Text, logo

Description automatically generated

**Analysing US Open Champions**

Submitted By- Kushal Ahuja

Table of Contents

[Data Exploration 3](#_Toc117033152)

[Data Preparation 4](#_Toc117033153)

[Missing Values 4](#_Toc117033154)

[Win Rate 5](#_Toc117033155)

[Parallel Coordinates 8](#_Toc117033156)

[Advantages of Parallel Coordinate 10](#_Toc117033157)

[Disadvantages of Parallel Coordinate 10](#_Toc117033158)

[Treemaps 10](#_Toc117033159)

[Advantages of Treemaps 11](#_Toc117033160)

[Disadvantages of Treemaps 12](#_Toc117033161)

[Geographic Map 12](#_Toc117033162)

[Advantages of Geographic Map 13](#_Toc117033163)

[Disadvantages of Geographic Map 13](#_Toc117033164)

[Visual Analytics: Champions Performance 13](#_Toc117033165)

[Executive Summary and Conclusion 15](#_Toc117033166)

[References 17](#_Toc117033167)

# Data Exploration

Data Exploration is the part and among the first step of data preparation; it consists of interpreting the dataset regardless of how big the data is and getting brief insights about the dataset through various data visualisation techniques; this process tells the users about the data types of each column and some specification about the attributes of a dataset, it also helps the users in the descriptive analysis as it makes it easier to perform the descriptive analytics method on the dataset. This method is used to get brief insights or information only and not for detailed granular data or information.

The data type and information of all the columns/attributes are given in the table below –

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| Year | Year column contains data in YYYY format with numerical data | Year of the US Open tournaments |
| Gender | Categorical data | Details about the gender of players |
| Champion | Categorical data (String) | Names of the players from the country |
| Champion Nationality | Categorical data (String) | States the nationality of the player |
| Champion Country | Geographical data | States the country or the citizenship of the player |
| Score | Numerical data | Reveal the score of the player in the tournament |
| 1st Won – 5th Won | Numerical data | States the score of the win |
| 1st Loss – 5th Loss | Numerical data | States the score of the loss |
| Win-rate | Numerical data (Percentage) | States the win-rate of the players in the matches |
| Runner- up | Categorical data | States the name of the runner-up |
| Runner-up Nationality | Categorical data (String) | Reveal the code of the players nationality |
| Runner-up Country | Geographical data | Name of the country or a citizenship of a player |

Findings

After implementing the brief data exploration, the findings are given as below-

* Most of the champions are from the United states.
* Canada is among the highest win-rate

# Data Preparation

The procedure of getting a dataset ready for further evaluation and processing is known as data preparation. The primary operations involve gathering, classifying, and converting data into a form that can be used for visualisation. (AWS, n.d.)

The following processes are generally included in the data preparation process, however, it might vary based on the work and output needed by the stack holder –

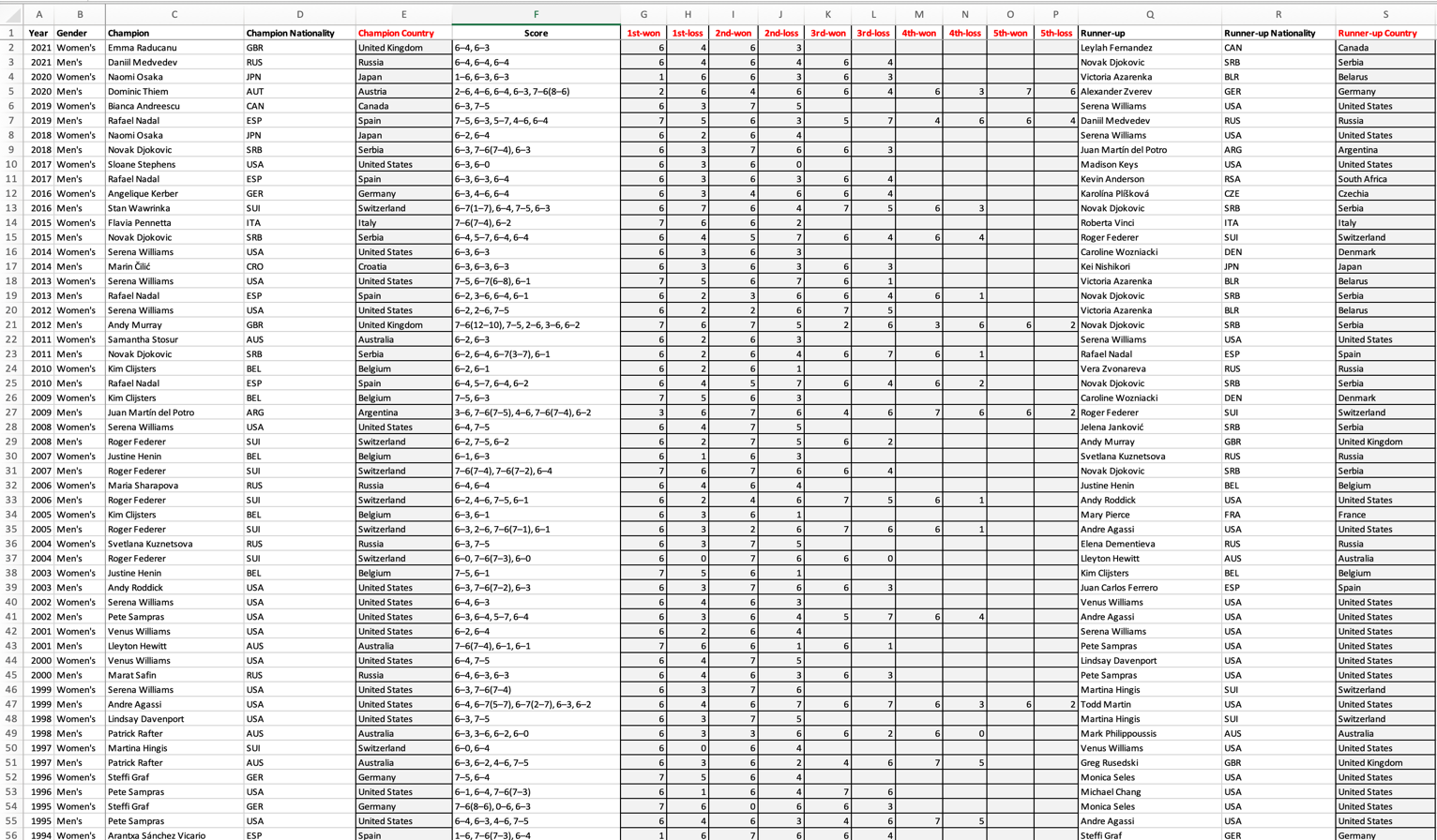
* Acquiring data- This step is for gathering the data and analysing which data would be efficient and important to perform the further task.
* Cleaning data – This step is considered among the most important part as it makes the further analysis and visualisation of actual use; cleaning data ensures less error or error-free results and also deal with the missing values and redundant data resulting in only getting and working on required data. Professionals have to make sure the data is cleaned correctly as it affects proportionally to the end result.
* Exploring data – This step consists of determining the attribute’s data types and analysing the dataset distribution among the various variables for the brief introduction of the dataset and further analysis
* Transforming data – Transforming data consists of visualisation techniques which means getting meaningful insights from the data provided. (Alteryx, n.d.)

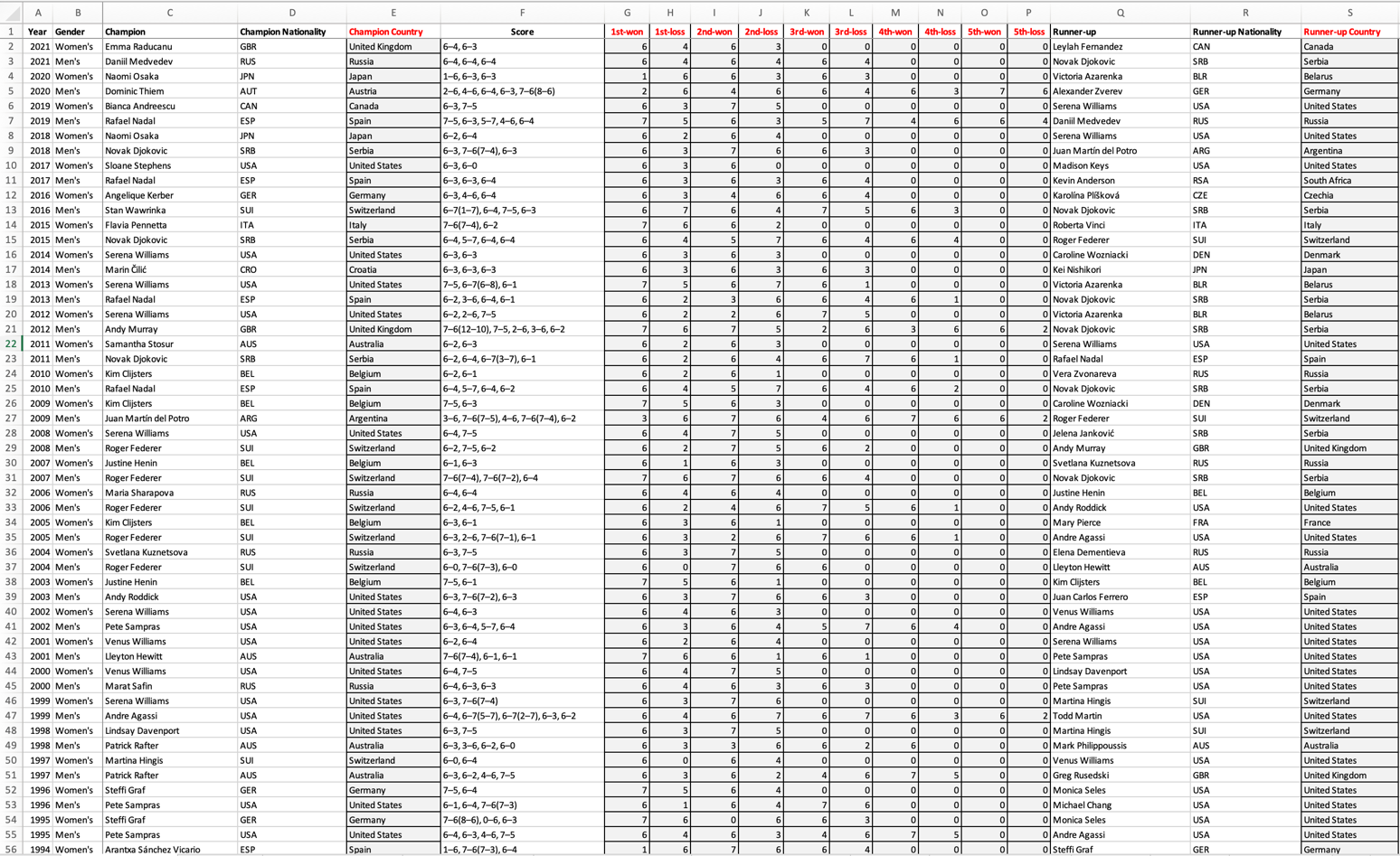
The data preparation techniques used in this assignment are as below-

## **Missing Values**

There can be any missing data in the real-world dataset due to the non-completition by humans, for example not inserting the data in the survey questions or can be by human-made errors. This missing data in the datasets are popularly known as missing values.

Dealing with the missing values is an important step of the data preparation as it makes the visualisation more precise and accurate; in the data set of the US Open Champions, there are many missing values from 1st Won – 5th Loss as depicted below –



To fill in the missing values, I have used Go to section and filled in the blanks with the value zero to avoid any discrepancy in the results and visualisation graphs. 

## **Win Rate**

Calculated the win rate by applying the formulas in excel and creating new columns –

* Created new columns named Win, Loss and Win Rate. Added the values of all the wins in the Win column using the sum() function and inserted them into the new column named Win, added all the values of loss and inserted them into the new column named Loss.
* Created a new column named Win-rate and calculated it by dividing the values of win column by the sum of Win and Loss and then changing the format of the values from numbers to percentage.
* Sorted the Champion name by A-Z to also use the count function required for task 7

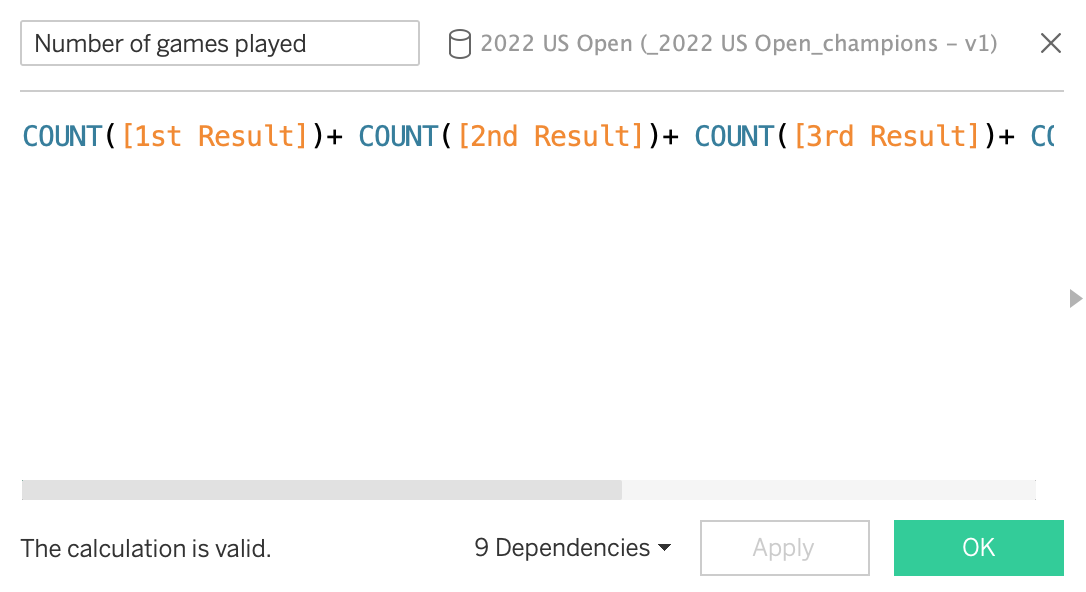


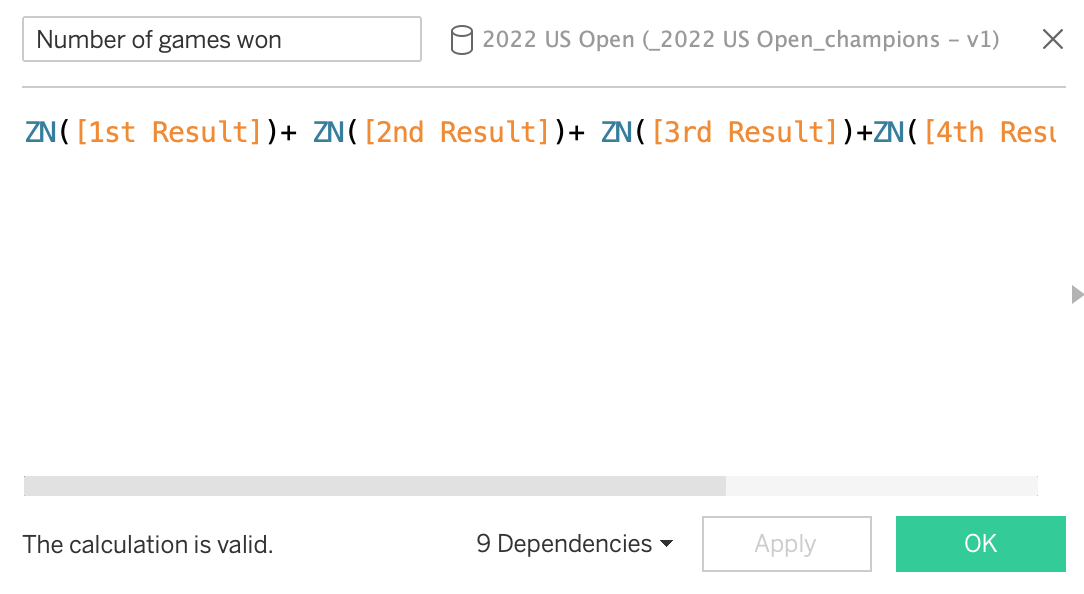
Moreover, also calculated the Win rate in Tableau by using the IFF function and formula

Graphical user interface, text, application

Description automatically generated

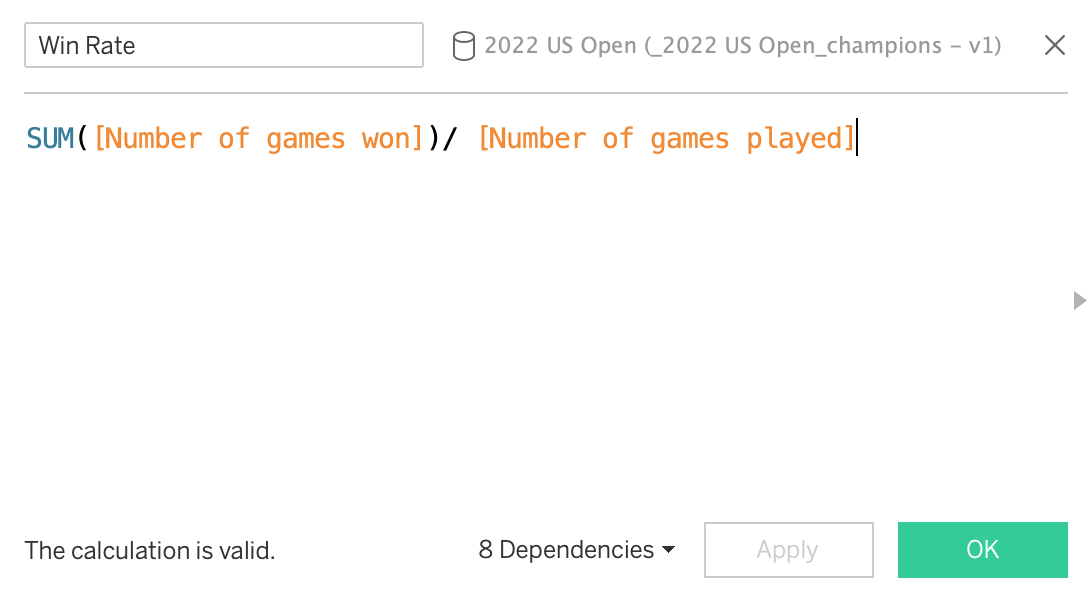
Similarly made 1st – 5th Result to calculate the desired Win-rate,





Calculated the Number of games played and number of games won by using Count() and ZN() function respectively.

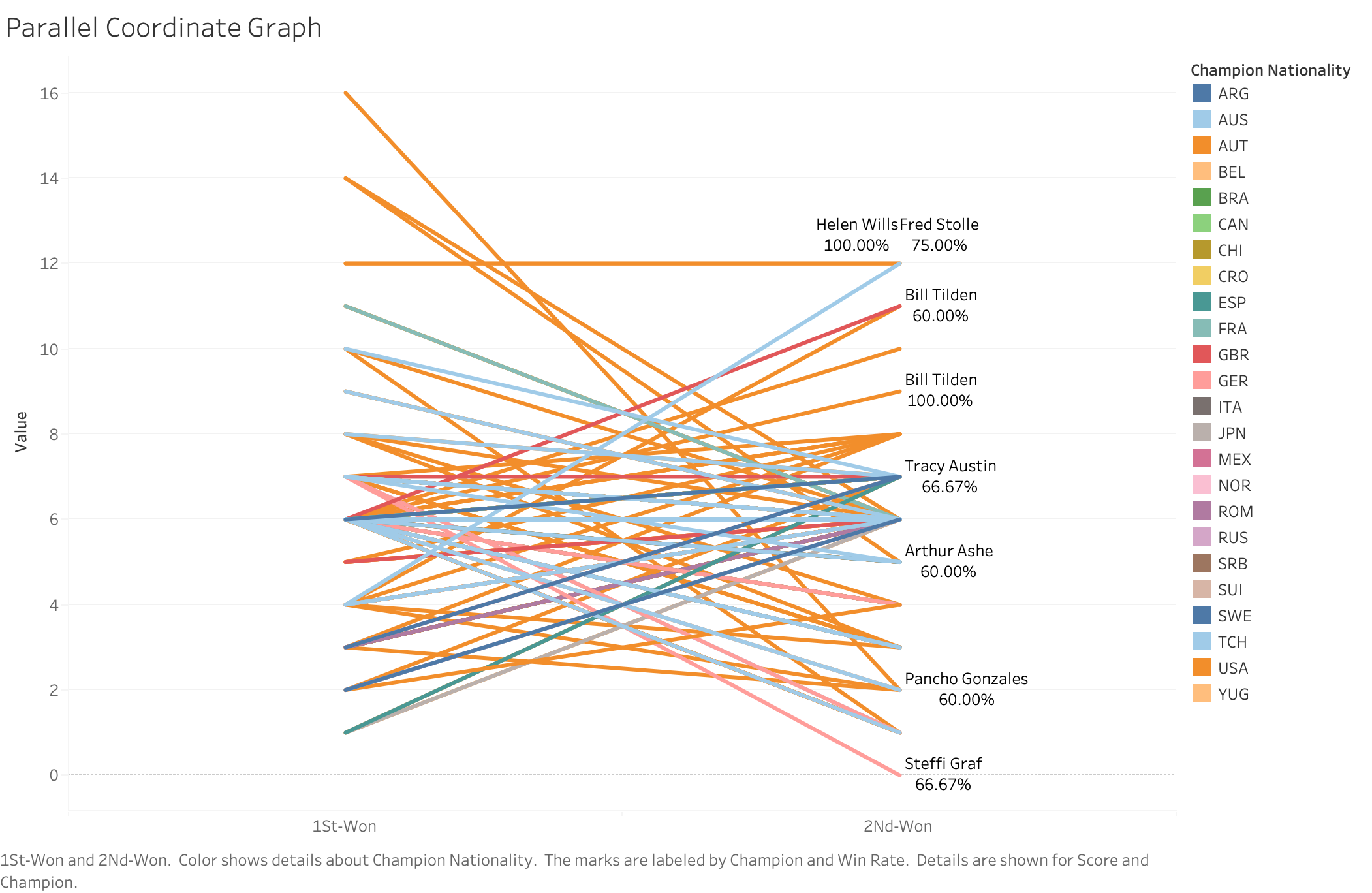
The next step is to divide the Number of games won by the Number of games played and using the sum() function.



# **Parallel Coordinates**

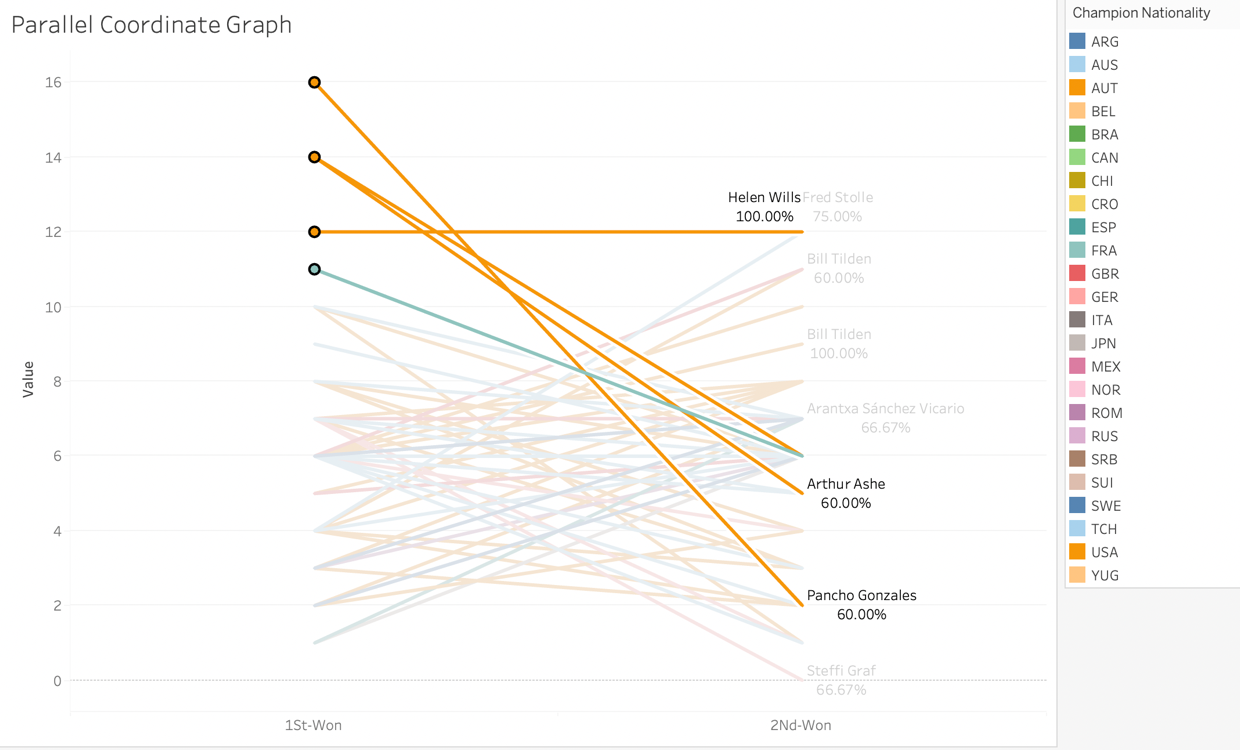
This technique of visualisation is used for multivariate and high-dimensional data, this visualisation method is one of the best ways to compare the different attributes and determine how they relate to each other in a single dataset. Each attribute or variable has its axis in this plot which is also parallel to each other, Although each attribute uses a distinct system of measurement, each axis can have a varying degree, or all the axis could be normalised to maintain a constant scale. This plot is also used to analyse the correlation between the different variables.

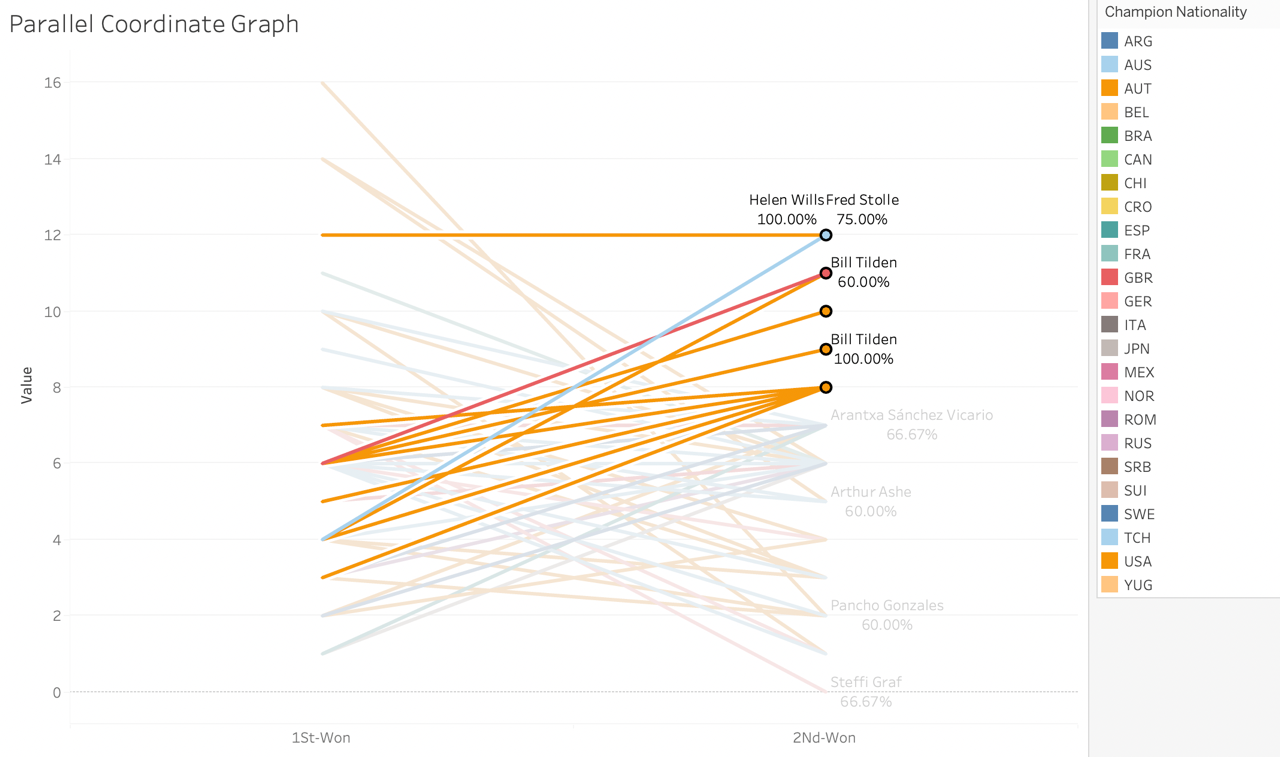
Interpretation of relationship between Champion, score, win rate , 1st Win, 2nd Win and Champion Nationality.



In the above graph, it is evident that 1st Won and 2nd Won are the two axis and the player's performance is being analysed by their wins among the games played additionally showing the win rate.

Interpreting the insights from the visualisation – the user can access to the specific data by clicking on the axis and analysing the performance of players in various games, as shown in the below graph, it demonstrates that players who had more scores and the win in the 1st won has fewer won or a score in 2nd Won except Helen Wills, this method of interpretation is called Isolating pattern of interest**.**



Similarly, the visualisation can give insightsby analysing the other axis**,** asdemonstrated by the below graph, which shows that majority of the top players in the 2nd Win got less score or number of wins in the 1st Win compared to the 2nd Win. (Bill Tiden and Fred Stolle****

The Parallel Coordinate can also be normalised to get the attributes on a similar scale, the normalised parallel coordinate graph is with the win rate as an axis

## Advantages of Parallel Coordinate

* As a two-dimensional presentation of high-dimensional data, parallel coordinate offers a significant advantage. Using this it is easier to see the trend showing from data entries when it is displayed as a line.
* More feasible to see the specific data or the visualisation by selecting the axis or the points made in the figure.
* Analyse the attributes of a single column better by plotting in the same dimension as others making it convenient to compare. For example, comparing the performance of players with the 1st and 2nd Won.

## Disadvantages of Parallel Coordinate

* Overlaying of the line can be the major demerit for the parallel coordinate as it makes the visualisation to look denser; this can be solved by highlighting the line or the group of lines as so to analyse the data which you want and eliminating the unwanted popularly known as noise.

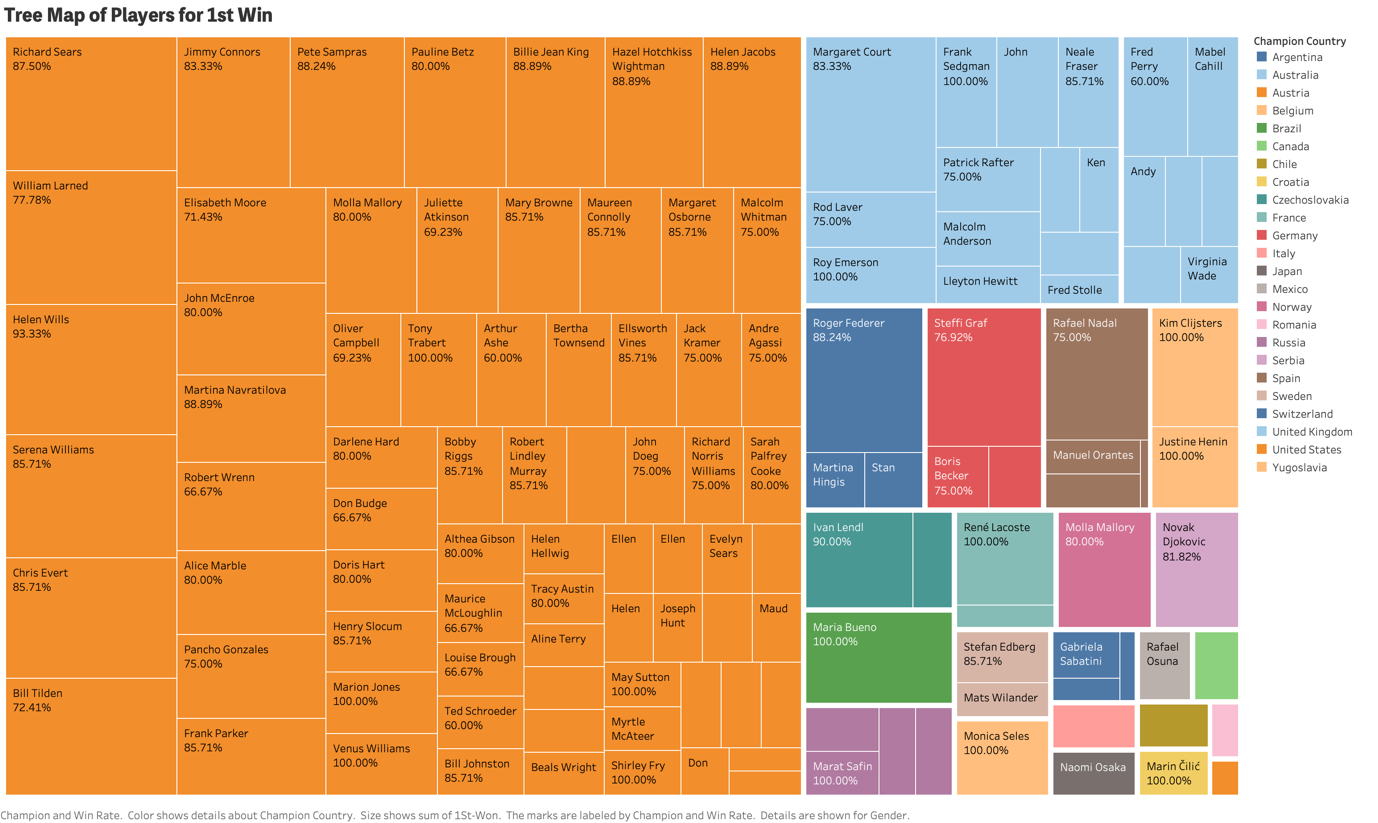
# **Treemaps**

Treemap is a way of visualising the Tree Diagram’s hierarchy, it displays the categories using rectangles; the size of the rectangles depends on the selected attribute and their values; higher values result in a bigger rectangle. Treemaps consist of different areas which makes various sub-categories in it, the area can be a domain of comparison or stating the common variable. For example Champion Country in the below figure. In the champion country area, there are various rectangular shapes that state the player's name thus arranging them according to the wins, more the wins resulting in a larger shape. Treemaps are considered to be space-efficient due to its capabilities to visualise the data in a compact rectangular format rather than making a hierarchal tree-like structure.

The treemaps given below are made using the attribute Champion country as a colour for ideal positioning and 1st Win and 2nd Win as size respectively as it is considered as a subcategory to make treemap in order. The champion attribute is used to label the subcategory shapes, additionally, Win rate and gender are added to the details (Win rate is added as a label also in the tableau file). Due to the limited screen size, some of the subcategories' shapes do not show the label.

The insights found by this visualisation are that the USA has the highest number of players who have top wins followed by Australia and Switzerland. The players having the most wins in these countries are Richard Sears, Margaret Court and Roger Federer respectively. It is evident from the treemap that even having a less win rate gets the player ranked up due to more wins than the player having more win rate. Various country’s and their player's performances can be analysed using this type of visualisation.

Chart, treemap chart

Description automatically generated

## Advantages of Treemaps

* Treemaps make the process of analysing the hierarchal study more convenient and easier to interpret.
* It categorises the data of different areas with subcategories as so the visualisation doesn’t look dense or overlaying over one another.

## Disadvantages of Treemaps

* Treemaps don’t display the hierarchical structures in an efficient way compared to the other graphs like Sunburst Diagram and Tree Diagram.

# **Geographic Map**

A geographic map is the visualisation method to analyse, interpret and display the geospatial data in the form of a map, it can be displayed in a different way such as country, continent, State and even more granular parts of the states and cities. This map makes it convenient to compare the performance of the specific attribute by the regions which has to be demonstrated or compared.

The geographic map for analysing the performance of players in the Wimbledon championship is given below with the appropriate legend.

Map

Description automatically generated

This visualisation is created according to the champion's nationality which is converted to the country’s name and then added the attributes – Score, Champion, 1st Win, 2nd Win and Win-Rate into the detail.

The insights found by this visualisation are that Russia, Canada, Spain, Australia, Brazil and Sweden have a larger win rate than other nations while USA and Australia have the most number of 1st Win and 2nd Win.

## Advantages of Geographic Map

* Efficiency in decision making – the geographic map makes the process of decision regarding the region seamless as it is so easy to interpret the required information from this type of visualisation as the stakeholders like the idea of data being in the demographic and the geographic state. So they can take the decision regarding how they want to proceed with the particular region.

## Disadvantages of Geographic Map

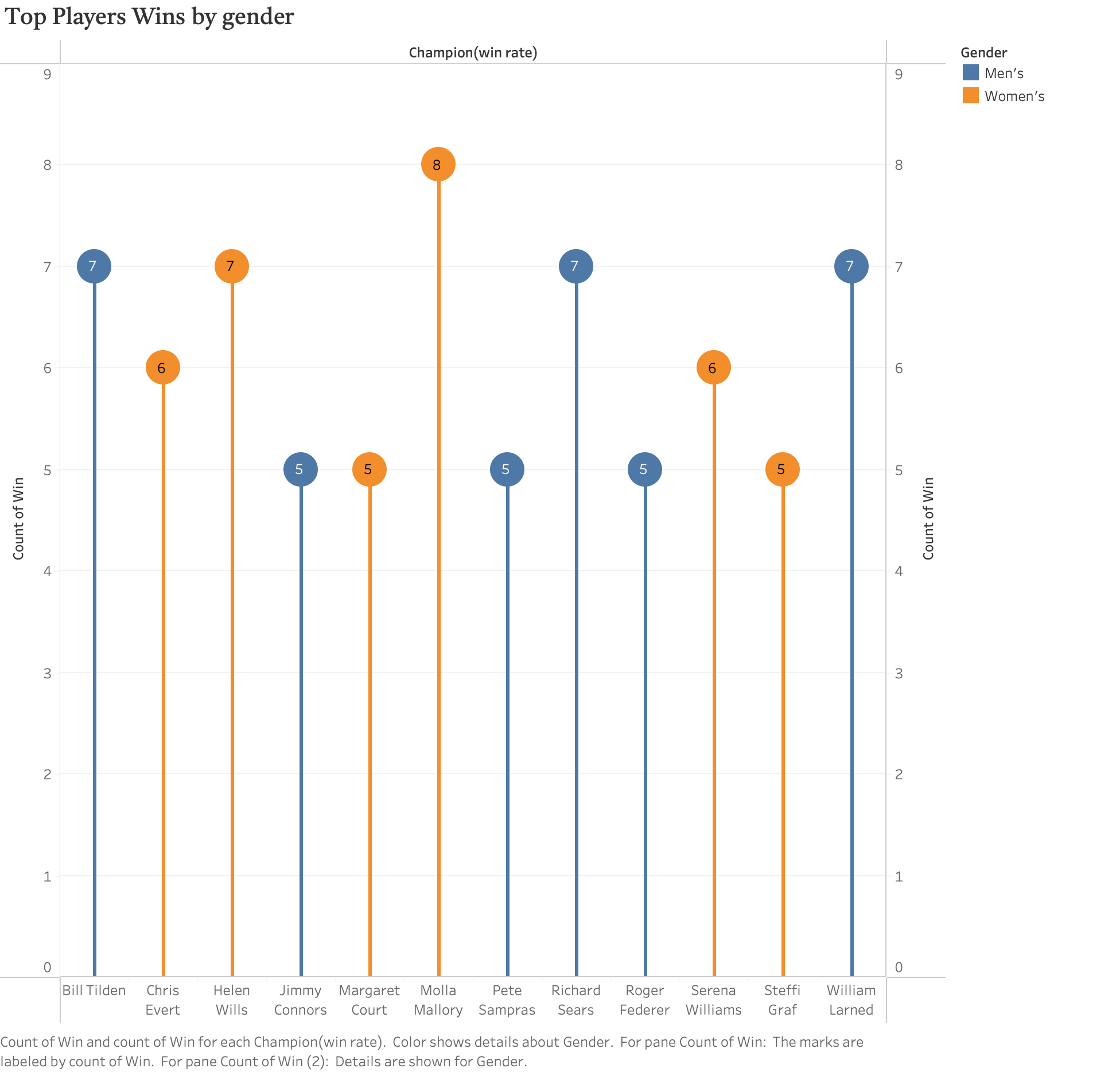
* Limitation of two-dimensional visualisation – the map can be represented in the two-dimensional form making the map sometimes to be mis-shaped in respect of the attributes associated with it.

# **Visual Analytics: Champions Performance**

Tableau and Excel are used to find out and interpret the players who have won five or more times in the Wimbledon Championship. After analysing the data, it is found that 12 players have won more than 5 times which consist of 6 men and 6 women.

The following visualisation concludes this information 

The visualisation is created using the attributes– Champion (Win Rate) which is being calculated in excel and tableau, Win-rate by size and the Runner-up, Score and Year into the details. The above diagram is the circles that illustrate the champion win-rate against the runner-up in respect of the champion's gender which is in order. Moreover, when selected at the dots, it displays additional information about the year and the score. The runner-up is sorted ascending by the sum of the win-rate. Some of the points overlapping over one-another due to the similar win-rate of the champion. This visualisation makes the comparison of the top player’s performance easier to interpret and understand.



The second visualisation created to analyse the champion's performance is the lollipop chart; this chart is similar to the bar chart but with a circle at the end which displays the value associated with it which makes it more convenient to use and to present to the stakeholders or useful for the storytelling. This chart is created by making the dual axis of the wins in the rows section and then reducing the size of bars while increasing the size of circles. The circles and bars here represent the same attribute, i.e. wins. This chart is then analysed in order of the gender of the champion.

# **Executive Summary and Conclusion**

The visual pattern for extracting the player's performance has been made by calculating the win rate and also analysing the wins and losses. The visual patterns have made it possible to extract the required or desired information from the dataset and then create the insights which can be used in various ways such as analysing the top player's performance and also analysing the performance of champions by their regions.

The performance patterns of the players have been analysed by the visualisation techniques resulting in the making of Parallel coordinate, Treemaps and Geographic map. Additionally, the top player's performance has been compared in the tableau by extracting the data calculated from excel.

Some of the performance patterns show that few of the players have a significant win over one season and then more losses over the next season, while some of the players have not performed well in the past but has been improved over the wins for the next few years such as Pancho Gonzales and Arthur Ashe with decline in their win rate and the wins.

I have dealt with the high dimensional data by creating the visualisation like parallel coordinates and then analysing them by specific section, which can be done by selecting the specific or targeted players for which the performance has to be analysed and then generating truthful insights about the player's performance over the years or a period of time or else can be compared with another players performance.

The multi-dimensional data has been visualised through the treemap as it consists of the categories and sub-categories inside it thus making it beneficial for better understanding or making it as an acceptable form of conveying the multi-dimensional data insights. This type of data can also be visualised and display through the various visualisation techniques available in tableau.

The graphic attributes design which are used for making this visualisation charts are Size, colour and details, for example, size is used to determine the attribute for which the graph should be presented like in treemap the 1st win and 2nd win are used as a size. The colour has been used in every representation as it is the only way to differentiate the attributes in a significant way. It is helpful in comparison and the identification of all the variables. The labelling technique is used in every visualisation as it shows the attribute which is being analysed or compared. For example, Champion and win-rate as a label in parallel coordinate, win-rate in a treemap and champion country in the geographic map.

They have impacted readability and storytelling in a huge way as they have enhanced the graphs and made it more convenient to understand. The storytelling is the main feature of data visualisation which includes the findings from the data , without these attributes and labels it is much more difficult to analyse the insights and even the actual data.

The breakthrough found from the top player's visualisation is that overall 12 players were found to have a win of 5 or more consisting of 6 males and 6 females.

Top female players have more wins than the top men players in the Wimbledon championship.

Advantages of tableau which I found while working on this dataset are –

* Ease of Implementation
* Instantly creating the interactive charts
* User interface – easy to use
* Use of calculated field- can write the commands and analyse the data in tableau itself without depending on the platforms such as Python, SQL or R.

# **References**

*What is Data Preparation - Data Preparation Explained - AWS*. (n.d.). Amazon Web Services, Inc. Retrieved October 09, 2022, from <https://aws.amazon.com/what-is/data-preparation/>

*What Is Data Preparation? | Alteryx*. (n.d.). Retrieved October 09, 2022, from <https://www.alteryx.com/glossary/data-preparation>

*Parallel Coordinates Plot - Learn about this chart and tools*. (n.d.). Retrieved October 11, 2022, from <https://datavizcatalogue.com/methods/parallel_coordinates.html>

*InfoVis CyberInfrastructure- Parallel Coordinate Visualization*. (n.d.). Retrieved October 14, 2022, from <http://iv.cns.iu.edu/sw/parallel.html>

*Treemap - Learn about this chart and tools to create it*. (n.d.). Retrieved October 14, 2022, from <https://datavizcatalogue.com/methods/treemap.html>