Executive Summary:

In order to analyse a high dimensional dataset from US Tennis Open match records, this paper employs advance data visualisation techniques. The dataset has 22 characteristics and 276 example records. Using the analytical visualisation application Tableau, visualisation techniques like TreeMap, Parallel Coordinate, Geographic Mapping, and others are applied to the dataset. The report provides information on the data properties, data pre-processing procedures, and analysis of the various approaches used. There is also a conclusion that summarises the main findings.

Because it divides the data into proportions for each category or subcategory using different colours and sizes, the TreeMap approach is useful for visualising high-dimensional data. This may immediately provide a summary of the percentage or size of each category in the dataset, which is useful for fast analyses. The difficulty of precisely comparing the smaller squares in the TreeMap is one of its drawbacks, though. Despite having what seem to be comparable sizes, they have different values. As a result, if one wish to appropriately represent these data, different method is required.

The Parallel Coordinate chart is another method that is employed. The ability to represent numerous variables concurrently in a single graph makes this method beneficial when working with multivariable data. It helps to illustrate the link between the variables, and the use of colour makes this relationship stand out sharply. However, it has a flaw when a lot of data needs to be displayed on a single graph. Too much clustering might make it difficult to understand or spot any trends. To study the intricate interactions between each property, a different visualisation approach needs to be utilised.

Finally, geographic mapping is used to quickly and easily visualise data, particularly geographical data. This method makes it simple to find and analyse trends between various nations on the map by allowing us to access information on a global scale solely through a map. Similar to the two before graphs, this approach still does not provide a detailed visualisation of the attribute information.

Sample Dataset:

The US Tennis Open event, one of the most esteemed tennis competitions in the world, and the fourth and final Grand Slam competition of the year, served as the dataset for Assignment 2. It has 141 years' worth of men's and women's championship games from 1881 to 2021.

276 sample records and 22 attributes make up the dataset's original structure. Even though there are 22 attributes, **Champion Seed**, **Mins** and **Runner-up Seed** are not provided for this assignment. The available characteristics are:

- **Year**: This property provides the year as the timeframe for the data that has been recorded, and for each year, it comprises two types of data about the Men's and Women's Champions.
- **Gender**: Indicate the player's gender (Men or Women).
- **Champion**: The name of the champion from each year is listed in this attribute.
- **Champion Nationality**: This characteristic holds the winning nation's three-letter name, for example: (USA, GER, ESP, AUS and so on).
- **Champion Country**: This characteristic contains the name of the winning nation, such as the United States, Australia, Germany, etc.
- **Score**: The score of the set that was played in each final match is a characteristic that provides the score. For instance, the score of 6-3 indicates that the winner of the game won since the winner set score is 3 points more than loser set score.
- **1st to 5th sets won and loss**: The score for each set of 1 to 5 sets. And whoever won the most sets win the match.
- **Runner-Up**: The name of the player who finished second is shown under Runner-Up.
- **Runner-Up Nationality**: This characteristic, which is identical to Champion Nationality, holds the runner-nationality ups in three-letter format.
- **Runner-Up Nation**: The name of the runner-country up's is kept in this attribute.

Data Preparation:

Data exploration is an important step before data preparation and transformation. It makes it easier to navigate the data and spot patterns and trends while facilitating a better knowledge and interpretation of the dataset.

Data Format:

Categorical data (String data)

Categorical data, such as a country's name or a person's name, may be distinguished from one another. Gender, Champion Name, Nationality, and Runner-up Name are among the features in the dataset that solely include categorical data.

Interval data

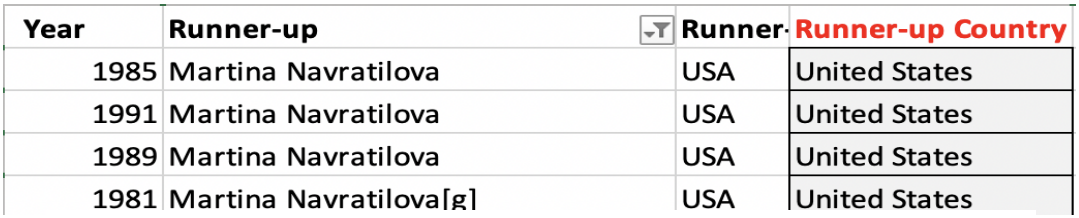
In addition to representing data with significant disparities between each value, interval data also has a relevant rating for the data it represents. The Year attribute is the only one with this property.

Text Data

Because the attribute Score contains numerical scores for each participant, the data type should be Interval or Ratio. However, it is in Text format since one individual gets numerous scores from each round, which are separated by delimiters such as (,) and (-). This property requires data cleansing and modification before it can be calculated and analysed further.

Data Cleaning:

Wrong Spelling Name

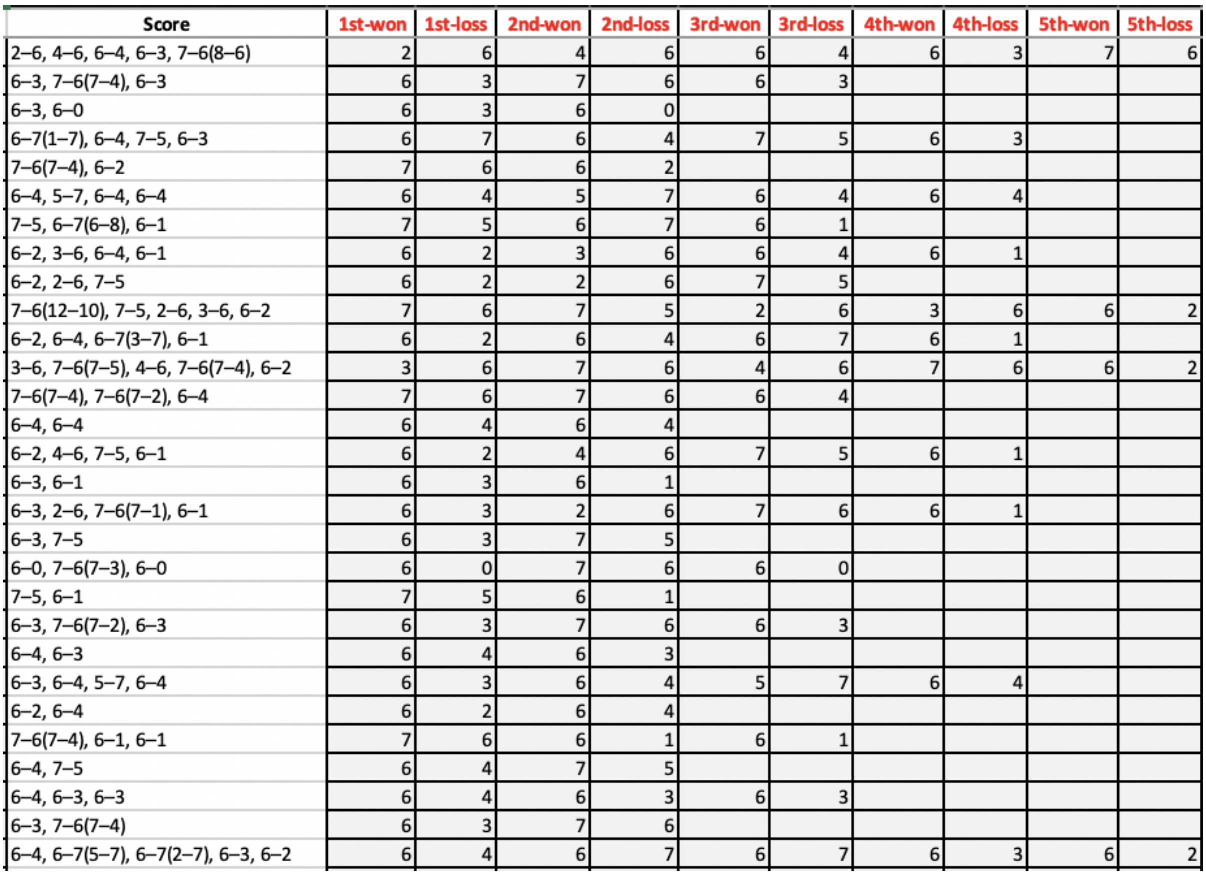


Year	Runner-up	Runner-	Runner-up Country
1985	Martina Navratilova	USA	United States
1991	Martina Navratilova	USA	United States
1989	Martina Navratilova	USA	United States
1981	Martina Navratilova[g]	USA	United States

Figure 1

From Figure 1, we can see that there are four names such as Martina Navratilova and Martina Navratilova[g]. We can clearly notice there seem to be error in fourth name. Her full name is Martina Navratilova, so all three names will be changed to Martina Navratilova by using Replace Function in Excel.

Anomaly Score



Score	1st-won	1st-loss	2nd-won	2nd-loss	3rd-won	3rd-loss	4th-won	4th-loss	5th-won	5th-loss
2–6, 4–6, 6–4, 6–3, 7–6(8–6)	2	6	4	6	6	4	6	3	7	6
6–3, 7–6(7–4), 6–3	6	3	7	6	6	3				
6–3, 6–0	6	3	6	0						
6–7(1–7), 6–4, 7–5, 6–3	6	7	6	4	7	5	6	3		
7–6(7–4), 6–2	7	6	6	2						
6–4, 5–7, 6–4, 6–4	6	4	5	7	6	4	6	4		
7–5, 6–7(6–8), 6–1	7	5	6	7	6	1				
6–2, 3–6, 6–4, 6–1	6	2	3	6	6	4	6	1		
6–2, 2–6, 7–5	6	2	2	6	7	5				
7–6(12–10), 7–5, 2–6, 3–6, 6–2	7	6	7	5	2	6	3	6	6	2
6–2, 6–4, 6–7(3–7), 6–1	6	2	6	4	6	7	6	1		
3–6, 7–6(7–5), 4–6, 7–6(7–4), 6–2	3	6	7	6	4	6	7	6	6	2
7–6(7–4), 7–6(7–2), 6–4	7	6	7	6	6	4				
6–4, 6–4	6	4	6	4						
6–2, 4–6, 7–5, 6–1	6	2	4	6	7	5	6	1		
6–3, 6–1	6	3	6	1						
6–3, 2–6, 7–6(7–1), 6–1	6	3	2	6	7	6	6	1		
6–3, 7–5	6	3	7	5						
6–0, 7–6(7–3), 6–0	6	0	7	6	6	0				
7–5, 6–1	7	5	6	1						
6–3, 7–6(7–2), 6–3	6	3	7	6	6	3				
6–4, 6–3	6	4	6	3						
6–3, 6–4, 5–7, 6–4	6	3	6	4	5	7	6	4		
6–2, 6–4	6	2	6	4						
7–6(7–4), 6–1, 6–1	7	6	6	1	6	1				
6–4, 7–5	6	4	7	5						
6–4, 6–3, 6–3	6	4	6	3	6	3				
6–3, 7–6(7–4)	6	3	7	6						
6–4, 6–7(5–7), 6–7(2–7), 6–3, 6–2	6	4	6	7	6	7	6	3	6	2

Figure 2

From Figure 2, there are different formats in score showing in brackets. The scores shown in brackets are not included in 1st to 5th sets. Therefore, the data in score attributes is cleaned and reformatted according to 1st to 5th sets. Investigation reveals that Helen Jacob retired in the beginning of the fourth women's match in 1933, and for improved data quality, the retire text will be removed from the score attribute.

Visualisation Method 1. Parallel Coordinate:

Plotting the pattern of a dataset with several attributes is done using the parallel coordinate data visualisation approach. Only two types of attributes may be plotted using a scatter plot. The Parallel coordinate, on the other hand, can plot as many qualities as are necessary for the discovery. As a result, rather than training with every characteristic, data scientists frequently utilise it in the actual world to identify the attributes that are useful for doing categorisation or advanced analytics.

We make a Parallel Coordinate chart using Tableau that includes axes for score, win rate, first to fifth won and loss. As seen in Figure 2, each line represents a champion and is coloured to indicate their nationality . In addition, information on the player's name, gender, score, and participation year is provided.

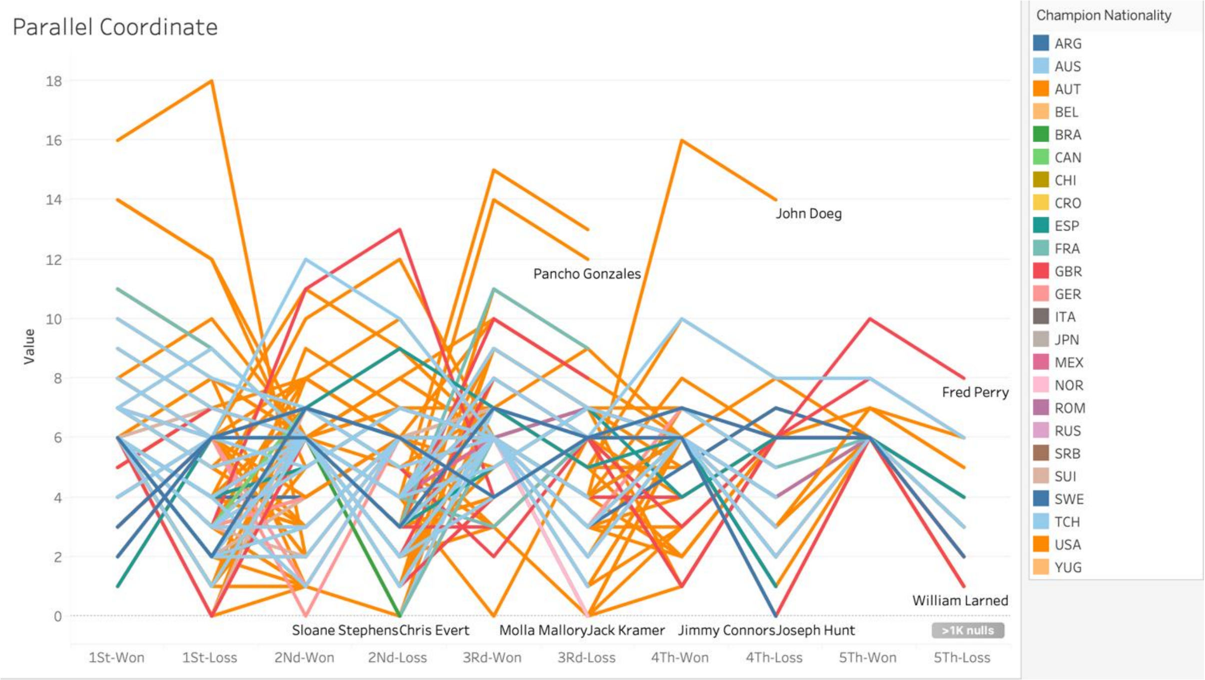


Figure 3 Parallel Coordinate

With an average win percentage of 60% in sets 4 and 5, the win rate decreases as the match goes on. Only champions from major countries like the UK, USA, Australia, and Spain notably advanced to the fifth set. Other country champions finish the match in the second or third sets. This may also contribute to the regional strategy in approach. Also except from three female champions from 1890 to 1899, only male champions got to 5th sets.

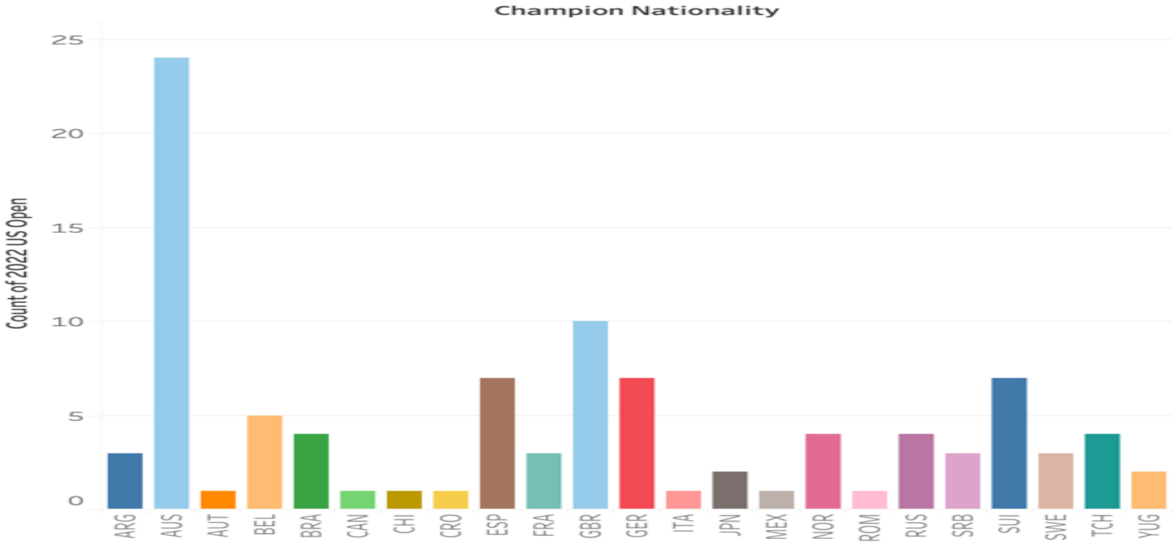


Figure 4 Top Three Countries except from USA

In order to establish a clear relationship between the number of sets played and the country's performance, the top 3 countries with the highest number of champions are chosen based on this data, with the exception of the USA because it is a US Open match and the USA has a significant difference in champions. The graph shows that these top countries have matches that last till 4th or 5th sets than other nations, suggesting that these players may frequently have moderate to long matches. Players whose bouts go longer than the average amount of time may indicate that their opponents are fiercely competitive.

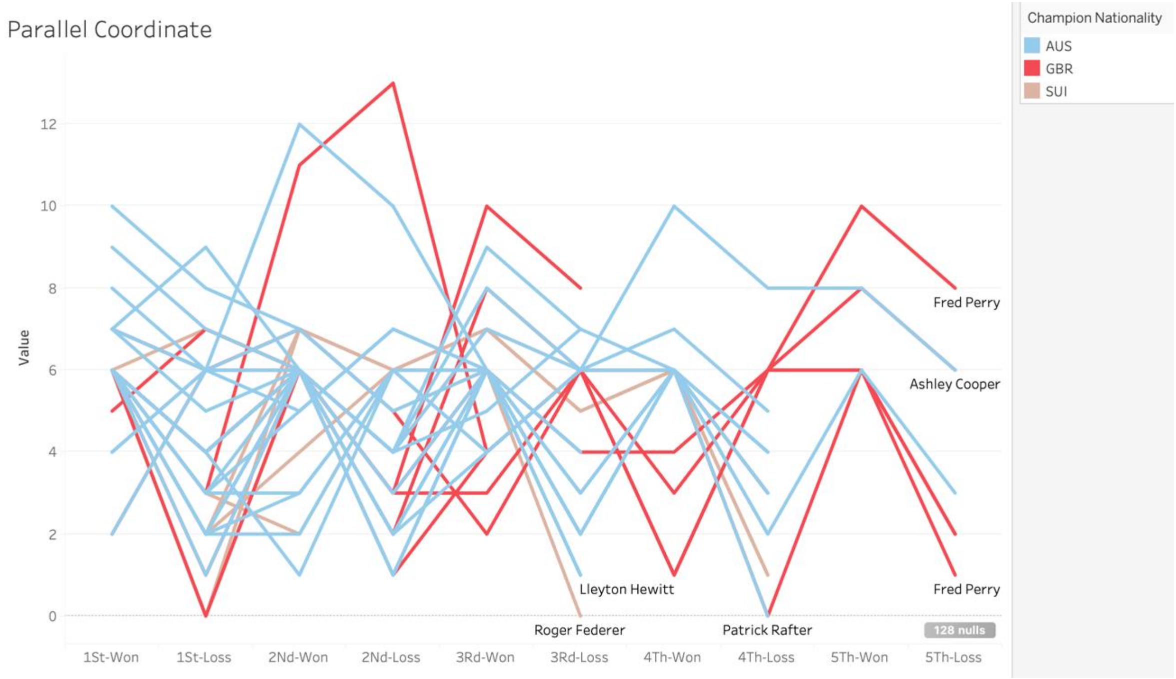


Figure 5

The Parallel Coordinate method is advantageous since it enables us to visualise several variables in a single graph. The correlations between the nationalities of the champions and match time are made more obvious by the graph's vibrant lines. However, because so many attributes and values are displayed simultaneously using this approach, the chart is complex and difficult to analyse. Figure 3 demonstrates this by showing how all the lines are grouped together. Despite the use of colours, it is obvious that it is difficult to tell one line from another. It is necessary to exclude certain data by simply choosing the nations of interest, as shown in Figure 5, in order to find intriguing trends.

Visualisation Method 2. Treemaps:

A data visualisation approach called TreeMap separates the data into layered rectangles. Rectangle sizes correspond to data values, while colours indicate various categories. It is done by comparing the amount of the sum data for each type of category using the total of all the values for that attribute. When dealing with hierarchical data, especially when there are several components in the dataset, the TreeMap approach is used to depict the compositions of the data.

The champion totals for each nation are shown in the treemap above. The tree map compares the number of champions each player has won inside each nation. This treemap reveals that the United States won nearly half of the championships. The champion of 23 nations' combination makes up the other half. Australia won around 20% of the medals out of the 23 nations. However, if we focus on each nation individually, we can observe the boundary between the men's and women's competitions. In men's competitions, Australia has more success than in women's tournaments. While in the USA and UK, men and women on the size side were equally split.

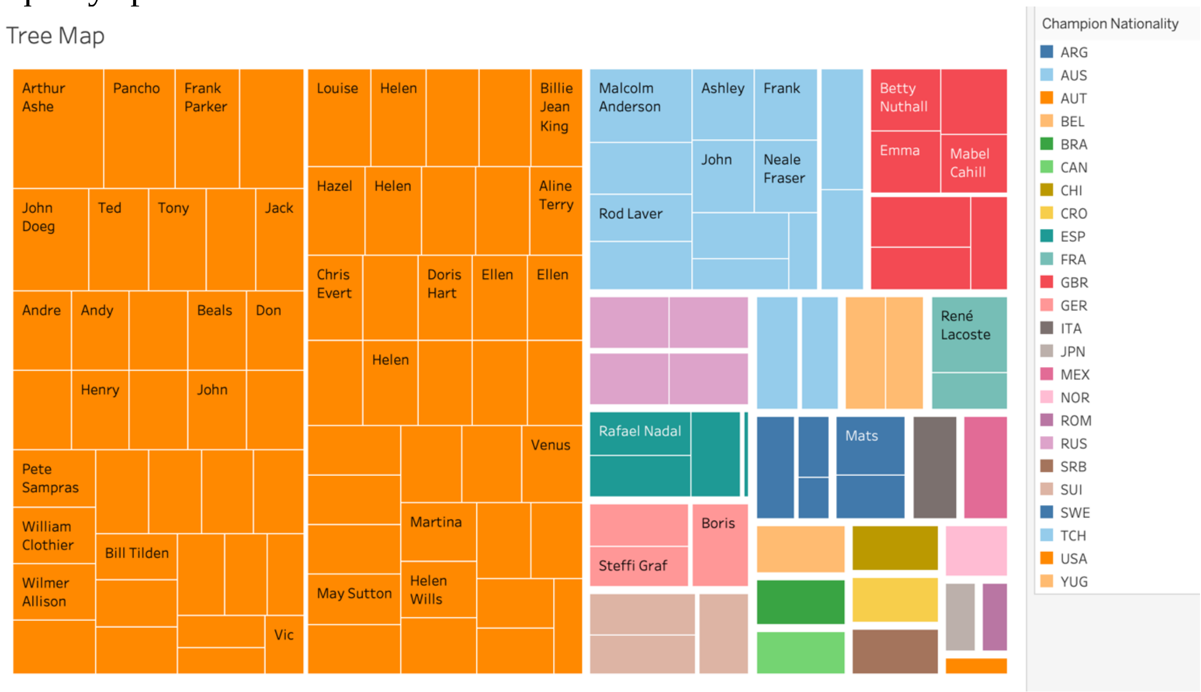


Figure 6 Tree Map

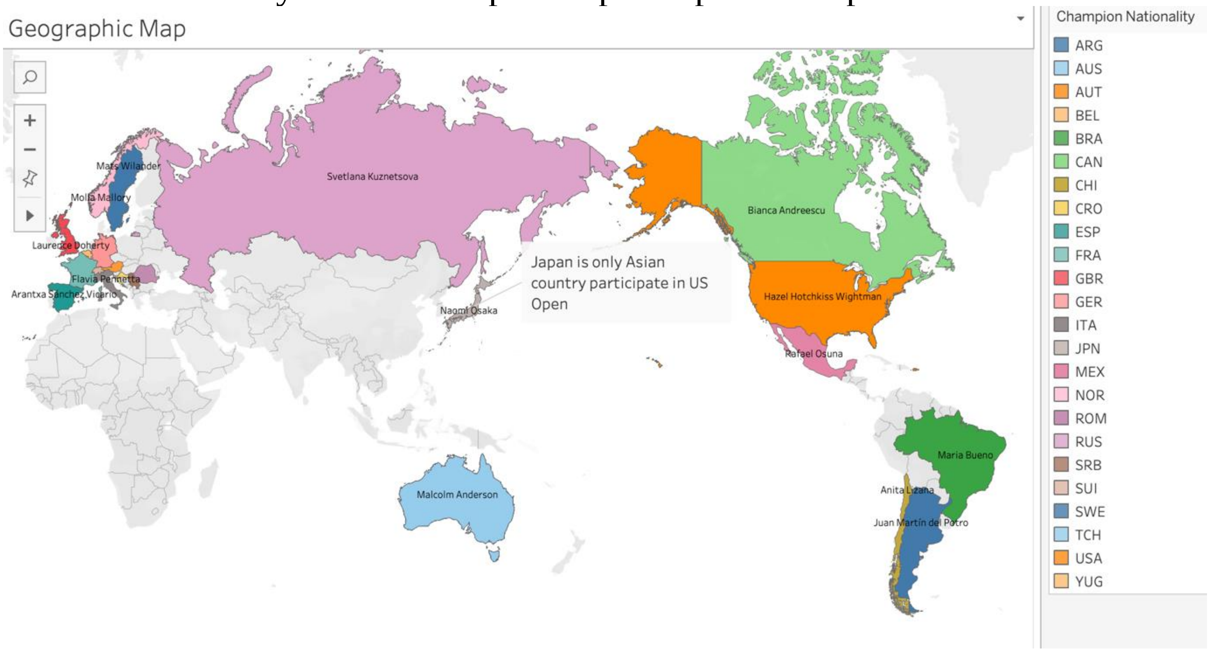
Using colours and sizes, the TreeMap approach enables us to efficiently visualise the percentage of the champions' nationalities in the dataset. Because it can be used to large dimensional data with numerous properties in the same graph, it is practical. The further separation of gender from the nationality proportion of the champions provides evidence of this. As illustrated in the information box, other data can be added to the graph, including score, win rate, 1st to 5th sets win, and loss.

We can see that this method still has certain drawbacks, though. Comparing the smaller squares, for instance, will be challenging if we wish to further analyse the trend of the champions' average match time. Therefore, to visualise this feature and truly reveal any patterns, we will need to use another visualisation approach.

Visualisation Method 3. Geographic Mapping:

The analysis and presentation of spatial and geographically linked data in the form of maps is done using the eye-catching visualisation approach known as the geographic map. A city, state, country, or even a whole continent might be displayed on the map depending on the geographic information provided in the dataset. Categorical Maps are the most common sort of geographic data utilised in Geographic Mapping, which is a method developed from Geographic Information System (GIS) architecture. The categorical values, which show the locations of each value on the map, will be shown on top of the map.

As demonstrated in Figure 7, we use Tableau's produced longitude and latitude as columns and rows, respectively, to construct dots at the relevant geographic location. This categorisation map illustrates how the nationalities of the champions are divided into groups based on their colours and sizes and each country is distinguished by a distinct colour. Figure 7 also shows that the majority of the world's nations are situated in Europe, suggesting that tennis may be a more well-liked sport in Europe than on other continents. We can also note that there is no Asian country win in championship except from Japan.



Japan is only Asian country participate in US Open

Figure 7 Geographic Map

Geographic mapping has the benefit of being a simple method for visualising geographic data. It is easier to compare and analyse patterns across the various nations since it gives us a complete perspective of the global scale via a map. This method, however, can only offer an overview of the data as a basic idea. Thus, geographic mapping is not appropriate if we wish to analyse each participant more thoroughly.

Visual Analytics: Champion performance

Excel's "Filter" and "Count" features allow us to identify the top 12 performers who have won the title five or more times. These players include Bill Tilden, Chris Evert, Helen Wills, Jimmy Connors, Margaret, Molla Mallory, Pete Sampras, Richard Sears, Roger Federer, Serena Williams, Steffi Graf and William Larned.

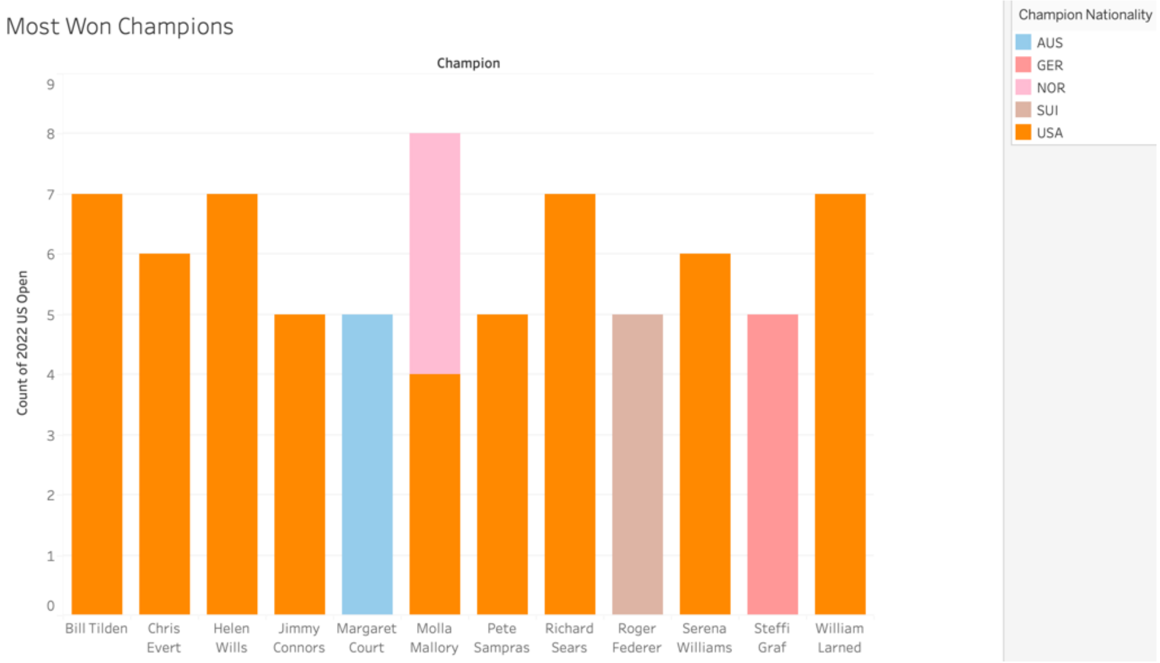


Figure 8 Most won Champions

Figure 8 demonstrates that the United States continues to have the most champions among the top 12. The four US champions—Bill Tilden, Helen Wills, Richard Sears, and William Larned—each have seven victories to their credit. The highest record in this dataset was made by Molla Mallory, a player, who has won up to 8 times. The fact that four of Molla Mallory's eight victories represent the United States and the other four represent Norway is also noteworthy. Australia, UK, Switzerland only has one champion each. The champions have minimum 72% win rate in their victories, showing the superiority.

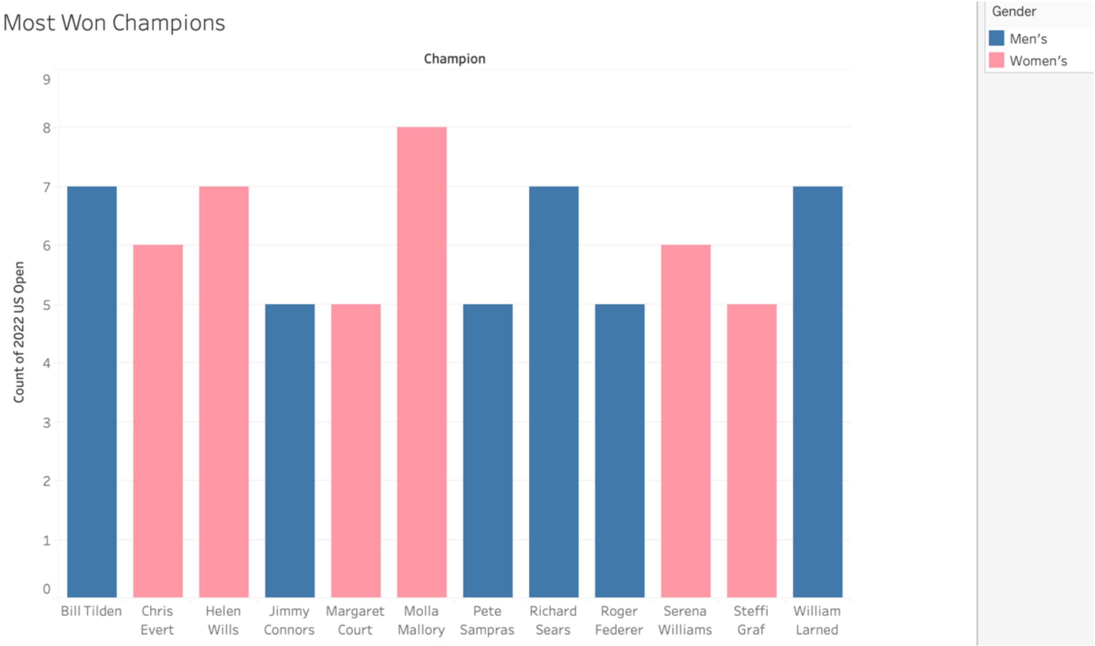


Figure 9

Following a summary of the nationalities and winning percentages of the best players, we now go deeper into the player's specifics, like gender and winning percentage over time, as shown in Figure 9. Six male and six female players are represented among the top 12 titles. Molla Mallory, a female champion, is the top achiever with eight championships, while three male players and one female player are the runners-up with seven championship victories each.

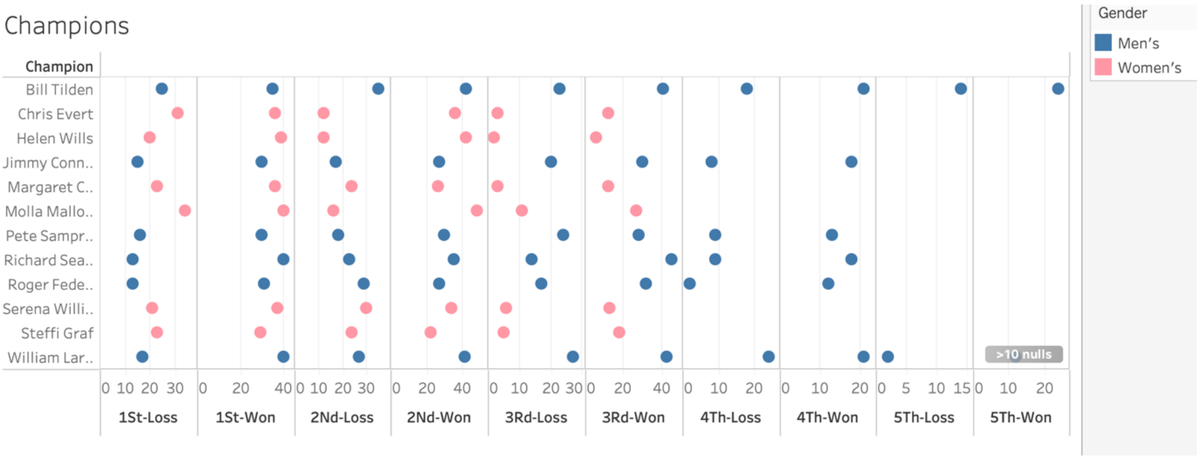


Figure 10

The championship year and winning percentage of the top performers also provide us with an intriguing perspective. We identify the players' potential opponents by carefully examining the time frame during which they won their titles. For instance, the two players, Molla Mallory and Helen Wills, cluster between 1920 and 1930. They do, however, finish their match in the second or third set when they play each other. Since they both had excellent win rates, we are forced to wonder if they are actually rivals or not because their matchups weren't competitive.

Conclusion

The high dimensional Assignment 2 dataset has been analysed using various data visualisation techniques. One of the most popular tools for advanced data visualisation is Tableau. With the use of Tableau, we can apply several visualisation methods like TreeMap, Parallel Coordinator, Geographic Mapping, and other approaches to analyse trends concerning the performance of the champions.

The following are some noteworthy trends:

- There are 12 players classified as the top achievers for winning 5 or more championships with six male players and 6 female players equally divided.
- The 12 champions have minimum 72% win rate in their victories, showing the superiority.
- The USA has the highest number of champions in US Open with 177 people, following by Australia at 24 and UK at 10 champions. These 3 countries are classified as the top countries.
- The highest win among champion is Molla Mallory and she won 8 championships. while three male players and one female player are the runners-up with seven championship victories each.
- Among Molla Mallory's 8 championships she got four representing USA and other four are for Norway.
- The champions of the best nations typically play extended sets than the majority, showing that their matches are intensely competitive with formidable opponents and that they are among the best performers.
- Figure 7 also shows that the majority of the world's nations are situated in Europe, suggesting that tennis may be a more well-liked sport in Europe than on other continents. We can also note that there is no Asian country win in championship except from Japan.
- In men's competitions, Australia has more success than in women's tournaments. While in the USA and UK, men and women on the size side were equally split.
- With an average win percentage of 60% in sets 4 and 5, the win rate decreases as the match goes on.
- The two players, Molla Mallory and Helen Wills, cluster between 1920 and 1930 seem to be rival. However, they finish their match in the second or third set when they play each other. Since they both had excellent win rates,

we are forced to wonder if they are actually rivals or not because their matchups weren't competitive.