Comparing Tennis Player Visualisations for Casual Online Bettors

Yosef Ardhito Winatmoko (2032342)

ABSTRACT

In this report, we introduce a case of casual bettors who need information about top tennis players statistics. We propose and analyze the strong and weak points of six visualizations to help a fictional character make better decisions. In the end, we rank the proposed diagrams and argue which is the most suitable for the case.

1 INTRODUCTION

In the first assignment of the Data Visualization course, we analyze the strengths and weaknesses of types of visualization for a specific dataset and use case. Among the three options, we are using tennis player data based on Wikipedia. The data contains information about the current top-5 tennis players in the world. There are two categories: track record and head-to-head data. The former consists of personal data, such as birthdate and height, as well as professional data, including the number of won/lost matches, the number of titles, and when they start becoming a pro tennis player. On the other hand, the head-to-head data stated how many times every player has won against the four other players.

The organization of the report is as follow. First, we determine the audience and also define a persona. Next, we discuss the issue addressed in the use case. Moreover, we have to take into account the additional requirements of the user when we design the visualization. We propose the graphs and the visual elements description afterwards. There are three charts for each track record and head-to-head data. Based on the alternatives, we elaborate on the plus and minus of each visualization. Finally, we conclude which representation is the most suitable for the defined use case.

2 USE CASE SCENARIO

A visualization can be regarded as useful if it helps a certain audience make a decision. Therefore, the first step is to define the audience for whom the information is going to be presented. For the chosen dataset, we can determine that the audience will be the general public that has some sort of interest to tennis. As an example, one option is to help people who regularly watch tennis matches and discuss with their friends afterwards. For such an audience, the visualization would serve as mere statistics to bring up during a conversation. For the purpose of measuring the usefulness of the produced visualization, we need a more appealing type of audience.

One of the most apparent users of sports information to make a decision is the online gambling community. Sports betting is one of the most popular forms of gambling and tennis is not an exception. As an illustration, at least 45% of adults in the United Kingdom (UK) have participated in any form of gambling in 2017 [1]. For online gambling, the highest age group proportion is between 25-34 years old. Looking at the age group, it is not surprising that laptops remain the most popular device for gambling. However, 51% of the respondents have used either mobile phones or tablets for gambling and the trend is going upward. In terms of the reason, in general, they want to earn profit but also for enjoyment.

2.1 Persona

Based on the survey conducted in the UK in 2018, we establish a persona as a representation of a group. We call the group as "The Casual Bettors", and the persona is a fictional character named Zach Wood. He is a 30-year old project manager of a UK-based software company. He comes from Birmingham city but currently lives in an apartment in downtown London. The company he works for has footprints throughout Europe, with more than 2000 employees in 3 different countries. His yearly income is 45.000 euro and he loves to bet on sports because he is confident with his knowledge regarding a variety of sports and has earned profit from sports betting so far. However, he never gambles a significant amount of his money, 100 euro on average and at most 400 euro per month. He can afford to gamble due to his status as a single.

As a business graduate who works in a software company, Zach is a tech-savvy person. He owns the latest version of gadgets and organizes his professional life through his laptop. He bets with applications in his mobile phones after work or on the weekend. He has multiple accounts in various betting sites. He is fond of betting not only in tennis but also in football and basketball. In the office, he often discusses with his fellow co-workers on who to bet for. In addition, Zach and his colleague also play sports together. However, tennis is not the most popular sport they play, only once or twice every year. Instead, they play basketball or football regularly every week. Nevertheless, Grand Slam tournaments are always a big theme of conversation when it is on.

2.2 Requirement

With regards to the news source, Zach relies on popular sports sites such as BBC sports and ESPN. He never misses any updates because he checks almost every moment he has to wait or has nothing to do for a few minutes. Because he checks through his phone, this introduce a requirement for the visualisation: it has to be readable even on a smartphone screen. Moreover, since Zach frequently skims the articles, the graph has to catch his attention instantly. The goal for him is to make a profit out of sports betting, thus the visualizations should clarify where to put his money on for the next tournaments. The detail of the value from the visualization is discussed in the following section.

2.3 Value

In general, Zach needs the player's data to help him decide if he should participate in sports betting. As a busy project manager, he does not have the time to digest the table as-is. Although the data contains only the top-5 players, the track record data has 16 columns. The head-to-head is less complicated, yet it is still time-consuming to discover which player to bet on by looking on the table directly. Therefore, the visualization should help Zach understand the player's quality better and easier.

For the sake of the scenario, assume the release date of the visualization would have been January 14, 2020. One of the Grand Slams, the Australian Open, will be starting in a week. The information about player's record will be valuable for Zach to determine if he should consider betting on one of the player and, if he does, which one. Therefore, the value of the visualization is to help Zach make decision on betting for the Australian Open. In particular, we define the following five questions:

1. Who will emerge as the champion of the Australian Open for the men's single category? (Q1)

2. Who is the best player in Australian Open historically? (Q2)

3. Which player is leading the head-to-head among the Top 5? (Q3)

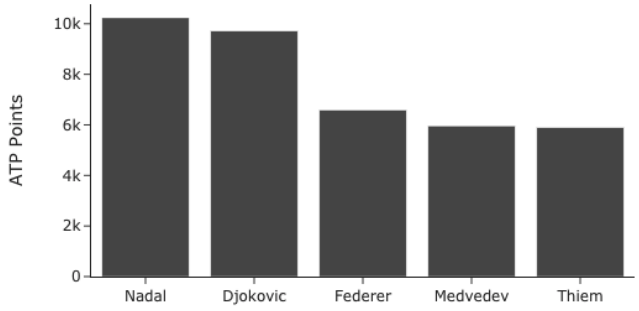

Figure 1: ATP Points of each player

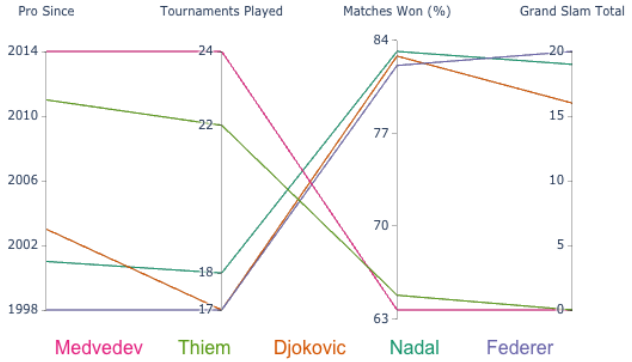

Figure 2: Parallel coordinates plot of player's track record

4. As Australian Open will be held in a GreenSet court, which player is good at this type of court? (Q4)

5. Are there any potential new challengers for the men's single? (Q5)

3 Track Record Visualizations

The challenges with the track record data are the number and variety of the columns. We have birthdate and pro since, which show date. On the other hand, there are also counts of title, including the breakdown by each Grand Slam tournaments. Of course, not all the columns are relevant for the use case. We start with the minimal data which already contain essential information and build upon more columns.

3.1 ATP Points Bar Chart

ATP Points is a summary number that shows the performance of each player for the past 52 weeks. This number is also the base of ATP ranking. By itself, ATP Points already contain a meaningful indication of a player's current form. A player earns points by winning prestigious tournaments, stratified by the scale. Fig. 1 shows a simple bar chart of the ATP Points sorted from the highest to the lowest. The only mapping is the ATP Points to positions. We do not need to use any other visual channels because we want to show the ATP Points only.

Strength The most obvious reason for using this chart is simplicity. By showing only ATP Points, the audience requires only a few seconds to comprehend the chart. It will also work as well in small screens, such as mobile phone. Bars are easy to digest because humans are good at perceiving positional differences. As an illustration, we can immediately see that the gap between Nadal and Djokovic is smaller than Djokovic to Federer. In a similar vein, we can immediately understand that Medvedev's and Thiem's points are more or less on the same level.

Weakness Out of the available 16 columns, here we only use one, albeit it might be the most representative one of a player's performance. If the viewer wants to observe which player is the best in general, this chart might suffice. However, ATP Points only includes the last 52 weeks. It is impossible to see a player's whole track record from this single measurement. Besides, the simplicity of the chart can be too monotonous, thus fail to attract his attention.

3.2 Professional Record Parallel Coordinates

Fig. 2 shows four columns of track record in one parallel coordinates chart. Each line represents a player, distinguished by the colour. We pick the colours using palettes from ColorBrewer[1]. One key decision is how to indicate which line represents which player. In the end, we decided to place the legend at the bottom for the reader to observe the plot first and then they can check the names. Additionally, we do not put additional small coloured lines for legends because the font colour is already distinguishable and we refrain from adding additional unnecessary elements to the chart.

Strength By representing four columns in a plot, one can make a better judgement of which player is performing well using more comprehensive criteria. The strong point of a parallel coordinates plot is to surface the correlation between attributes. In summary, we can see that there are at least two groups: the experienced and newcomers. The former consists of Djokovic, Nadal, and Federer. Their pro career started between 1998-2004, but they play fewer tournaments compared to the junior player. Although there is no additional information provided regarding this, we assume this number only counts recent tournaments instead of all-time. Their seniority is proven by the high won percentage. Consequently, we can see that the three senior players have been dominating the Grand Slams with no title for Medvedev and Thiem.

Weakness The rich elements might hinder the readability of the graph if the reader uses a small screen. One aspect is the font size of the axes, which can be too small. Regarding the colours, the perceived difference can diminish because there are five players, for example between Thiem and Nadal. We can try to use other colours, but five steps present a challenge. The legend at the bottom also introduces additional time for the reader to match with the lines.

3.3 Grand Slam Treemap

An important aspect of the use case is the upcoming Australian Open. With a treemap shown in Fig. 3, we use two visual cues: the colour mapped semantically to court type and the area of the rectangles represent a number of titles won by each player. The saturation shows the three different levels, i.e. court type, then grand slam tournament, and lastly the player names.

Strength For every grand slam tournament, it is trivial to define the dominating players. For instance, Nadal is a standout in the French Open but not doing well in the Australian Open. We can also see the tournament group based on the court type. We match the colour according to the real-world court, which makes it easier to understand the chart [2].

Weakness Comparing areas of similar rectangles might be tricky. For instance, Djokovic has won one more Australian Open title than Federer, but it is not apparent in the chart. Regarding the labels, if a player name does not fit in the rectangle, the name is omitted. It is easy to deduce the names for the empty boxes, but it reduces the readability to some extent. The treemap also does not show absolute values. Readers can only compare the number relatively between players and tournaments, but no way of knowing the absolute value.

[1] http://colorbrewer2.org

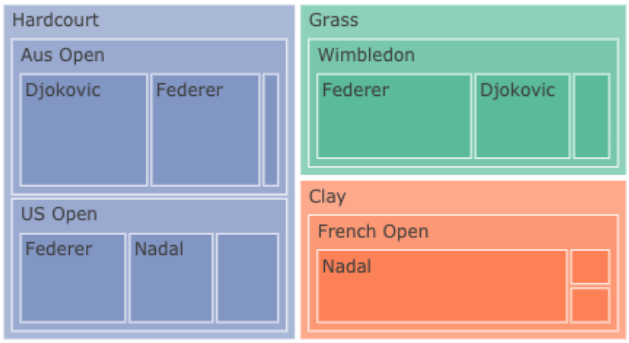

Figure 3: Number of Grand Slam titles per player by court type

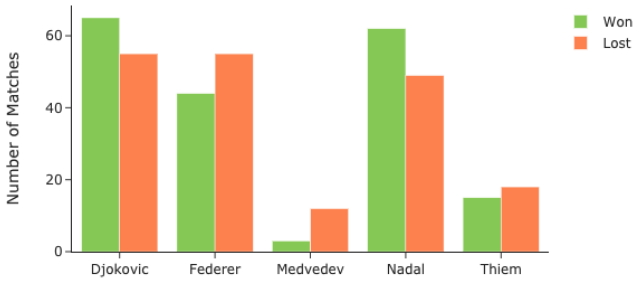

Figure 4: Accumulative number of matches played by each player against the rest of the top-5

4 Head-to-head Visualizations

For the head-to-head data, we compare the number of matches won and lost between every player of the Top-5. The challenge is to aggregate the data and make it easy to see the relative strengths of each player. For the aggregation, we introduce "Won Percentage", which indicates the percentage of matches won from all match played between two players.

4.1 Top-5 Head-to-head Bar Chart

Fig. 4 shows an accumulative number of won/lost matches of each player by all other players in the top-5. The number of matches is mapped to the bar positions and the green/red colour represents the result of the matches. X-axes for the won and lost matches of the same player are grouped.

Strength Comparing who has won the most or lost the least with this chart is trivial because it shows the absolute number. Subsequently, we can compare for each player whether they have won or lost more matches against the top-5. The colour makes it easy to digest as well because green means something positive and red is negative. The chart can also be shown equally well on a smaller screen.

Weakness The head-to-head information between each player is no longer available. Comparison between player becomes harder because we have to take into account two If we want to know what is the won percentage of every player, we have to calculate by ourselves.

4.2 Top-5 Head-to-head Relationship Mapping

In the mapping shown in Fig. 5, we can see the won percentage of every player against each other players in the top-5. The comparison of players is presented as the position, y-axis as the player and x-axis as the compared opposition. The won percentage value is represented by a colour scale: 0% as red, 50% as yellow, and 100% as green. 50% is yellow because it means both players have won the equal number of matches between each other. Any number lower than 50% means higher lost number and expressed with the orange-to-red scale, whereas yellow-to-green means won percentage above 50%. We plot only half of the map because the other half is just an inverse.

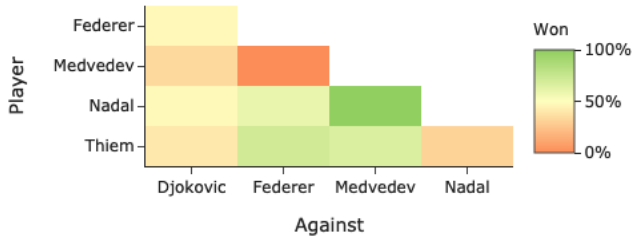

Figure 5: Won percentage mapping between each player

Strength With the mapping, we can see the won percentage of every pair of players. Multiple yellow-to-green colours in the same row mean the designated player tend to be better than other top-5 players. As we only show half of the mapping, there is no redundancy of information. Similar to Fig. 4, being consistent with the common norm by using green to indicates something wanted and red for unwanted. In terms of the red-to-green gradient, which is inexistent in human minds, we put yellow in-between to help with understanding the scale.

Weakness Using a half mapping is by no means trivial. The reader needs to first understand what are the axes represent. They can make a critical mistake for misunderstanding the meaning of the position. Additionally, we need more time to analyse the performance of players that appear on both axes. To illustrate, if we want to see all Federer's head-to-head, we need to check both: horizontally, where green is preferable, and vertically, where red means better. Using hue as the channel also presents a challenge to compare between two won percentages. We know that Djokovic has above 50% won percentage against Nadal and Medvedev, but what is the relative difference between the two? Subsequently, the yellow colour in the mid-point might also introduce non-linear gap, which makes it even harder to compare between two blocks.

4.3 Top-3 Head-to-head Grouped Bar Chart

Not all of the top-5 players have won any grand slam. Fig. 6 presents the head-to-head exclusively between players who have won any grand slam title. However, as oppose to Fig. 4, we do not accumulate the number of matches. Instead, we plot the won percentage of every player with the other two possible oppositions. The colour is an indication of the player that is represented. Furthermore, we put a line at 50% as the "Equal" line, which is the threshold of whether the player acquires more wins or loses from the opponent.

Strength Compared to Fig. 5, it takes less effort to comprehend this chart because it uses position as a channel as opposed to hue or saturation. In terms of head-to-head, Djokovic is the winner because both bars passed the helper "Equal" line. Nadal vs Federer also pops out immediately, where the former seems to get the best out of the latter by a margin. With the "Equal" line, there is a clear takeaway for the reader.

Weakness In favour of readability, this chart introduces redundancy because every pair is presented twice. Another thing is the fact that this chart only includes head-to-head between three players and adding more player will bloat the number of bars exponentially. The use of different colours for every player may also be redundant with the position of the bars.

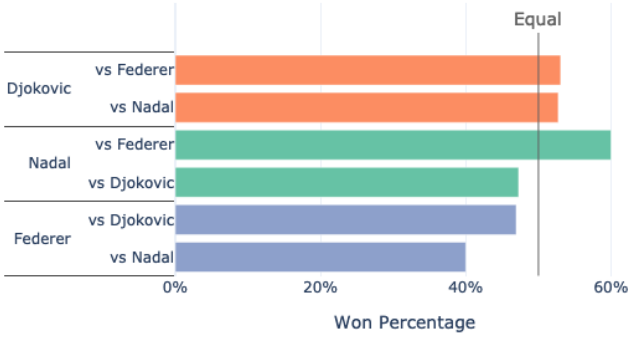

Figure 6: Head-to-head won percentage of top-3 players

Table 1: Comparison of usefulness for each figure. Q=Question, RR=Readability Rank

	Q1	Q2	Q3	Q4	Q5	RR
Fig. 1	✓				✓	1
Fig. 2	✓				✓	6
Fig. 3	✓	✓		✓		**4**
Fig. 4			✓		✓	**2**
Fig. 5			✓		✓	5
Fig. 6			✓			3

5 Discussion

Based on the strong and weak points of each figure, we argue which charts are the best in the case of Zach. First, we shall define general criteria of a high quality visualization. A good visualization has at least two characteristics: it has to support the user in performing their task and take into account our perceptual measure [3]. For the sake of simplicity, we call the former as "usefulness" and the latter as "readability". For the usefulness, we measure every figure by the ability to answer the questions of the user. There are five questions regarding Zach's scenario. Table 1 compare the usefulness of every figures according to the questions.

In terms of usefulness, we see that Fig. 3 can answer the most number of questions. It is also the only visualization that can answer Q2 and Q4. The worst among the six charts is Fig. 6, because it can only answer Q3, while Fig. 4 and Fig. 5 answer the same question along with Q5. Fig. 1 and Fig. 2 can address Q5 as well, although we can argue that each can answer to a different extent. Thus, by combining Fig. 3 with either Fig. 4 or Fig. 5, we can help Zach with all his five questions with no redundant information.

When we compare according to the readability, the ranking is less obvious. There are many visual aspects that we need to consider. We based our criteria on the requirement of the use case. As Zach uses mobile phones to read the news and does not have much time, the visualization should fit on a small screen and also does not take more than approximately 5 seconds to comprehend ("5 Seconds Rule"). All of the analysis is going to be based on hypothesis as we do not experiment to verify them. The rank of the figures based on the readability (RR) from the scenario is also summarized in Table 1.

Based on the mobile phone medium requirement, Fig. 1 and Fig. 4 can definitely work with smaller screen. Although, the font size of Fig. 4 need to be adjusted large enough to make it readable. Fig. 3 might encounter an issue due to relying on colour as value indicators. When the screen is small, it will be harder to spot the difference between two different blocks. Finally, Fig. 2 will not fit well on a small screen due to the text size and the usage of colour as well for the player names. On the other hand, Fig. 5 and Fig. 6 might work because the values are mapped to area and position as opposed to colour, which is used only for grouping similar elements.

When we consider the "5 Second Rule", again Fig. 1 is the most obvious winner. Other bar charts, Fig. 4 and Fig. 6, are also easy to comprehend. However, we have to check whether the "Equal" line in Fig. 6 helps the user understand the message better or not. Fig. 5 also requires more time to understand due to the axes and colour scale used. The user needs to grasp the player names location and the fact that some players only appear in one axis. Although not as powerful as position, area comparison should be sufficient to deliver the main message, which is the dominant players. Therefore, Fig. 3 should fulfil the rule as well.

6 Conclusion

In this report, we propose six alternative visualizations for an audience called "The Casual Bettors". They are the user of online sports betting platform, but they do not rely financially on the profit and play it also for the enjoyment. We introduce our fictional character, Zach Wood, as a persona. Zach's question and requirement are the basis of justifying which figures are the most suitable for the use case.

According to two criteria, namely usefulness and readability, we can conclude that one figure will not be sufficient. We require a composition of at least two charts out of the proposed six. The best visualization is a combination of Fig. 3 and Fig. 4. The former is mandatory to answer Q2 and Q4, as no other visualization is capable of answering those questions. The treemap form also fulfills the additional requirements from Zach. In contrast, Fig. 4 and Fig. 5 are addressing the same set of questions. Therefore, in terms of usefulness, they are interchangeable. However, Fig. 4 is better at the readability due to relying on position to encode the values instead of colour scale in Fig. 5. As for the format, Fig. 4 and Fig. 5 can be arranged top and bottom, respectively. We want Fig. 4 as the top one because it includes more players and Fig. 5 can act as further analysis of the top-3 from the bar chart.

7 Reflection & Future Work

We look into reflection and possible improvement from three perspectives: the visuals, the data, and the scalability. For the visuals, we have not considered colour-blind people so far. It is better to refrain from using red and green combination as used in Fig. 5 and Fig. 4.

Regarding the data, we should include more data about the tournaments, as it is the point of interest as shown in the questions. Because there is only one visualization, Fig. 3, that uses the tournament hierarchy, there is no alternative to answer Q2 and Q4. We believe by adding tournament information and time dimension to the data, they will increase the choices on how to visualize the data significantly. For instance, we can visualize when the matches between the players occur, make a timeline, or show only one tournament, in this case the Australian Open.

Finally, we discuss the scalability of the proposed visualization. For Fig. 3, we can add more tournaments as well and then we can see more player involved in the visualization and also compare with using Grand Slam vs non-Grand Slam rather than the court type. The more players included in Fig. 4, the harder it is to fit the x-axis. To tackle this, we can convert it into a horizontal bar chart, making available more space for more players.

References

[1] J. Barnfield-Tubb and L. Harris. *Gambling participation in 2017: behaviour, awareness and attitudes*. UK Gambling Commission, 2018.

[2] C. G. Healey. Choosing effective colours for data visualization. In *Proceedings of Seventh Annual IEEE Visualization'96*, pp. 263–270. IEEE, 1996.

[3] H. Rushmeier, H. Barrett, P. Rheingans, S. Uselton, and A. Watson. Perceptual measures for effective visualizations. In *IEEE Visualization*, pp. 515–517, 1997.