Executive Summary

This report primarily focuses on exploring the Wimbledon Championships' dataset provided. The Wimbledon Championship is internationally known as the oldest tennis tournament played every year in London at Wimbledon and is widely regarded as the most prestigious tennis tournament. This dataset holds 133 years of championship match record for both men and women between 1877 and 2019 (258 matches in total). It contains 12 columns; each column represents an attribute that stores some important information about the match record.

After using Excel and Tableau to count the number of titles won by each champion, 15 top players who have won 5 or more titles are found:

- **9 women:**
- **6 men:**

Through in-depth interpretation and analysis of the highlights, patterns and trends identified in the visualisations, the following key conclusions are drawn:

-
-
-
-
-
-
-
-
-
-

Data Exploration

Data exploration is the initial step in data visualisation and analysis. It facilitates deeper understanding and interpretation of the dataset, making it easier to navigate data and identify patterns and trends. The type and description of all attributes are thoroughly investigated and summarised in the table below:

Attribute Name	Type	Description
1. Year	4-digit format in YYYY, quantitative (interval)	Year of the tournament
2. Gender	Binary data, categorical nominal	Divides tournaments by gender, either Men's or Women's tournament
3. Champion	String format, categorical nominal	The name of the tournament's winner
4. Champion Nationality	3-letter string format, categorical nominal	The 3-letter (**ISO Alpha-3**) country code of the champion's nationality
5. Champion Country	String format, categorical nominal	**Geographical data**, the full name of the champion's country
6. Champion Seed	Numeric (integer) format, quantitative	The preliminary ranking of the champion for the purposes of the draw

7. Mins	Integer format, quantitative (ratio)	The length of the championship match in minutes
8. Score	String format, pairs of integers (e.g., 6-4) separated by commas	The score results for each set of games in the championship match
9. Runner-up	String format, categorical nominal	The name of the runner-up in the tournament
10. Runner-up Nationality	3-letter string format, categorical nominal	The 3-letter country code of the runner-up's nationality
11. Runner-up Country	String format, categorical nominal	Geographical data, the name of the runner-up's country
12. Runner-up Seed	Numeric (integer) format, quantitative	The preliminary ranking of the runner-up for the purposes of the draw

Interesting Findings

After exploring the dataset, some interesting findings and anomalies are discovered in the data:

-
-
-
-

Note: A player that has a greater chance to win the tournament will have a numerically lower seed. Player who is seeded #1 is mostly likely to win

Data Preparation

Data preparation is an essential step to perform before creating data visualizations. Clean, complete, and consistent data with no errors will be much easier to visualize. Starting from a clean dataset allows the analyst to focus on creating effective visualizations rather than trying to diagnose and fix errors while creating visualizations.

Inconsistency in Player's Names

Inconsistent naming conventions existed in the columns of *Champion* and *Runner-up* can cause errors and confusion, each player's name should be recorded consistently throughout the dataset. Players that are found to have their names in more than one formats need to be fixed:

Player's Full Name	Different Formats	Fixed
Angelique Kerber	A.Kerber, A. Kerber	A. Kerber
Blanche Bingley	B Bingley, B. Bingley	B. Bingley
Garbiñe Muguruza	G. Muguruza, G.Muguruza	G. Muguruza
Henry Ellsworth Vines	H.E. Vines, H.E.Vines	H.E. Vines
Joshua Pim	J.Pim, J. Pim	J. Pim
Lindsay Ann Davenport	L. Davenport, L.A. Davenport	L.A. Davenport
Maud Edith Eleanor Watson	M.E.E. Watson, M.E.E Watson	M.E.E. Watson

Petra Kvitová	P. Kvitová, P.Kvitova	P. Kvitová
Serena Jameka Williams	S. Williams, S.J. Williams, S.J. Williamsms	S.J. Williams
Venus Ebony Starr Williams	V.Williams, V.E.S. Williams	V.E.S. Williams

Anomaly in Score

One anomaly in the *Score* column is detected on Row 198. It has an unusual "retd" at the end. This anomaly, if not handled, will cause an error when transforming it into numeric format. After investigation, it is found that, in 1911 Men's Singles final, the runner-up H.R. Barrett had to retire at the start of the fifth set. Hence, "retd" stands for retired, it will then be removed as the fifth set was not even played.

Errors in Country Names

All country names should be spelled correctly, errors in the *Runner-up Country* column are detected:

- "Croatiatia" on Row 7. It will be modified to "Croatia".
- "SweDenmark" on Row 62 and 78. They will be modified to "Denmark".

Note: These modifications are just for practice, but not necessary as they will not be used in the visualisations.

Missing Values

Missing values are a common occurrence, especially in a dataset that holds historical time series data like this one. The isnull() and sum() function in Pandas Python are used to get a quick summary of the missing values in this dataset. According to the table on the right, three columns are found to contain missing values: *Champion Seed*, *Runner-up Seed* and *Mins*. For *Champion Seed* and *Runner-up Seed*, the values are missing not at random. Before 1927, they are missing because the seeding system was not introduced at Wimbledon. After 1927, they are missing because the player was unseeded. For *Mins*, the values are missing because they were not collected at the time and too old to be traced, so they become irretrievably lost.

Column name	No. of Missing Values
Champion Seed	88
Mins	64
Runner-up Seed	100

Clearly, it is not a good idea to delete the entire record that contains missing values, as it holds other important information that should be kept. Also, statistical imputation does not necessarily give better results in this case. Note that Tableau can recognize null values and automatically exclude them from the view and aggregate calculations. Hence, the missing values in this dataset do not require special treatment and will be left null.

Note: Replacing them with 0s is a bad idea as it will cause serious errors when using Average function.

Attribute Transformation - Score to Win Rate

The *Score* column stores the text data in string format that cannot be used in visualisation. In order to extract the valuable information from the *Score* column and facilitate further data analysis, it is necessary to transform it into numeric format and create a new attribute as win rate using Excel.

Tree map

A tree map is a rectangle-based visualisation that displays hierarchically structured data using a set of nested rectangles of varying size and colour. In a tree map, each category is represented by a rectangle, and each rectangle can be further divided into smaller rectangles that represent subcategories. Meanwhile, the size of each rectangle is proportional to the quantitative value of the measure they hold, and the largest rectangle is placed in the top left and arranged in order of decreasing size down toward the bottom right. Also, it is important to note that the color-coding in tree maps can be used in two ways; one is for denoting the categories by using different colours like pie charts, the other use is to express a second quantitative measure by using the shade or intensity of a single colour.

Interpretation and Pattern Analysis

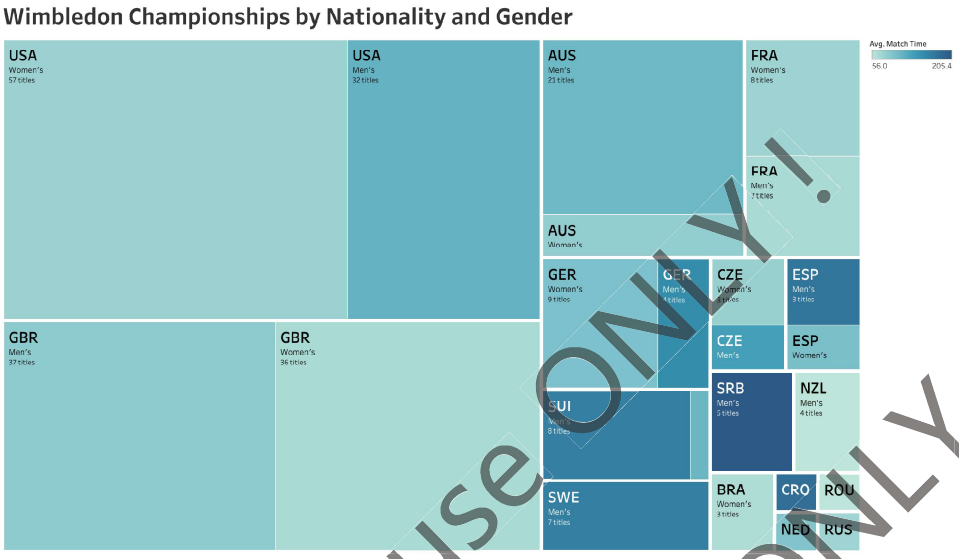


Figure 1: Tree map, Wimbledon Championships by Nationality and Gender

Two tree maps are created and elaborated respectively. The first tree map is configured with 2 levels of hierarchy (Champion Nationality and Gender), and 2 quantitative measures (number of titles and average match time) as *Figure 1* illustrates. The number of titles measure is used to determine the size of each rectangle. The average match time measure is used to determine of the shade of the colour for each rectangle as the legend depicts in *Figure 1*. A darker shade of blue implies a higher average match time for that rectangle. At the first level of hierarchy, each rectangle represents a nationality, the 3-letter country code is included in the labels at the top left corner of each rectangle.

Comparing to the first tree map, the second tree map is configured with one more level in hierarchy – Champion.

The Advantages and Disadvantages

The use of the tree map visualisation technique for this dataset has many advantages. Firstly,

The tree map also has its limitations.

Parallel Coordinate

Parallel coordinate plot is a powerful visualization technique for analysing high-dimensional data. In a parallel coordinate plot, there are multiple axes that are placed parallel, vertical, and equally spaced, each vertical axis represents a quantitative variable (measure) and often has its own scale and even different units of measurement. In that case, all axes can be normalized to keep the scales uniform. Meanwhile, each data record from the dataset is mapped into a series of lines connected across each axis. In addition, since there is no natural order among the axes, the order of the axes becomes crucial because it can strongly affect the readability of the plot. One reason for this is that the relationships between adjacent variables are easier to perceive than non-adjacent ones. Hence, re-ordering the axes can help in discovering patterns and correlations across variables more easily. When most of lines between two adjacent axes are somewhat parallel to each other, it suggests a positive correlation between these two dimensions. On the other hand, when there are lots of crossing lines (X-shapes), it indicates a negative relationship.

Interpretation and Pattern Analysis

The parallel coordinate technique is used to analyse the relationship between champion's nationality, gender, year, and match time.

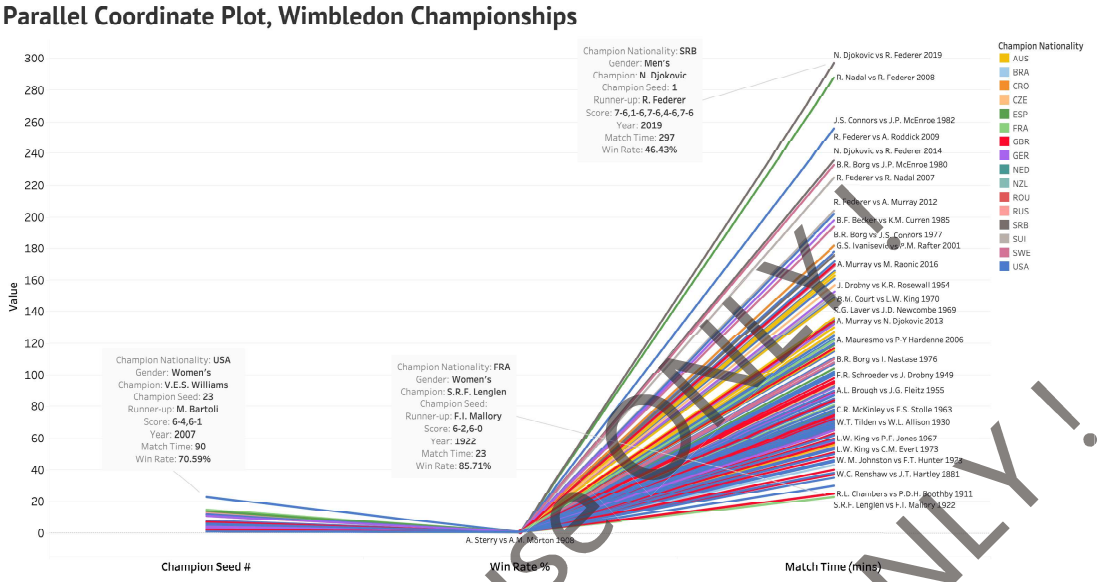
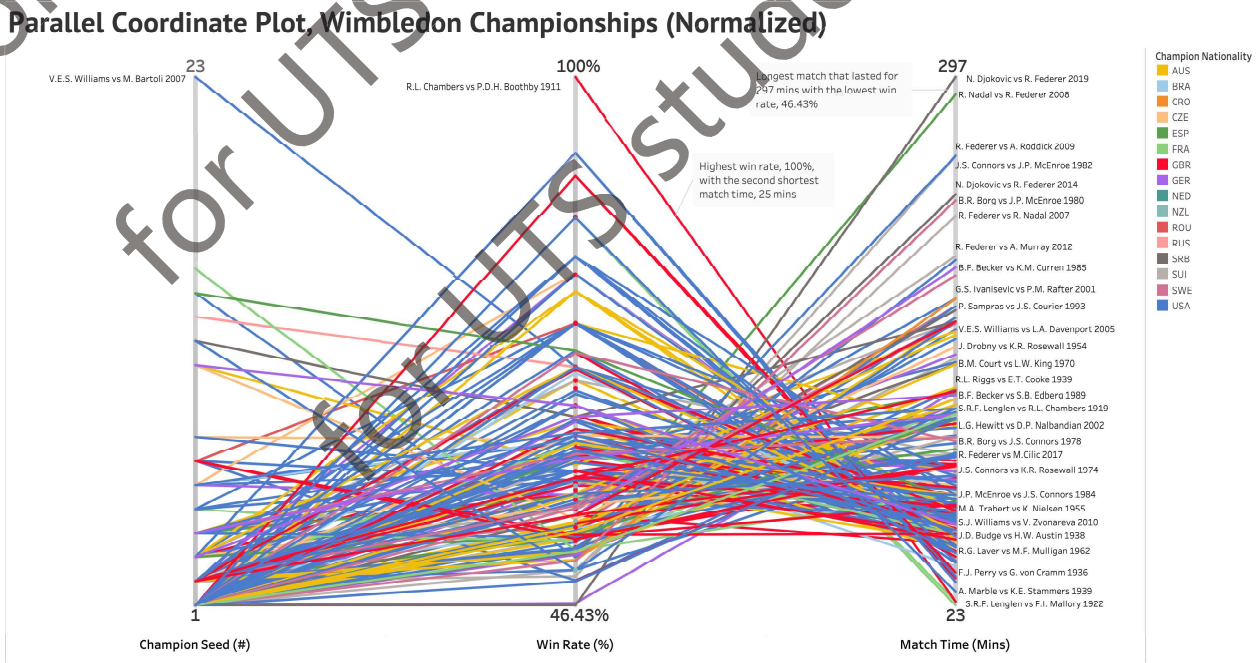


Figure 3: Parallel Coordinate Plot (Champion Seed, Win Rate, and Match Time)

As *Figure 3* above illustrates, the first parallel coordinate is made of 3 axes (variables): *Champion Seed, Win Rate* and *Match Length*. In this chart,



The second parallel coordinate plot is an improved version of the previous one by utilizing the scaling technique (the min-max normalization).

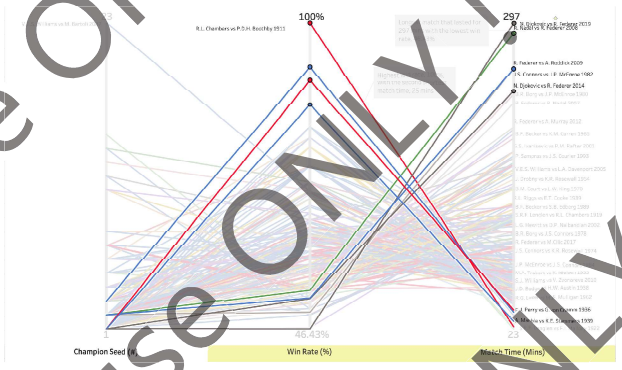


Figure 5: X-shapes between Win Rate and Match Time

The Advantages and Disadvantages

The main advantage offered by the parallel coordinate is

The disadvantages of the parallel coordinate are also obvious.

Geographic Map

Geographic map is an eye-catching visualisation technique used to analyse and display the spatial and geographically related data and present it in the form of maps. According to the geographic data given in the dataset, the map can be displayed by district, city, state, country, or even continent. There are many types in map visualisation, such as filled (choropleth) map, symbol map, heatmap, etc. In a filled map, the geographical areas are coloured or shaded in proportion to a quantitative variable being displayed over the map. Meanwhile, in a symbol map, each geographic location is represented by a circle symbol on the map. The circle symbols provide two visual cues to help the audience analyse the data on the map: size and colour. The size of the circle is proportional to the value of a quantitative measure. The use of colour is similar to the colour-coding for rectangles in tree maps, it can be used to either represent a second quantitative measure or a categorical variable.

Interpretation and Pattern Analysis

Geographic Map, Wimbledon Champions across the World

Figure 6: Symbol map, Wimbledon Champions across the world

The geographic role is assigned to the *Champion Country* field, so Tableau can then recognize each country name and automatically generate the corresponding latitude and longitude. The first geographic map is an example of symbol map as shown in *Figure 6*.

Dual-Axis Geographic Map, Wimbledon Champions over the World

The second geographic map is a dual-axis (layered) map created by stacking a symbol map onto a filled map in Tableau. As shown in *Figure 7*,

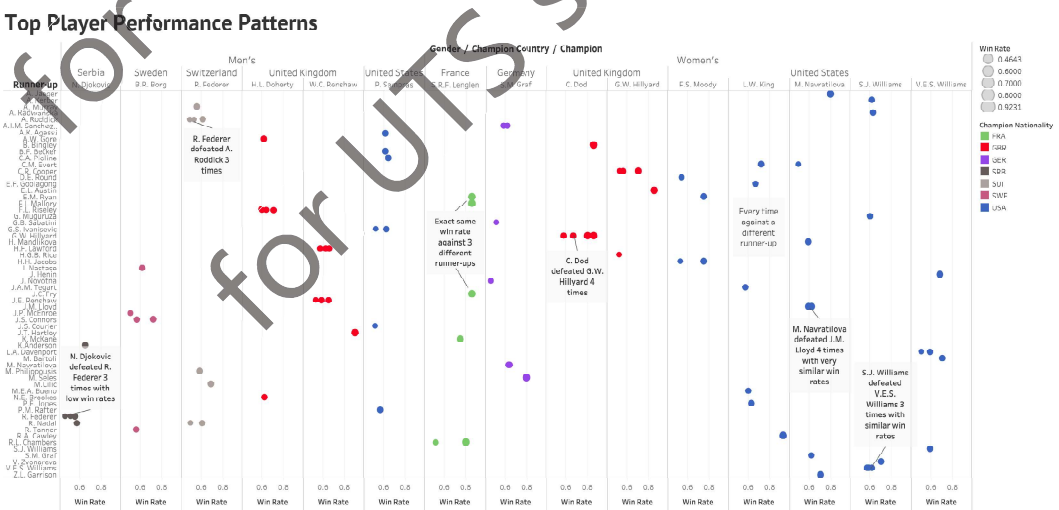
The Advantages and Disadvantages

Comparing to the tree maps, the geographic map is clearer and more intuitive.

The disadvantages of the geographic map are similar to the tree maps.

Top Player's Performance Visual Analysis

After cross checking the results from Excel and Tableau, 15 players are found to have won 5 or more Wimbledon titles.



To further explore and study the top player's performance, an interactive visualisation is created to reveal the underlying patterns in the data. As *Figure 8* illustrates, the dashboard visualises the top player's win rate against each runner-up they have played against as circles.

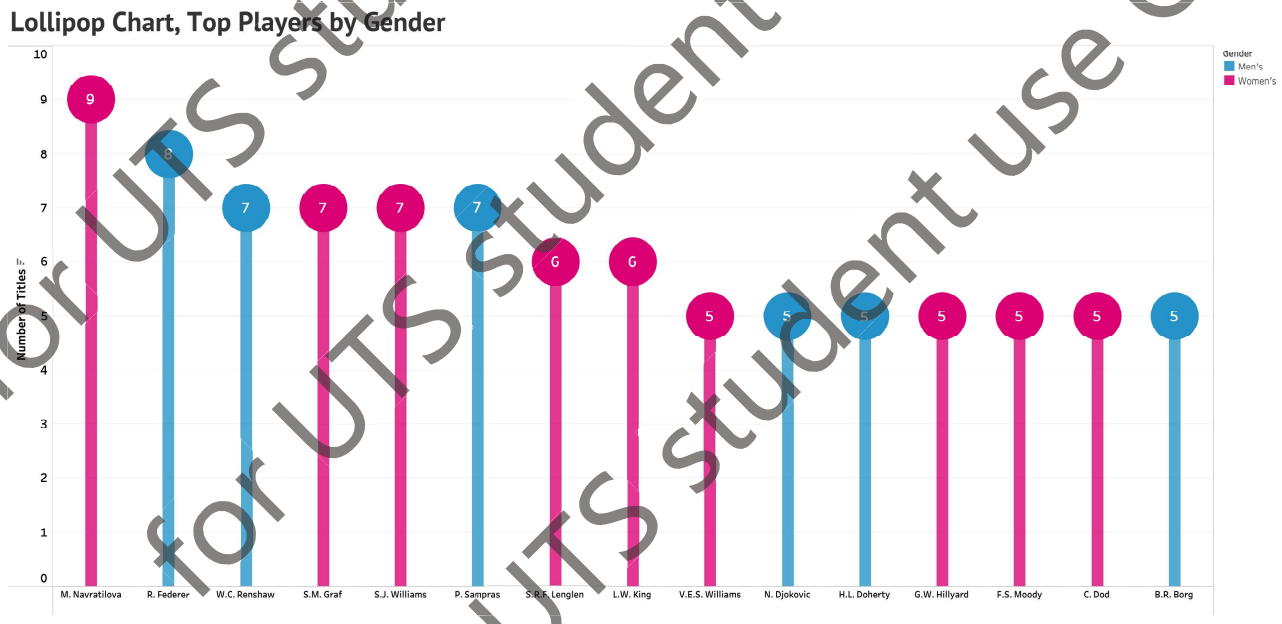


Figure 9: Lollipop Chart, Top Players by Gender

The second visualisation created is a lollipop chart as shown in *Figure 9*. The lollipop chart is basically a useful variation of a bar chart where the bar is replaced with a line and a circle at the end. It is often claimed to be more visually attractive and clear, compared to a normal bar chart. In this chart, each circle is also labelled with the number of titles won by that specific champion.

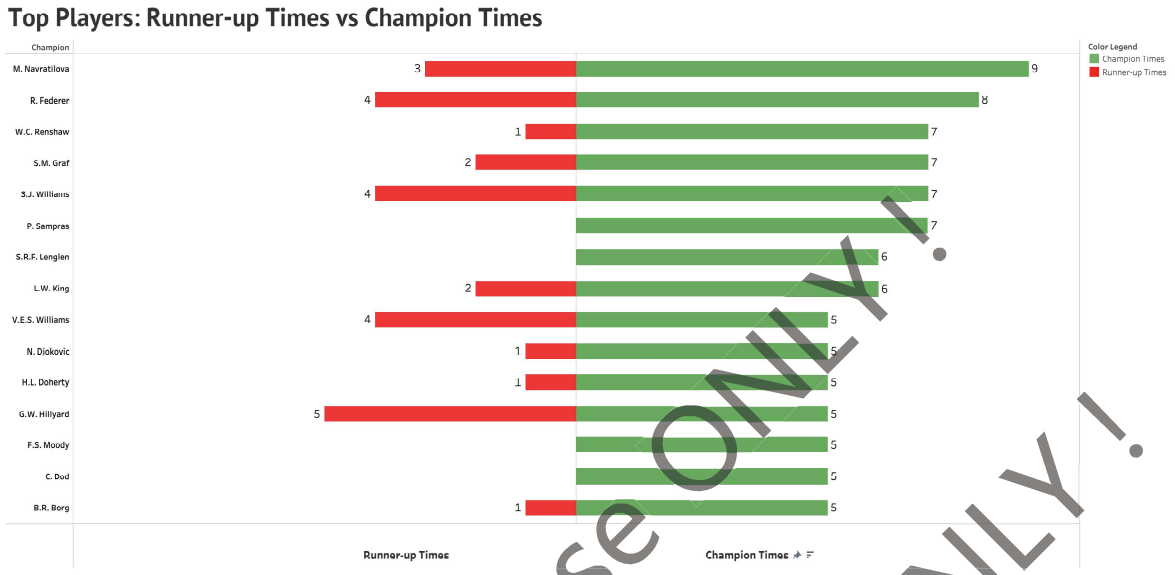


Figure 10: Bi-directional Bar Chart, Runner-up Times vs Champion Times

The third visualisation created by Tableau is a bi-directional bar chart as shown in *Figure 10*. It is mainly used for comparing two sets of data side by side along one vertical axis. In this visualisation, the number of titles won by each top player is displayed as green bars on the right side of the vertical axis, while the number of runner-up times for each top player is displayed as red bars on the left side.

Figure 11: Scatter Plots for Time Series Analysis

Lastly, two scatter plots are crafted to visualise the trends existed in the top player's average match time and win rate over time as *Figure 11* shows.

Conclusion

In conclusion, the data visualisation techniques have effectively uncovered highlights, hidden trends, and patterns in the Wimbledon Championship dataset. Throughout the in-depth analysis of all the visualisations, the following valuable insights and conclusions are drawn: