10/16/2022

Assessment Task 2

Advanced Data Visualisation

Yangyang Jin
13647716

Table of Contents

1.0 Executive Summary

The US Open (tennis), founded in 1881, is one of the oldest tennis tournaments in the world, originally known as the US National Championships (2022, Wikipedia). This report focuses on the dataset provided by the US Open Tennis Championships, with 276 men's and women's championship matches from 1881 to 2021. The purpose of this report is to summarize the performance of these players through a data visualization analysis of this dataset, as well as analysis of the trends found.

After using Tableau to count the number of championships, 12 top players with 5 or more championships are founded:

Six men： William Larned, Richard Sears, Bill Tilden, Roger Federer, Pete Sampras, Jimmy Connors.

Six woman： Helen Wills, Serena Williams, Chris Evert, Steffi Graf, Margaret Court, Molla Mallory

Through the visualization and in-depth analysis of the dataset, the following main conclusions are drawn:

- In those 276 races, the United States has won 177 titles, almost two-thirds of the total, followed by Australia with 24 titles.
- All games have a win rate ranging from 45.45% to 92.31%, with 9 of them having a win rate above 75%.
- Players with a high win rate do not necessarily have more championships. Among the players with a win rate of 75% or more, almost all have won the championship only once, and only two players have won the championship twice.
- Of the nine players with a 75% or higher win rate, eight are women and only one is a man.
- Across the 276 races, the women's race seemed to consistently outrank the men's race, but the men's race won more titles than the women's.
- Molla Mallory has the most U.S. Open singles victories, eight times, but has represented different countries.
- Richard Sears is the only player in U.S. Open history to win seven consecutive years and has never lost in the final.
- During the 140-year history of the U.S. Open, both men's and women's winning rates have declined to varying degrees
- Among the 12 top players, the most one has won 8 championships, four players have won 7 championships, two have won 6 championships, and the remaining 5 have won five championships each.
- There is a certain negative correlation between total score and win rate. Those with higher scores usually have lower winning rate.

- Starting in 1950, more and more players from other countries joined the ranks of champions. But even so, the United States still puts considerable pressure on other countries at the top of the rankings.
- The champions are mainly distributed in Europe, America and Oceania. It is worth thinking that there are no champions from Asia and Africa except Japan.

2.0 Data Understanding & Preparation

2.1 Data Understanding

This dataset is from the US Open dataset and contains 22 columns of attributes and 276 match records. It is important to have a clear understanding of the dataset before visualizing and analyzing the data. The following table gives a detailed description of the attributes and types of all the data in the dataset:

Attribute Name	**Type**	**Description**
Year	Interval - quantification that allows sorting and addition and subtraction	Year of the tournament
Gender	Nominal – category attribute	Divides tournament by Men's or Women's tournament
Champion	Nominal – category attribute	The name of tournament's winner
Champion Nationality	Nominal – 3-letter string format, category attribute	An abbreviation for the champion's countries
Champion Country	Nominal – category attribute	The full name of the champion country
Champion Seed	Ordinal – can be sorted	Champion Initial Ranking
Mins	Interval – can be added or subtracted	The length of the championship match in minutes
Score	Nominal – String format, separated by commas	The score results for each set of games in the match
1st-won	Interval – integer format	Winning score of the first game
1st-loss	Interval – integer format	Losing score of the first game
2nd-won	Interval – integer format	Winning score of the second game
2nd-loss	Interval – integer format	Losing score of the second game
3rd-won	Interval – integer format	Winning score of the third game
3rd-loss	Interval – integer format	Losing score of the third game
4th-won	Interval – integer format	Winning score of the fourth game
4th-loss	Interval – integer format	Losing score of the fourth game
5th-won	Interval – integer format	Winning score of the fifth game

5^{th}-loss	Interval – integer format	Losing score of the fifth game
Runner-up	Nominal - category attribute	The name of runner-up in the tournament
Runner-up Nationality	Nominal – 3-letter string format, category attribute	An abbreviation for the runner-up's countries
Runner-up Country	Nominal – category attribute	The full name of the runner-up's country
Runner-up Seed	Ordinal – can be sorted	The runner-up's initial ranking

2.2 Data Preparation

Data preparation is also an essential step before creating data visualizations. In this dataset, since champion seeds, mins and runner-up seeds were not recorded before 1945, these three columns were removed from the dataset. In the scoreboard field, since this Score is in a text format that cannot be compared, in order to extract valuable information from the Score column, 10 new columns in numeric format are converted to represent the win or loss of each group.

In addition, in order to better represent the performance of the players, I added the total points won and lost in a single game and their difference, and calculated the winning rate of each champion through the formula. I also created a Top players sheet, which contains the number of championships and runners-up won by the top players.

When drawing the parallel coordinate chart, I normalized the data of Wins, Lost and Win Rate due to the different dimensions of the data. The purpose of normalization is to place the range of data of different dimensions in a small specific interval. Using normalized data for data analysis can eliminate dimensions and increase the readability of the data.

Finally, here are some data to be aware of when doing visualization plots:

- Czechoslovakia had the lowest average win rate among women's champions and was peacefully dissolved on 31 December 1992 (Wikipedia, 2022), so it cannot be shown on the Geographic map. A similar situation was observed in Yugoslavia, which officially changed its name to the Kingdom of Yugoslavia on October 3, 1929 (Wikipedia, 2022).
- In the visualization of the top players, there are two players with the same name but from two countries. In fact, the two are the same player. According to (Wikipedia, 2022), Molla Mallory was a Norwegian tennis player who became a naturalized American and continued to play on behalf of the United States (Wikipedia, 2022)

- There are some empty values in the newly created 10 columns of numerical scores, but these are not missing values. That's because of the competition system, there is no need to play the following games, and the champion can already be determined.

3.0 Data Visualization

3.1 Parallel Coordinate

Parallel coordinate is a visualization method mainly used for analyzing high-dimensional data. It can compare multiple quantitative variables together, and look for patterns and relationships between them. In parallel coordinates, each vertical axis in the graph represents a feature, the point above represents the value of the feature, and each sample is represented as a line graph that runs through all vertical lines. In general, different colors are used to represent different categories, so that it is easy to see the impact of different features on the classification. However, high-dimensional data are often inconsistent in scale, and different measurement units lead to unfavorable observation of the relationship between different attributes. In this case, normalizing the data of different dimensions can solve this problem well. In addition, the order of the axes is critical for finding attribute features, as it affects the readability of the graph. Try a variety of reorders in the drawing diagram, which may also yield new potential information.

3.1.1 Observation

This parallel coordinate chart analyzes all wins, lost, and winning rate for the champions. As can be seen in the parallel coordinates plot below (Refer to Figure 1), the color of the lines is used to differentiate the different countries and is labelled with the champion's name, champion nationality, year, score, gender, and win rate. For example, it can be found that the champion with the most points is Pancho Gonzales from USA, who has 36 points in total, with a 53.73% win rate, and won that championship in 1949. Also from the United States, the player who lost the most points was John Doeg, who lost 32 points in that game in 1930. Nonetheless, John Doeg still defeated his opponent Frank Shield in that event and took the title. However, due to the different dimensions between winning rate and winning points, it is difficult for us to find the correlation between them. So I normalized these three data.

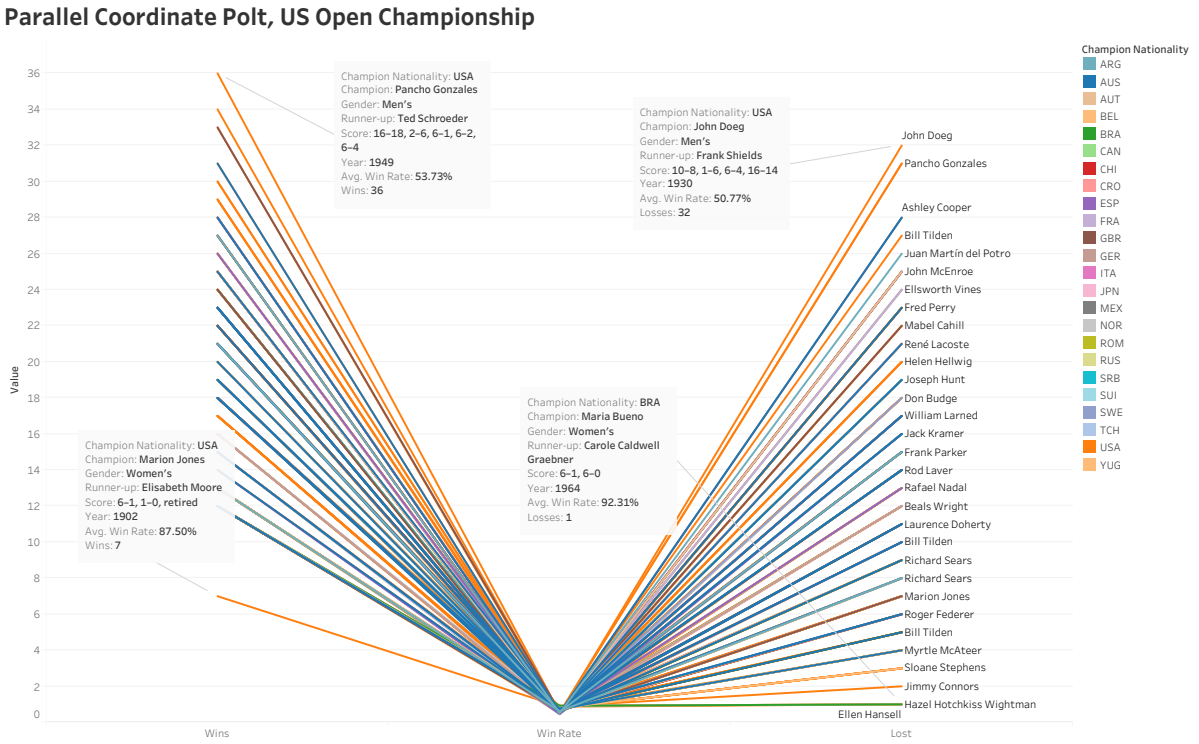

Figure 1 # Parallel Coordinate by Wins, Win Rate and Losses

The following figure is the normalization of these three data through scaling technology (Refer to Figure 3), so that they can be analyzed on the same scale. The parallel coordinate graph after data normalization can make it easier to find the connection between them. Firstly, it can be found that there is a lot of intersection between these three attributes. This precisely also shows that there is a certain **negative correlation** between score and win rate. For example, as can be seen from figure 2, if a player has a higher win point, then they will also lose a higher number of points, so that the player won't have a very high win rate. Conversely, if the player has a high win rate, he/she also won't need too many points to win the game. Frankly, champions with more points tend to have more tortuous games (usually narrow wins), while champions with fewer points may be smoother. Moreover, by looking at the labels, it can be found that several championship years with higher win rates are earlier, which seems like a question worth pondering.

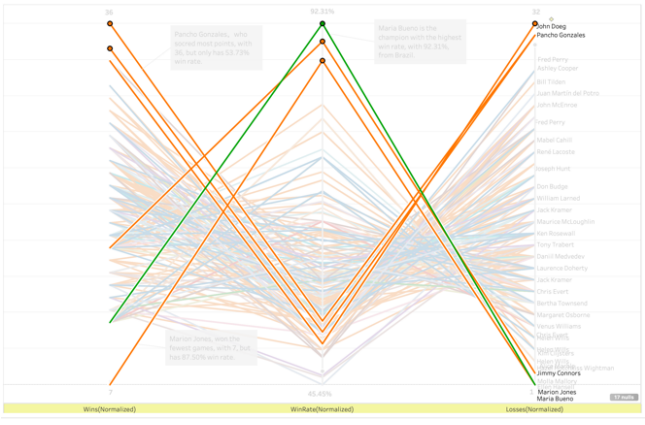

Figure 2 # X-shapes between Wins, Win Rate and Lost

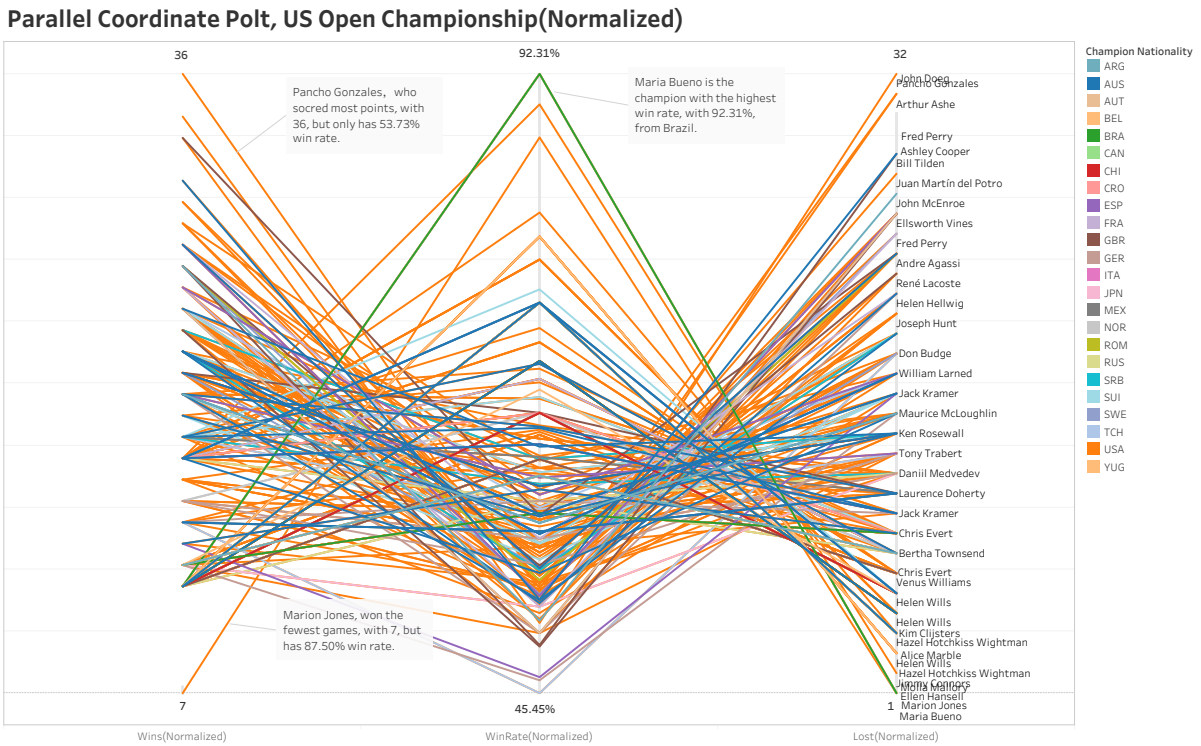

Figure 3 # Parallel Coordinate by Wins, Win Rate and Losses(Normalized)

3.1.2 The advantages and disadvantages

The main advantage of parallel coordinates is that it can represent high-dimensional data as a two-dimensional visualization. Parallel coordinates plots are great for comparing multiple variables and seeing how they relate to each other. It can easily find "drop points" on each axis to connect them to form a polyline. As data grows and lines are stacked, it becomes easier for analysts to spot trends and changes displayed across multiple variables, providing viewers with a more intuitive data visualization. However, the disadvantage of parallel coordinate plots is that due to overlapping data lines for common data values between data entries, they can become too cluttered and thus difficult to discern when the data is very dense. The best way to deal with this is through interactivity and a technique called "brushing". Brush highlights a selected line or collection of lines while fading out all other lines. This can highlight parts of the plot of interest while filtering out noise.

3.2 Treemaps

Treemaps are ideal for displaying data that requires a large amount of hierarchical structure. It's a rectangle-based visualization. In the tree diagram, the space in the visualization is divided into rectangles of different sizes, and each rectangle can also be subdivided into several small rectangles, which represent its subcategories. The size of the rectangles in the treemap ranges from the upper left corner to the lower right corner of the visualization, with the largest rectangle in the upper left corner and the smallest rectangle in the lower right corner. Also, the color of the dendrogram plays an important role. It can represent the use of

different colors to distinguish categories, or the intensity of a unit of measure can be represented by shades of color.

3.2.1 Observation

In this case, two treemaps are created. The first treemap shows the respective number of champions for each country's men's and women's competitions and their respective average win rates. As shown in the treemap below (Refer to Figure 4), in the first layer of rectangles, each large rectangle represents a country, and then the rectangles represent the men's and women's games of that country respectively. The colors of the rectangles show the different winning percentages of the men's and women's competitions in each country. Firstly, it can be found that in the US Open, the United States has the largest number of champions in both the men's and women's competitions. Among them, the women's competition has won 92 championships with a winning rate of 64.86%, and the men's competition has won 85 championships with 61.71% win rate. It was followed by Australia with 18 men's titles and six women's titles. By looking at the depth of the matrix, we can see that some of the highest win rates are from women's races. These include Switzerland with an average win rate of 75%, Brazil with 74.25% and Norway with 72.20%. However, it is also interesting to note that for most countries the women's matches are darker than the men's matches, which indicates that the women's average win rate is higher than the men's average win rate.

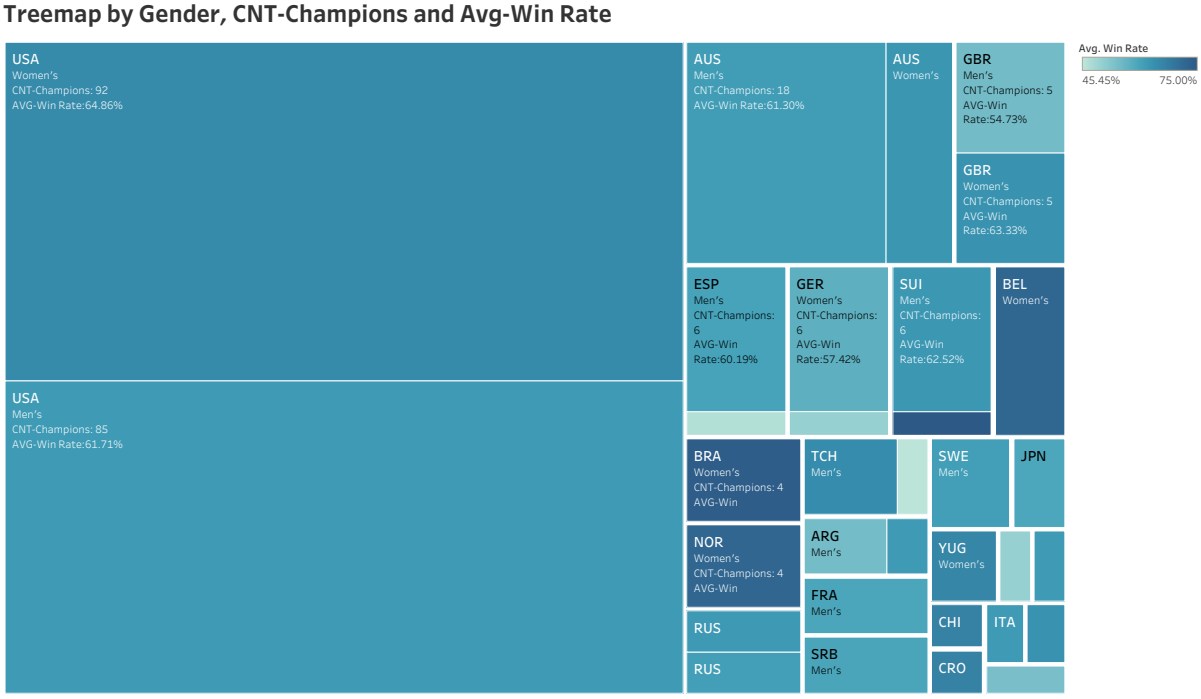

Figure 4 # Treemap by Gender, CNT-Champions and Avg-WinRate

In the second treemap (Refer to Figure 5), I added the champion label to the previous one. The previous matrix is further divided into smaller rectangles, and the size of the rectangle represents the number of champions each champion has won. Unlike the first treemap, colors

are no longer used to represent units of measure, but categories. Different colors represent different countries, so that the audience can have a clearer view of the distinction between countries. In this treemap, you can easily find the top players in each country. For example, Helen Wills, Bill Tilden, Richard Sears and William Larned from the United States have each won 7 championships, of which Helen Wills has the highest average win rate with 72.58%. Margaret Court from Australia, Steffi Graf from Germany and Roger from Switzerland Federer all had a similar situation, winning five titles each to become the top players in their respective countries. It is worth noting that some countries such as Australia, Canada, Chile, Croatia, Italy, Mexico and Romania have only one and only one championship.

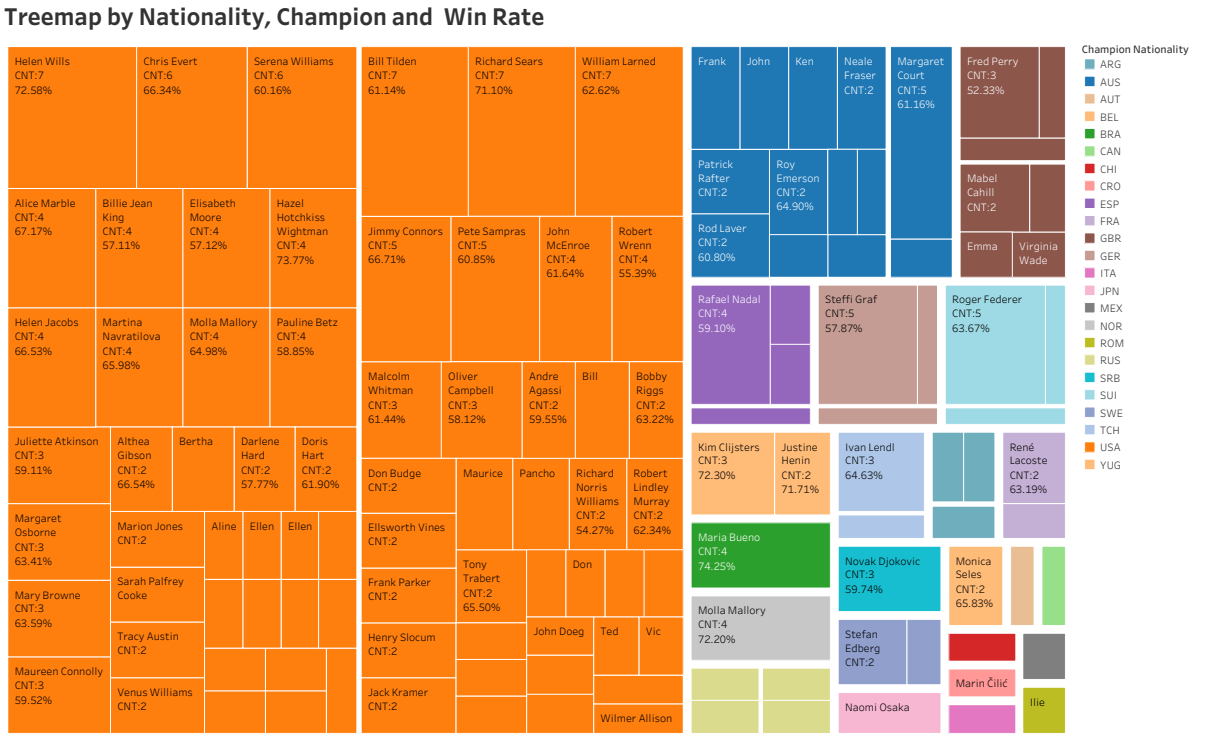

Figure 5 # Treemap by Nationality, Champion and Personal Win Rate

3.2.2 The advantages and disadvantages

The advantage of a treemap is that it provides a hierarchical view of the data, allowing viewers to easily view and compare each section on different hierarchies. For example, at the national level, viewers can easily find that the United States is the country with the most championships. The different winning rates of male and female players can also be compared by the depth of color and gender distinction. Compared to the parallel coordinate map, the treemap can visualize all the champions by different categories in an attempt without being too cluttered or difficult to read. However, treemap also has its limitations. Although treemap can represent classification by matrix size and color depth, when some values are too small, the size of the matrix will become relatively small. This may directly reduce the readability of the visualization for automatically hiding part of the label's information. However, in Tableau

we can hover over the rectangle to see the complete information, including country, gender, win rate and number of champions.

3.3 Geographic Map

Geographic maps are mainly used to analyze and display data related to geography and present them in the form of maps. Using the form of a map to reflect the data related to the geographic location can display the data more intuitively and vividly. However, geographic maps can only be used when geographic data is involved, and the data must be geographic dimension data, with only one dimension and measure. With geographic maps, analysts can visually see the distribution or proportion of data in each area. There are many types of geographic map visualizations, such as area maps (also known as filled maps), heat maps, statistical maps, trajectory maps, and interactive maps.

3.3.1 Observation

The first geographic map was a dual axis chart with a pie chart and a filled map overlaid into one piece via tableau. As shown in the geographic map below (Refer to Figure 6), I used Champion Nationality as a color fill to all the countries that have won the championship. The size of the pie chart represents the number of championships the country has won. Through the color ratio of the pie chart, we can clearly see the proportion of male and female champions in the country. As can be seen from this geographic map, there is no doubt that the United States is the strongest country in both the men's and women's competitions,

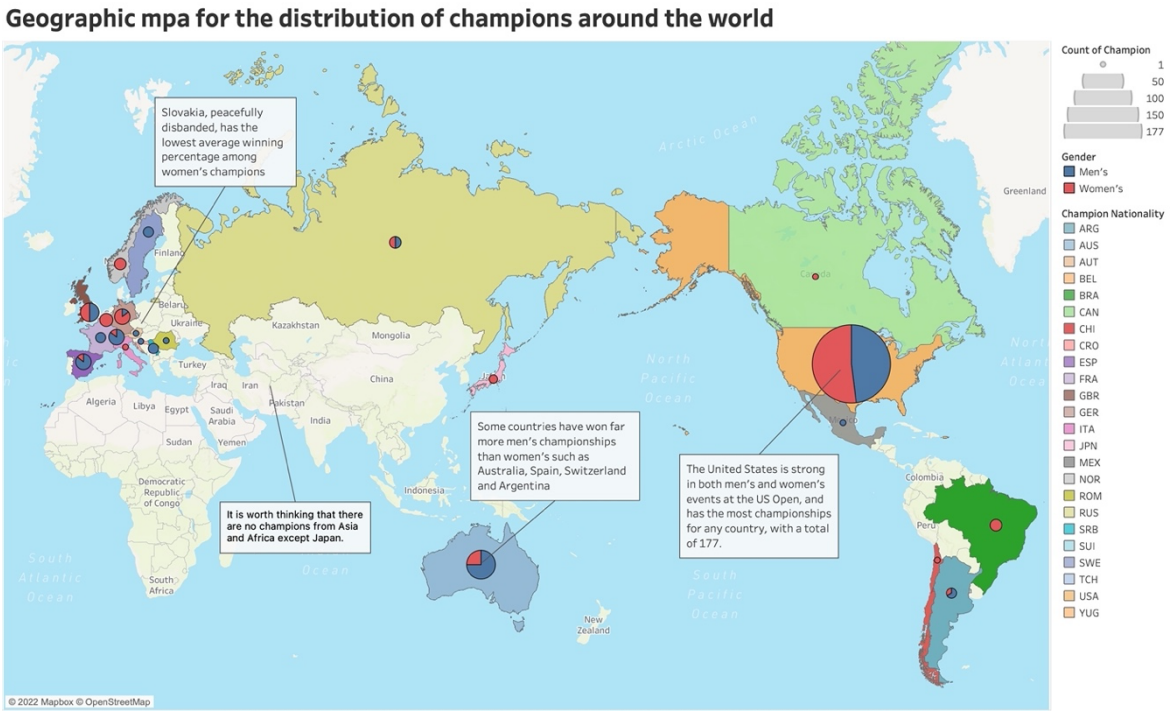

Figure 6 # Geographic Map for the distribution of champions around the world

winning a total of 177 championships. However, Canada and Mexico, both in North America, have only won the championship once. Interestingly, some countries have won far more men's titles than women's such as Australia, Spain, Switzerland and Argentina. In contrast, the number of champions in Germany is dominated by women's. It is worth noting that the Asian and African countries have never won the championship except Japan. According to (Marvitr, 2021), the author mentions that this seems to be a historical problem. The main reason is that tennis originated in Europe and developed in the United States. Although Asia has also developed, the construction of tennis courts has also been lagging behind compared with Western countries.

As mentioned earlier, tennis originated in Europe, but Europe's performance in the US Open is not very prominent. In any case, in order to have a clearer comparison of the differences between the European countries and the Open Tennis Championships, I filtered out the two major championship-producing countries - the United States and Australia. As can be seen in this geographic map (Refer to Figure 7), I used the number of champions as the fill color between countries. The shades of color on the pie chart for each country represent the different winning percentages of the men's and women's games. By hovering over the pie chart in Tableau, we can see better and more detailed information, such as country name, Gender, and the average win rate. In this geographic map, we can find that after the United States and Australia, the United Kingdom is the country that has won the most championships, with a total of ten, five men and five women. In addition, some European countries, such as Britain, Switzerland and Spain, have a large gap in the winning rate of the men's and women's

Geographic mpa for the distribution of champions in Europe

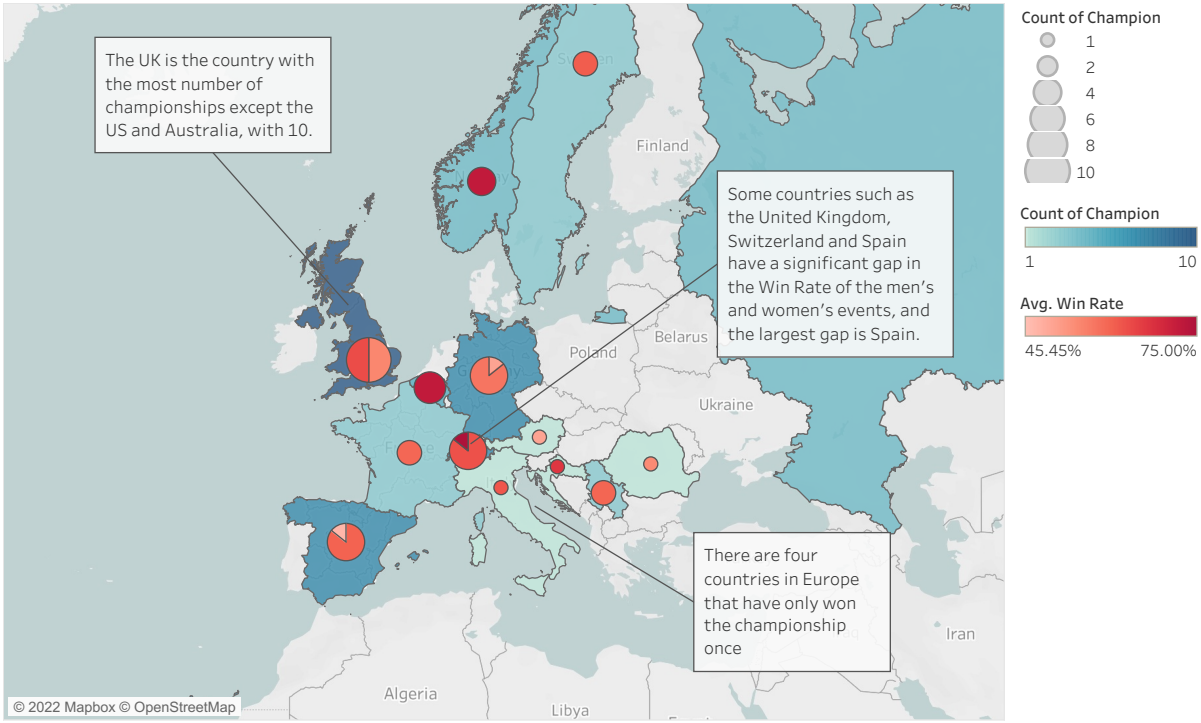

Figure 7 # Geographic Map for the distribution of champions in Europe

champions. The biggest gap is Spain, where the average winning rate for men's champions is 60.15%, compared with 46.67% for women.

3.3.2 The advantages and disadvantages

Compared with the tree map, the geographic map can more intuitively view the distribution of data in the world. It's also great for comparing distributions across different countries. For example, through the geographic map we can find that the champions are mainly distributed in the Americas and Europe, and Asia and Africa have hardly produced the champion of the US Open. In addition, the geographic map can also be displayed in combination with some pie charts or three-dimensional column charts. This can effectively convey more information to the viewer, and can also make the visualization more beautiful. However, the disadvantage of geographic map is that it can only be applied to data with geographic latitude, otherwise the information in the geographic map will not be displayed.

3.4 Top player performance

Through the statistics of the top champions, it is found that there are a total of 12 champions who have won five or more U.S. Open championships. They are: Roger Federer, Bill Tilden, Jimmy Connors, Pete Sampras, Richard Sears and William Larned, Margaret Court, Steffi Graf, Chris Evert, Helen Wills, Serena Williams and Molla Mallory.

As shown in the graph below (Refer to Figure 8), this graph shows the top players who have won five or more, and the runner-up they played against. The size of the circles represents the winning rate of the players, and the colors represent different countries. When the mouse

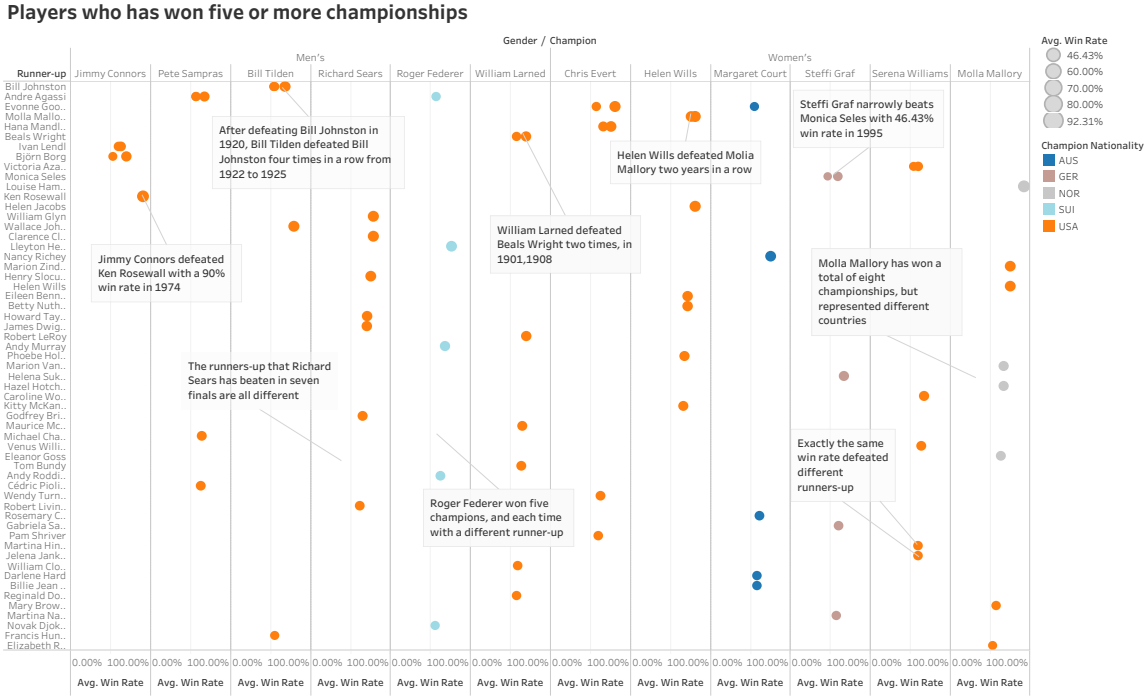

Figure 8 # Scatter for Players who has won five or more championships

stays over the circle, more detailed information is displayed, such as: gender, runner-up name, score, year of the game, etc. In addition, a lot of potential information besides the basic information of the players is also found in this figure. Firstly, 8 of the 12 top players are from the United States, which again proves the strength of American players at the US Open. It is worth noting that the graph shows Molla Mallory represented two different countries (Norway and the United States) during her championship winning career. As mentioned earlier, Molla Mallory was born in Norway, but became a naturalized American. She won a total of eight U.S. Open championships, and she belongs to the legend of that era. Secondly, these top players often encounter the same runner-up in the finals. For example, Bill Tilden beat Bill Johnston four times in a row in 1922-1924, and before that, Bill Tilden had already beaten Bill Johnston once in 1920. Helen Wills beat Molla Mallory in the final two years in a row with a high win rate. In addition, Jimmy Connors crushed runner-up Ken Rosewall with a high win rate of 90% in 1974. By contrast, Stef in 1995 had a very tough win, narrowly beating Monica Seles with a 46.43% win rate.

In addition, the following bar chart ranks the number of championships won by the top players (Refer to Figure 9), and distinguishes the gender of the top players by color. As can be clearly seen from the picture below, Molla Mallory is the player who has won the most championships in the history of the US Open, winning a total of 8 championships. Next are William Larned, Richard Sears, Helen Wills and Bill Tilden with seven titles each. Serena Williams and Chris Evert have six titles each and the remaining five have five titles each. Additionally, six of the 12 top players are male and six are female, with the highest average win rate being Helen Wills, with 72.58%.

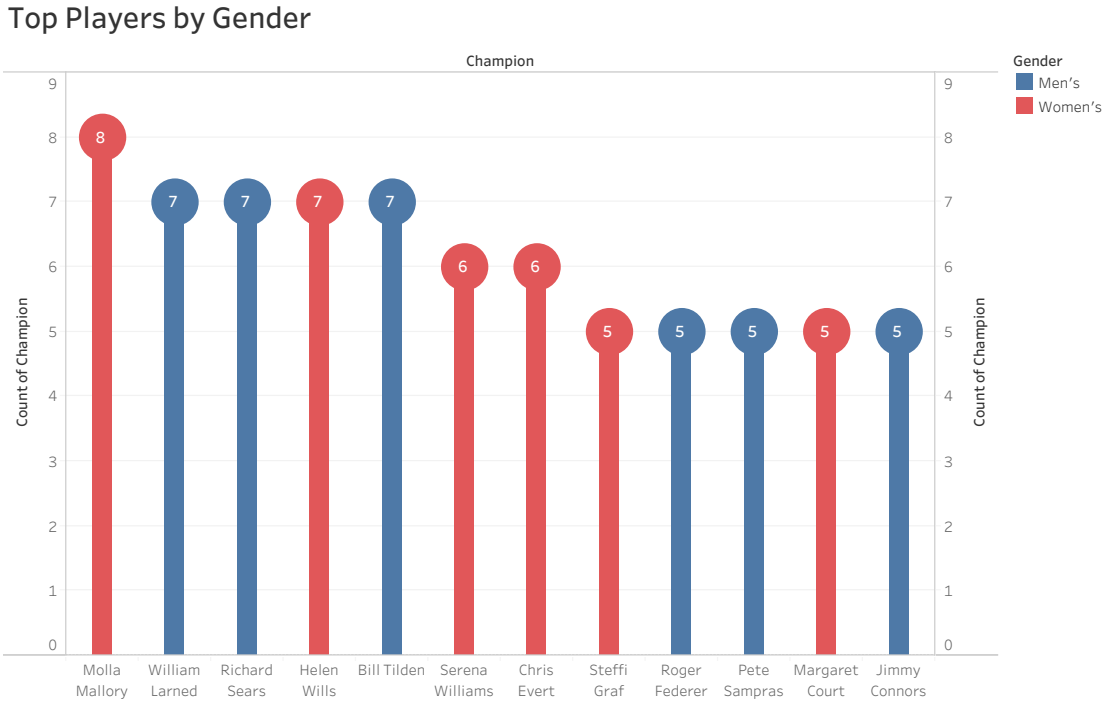

Figure 9 # Top Players by Gender

The third visualization created is a bidirectional bar chart. As shown in the figure below (Refer to Figure 10), this two-way bar chart shows the number of champions for the top players meanwhile showing the number of runners-up. The red part on the left represents the number of times the top player has won the championship at the US Open, while the green part on the right represents the number of times the player has won the runner-up at the same time. With this visualization, it's easier for viewers to compare them. Firstly, 10 of the 12 top champions have also won runners-up. Among them, Helen Wills, who has the smallest number of runners-up, has only won a runner-up once, and Serena Williams has the most, who has won 4 times. Additionally, both Richard Sears and Margaret Court have never been runners-up at the U.S. Open, which means they never lost in the final.

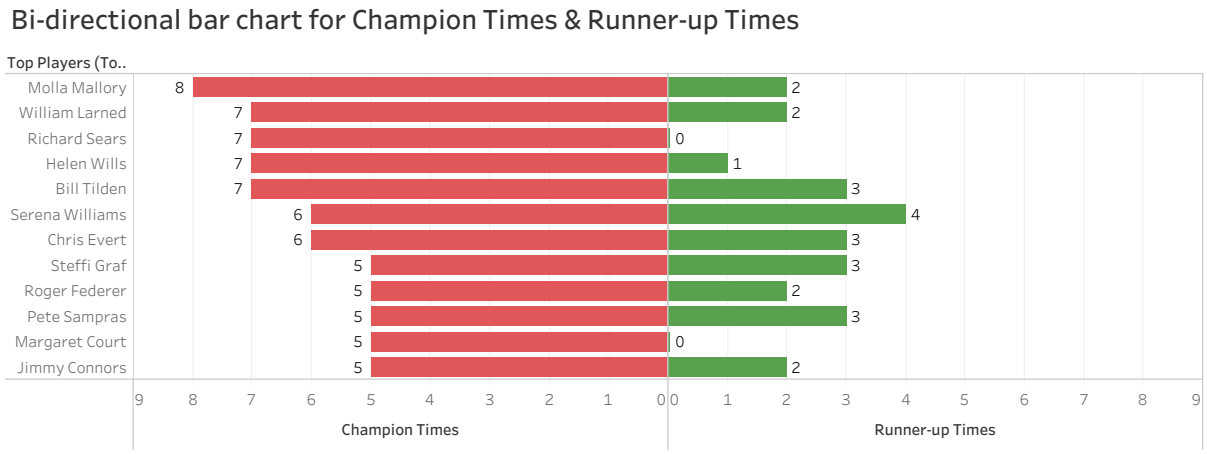

Figure 10 # Bidirectional bar chart for Champion times & Runner-up times

In addition, by using a scatter plot I visualized the change in winning rate between the top players over time. I also analyzed the changes for all champions so as to better compare the changes between the top players and all champions (Figure 11&12). In the following two time series analysis graphs, firstly, we can find that during the whole period of the US Open, the winning rate of both men's and women's matches has decreased to different degrees.

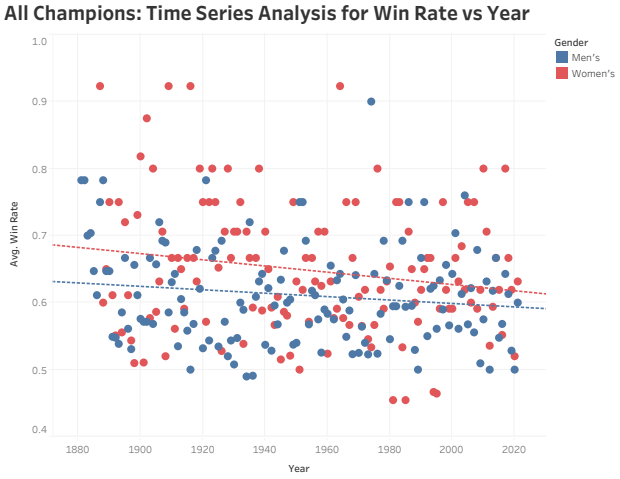

Figure 11 # All Champions for Win Rate vs Year

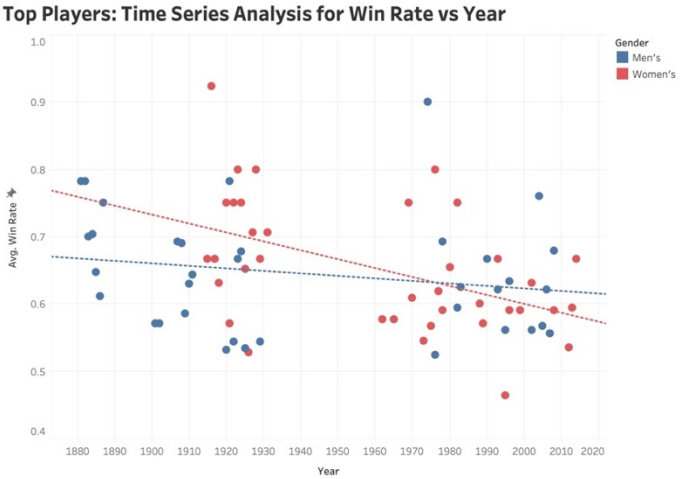

Figure 12 # Top Players for Win Rate vs Year

Although the women's win rate seems to be consistently higher than the men's, the decline seems to be greater for the women than for the men. In particular, among the top players, the women's win rate has been steadily lower than that of the men's since 1980. It's worth noting that two people have made it to the top tier with a winning rate of more than 90%: Molla Mallory and Jimmy Connors.

However, for this case, I also performed a time series analysis of the total score among the top players. It can be found that among these top players, there is a more obvious increase in the total score of the women's competition (Refer to Figure 13). This seems to indicate that the chances of winning the women's competition in the final have also become more and more difficult. As shown in Figure 14, starting in 1950, more and more players from other countries also joined the ranks of champions. As the game of tennis has become more technically sophisticated, the competition between the champion and the runner-up has become more intense. Especially for the women's game, the competition pressure seems to be greater than the men's game.

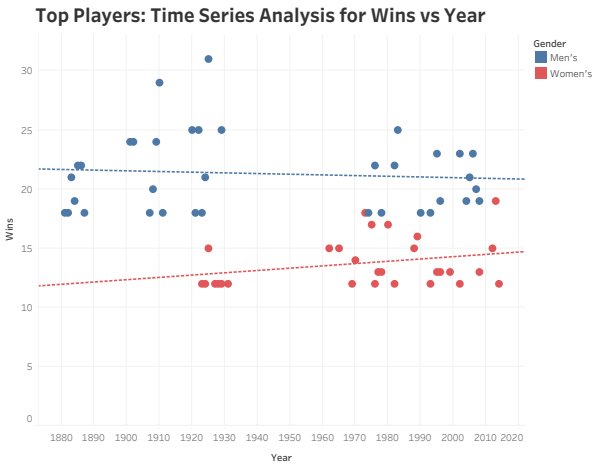

Figure 13 # Top Players for Wins vs Year

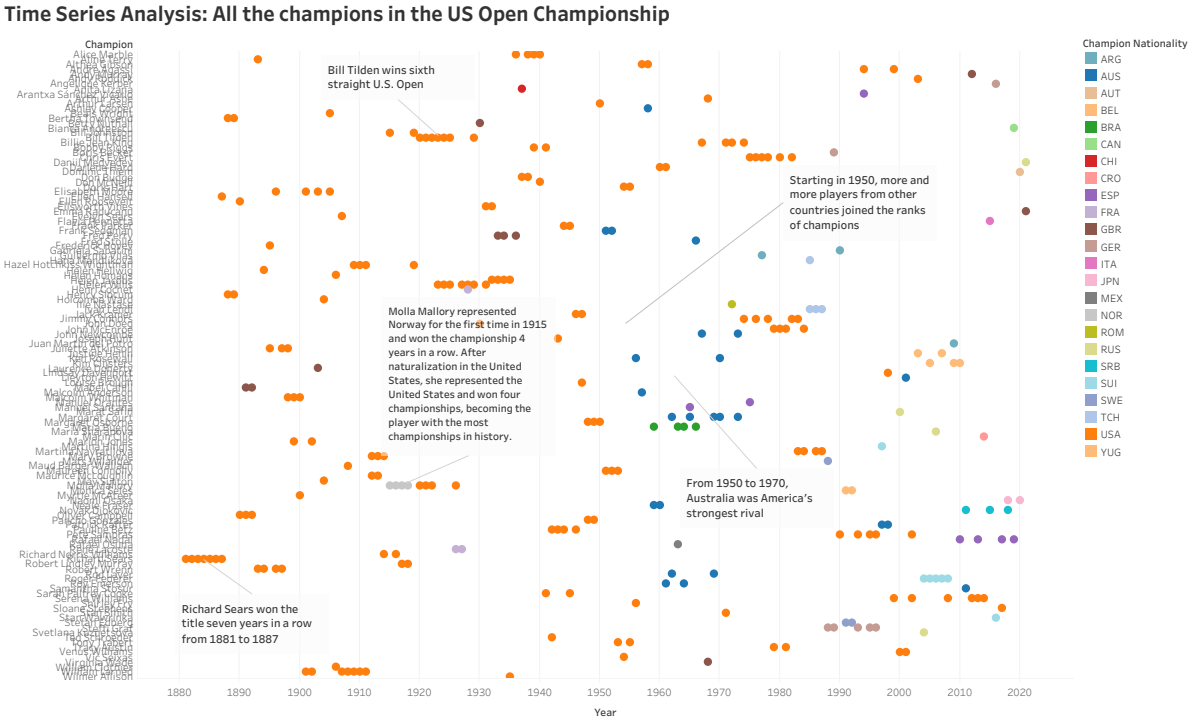

Figure 14 # Time Series Analysis: All the champions in the US Open Championship

4.0 Summary and Conclusion

4.1 Summary

Through the visualization and in-depth analysis of the US Open data, we found a lot of value behind the players, as well as potential information and trends, which are summarized as follows:

- The United States is without a doubt the most successful country in the entire U.S. Open period with 177 titles, followed by Australia with 24.
- All games have a win rate ranging from 45.45% to 92.31%, with 9 of them having a win rate above 75%.
- Players with a high win rate do not necessarily have more championships. Among the players with a win rate of 75% or more, almost all have won the championship only once, and only two players have won the championship twice.
- Of the nine players with a 75% or higher win rate, eight are women and only one is a man.
- Women consistently seem to have a higher win rate than men in those 276 matches, but men have won more titles than women.
- Molla Mallory represented Norway for the first time in 1915 and won the championship 4 years in a row. After naturalizing in the United States, she represented the United States and won four championships again, becoming the player with the most championships in history.
- Richard Sears was a champion for seven consecutive years from 1881 to 1887, and never lost a final, belonging to an era of legend.
- In these 140 years, the winning rates of men's and women's competitions all have declined to varying degrees.
- Among the 12 top players, the most one has won 8 championships, four players have won 7 championships, two have won 6 championships, and the remaining 5 have won five championships each.
- Ten of the 12 top players have also been runners-up.
- Of the 12 top players, the women's competition appears to be under more pressure in the final than the men's competition.
- There is a certain negative correlation between total score and win rate. Those with higher scores usually have lower winning rate.
- Since 1950, more and more players from other countries have joined the championship queue. But despite this, the United States still puts considerable pressure on other countries to be ranked among the top players.
- Between 1950 and 1970, Australia was America's most formidable adversary.
- Champions are mainly distributed in Europe, America and Oceania. It is worth thinking that there are no champions from Asia and Africa except Japan.

4.2 Visualization techniques

In data visualization, how to deal with high-dimensional data is very important. In this project, I used normalization to preprocess the data. As mentioned earlier in the parallel coordinates plot, before the data is normalized, the different data dimensions make it difficult to observe the relationship between different attributes. After data normalization, the data of different dimensions will be uniformly scaled to the 0-1 interval, which not only eliminates the problem of different dimensions between the data, but also increases the readability of the data. In addition, I also added labels and annotations to some visual charts (such as coordinate charts, geographic charts, and scatter charts), which allows the viewer to see the key points or trends of the visual charts very intuitively and clearly.

In the visualization of the top players, I used different types of charts to visualize them (such as scatter, bar, and bidirectional bar charts, etc.). Use different visualizations for different content to discover more information or breakthrough analysis. For example: using a scatter plot can have a clear display of the performance of the top players by corresponding to different x-axis and y-axis. A bidirectional bar chart is an easy way to show the number of simultaneous champions and runners-up among the top players. And through time series analysis, we can have an intuitive understanding of all champions during the US Open.

Compared to excel, I think Tableau is better for data visualization. Tableau has put a lot of academic energy into researching what kind of charts people like and how to bring the ultimate experience to users both operationally and visually. Its design, color, and operation interface give people a simple and fresh feeling. A beautiful chart Tableau can generate with one click. Tableau can manage large amounts of data. Tableau's promotion is that it can manage hundreds of millions of data, which is difficult to do in Excel. However, excel also has its advantages over Tableau. Excel installed capacity is large. If you use Excel to make an analysis report, you can easily send this report to the target group as long as they have Excel installed and they can open it. If it is Tableau's .twbx file, you need to install Tableau correctly to open it. And there is basically no cost to use Excel, but genuine Tableau is expensive.