Executive Summary

Summarise data characteristics

Data Name	Data Type	Data Description
Year	Numeric quantitative interval scale	This value describes the year the event took place
Gender	Categorical nominal	Represents whether mens or womens championship
Champion	Text, categorical nominal	This field is used for the name of the championship winner
Nationality	3 letter pairs represented in text format	Depicts the nationality of the player
Score	Integrer pairs represented in text format	Depicts the score of each game
Runner-up	String, categorical nominal	This field is used for the name of the championship runner-up
Runner-up country	String, text categorical nominal	The nation of the runner-up

Summary of the dataset

- Found that there were 128 number of distinct winners of the tournament throughout the 294 match history of the tournament
- Only 67 players have won 2 or more championships
- 33 players have won the tournament 3 or more times
- The most frequent nationality to win is the United States with a total of 177 victories
- 16% of matches were won in 5 sets for the men

Data Cleansing

The first step to cleansing a dataset is to locate any missing values. To fix this issue, the first solution would be to try and fix the value, otherwise the value will need to be removed.
The next step is to look for misspelt values, such as names of countries, or alternate spellings of player names.
The next step is to look for any values that may be defined as a string or decimal that should be an integer format. This could cause significant issues when trying to create visualisations.

Essentially, ensuring that the data is in the correct format so that it can be manipulated in the correct way.

Calculating Win Rate

To calculate the win rate of each player in Excel I followed the steps below:

- Create new sheet named *win rate*
- Create headings; *win rate, wins, loss*
- Calculate wins by adding 1st set wins, 2nd set wins, 3rd set wins, 4th set wins and 5th set wins
- Calculate loss by adding 1st set loss, 2nd set loss, 3rd set loss, 4th set loss and 5th set loss
- Calculate win rate with the following formula: $wins \ / \ (wins \ + \ loss)$

Additionally the runners up win rate will be calculated by following the same process as above. The same process is also performed using the players who have won 5 or more championships.

Treemap

A treemap represents structured hierarchical data as a collection of layered rectangles.
It is frequently used to determine the relative magnitude of data groupings. Larger items signify a higher level of importance. Colour can be used to distinguish items that outperform or underperform their counterparts in the same category. The largest box is in the upper left corner and gets smaller as it approaches the lower right corner. This makes distinguishing a hierarchical structure simple. The key idea of a treemap is to represent a hierarchical graph structure within rectangular space instead of using a typical tree structure.

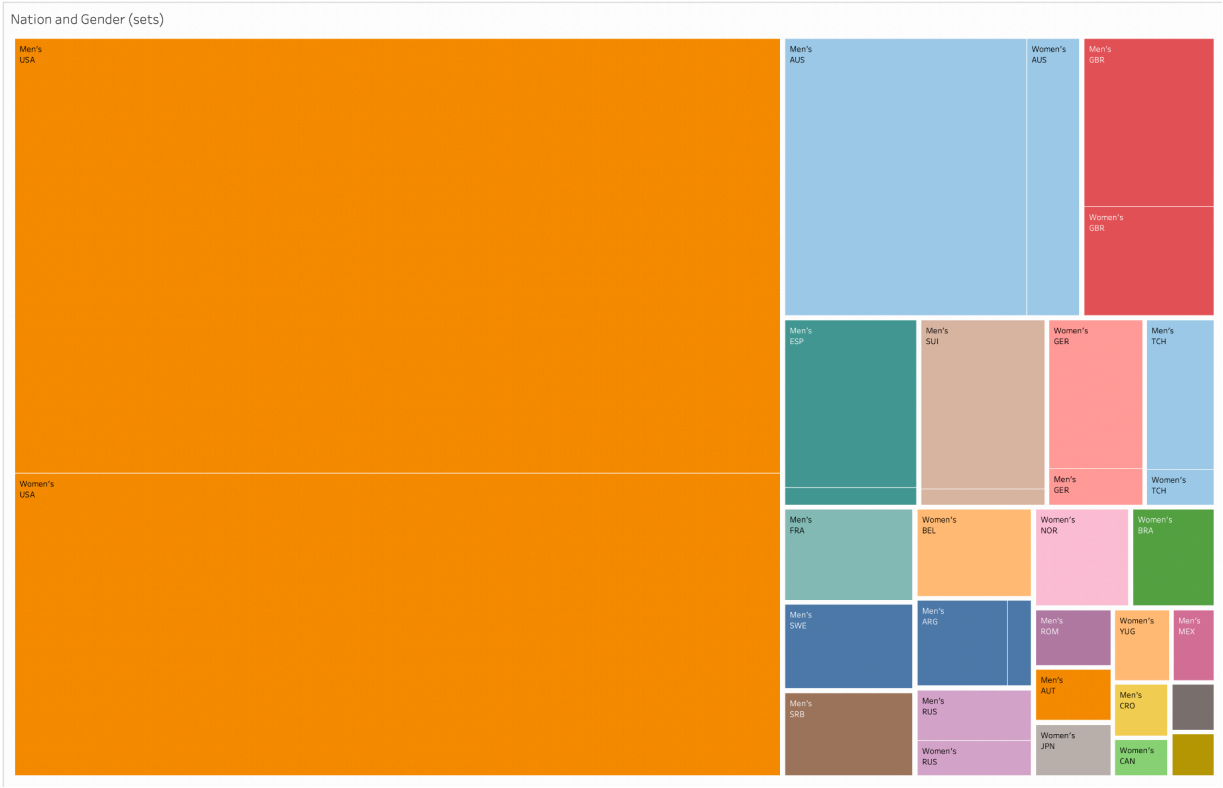

Figure 2.1 Nation and Gender (sets)

The treemap in figure 2.1 looks at the two variables gender and nationality in relation to the number of sets won at the US Open. The number of sets won is identified by the size of each tile. The nationality of each is shown at the top left of the tile, with the gender being displayed underneath nationality. It is clear to determine that the USA have won the highest number of sets, followed by Australia and Great Britain. Expanding a specific nation by clicking on the associated rectangle will reveal more detailed statistics. These details include the number of sets the nation has won as a whole further broken down into 1st set won, 1st set loss, 2nd set won and so on.

It is interesting to note that in almost all cases the nation's men's side have won more sets than the women's side. This is due to the fact that in the men's championship they play to the best of 5 sets, whereas the women play to the best of 3 sets. We will take a look at the win rate of each nation in figure 2.2 below. By looking at the win rate rather than the outright number of sets won, it will act to normalise the difference between the mens 5 set matches and the womens 3 set matches as win rate is calculated using a percentage. This will allow us to see whether the men's side or womens side perform better at the US Open.

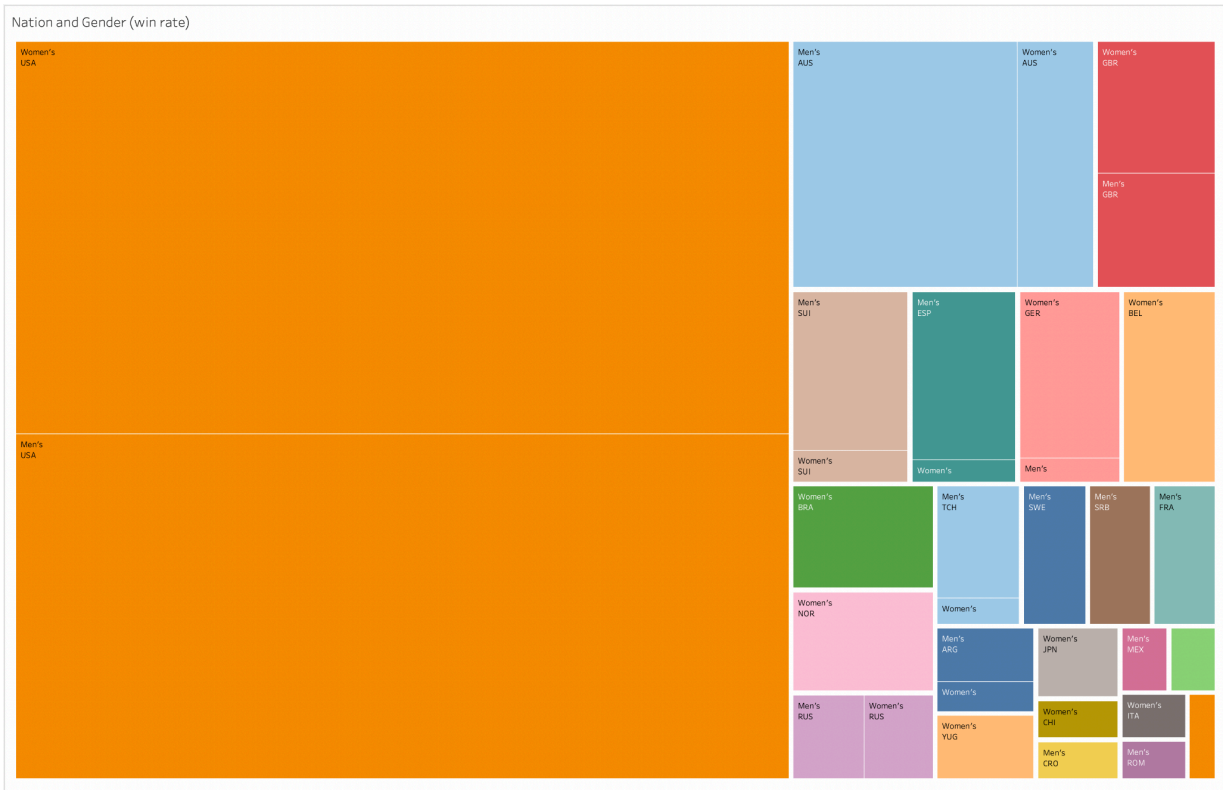

Figure 2.2 Nation and Gender (win rate)

In this treemap the champion's nationality is used for the colour feature of the tree-map. The size of each rectangle is set using the win rate of each nation. This win rate is calculated by taking the total number of sets won divided by the total number of sets played. Now you can see that for most nations the results between mens and womens match more closely than in the previous graph (*figure 2.1*). In particular, the win rate of the women representing the USA is 59.67 whereas the men's win rate is 52.45. Great Britain showcases a similar result to that of the USA.

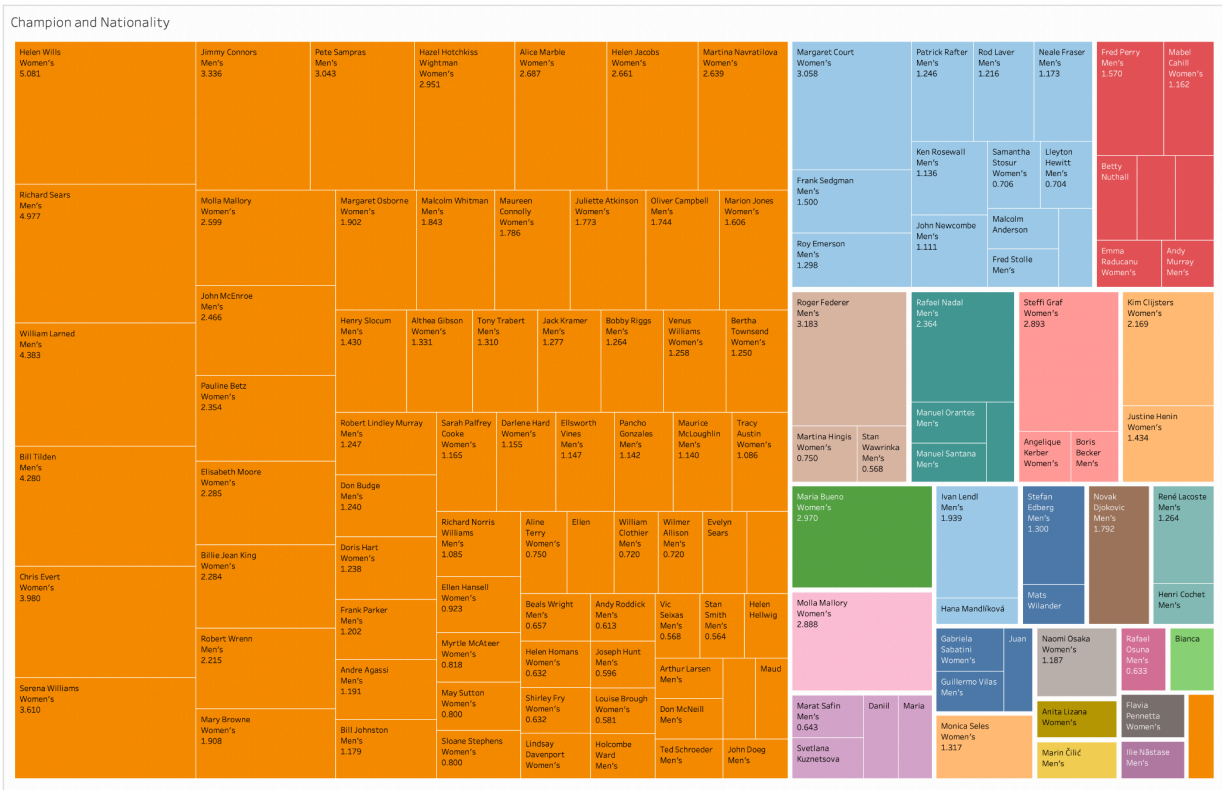

Figure 2.3 Champion and Nationality

In the Treemap above (*figure 2.3*), we can see the most effective and best players of each nation. This treemap is created using 3 hierarchical measures including champion, nationality and gender, The colour is defined by the champions nationality, while the size is defined by the win rate of each champion. Each nation is grouped making it easy to distinguish between separate nations. In the detail section the inclusion of win rate, gender, nationality and score makes it clear to see detailed statistics at a glance.

Advantages and Disadvantages

An advantage of utilising treemaps is that you can apply data labels in a more intuitive way than with parallel coordinates. For example, in my case I have labelled gender, nationality and win rate directly on each rectangular segment. Additionally, it is easy to group data, such as grouping nationality as well as grouping sub-groups such as champion win rate and gender. This makes it easy to distinguish between these subgroups. For example, in figure 2.3 nationality is colour coordinated, with orange representing champions who originated from the USA and blue representing Australia.

A negative of using treemaps is that other charts may be useful and may allow for a clearer understanding of how the data correlates with each other. It is easier to determine correlations between data when using parallel coordinates. I also ran into the issue of size distortion, wherein the rectangle box showcasing USA occupied more than half of the available space.That meant that other nations were very small and became lost in comparison. There is no way

around this as scale is one of the defining features of treemaps. Additionally, it can be difficult to determine the value of an item without the use of labels.

Parallel Coordinates

Parallel coordinates are used for higher dimensional data, which often means data with four or more variables. There are several axes that are all vertically aligned, and each axis has its own scale. It is an efficient method of plotting several data variables rather than constructing 3 or 4 dimensional plots, which are difficult to grasp. If lines between 2 axes run parallel with each other, this indicates some correlation. In contrast, the presence of perpendicular lines implies that there is no link between two variables.

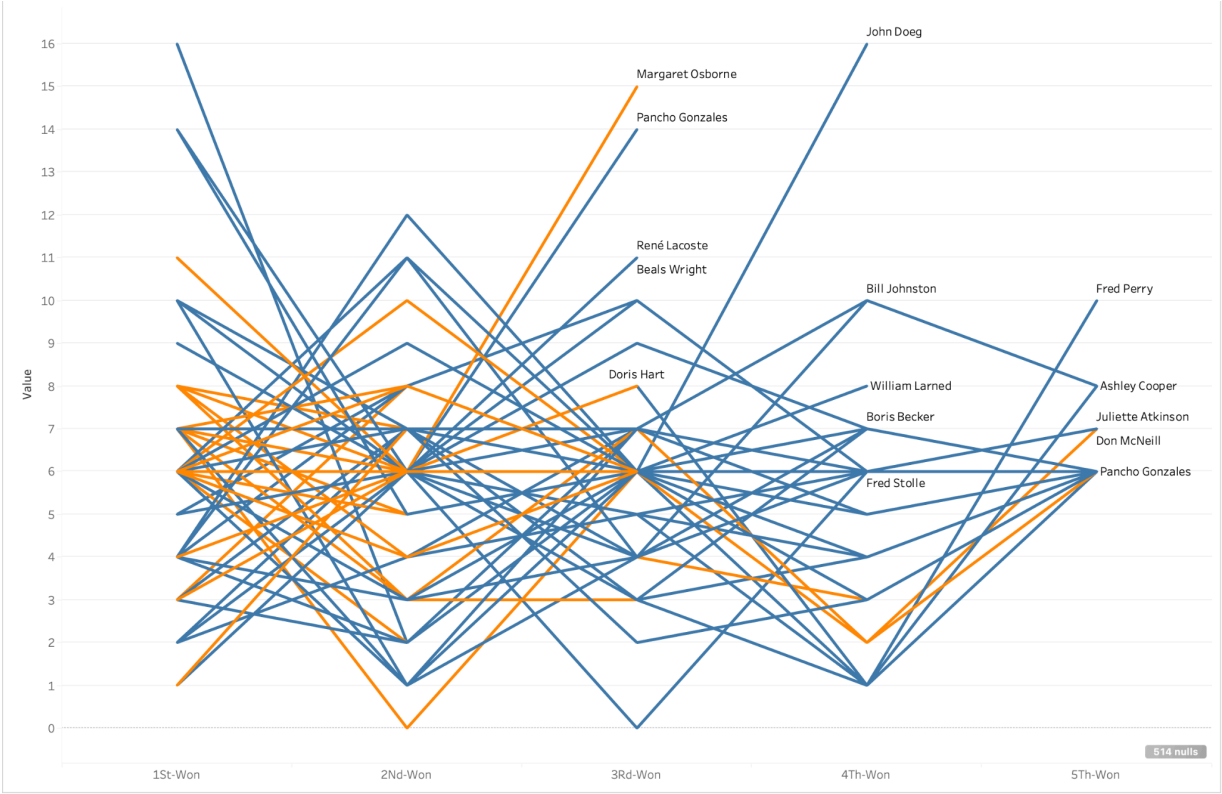

Figure 1.1 Parallel Coordinate Sets Won

In the above parallel coordinates, we are looking to see if there is any correlation between winning the first set and then going on to win the match. We can see an interesting trend, wherein the majority of players who won the first set go on to lose the second set, this occurs equally between both men and women. It seems to be a consistent theme amongst champions to fluctuate between winning the first, losing the second, winning the third, losing the fourth and then winning the fifth and the match. For example, mens champion Pancho Gonzales has won

16 first sets, and only 2 second sets. It is similar for John Doeg who won 10 first sets, then only 1 second set, before winning 6 third sets and 16 fourth sets.
From this parallel coordinate graph, we can also discover the players who have won the highest number of each set:

- Pancho Gonzales has won 16 first sets
- Fred Stolle has won 12 second sets
- Pancho Gonzales has won 14 third sets
- John Doeg has won 16 fourth sets
- Fred Perry has won 10 fifth sets

In terms of the women:

- Billie Jean King has won 11 first sets
- Darlene Hard has won 10 second sets
- Margaret Osborne has won 8 third sets

Advantages and Disadvantages

Unless using a more limited dataset, I found parallel coordinate graphs to be excessively cluttered. Understanding the relationship between multiple variables becomes more difficult if they are not thoroughly thought out. However, parallel coordinates allow you to display higher-dimensional data. This makes it more helpful than simple column or bar graphs, which can only display two-dimensional data.

Geographic Map

Geographic maps are used to create intricate visualisations of large quantities of geographically related data. They are commonly used in weather forecasting, real estate, and public health. Geographic maps can indicate significant trends and occurrences, particularly when it comes to a champion's nationality.

Figure 3.1 Champion by Country

The nationality of each champion is displayed on this geographic map (*Figure 3.1*). Tableau uses data containing a champion's nationality to calculate the champion's location. Geographical data is particularly useful for displaying champions nationalities and provides a distinct advantage when compared to other graphs such as treemaps. The size of the circles over a nation represent the number of champions of that nation's origin. Larger circles indicate a higher number of wins and vice versa. The colour of each nation appears as a different colour to help separate countries from each other. I have selected the pre-defined outdoors style for the map's background as I found this to be the most usable. I enabled state and country borders in tableau's background map layer options. The inclusion of annotations helps to depict the story of the US Open champions dataset, providing viewers with interesting findings within the dataset.

Interestingly, there have been 24 distinct nations to win a US Open championship with the US holding the most wins at 177. Australia holds the second most number of wins at 24 and Great Britain accounts for 10 wins. The majority of nations to have won a US Open championship other than the United States appear to be European. This is clearly seen in the above graph with a vast clustering of data points appearing in the European continent. No champions appear to be of African descent. It is also interesting to note that despite comfortably being the most populated nations, neither India nor China have won an US Open championship.

Figure 3.2 Runner-Up by Nationality

On this geographical map, the nationality of each runner-up is displayed (Figure 3.2). Tableau calculates a champion's location using data containing the champion's nationality. Again, the size of the circles around a country represents the number of champions from that country's origin. Larger circles represent more wins, and vice versa. The colour of each nation appears as a different colour to help separate countries from each other. I have re-used the pre-defined outdoors style for the map's background and enabled state and country borders in tableau's background map layer options. The inclusion of annotation draws provides interesting findings from within the dataset.

Interestingly, there have been 23 unique nations to achieve a runner up position at a US Open championship with the US appearing the most with 174. Australia appears 24 times and Great Britain appears 14 times. Once again, despite being the most populous nations, neither China nor India appear to have garnered a runner-up position at a US Open championship. However, South Africa have achieved a runner-up position on 3 occasions.

Advantages and Disadvantages

It is clear to see the nationality of the champion, as maps of the world are easier to understand than rectangular boxes. The ability to click on a specific country and uncover additional information is more simplistic while using a geographic map.

However, it is more difficult to group data like you are able to do using treemaps. For example, within a treemap I was able to group by nationality and additionally split each rectangle into gender. Moreover, it is difficult to uncover relations between data which you are able to do when using parallel coordinates.

Performance Patterns of top players

There were 12 players who won 5 or more championships with Molla Mallory on 8 being the highest. To take a closer look at the top players, a simple bar chart can be used. The number of wins is set on the x-axis and the name of the player is set on the y-axis. I found the bar chart conveyed this simplistic style of information in the clearest manner when compared to alternative graphing strategies. This chart (*figure 4.1*) can be seen below.

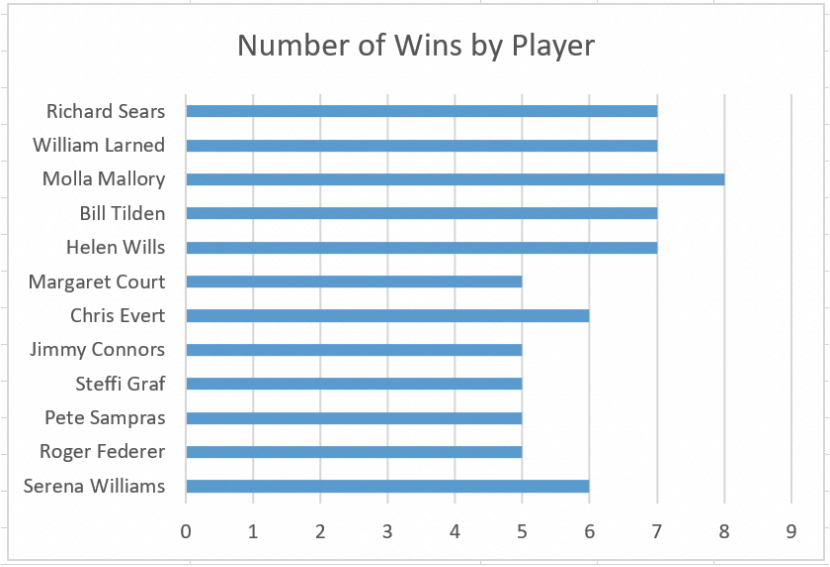

Figure 4.1 Number of Wins by Player

Additionally, I decided to use tableau to create a bubble chart (*figure 4.2*) as it draws your attention in a more dynamic way than regular bar charts. It also allows you to utilise the detail feature within tableau to allow more information to be uncovered when someone hovers over a bubble. The size of each bubble is directly related to the number of wins that player has achieved.

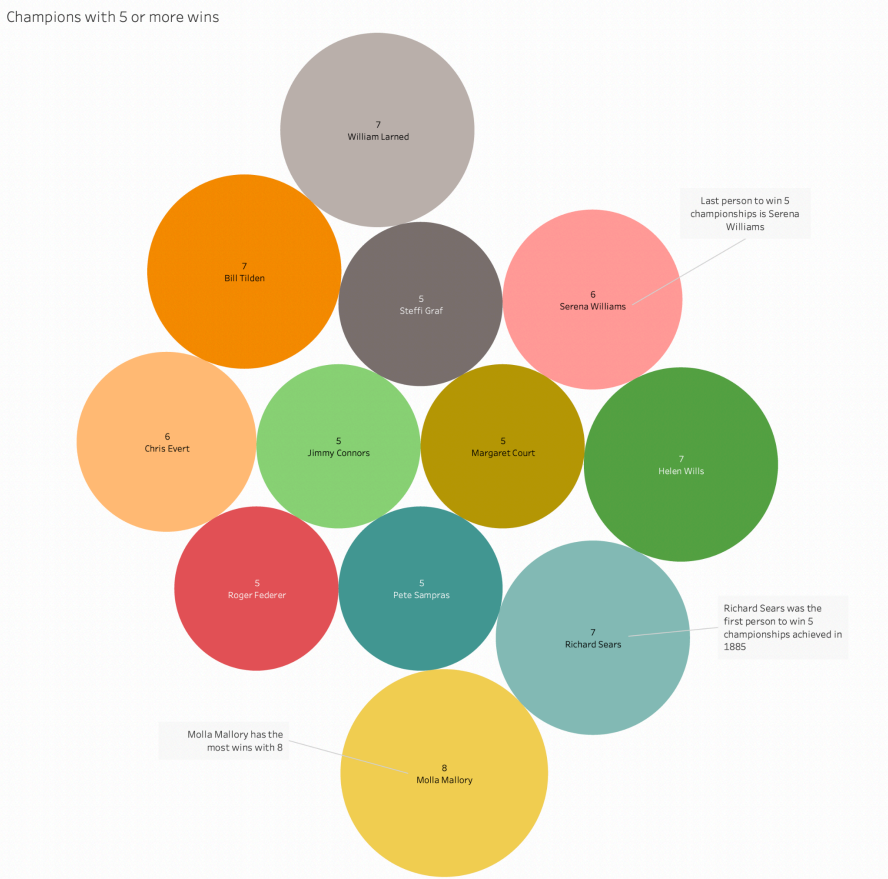

Figure 4.2 Champions with 5 or more wins

A few interesting finds found while looking at those who have 5 or more championships is that the highest number of wins in mens is tied between 3 individuals Richard Sears, William Larner and Bill Tilden. It is also interesting to note that Richard Sears won 7 titles in a row beginning from the inception of the championship. There are several players who have 3 or more consecutive wins. These include William Larned (1907-1911), Molla Mallory (1915-1918), Chris Evert (1975-1978), Roger Federer (2004-2008), Bill Tilden (1920-1925), Helen Wills (1923-1925, 1927-1929) and Serena Williams (2012-2014).

Only 5 nations have won 5 or more championships, those being USA, Norway, Switzerland, Germany and Australia. The most successful nation is the United States winning 74%, with the other nation winning 7% each. This can be seen clearer in figure 4.3, where a pie graph is used as there are few nations making it easy to determine the percentage of wins between nations.

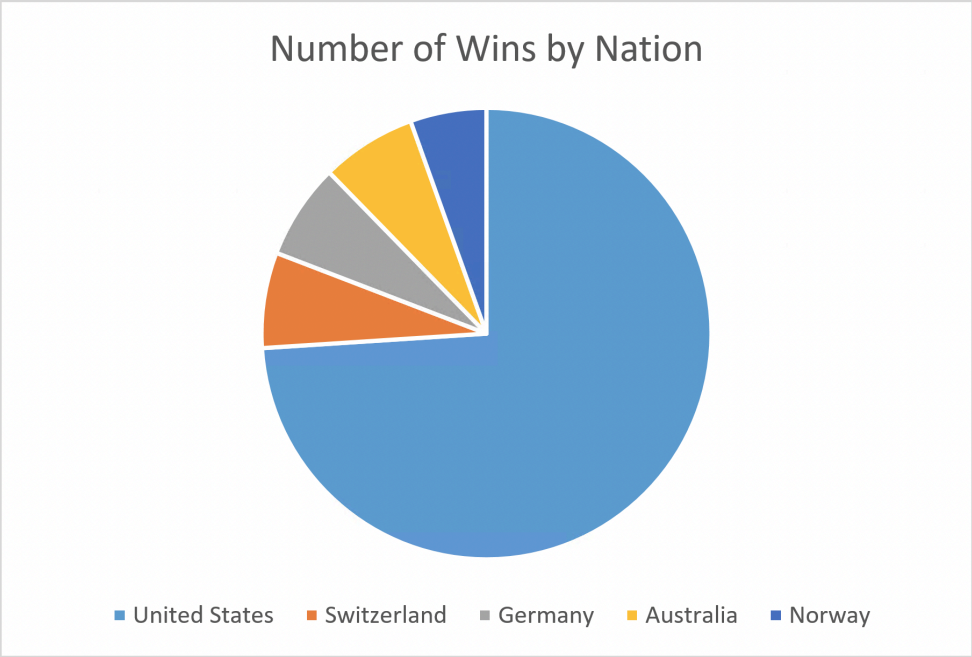

Figure 4.3 Number of Wins by Nations

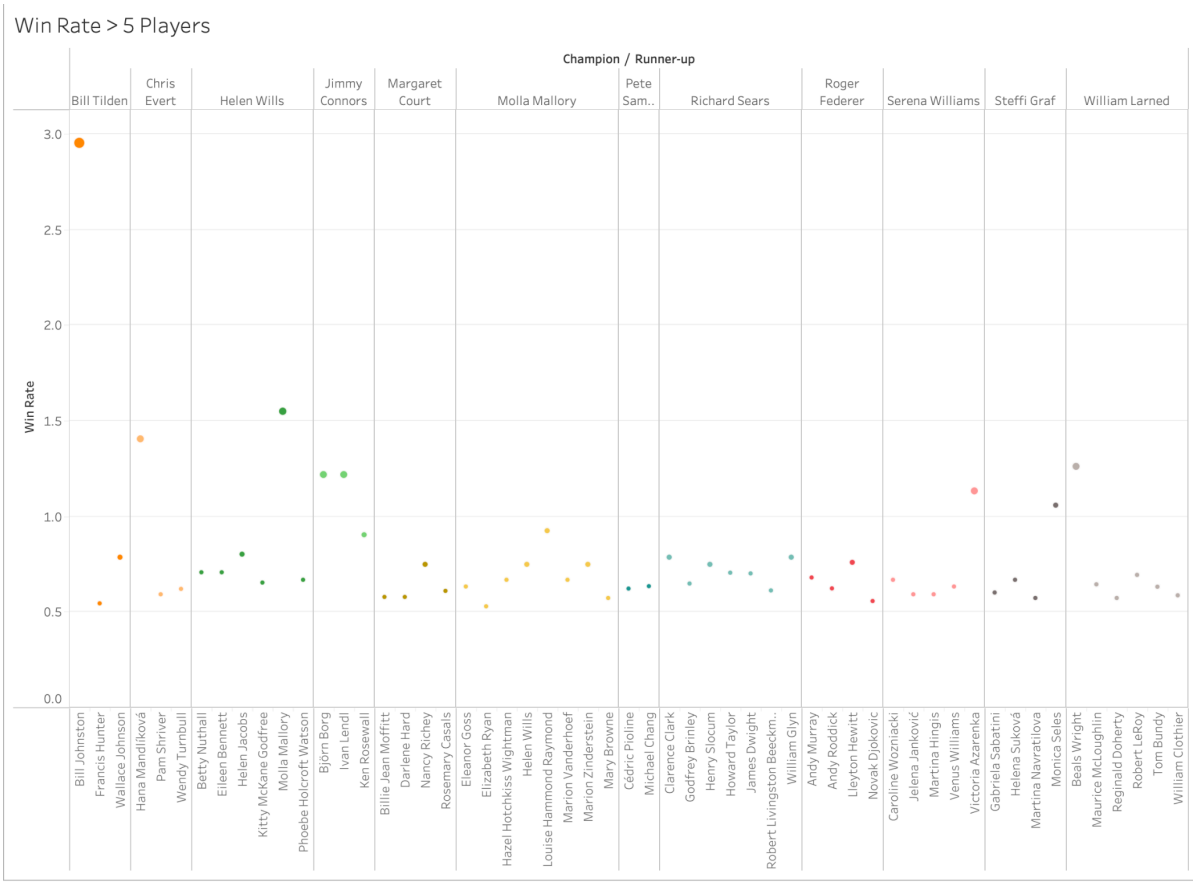

Figure 4.4 Top Players Match Performance

In tableau it is easy to create a visualisation depicting all the matches played by the top players. To do this, we create a standard bar graph using the columns *Champion* and *Runner-up* from the *Win Rate > 5* excel sheet. The colour is determined by the distinct champion making it clear to distinguish between each player. Size is determined by the win rate, larger circles indicating a more dominant performance in that match. The data in this case has some problems. The problem stems from tableau believing that wins against a repeat player count as a single match. This, for example, leads tableau to believe that Bill Tilden won 26 first sets in a single match versus Bill Johnston. This is due to Bill Tilden competing in three different finals against Bill Johnston. This is also the case for Helen Wills and Molla Mallory as well as Serena Williams and Victoria Azeranka. Adjusting to remove these specific games, we can find that the most dominant final match occurred between Molla Mallory and Louise Hammond Raymond where Molla won 6-0, 6-1.

Conclusion

There were a number of graphic attribute design and labelling techniques used in creating this report. These include the following:

- Made use of colour to help group different categories such as nationality and gender. The use of colour helped to tell the story as it made it easier to understand graphs. For example, in figure 3.2 I used colour to separate the nationality of the players. The USA is depicted in orange, and Australia is depicted in blue thus making it super clear to determine groupings.
- Moreover, utilisation of labelling to emphasise important data points. This aids in storytelling as the reader is directed towards the most important aspect of the graph, thus avoiding having to deduce information for themselves. For example, in the top players analysis in particular the bubble chart the use of annotations outline the key data.
- I also utilised the detail feature under the marks tab to display additional and relevant information on top of labels. For example, this allowed me to display gender and nationality as a label on each rectangle in figure 3.1, but display more information in each rectangle. This information included 1st set won, 1st set loss, score, win rate and more, accessible when the user hovers over a specific rectangle. It acts to encourage the viewer to discover additional information without crowding them on initial viewing

It is clear to see the United States dominance at their own championship throughout the years. They have won the most number of championships, and achieved a runner-up position on more occasions than any other nation. Players of American descent are the most likely to win based on the data, and moreover 74% of all sets won by top players have come from an American. There are some additional interesting findings that came from this dataset including:

- Richard Sears is the most dominant men's player with a win rate of 4.977
- Helen Wills is the most dominant women player with a win rate of 5.081
- In terms of the US, mens players have won more sets than women
- However, women have a higher percentage of sets own than men, 59% compared to 52%
- The United States have won a combined 3,286 sets at the US Open
- The next closest nation, Australia have only won 482 sets
- There are only 12 players who have won 5 or more championships
- No nation in the continent of Africa has won a championship
- Similarly, despite their immense population, neither India nor China have ever competed in a championship final
- Richard Sears has won 7 times in a row in the 1880’s, the most of any player
- Helen Wills won 3 consecutive championships twice in the 1920’s

It is also interesting to look at Australia’s performance at the US Open as the nation has performed exceptionally. For a relatively small nation, Australia has achieved 24 championships and 24 runner-up positions.

I will outline several of the key advantages of using Tableau:

- Tableau offers a vast amount of graphing techniques
- Makes it easy to label/annotate data points

- Excellent geographical mapping, taking geographical data from database and effectively creating a map from it
- Excellent way to display additional information by using the detail feature. This allows you to only display the most important information, with the ability to display additional relevant information by hovering over it