Appendix:

Data Exploration:

Data exploration is the process of examining and analyzing a dataset to understand its structure, content, and relationships between variables. It involves using statistical and visual methods to identify patterns, trends, outliers, and relationships within the data.
The goal of data exploration is to gain a deeper understanding of the data, uncover insights, and generate hypotheses for further analysis. This process is an essential step in the data analysis workflow and helps to guide subsequent steps such as data cleaning, feature selection, and model building.

Below is the data dictionary for the **2023 Australia Open champions** dataset:

Attribute Name	Type	Description
Year	4 digit format in YYYY, Quantitative	Year of Tournament
Gender	Binary data, Categorical Nominal	Divides Tournament by Gender, either Men's or Women's Tournament
Champion	String Format	Name of the Tournament Winner
Champion Nationality	3 letter string format	ISO Alpha-3 Country Code of Champion
Champion Country	String Format	The full name of Champion Country
Champion Seed	N/A	
Mins	N/A	
Score	String Format, Pairs of Integer, Separated by Commas	The Score Results for Each Set of Games
1st-won	Numeric	Winning the First Game of the Match
1st-loss	Numeric	Losing the First Game of the Match
2nd-won	Numeric	Winning the Second Game of the Match
2nd-loss	Numeric	Losing the Second Game of the Match

3rd-won	Numeric	Winning the Third Game of the Match
3rd-loss	Numeric	Losing the Third Game of the Match
4th-won	Numeric	Winning the Fourth Game of the Match
4th-loss	Numeric	Losing the Fourth Game of the Match
5th-won	Numeric	Winning the Fifth Game of the Match
5th-loss	Numeric	Losing the Fifth Game of the Match
Runner-up	String Format	Name of the Tournament Runner-up
Runner-up Nationality	String Format	ISO Alpha-3 Country Code of Runner-up
Runner-up Country	String Format	The full name of Runner-up Country
Runner-up Seed	N/A	

Additional Columns has been created for Analysis Purpose:

Attribute Name	Type	Description
Win	Numeric	Total number of Wins
Loss	Numeric	Total number of Losses
Win Ratio	Integer Format(Ratio)	Ratio of Total Wins by Total Games

Data Preparation:

Data preparation is an important step in the data analysis process, as the quality of the results obtained from data analysis depends heavily on the quality of the data used. Data preparation is the process of cleaning, transforming, and organizing raw data into a format suitable for analysis.
Any inconsistency in the data can greatly hamper the analysis and visualization of the data which may result in making bad decisions for the management.

Additional Attributes:

Three additional attributes has been added for analysis purpose.

- The Win Column is numeric data that contains the sums of wins from the 1st to 5th match.

- The Loss Column is numeric data that contains the sums of losses from the 1st to 5th match.
- The Win Ration Column is an Integer data that contains the Ratio of Total Wins by Total Games

Missing Values:

Missing Values are a common occurrence in the dataset. Sometimes, these values are replaced by mean, highest value, and smallest values or they are completely ignored for the analysis.
I used the Pandas library in Python to find out the number of missing values. According to the table below, all of the values for Champion Seed, Mins, and Runner-up Seed are missing. So, they can just be removed. While the missing values from 1st win to 5th loss can be left as null. We have one value in row 116 for the Win Ratio Column that is divided by 0. It is because the match was a walkover. Thus it can be replaced by 0.

Colums	Empty values
Champion Seed	208
Mins	208
Runner-up Seed	208
1st-won	1
1st-loss	1
2nd-won	1
2nd-loss	1
3rd-won	65
3rd-loss	65
4th-won	143
4th-loss	143
5th-won	184
5th-loss	184
Win Ratio	1

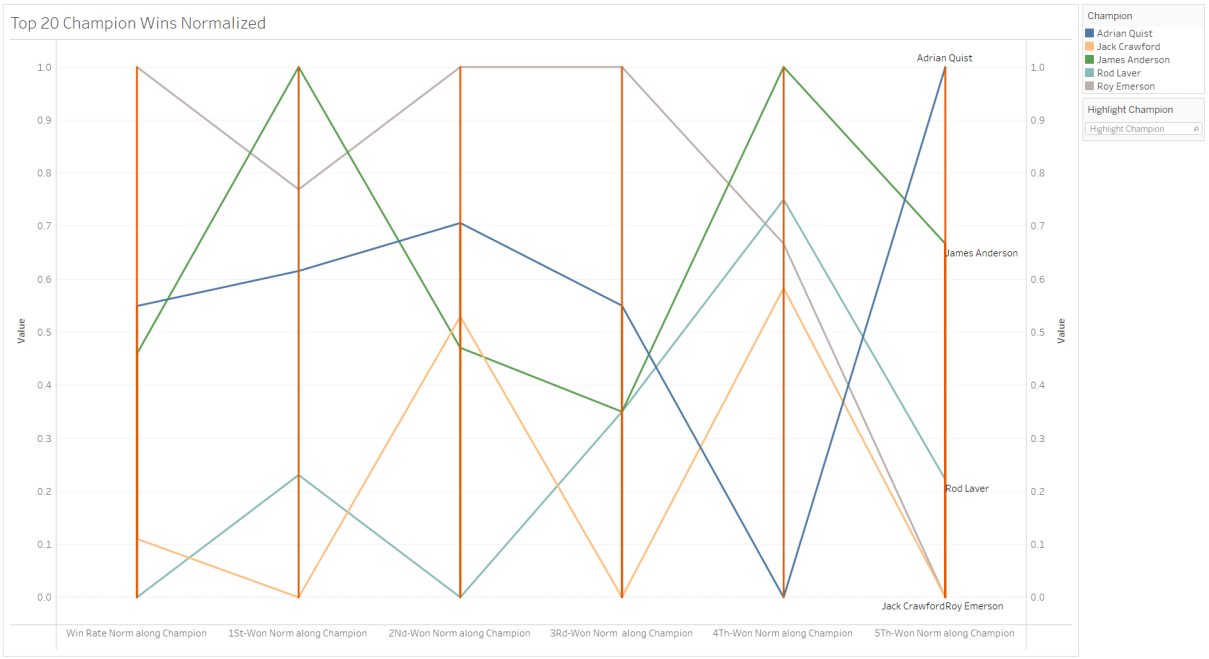
Visualization Techniques:

Parallel Coordinates:

Parallel Coordinate displays the data frequencies, relationships and aggregation patterns for multidimensional data. It is normally using a set of vertical axes represented as dimensions and polylines linked between axes at appropriate values. Parallel coordinates are useful for the interpretation of high dimension data and axis.

Interpretation and Pattern Analysis:

The Parallel Coordinate is used to analyze the relationship between top 20 champions Winrate, and all of the match wins. All of these values have been normalized for analysis purposes.



From the above graph, we can see that in order to maintain a high win rate ie above 0.5 the normalized wins for the player should also be higher than 0.5 on at least 3 wins.

The Advantages and Disadvantages of Parallel Coordinates are:

Advantages:

Allows for visualization of high-dimensional data: Parallel coordinates can effectively display data with a large number of variables or dimensions, making it easier to identify patterns and relationships in complex datasets.

Enables easy comparison of data across multiple variables: Parallel coordinates allow users to easily compare data across multiple variables or dimensions, as they can be displayed on the same scale and in the same visualization.

Can reveal correlations and outliers: Parallel coordinates can reveal correlations and outliers in the data, as they highlight relationships between variables and can show when data points fall outside of expected patterns.

Provides interactive exploration: Interactive parallel coordinates can allow users to interact with the data, selecting subsets of data to visualize or zooming in on particular data points for closer inspection.

Disadvantages:

Can be difficult to interpret: Parallel coordinates can be difficult to interpret for users who are not familiar with the visualization technique, especially when there are a large number of variables or dimensions.

Can be sensitive to data order: The order in which variables are displayed in parallel coordinates can impact how the data is interpreted, and changing the order can change the appearance of the visualization.

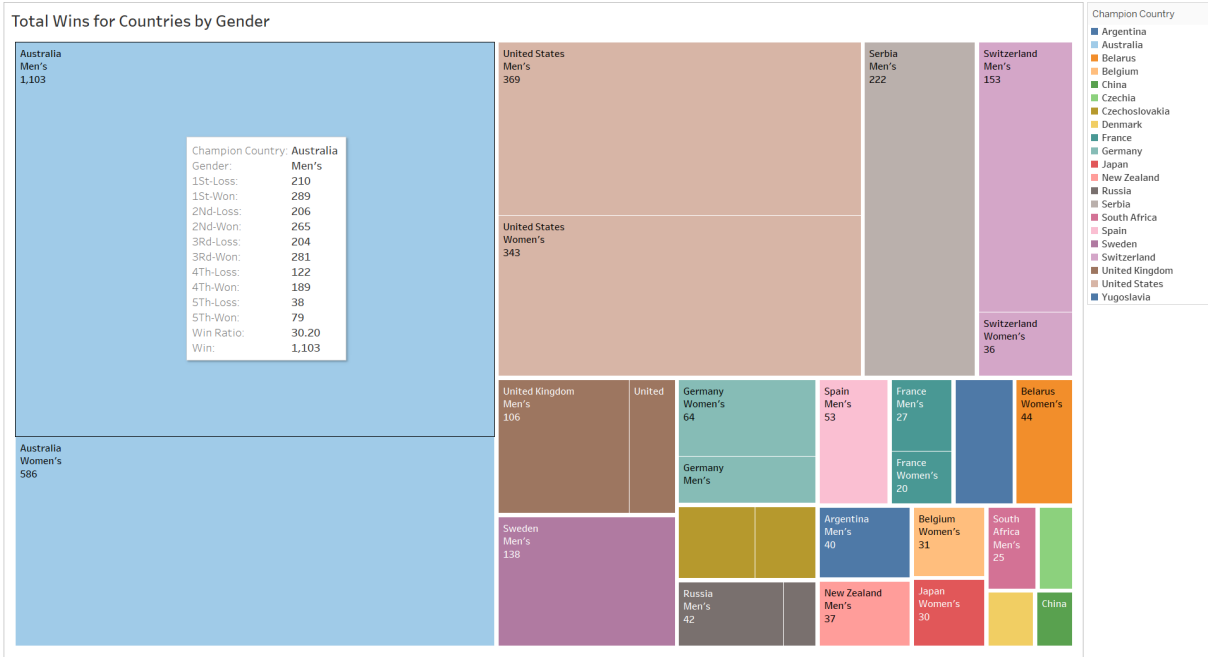
Can be affected by scaling and normalization: The scaling and normalization of data in parallel coordinates can impact how the data is displayed and interpreted, and different scaling methods can produce different results.

May not be suitable for all types of data: Parallel coordinates may not be suitable for all types of data, particularly when there are many categorical variables or when the relationships between variables are not linear.

Treemaps

A treemap displays structured hierarchical data as a set of nested rectangles (or leaves) by using different sizes and colours. The leaf size illustrates the data value and the colours show separate data categories. Higher leaf size equals higher values and vice versa.

Interpretation and Pattern Analysis:



The treemap is configured with 2 levels of hierarchy, Champion Nationality and Gender. The Win Column is used to determine the size of each rectangle. Labels for Champion Country Gender and Sum of wins have been added. Many details like Win Ratio, 1st Won, 1st Loss, 2nd Won, 2nd Loss etc have been added.

From the Map, we can see that Australia has the highest number of Wins followed by the United States and Serbia. Men have higher wins than female in almost all countries except Switzerland, Germany, Belarus etc

The Advantages and Disadvantages of Treemaps are:

Advantages:

Efficient use of space: Treemaps are designed to efficiently use the available space, allowing users to visualize a large amount of data in a relatively small area.

Can display hierarchical data: Treemaps are ideal for displaying hierarchical data, as they can represent the nested structure of the data using colour, size, and position.

Facilitates easy comparison: Treemaps enable easy comparison of data by using a colour scheme and size of the rectangles to represent the data.

Interactive: Interactive treemaps enable users to zoom in and out of the visualization, drill down into sub-trees, and hover over individual rectangles to view more information.

Disadvantages:

Complexity: Treemaps can be complex to read and understand, especially when there are multiple levels of hierarchy and a large amount of data.

Difficult to compare individual items: While treemaps facilitate easy comparison of groups of data, they can make it difficult to compare individual items within each group.

Can be difficult to colour-code: Choosing an appropriate colour scheme for treemaps can be challenging, as it can be difficult to find colours that are easily distinguishable from each other.

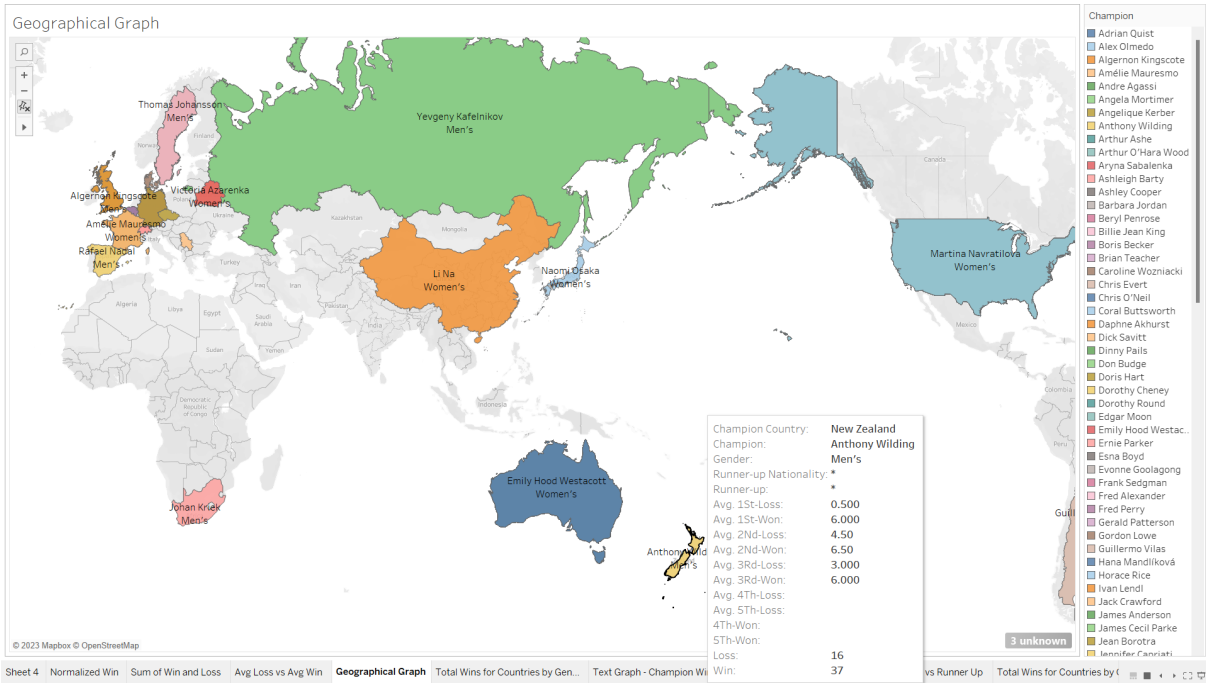
May require preprocessing: In some cases, data may need to be preprocessed before it can be visualized in a treemap, as the layout of the rectangles can be influenced by the order in which the data is presented.

Geographical Map:

3D Geographic data mapping is a method to visualise data on top of a map and illustrates categories, values and other data by using colour, sharp, and labels for comparison.

Interpretation and Pattern Analysis:

When provided with the country name, Tableau automatically generates the Longitude and Latitude feature. This can be used to create a Geographical Graph.



The Geographic Map is configured with 1 level of hierarchy, Champion Nationality. The different colour represents the different nation of the Champions. Labels for Champion Country Gender have been added. Many details like Runnerup, Runnerup Nationality, Average 1st Won, 1st Loss, 2nd Won, 2nd Loss etc have been added.

The Advantages and Disadvantages of a Geographical Map are:

Advantages:

Spatial analysis: Geographical maps enable a spatial analysis of data, making it easier to identify patterns and relationships between data points that are related to location.

Visualization of geographic trends: Geographical maps are ideal for visualizing geographic trends, such as population density, regional economic growth, and climate changes.

Easy to understand: Geographical maps are easy to understand, as most people are familiar with the geography of their local area and can easily interpret the location and density of data points.

Can help with decision-making: Geographical maps can help decision-makers, such as policymakers and urban planners, make informed decisions by visualizing the impact of various options on the local community.

Disadvantages:

Limited scope: Geographical maps are only useful for displaying data that is related to geography, such as population density or natural resources.

Difficulty with high-dimensional data: Geographical maps can be difficult to use with high-dimensional data, as it can be challenging to represent more than a few variables at a time.

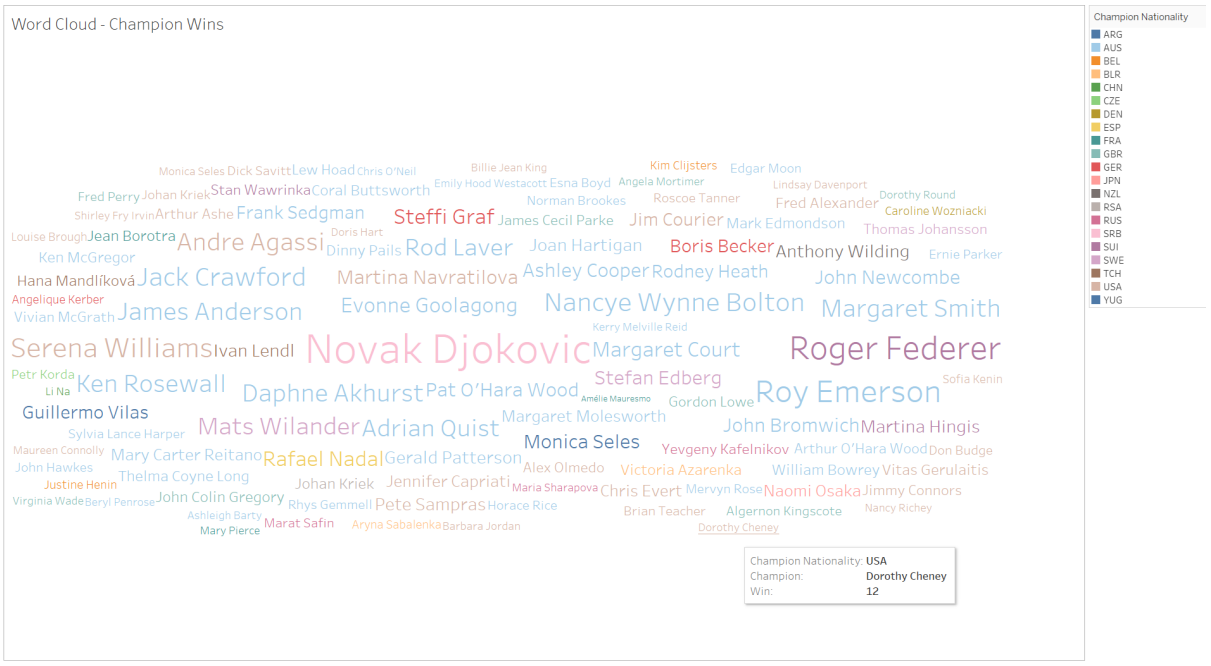
Sensitivity to spatial scale: The spatial scale of a geographical map can impact the way the data is displayed and interpreted, and different scales can produce different results.

Map projection issues: The projection used for a map can impact the shape and size of countries and regions, which can affect the way the data is displayed.

Word Cloud:

Word Cloud is a basic visualization method for text. Higher the number of words, the bigger the size and vice versa.

Interpretation and Pattern Analysis:



The Word Cloud is configured with 2 levels of hierarchy, Champion Nationality and Sum of Wins. The different colour represents the different nation of the Champions and the size of the text represents the total wins for the Champion. So, a higher number of wins equals the bigger size of the Champion. Label for Champion Nationality Total Wins and Champion has been added.

From the Word Cloud, we can see that Novak Djokovic has the highest number of wins followed by Roger Williams, Roy Emerson, Serena Williams etc.

The Advantages and Disadvantages of Word Cloud are:

Advantages:

Provides a quick overview: Word clouds provide a quick overview of the most common words or themes in a text, allowing users to identify key topics or trends at a glance.

Easy to create: Word clouds are easy to create using a variety of online tools, making them accessible to a wide range of users.

Provides a visual representation: Word clouds provide a visual representation of text data, which can help to make the data more engaging and memorable.

Facilitates exploration: Word clouds can facilitate the exploration of text data by allowing users to drill down into specific words and their associated context.

Disadvantages:

Ignores context: Word clouds do not take into account the context or meaning of the words they display, which can lead to misinterpretation of the data.

Limited to one-dimensional data: Word clouds are only useful for displaying one-dimensional data, such as word frequency, and may not be suitable for more complex data sets.

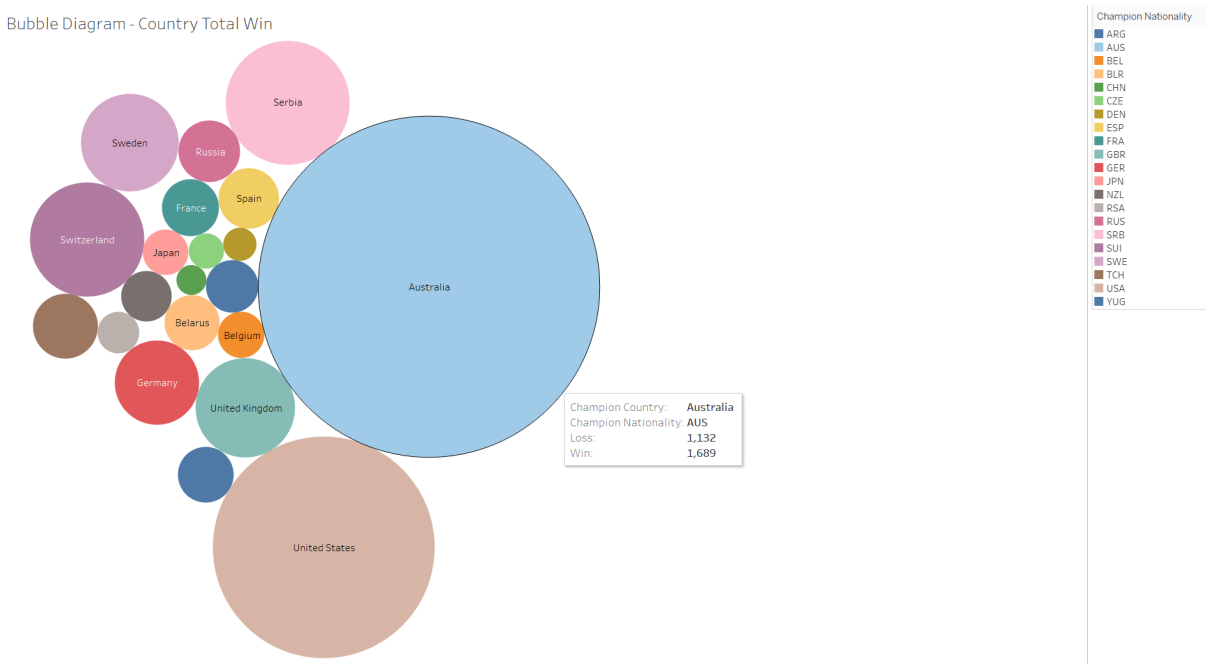
Limited customisation: Word clouds are often limited in terms of customisation options, such as font and colour choices, which can make them less visually appealing.

Can be misused: Word clouds can be misused to manipulate or misrepresent data, especially if they are not created and presented in an ethical and transparent manner.

Bubble Chart

Bubble chart displays three dimensions of data. Each entity (x,y,z) of associated data is plotted x and y location and the z size.

Interpretation and Pattern Analysis:



The bubble chart is configured with 2 levels of hierarchy, Champion Nationality and Win. The Win Column is used to determine the size of each Bubble and Champion Nationality determines the colour. Labels for Champion Country have been added.

From the chart, we can see that Australia has the highest number of Wins followed by the United States and Serbia.

The Advantages and Disadvantages of Bubble Chart are:

Advantages:

Can display three dimensions: Bubble charts can display three dimensions of data, with the size of the bubbles representing one variable and the position on the x and y axis representing two other variables.

Facilitates easy comparison: Bubble charts enable easy comparison of data points, as users can quickly compare the size and position of bubbles.

Provides a clear visual representation: Bubble charts provide a clear visual representation of data, making it easier for users to understand the relationships between variables.

Can handle large data sets: Bubble charts can handle large data sets, as they can display a large number of data points without becoming cluttered or difficult to read.

Disadvantages:

Limited to three dimensions: Bubble charts are limited to displaying three dimensions of data, which may not be sufficient for more complex data sets.

Limited to continuous data: Bubble charts are most useful for continuous data sets, and may not be as effective for displaying categorical or discrete data.

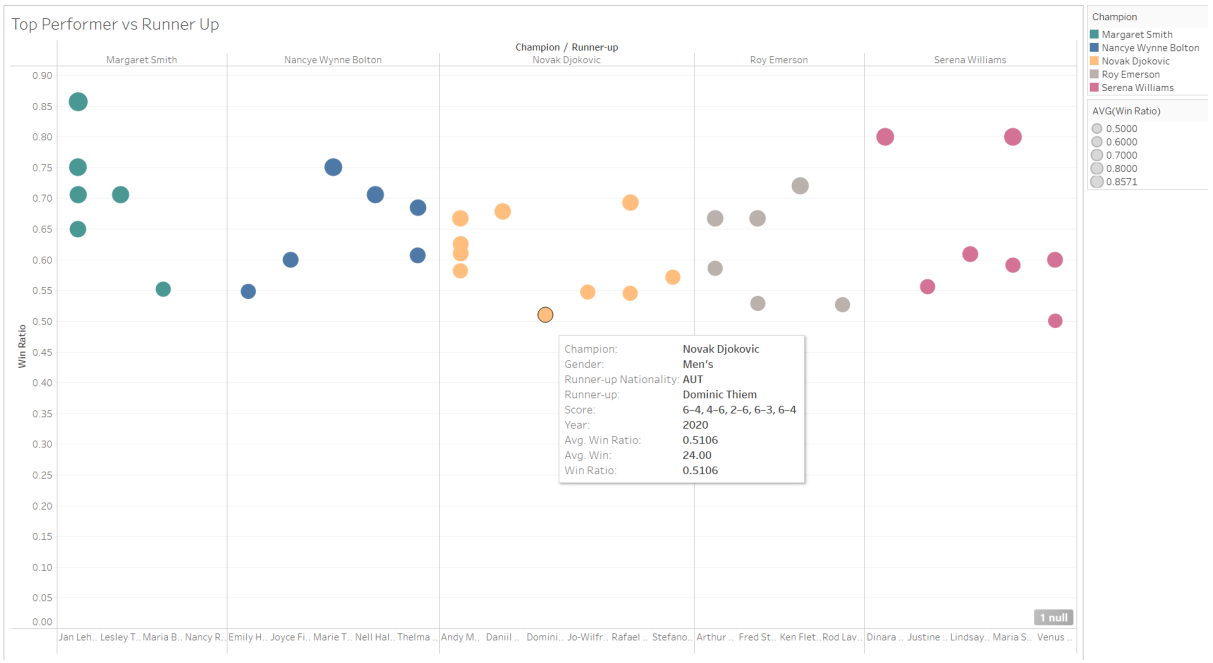
Limited customisation options: Bubble charts can be limited in terms of customisation options, such as colour and shape choices, which can make them less visually appealing.

Requires careful scaling: Bubble charts require careful scaling to ensure that the size of the bubbles accurately represents the data, and incorrect scaling can lead to misinterpretation of the data.

Scatter Plots:

The scatter plot displays two dimensions of data. Each entity (x,y) of associated data is plotted in x and y locations.

Interpretation and Pattern Analysis:



The Scatter Plot is configured with 3 levels of hierarchy, Champion, Runner-up and Win Ratio. The Win Ratio Column is used to determine the size of each plot and the Champion determines the color.

Details for Champion, Gender, Win Ratio, Runner-up Nationality, Score, Year etc have been added.

In the diagram, Champion is plotted on the higher x-axis, and Runnerup is plotted on the lower axis. The Win ratio is plotted on the y-axis. The filter has been applied to only select the top 5 winners.

From the diagram, we can see that Each champion has at least 4 runner-ups. There have been more finals for Novak Djokovic va Andy Murray and Novak Djokovic came out on top. Similarly, in Margaret Smith vs Jah Lehane where Margaret Smith came out on top.

The Advantages and Disadvantages of Scatter Plots are

Advantages:

Displays relationships between variables: Scatter plots can display the relationship between two variables, making it easier for users to understand the correlation between them.

Provides a clear visual representation: Scatter plots provide a clear visual representation of data, making it easier for users to identify patterns and outliers.

Suitable for both categorical and continuous data: Scatter plots can be used to display both categorical and continuous data, making them a versatile option for many types of data sets.

Easy to add additional variables: Scatter plots can easily incorporate additional variables through the use of colour or size, providing a way to display more complex data sets.

Disadvantages:

Limited to two dimensions: Scatter plots are limited to displaying two dimensions of data, which may not be sufficient for more complex data sets.

Can be misleading: Scatter plots can be misleading if the axes are not carefully labelled or scaled, which can lead to misinterpretation of the data.

Can become cluttered with large data sets: Scatter plots can become cluttered and difficult to read with very large data sets, making it hard for users to identify patterns or relationships.

Requires careful interpretation: Scatter plots require careful interpretation to identify relationships or patterns, and may not be suitable for users who are less familiar with statistical concepts.

Executive Summary:

The report focuses on exploring the Tennis dataset of Australia Open championship matches, which includes 118 years of championship matches, both men and women, between 1905 and 2023 (208 total matches). It contains 209 rows and 23 columns. Additional columns have been created for analysis purposes which is discussed later.

The tools I have used for exploratory and visualization purposes are mainly Excel and Tableau. The different charts and graphs that help us analyze the given datasets are:

1. Parallel Coordinates
2. Treemaps
3. Geographical Map
4. Word Cloud
5. Bubble Chart
6. Column Chart

Conclusion:

After the analysis of the above-mentioned chart, we can come to the following conclusions:

- In order to maintain a high win rate ie above 0.5 the normalized wins for the player should also be higher than 0.5 on at least 3 wins.
- Australia has the highest number of Wins followed by the United States and Serbia.
- Men have higher wins than female in almost all countries except Switzerland, Germany, Belarus etc
- Novak Djokovic has the highest number of wins followed by Roger Williams, Roy Emerson, Serena Williams etc.
- Each top 5 champion has at least 4 runner-ups. There have been more finals for Novak Djokovic va Andy Murray and Novak Djokovic came out on top. Similarly, in Margaret Smith vs Jah Lehane where Margaret Smith came out on top.