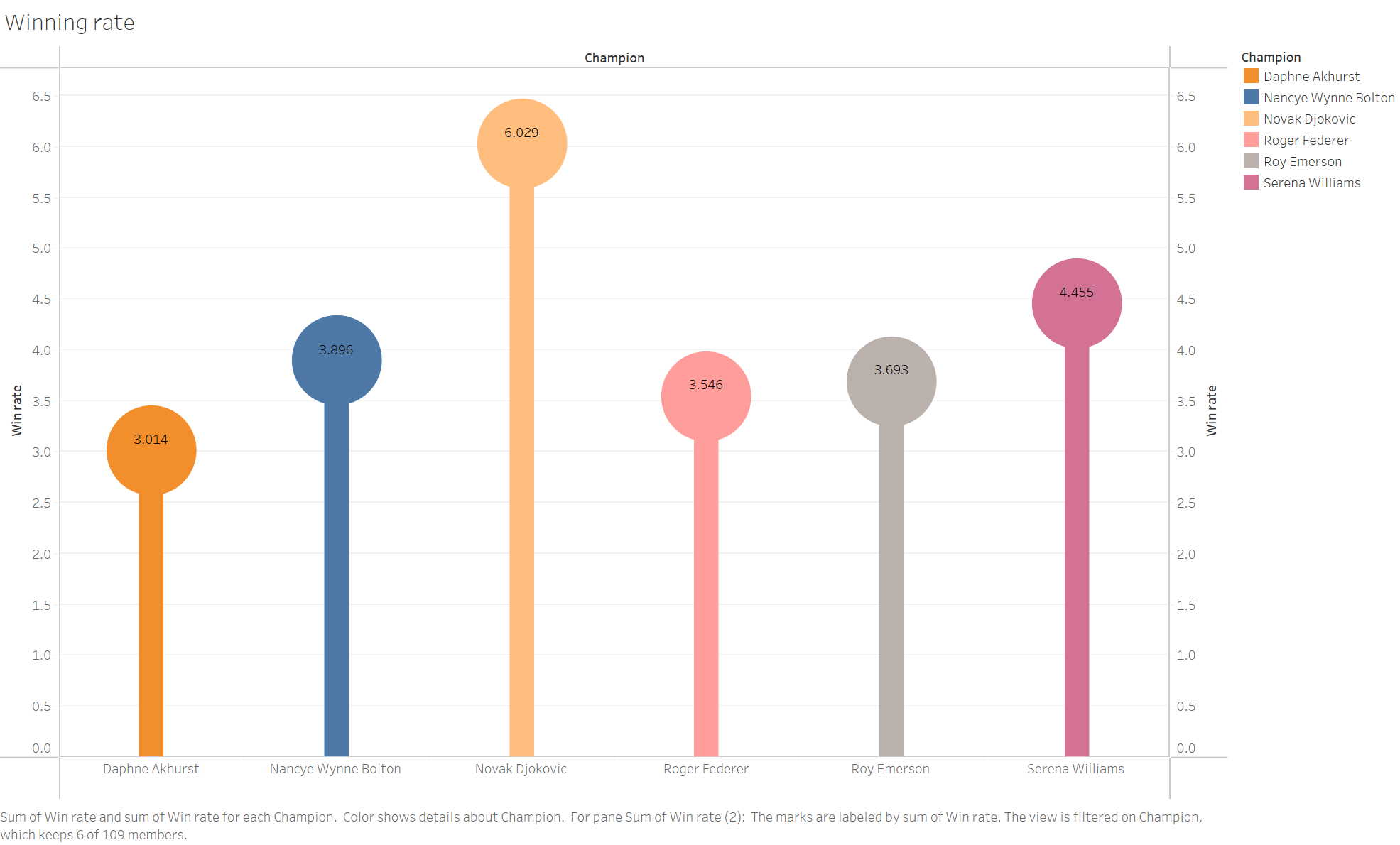
Visual Analytics: Champion performance

After exploring the dataset provided, there are a total of only 6 players who have won the tournament 5 times or more throughout 1925 to 2023.



Figure

From figure 7 above we can see that there are 6 different champions who have win the championship 5 times or more. Furthermore, figure 7 shows that the champion Novak Djokovic has highest wins out of all the champions in the chart with total wins of 10.



Figure

Figure 8 above illustrates the win rate of each player in the chart. Using the lollipop chart to illustrate the data is more visually appealing and easy to understand since the chart has labels on each circle to show the exact value of the win rate. Furthermore, colours are also assigned

Conclusion

In the process of accomplishing this assignment, there are a few obstacles that were faced such as figuring out how to create the lollipop chart. However, after looking at the Tableau website and looking at other resources I was able to create the chart. Dealing high dimensional data is quite complicated, however playing around with the worksheet in Tableau helps with understanding the ways to interpret the data such as using the Parallel coordinate chart that was introduced in the workshop from previous weeks. By assigning specific variables as the identifier such as colours and labels, it improves the readability of the chart used throughout the assignment. Tableau is one of the most compatible software for data visualization, due to the various options that can be utilizes when it comes to interpreting high dimensional data. Furthermore, it also helps with understanding of high dimensional data since it allows the visualization in a more familiar manner.