320146 Assessment Task 2

Advanced Data Visualization

Laila Lima Alves

—

Student ID 14344509

—

Workshop 02- MON- 10:30am

Executive Summary

The Report is focused on the exploration of the dataset for the Tennis US Open championship matches.
The US Open is a traditional tennis tournament which has a history of over hundred years. The tournament includes men and woman inputs between 1881 and 2021. There are 276 observations (data incidents) and 22 attributes (dimensions) in total.
The attributes have been manually modified by the instructor to allow an easier manipulation and visualization in Tableau. Some variables were excluded as they do not provide qualitative values, the ones used include: year, gender, champion's name, nationality, champion seed, match time, runner-up's name, nationality, and score.

Data Exploration

Characteristics

Original file contains 22 columns and 276 rows. Column names and characteristics are:
Year → numerical value
Gender → String
Champion → String
Champion Nationality → String
Champion Country → String, manually adapted by instructor
Champion Seed → String
Mins → String
Score → String
1st-won → Numeric, manually adapted by instructor based on the scores column
1st-loss → Numeric, manually adapted by instructor based on the scores column
2nd-won → Numeric, manually adapted by instructor based on the scores column
2nd-loss → Numeric, manually adapted by instructor based on the scores column
3rd-won → Numeric, manually adapted by instructor based on the scores column
3rd-loss → Numeric, manually adapted by instructor based on the scores column
4th-won → Numeric, manually adapted by instructor based on the scores column
4th-loss → Numeric, manually adapted by instructor based on the scores column
5th-won → Numeric, manually adapted by instructor based on the scores column
5th-loss → Numeric, manually adapted by instructor based on the scores column
Runner-up → String
Runner-up Nationality → String
Runner-up Country → String, manually adapted by instructor
Runner-up Seed → String

The File has been cleaned and adapted for the analysis and do not contain missing values, anomalies or double counts. Players names are consistent and follow equal formatting and standards. Country names are aligned in the same spelling and geographical standards for Tableau utilization.

Data Manipulation

Following manipulation was done in order to pursue a deeper analysis of the dataset:
Calculate the win-rate for each champion⟶ Based on the numbers of games won, divided by the number of games played from the first three rounds, the win rate for each game can be determinate.

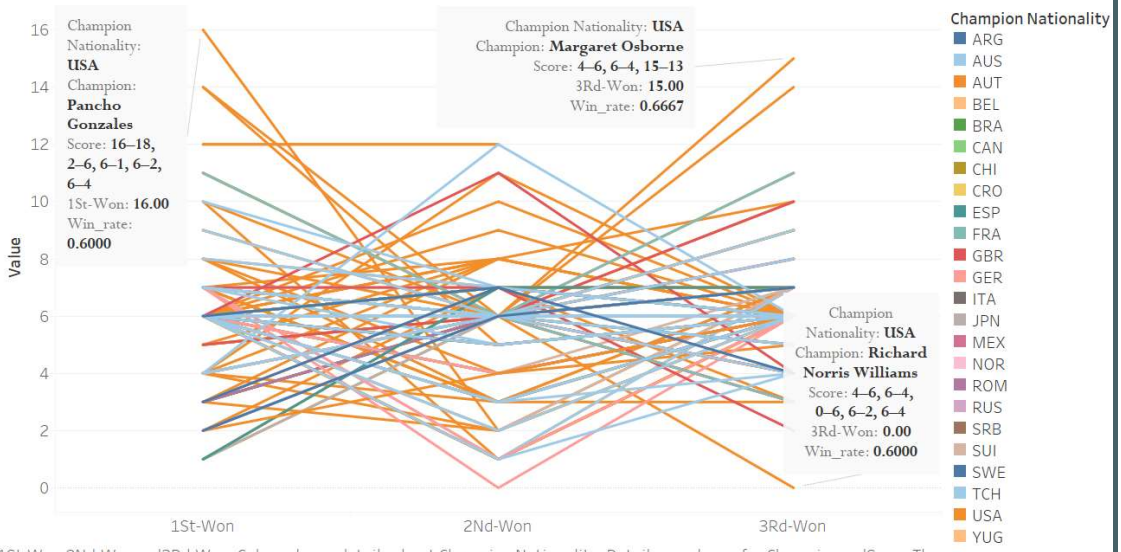
Visualization Method 1. Parallel Coordinate

Task 4. Create a parallel coordinate graph to analyse the relationship between champion, score, win rate, 1st win, 2nd win, and champion's nationality, and write a summary of the parallel coordinate graph telling the story of uncovering the player's performance and highlighting the story you find. I also added the 3rd Win to the graph in order to show evolution for longer games.

Parallel Coordinate are very good visualization to show evolution of different observations over time. With Tableau it is also possible to include an additional dimension related to the nationality country of player.

Parallel Coordinate, US Open games won from 1881 until 2021

Winners name, winners country Score, average win-rate



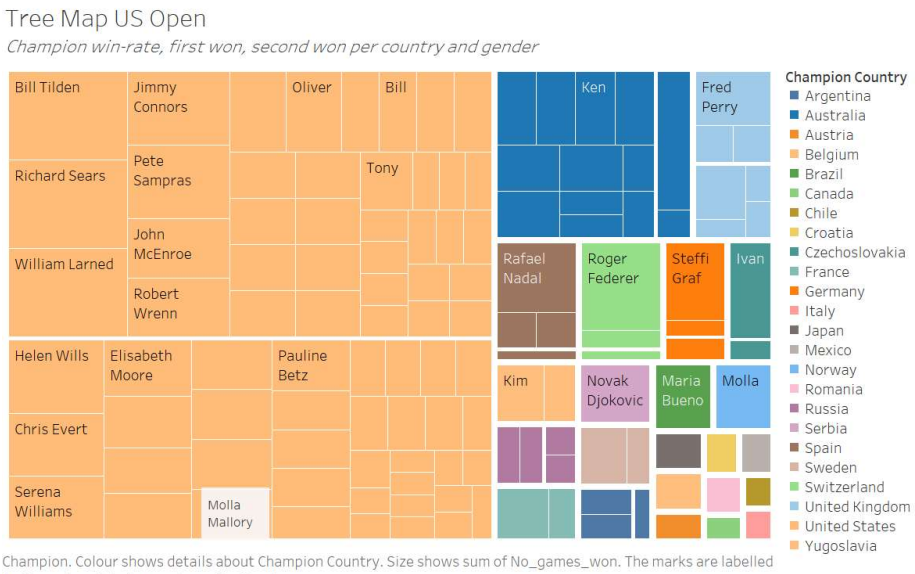
1St-Won,2Nd-Wonand3Rd-Won. Colour shows details about Champion Nationality. Details are shown for ChampionandScore.The view is filtered on Champion, which keeps 128 of 128 members.

Parallel Coordinates show a high number of orange lines, which indicates that most games won are from orange color countries(USA mostly). It also highlights extreme values like the scores shown in the Annotation marks. Values between 3 and 8 have the highest concentration of lines, it highlights the most common scores during the sets won. The overview also indicates that during the second set, results were not characterized by extreme values, with the highest being 12.

Visualisation Method 2. Treemaps:

Task 5. Create treemaps by using different visualisation tools – treemap to analyse players' performance based on the player's nationality, gender, win rate, 1st win and 2nd win. And write a summary of the treemap graph for analysing and uncovering any story you find.

Three maps are good visualizations to show hierarchical structures by using different sizes and colours. It is visually very effective in highlighting more than one dimension (Gender and Country as seen below).

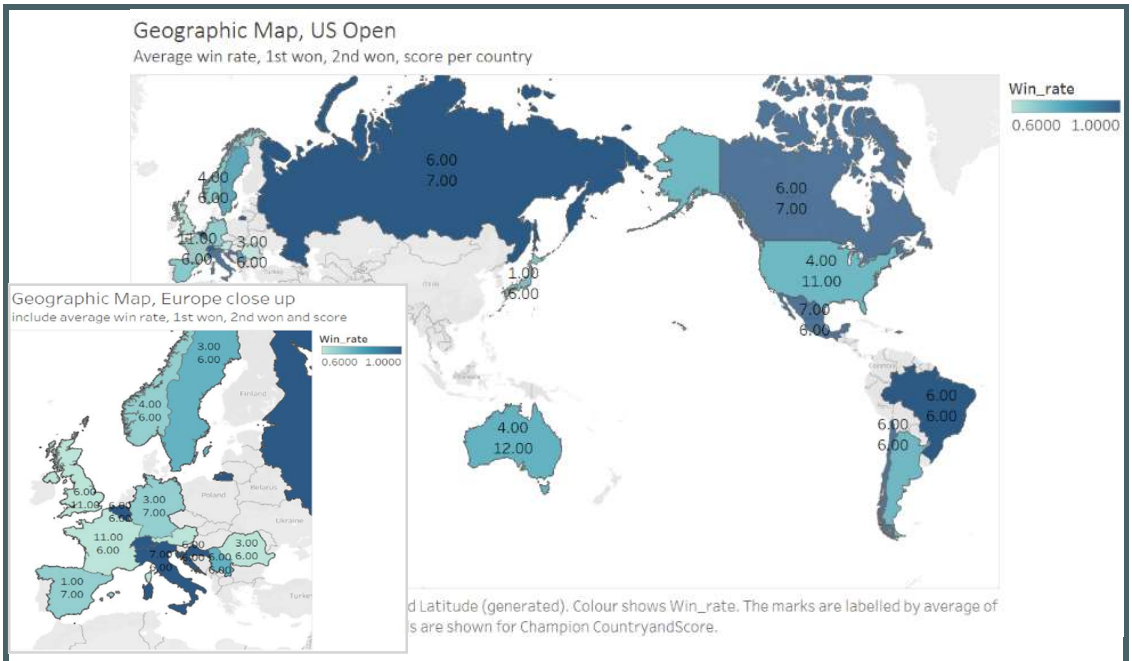


United States is leading country of winners both in male and female categories. Australia is the second highest winner in the male category, while the female category is very well diversified. It is also possible to see that some countries have only one player that won the tournament. Molla Mallory is included in two different countries (USA & Norway) because she changed nationality during her career, as she is the biggest winner of US Open, half her tittles went to USA and the other half to Norway.

Visualisation Method 3. Geographic Map:

Task 6. Convert the player's nationality code to the country's name (done by the lecturer), then create a geographic map combining the champion's nationality, score, win rate, 1st win and 2nd win. And Write a summary of the player's performance based geographic map for analysing the player's nationality and score.

Geographic map is a very effective visualization tool to see spatial data and geographical data in a form that is widely known by the public (Maps). These visualizations can be done at local or global level, provided that the location coordinates (names, latitude and longitude) are available and properly included in the original dataset.



The Map for US Open shows that win rates are higher in countries with lower number of players. Due to its calculation methods, win rate penalizes countries/players that have more sets played without a consistent number of positive results. Because Win rate calculations strongly focus on positive results/wins within the total, countries that have one/few successful player(s) within the tournament will have higher ratings (Switzerland, Italy, Serbia, Brazil, etc.).

Visual Analytics: Champion performance

Task 7. Find the players who have won 5 and more times championships for both men and women, then create visual patterns to analyse the champion's winning rate, which involves runner-up, score, and year.

The Visualization for Task 7 include two different options:

(1) A detailed table with all characteristics required : Players names, runner-up name, winning rate, score, year and gender. A table visualization allows a very detailed overview of the variables chosen, as for example the players names and runner-up names. In a graphical visualization, this level of detail would not be available. Furthermore, I can check the consistency of the output, if the system really considered players who won more than five times.

(2) The second visualization has no metrics available, but it is also as long as the table. The information provided is exactly the same as the table above, though it is more visual appealing than metrics oriented.

Both visualizations have their strengths and weakness, but due to the high amount of information provided on both visualizations, some would prefer the table rather than the colorful bullets (myself).

Tables with metrics usually requires more human recognition effect as all information have the same format and visual weight. Meanwhile, visual effects including colors, sizes and forms are more appealing to the brain and easier to assimilate.

US Open players who have won 5 and more times championships

Gender, Champions name, Champions runner-up, Average win-rate, Champion country

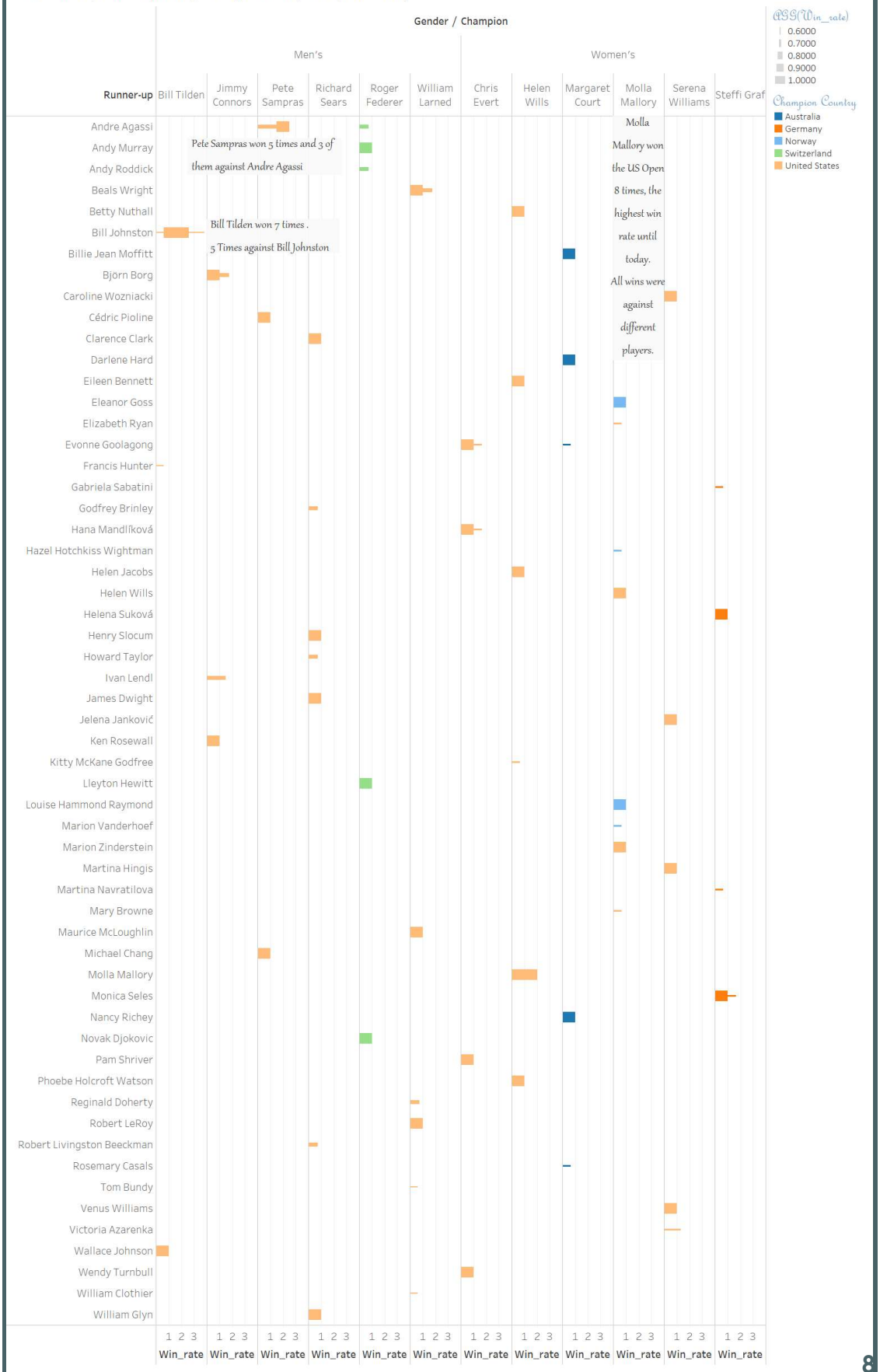
Gender	Champion	Runner-up	Champion Country	Year	Win_rate
Men's	Bill Tilden	Bill Johnston	United States	1920	0.6000
				1922	0.6000
				1925	0.6000
				1923	1.0000
				1924	1.0000
		Francis Hunter	United States	1929	0.6000
		Wallace Johnson	United States	1921	1.0000
	Jimmy Connors	Björn Borg	United States	1976	0.7500
				1978	1.0000
		Ivan Lendl	United States	1982	0.7500
				1983	0.7500
		Ken Rosewall	United States	1974	1.0000
	Pete Sampras	Andre Agassi	United States	1995	0.7500
				2002	0.7500
				1990	1.0000
		Cédric Pioline	United States	1993	1.0000
		Michael Chang	United States	1996	1.0000
	Richard Sears	Clarence Clark	United States	1882	1.0000
		Godfrey Brinley	United States	1885	0.7500
		Henry Slocum	United States	1887	1.0000
		Howard Taylor	United States	1884	0.7500
		James Dwight	United States	1883	1.0000
		Robert Livingston Beeckman	United States	1886	0.7500
		William Glyn	United States	1881	1.0000
	Roger Federer	Andre Agassi	Switzerland	2005	0.7500
		Andy Murray	Switzerland	2008	1.0000
		Andy Roddick	Switzerland	2006	0.7500
		Lleyton Hewitt	Switzerland	2004	1.0000
		Novak Djokovic	Switzerland	2007	1.0000
	William Larned	Beals Wright	United States	1901	0.7500
				1908	1.0000
		Maurice McLoughlin	United States	1911	1.0000
		Reginald Doherty	United States	1902	0.7500
		Robert LeRoy	United States	1907	1.0000
		Tom Bundy	United States	1910	0.6000
		William Clothier	United States	1909	0.6000
Women's	Chris Evert	Evonne Goolagong	United States	1975	0.6667
				1976	1.0000
		Hana Mandlíková	United States	1980	0.6667
				1982	1.0000
		Pam Shriver	United States	1978	1.0000
		Wendy Turnbull	United States	1977	1.0000
	Helen Wills	Betty Nuthall	United States	1927	1.0000
		Eileen Bennett	United States	1931	1.0000
		Helen Jacobs	United States	1928	1.0000
		Kitty McKane Godfree	United States	1925	0.6667
		Molla Mallory	United States	1923	1.0000
				1924	1.0000
		Phoebe Holcroft Watson	United States	1929	1.0000
	Margaret Court	Billie Jean Moffitt	Australia	1965	1.0000
		Darlene Hard	Australia	1962	1.0000
		Evonne Goolagong	Australia	1973	0.6667
		Nancy Richey	Australia	1969	1.0000
		Rosemary Casals	Australia	1970	0.6667
	Molla Mallory	Eleanor Goss	Norway	1918	1.0000
		Elizabeth Ryan	United States	1926	0.6667
		Hazel Hotchkiss Wightman	Norway	1915	0.6667
		Helen Wills	United States	1922	1.0000
		Louise Hammond Raymond	Norway	1916	1.0000
		Marion Vanderhoef	Norway	1917	0.6667
		Marion Zinderstein	United States	1920	1.0000
		Mary Browne	United States	1921	0.6667
	Serena Williams	Caroline Wozniacki	United States	2014	1.0000
		Jelena Janković	United States	2008	1.0000
		Martina Hingis	United States	1999	1.0000
		Venus Williams	United States	2002	1.0000
		Victoria Azarenka	United States	2012	0.6667
				2013	0.6667
	Steffi Graf	Gabriela Sabatini	Germany	1988	0.6667
		Helena Suková	Germany	1993	1.0000
		Martina Navratilova	Germany	1989	0.6667
		Monica Seles	Germany	1995	0.6667
				1996	1.0000

Win_rate: 0.6000 1.0000

Win_rate broken down by Gender,Champion,Runner-up,Champion CountryandYear. Colour shows Win_rate. The marks are labelled by Win_rate. Details are shown for Score.The view is filtered on Champion, which keeps 12 of 128 members.

US Open players with 5 or more championships

Winners name, Runner-up name, Average win rate, Champion country



Win_rate for each Runner-up broken down by GenderandChampion. Colour shows details about Champion Country. Size shows Win_rate. Details are shown for Year,GenderandScore.The view is filtered on Champion, which keeps 12 of 128 members.

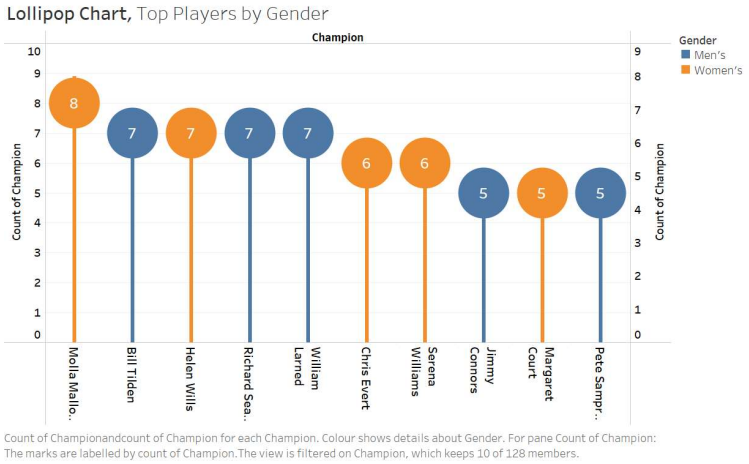
Some interesting outcomes from a detailed overview include, competition between some players were historical, like Andre Agassi and Pete Sampras or Bill Tilden and Bill Johnston, which have played against each other several times. These battles increase audience interest in the tournament and the specific matches (games).

Task 8. Write a summary of the visual patterns for extracting each player's performance and its performance patterns.

US Open top winners (=> 5 time)

Champion	
Molla Mallory	8
William Larned	7
Richard Sears	7
Helen Wills	7
Bill Tilden	7
Serena Williams	6
Chris Evert	6
Steffi Graf	5
Roger Federer	5
Pete Sampras	5
Margaret Court	5
Jimmy Connors	5

Count of Champion broken down by Champion. The view is filtered on Champion, which keeps 12 of 128 members.



The graphs above provide the same information in different ways. The Table on the left provides a list with all Top players including their names and the number of Titles received during their career. The left visualization provides exactly the same information but with a visual appealing that highlights the numbers in the circle, rather than the information itself. As already described in the task above, both visualization have pros and cons.
In this case the Graph on the right would be more effective, since it has low amount of information to assimilate and visual appealing is more effective due to colors, sizes and ordering.

Task 9. Write a report to explain how you dealt with high-dimensional data in your data visualisation, particularly in combining multi-dimensional data. Describe the graphic attribute designs and labelling techniques used in your data visualisation and how they enhanced the readability and storytelling of the visualisation. Highlights any trends and breakthrough analysis you have discovered through the data visualisation process, particularly the top player's performance patterns in the visual comparison chart. Concludes and summarises the advantages of Tableau or other visualisation apps you have used.

To deal with high-dimensional data, first I defined which dimensions (attributes) should be included in the graph. Depending on the characteristics of each dimensions, different graph choices are available. While choosing the numbers of dimensions to use it is

important not to overcrowd the graph with information, because it will loose clarity and focus. Therefore dimensions should be categorized by the priority level you want or what is the message you want to convey with the graph. Once priorities are settled, it is possible to define which measures goes to which graph location (axis).

Trends and individual incidents should not be graphically available as they strongly increase granularity, these should be included in Highlights, while hovering the mouse above these variables.
Tableau allows a very detailed overview of the incidents in large data sets while summarizing the main characteristics in a variety of graphical possibilities. Additionally it has the option to include up to four dimensions in a graph, without much coding effort. It is easy to use and learn, providing accessibility to beginners and non-specialist in visualization.