DWBI Final Project Report

Total Team Members – 1 (Rahul Ohri)

Topic - Analysis of the ATP world tour tennis dataset

1. Describe the scenario. What is the problem you are trying to solve?

The ATP (Association of Tennis Professionals) is the governing body of the men's professional tennis circuits - the ATP Tour, the ATP Challenger Tour and the ATP Champions Tour. With 64 tournaments in 31 countries, the ATP Tour showcases the finest male athletes competing in the world's most exciting venues. From Australia to Europe and the Americas to Asia, the stars of the 2020 ATP Tour will battle for prestigious titles and FedEx ATP Rankings points at ATP Tour Masters 1000, 500 and 250 events, as well as Grand Slams (non-ATP events).

The aim of my final project is to deep dive into the ATP world tour dataset and find out the correlation between a few key data points that are being captured. Being an avid Tennis fan, I have always had a few key questions which I wanted to explore and find more information on. Today, with ATP maintaining up to date match stats on their website, obtaining such information is very easy, however finding trends between the same requires some additional effort. My end goal is that using my Tableau dashboard, instead of referring to the ATP website and referring multiple pages to find data, there is a single place where one can refer to for the statistics.

- **Who is the intended audience?**

  The intended audience for this tableau dashboard is any tennis fanatic. This dashboard captures some basic metrics that are used during games which catches the eye of the viewers as well as the audience like beak point conversion or the number of aces served as well as gives a very high level overview of statistics that showcase the performance of each individual players and specially the Big 4 players in tennis history i.e. Nadal, Djokovic, Federer and Murray.

- **Where did your data come from?**

  I will be referring to 2 datasets which will contain information about various metrics pertaining to ATP tournaments, match scores, match stats, rankings and players overview . For this project , I will not be utilizing the rankings data since that is not a part of my analysis. There are roughly 7-10 csv files (Around 60MB in total) which I will be using for my analysis.

References for the dataset

- https://datahub.io/sports-data/atp-world-tour-tennis-data#pandas
- https://github.com/serve-and-volley/atp-world-tour-tennis-data
- https://www.atptour.com/en/stats
- https://tt.tennis-warehouse.com/index.php?threads/ranking-of-players-by-height-and-aces-match.605996/

2. Describe the steps taken in the project

1. I had decided that I want to work on a Tableau dashboard during the course itself and thus the first step for me was finalizing on a topic and understanding what my project end goal is and this is where the objectives played a crucial role . My end deliverable can be used for analyzing questions like
    - Do taller players have a better chance of serving aces based on historic data?
    - How many double faults on an average are made by the losing player in a game?
    - What percentage of break points are converted successfully by the winner of a match?
    - Roger Federer has won the maximum number of Wimbledon Titles. Is this because he has played more tournaments on grass surfaces? Similarly for Nadal with respect to the French Open. Is there a correlation between the number of tournaments won vs the number of grand slams being won on the same surface?
    - For the top 4 players, what is the win percentage in ATP events vs grand slams? Do players perform better in best of 3 or best of 5 set matches?
    - Does fatigue affect the win percentage of players? i.e. does a longer match time generally affect the overall performance of a player?

2. Once that was streamlined, focusing on a dataset was not very difficult in my case since the ATP website themselves maintain all the data. However finding a repository where all the datasets have been preprocessed was a blessing in disguise as it saved me hours of work. I finalized on 2 datasets and then proceeded to the next step which was loading the data in tableau

3. Since a lot of my datafiles had the same structure (**e.g. match scores and match stats**) but were spread across multiple years, I had the opportunity to explore the Tableau feature of data **union** which made data loading very easy for me.

4. Another great learning for me was making use of the **data blending** . I had to make use of a lot of data that was spread across other data sources and this feature helped me generate many visualizations and was crucial for my analysis. Understanding how to link the field for data blending was also very intriguing for me.

5. Finally I was able to create 2 dashboards, one which gives stats about players in general, and the other one that focuses on the Big 4 players.

6. To improve the performance of data loading, I made use of Tableau Extracts, which significantly boosted the data loading for me. I also understand that this feature is of great importance in the case of big data.

3. Describe the discoveries (what did you learn from the data, etc.)

1. While creating the dataset for player height and their corresponding aces, I had to copy data from the web. This is when I came across excel' s feature of copying data directly from web into tables. It saved so much manual effort for me and I found this very similar to data web scraping which we do in python using packages like BeautifulSoup . I was really taken aback by the accuracy of the data being scraped.

2. In the **Breakpoint conversion percentage by Year** chart it is very interesting to see that the despite there being a significant increase in the number of matches being played post 2006, the average breakpoint conversion percentage remains the same as before between the range of 55.5 % - 56 %.

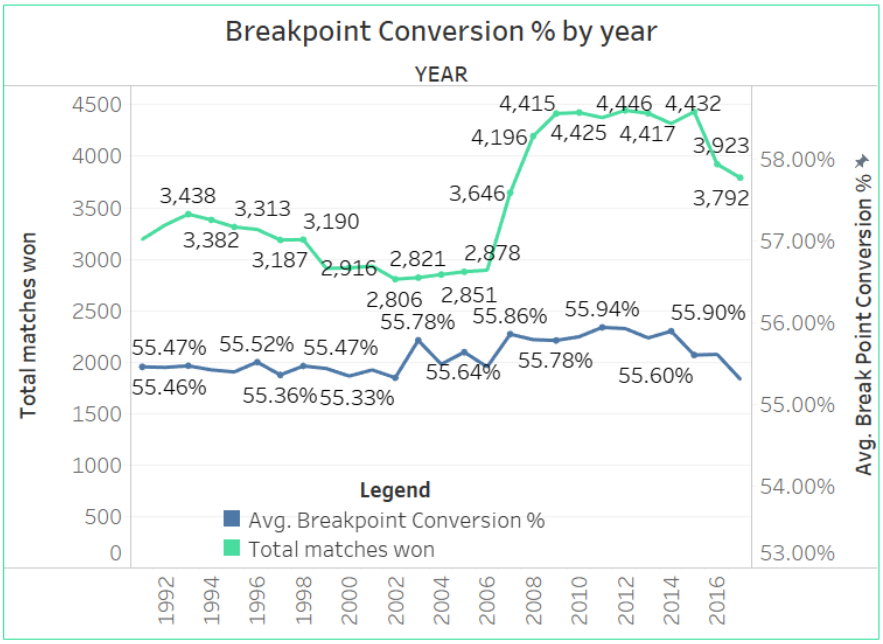

Figure 1

3. For the chart which highlights the **Percentage of singles tournaments won by countries since 1877,** I was really taken aback by two findings – the first one was that USA has produced the maximum number of tennis players on an average which is almost 2 times their next competitor which is Australia and Great Britain. The second was that Spain has produced only

272 players, but they have won almost 1.5 times the number of tournaments which means that a single player on an average has won at least 1 tournament which is commendable. An e.g. can be Nadal winning 10 Roland Garros titles himself.

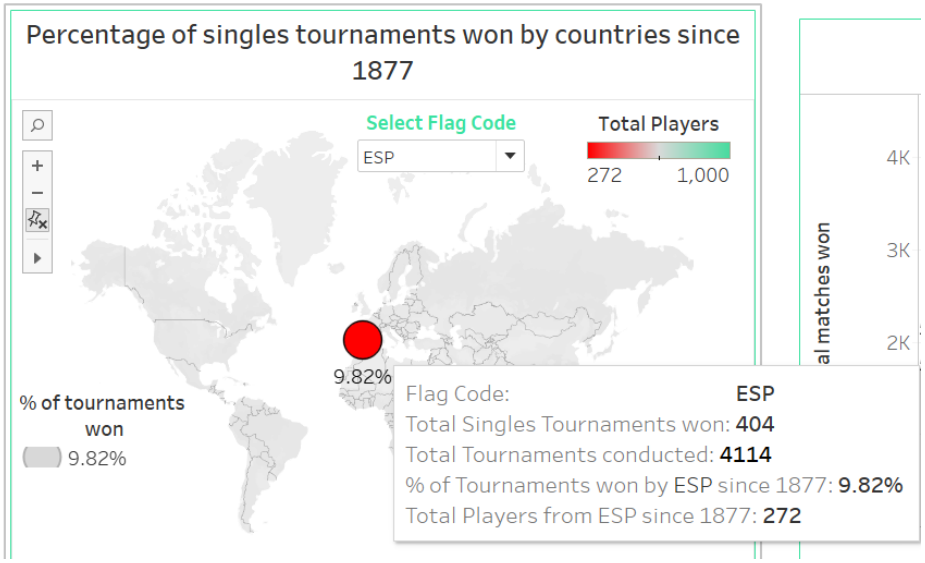

Figure 2

4. Moving on to the Big 4 analysis dashboard **(Figure 7)**

   - In the **Analysis of Player win % in Grandslams vs ATP events** chart, it was interesting to see that all the BIG 4 performed much better in Grandslam events as compared to ATP events . One big reason for the higher win % could be because in a year there are only 4 Grandslams whereas on an average they play around 15-20 ATP events.

   - In the **Analysis of the win % by surface in Grandslams & ATP events** chart, we see that each player has their own strength relative to the surface on which they play. All players except Nadal perform very well on a Hard surface whereas as Nadal who is know as the King of Clay for a reason, has outperformed everyone on the clay courts

   - The bottom two graphs show how fatigue affects the overall performance of the Big 4 and we see that out of the matches lost, all of them player longer matches well over 2 hours as compared to the times when they won matches. One reason for the longer match time I feel is that most of them go to the last deciding set before losing the matches (e.g. For a win you need 2/3 or 3/5) and hence the overall match duration increases. However an alternative reason could be that they might not be going into the decider but might be losing the other sets at a tiebreak which accounts for a longer match duration as well.

4. Describe any challenges encountered and how you resolved the challenge

Throughout the project I did come across a different set of challenges which included data transformation activities and making use of additional datasets from the web

1. One of the major adjustments that I had to make was that from the web I had to download additional datasets. E.g. For the visualization that shows a correlation between player heights and aces, I had to manually prepare an additional excel file. The file gives information about how tall the players are, the total number of matches they have played, the number of aces they have per match and the total number of aces they have served till date.

2. For the visualization which shows the **percentage of singles tournaments won by countries since 1877**, I had to make use of the internet to refer to country flag codes. The player details dataset has the flag code column; however tableau could not recognize a lot of countries because the format was in a format that was used specifically for Olympic games . The dataset had reference to the International Olympic Committee (IOC) which uses three-letter abbreviation country codes not recognized by Tableau. So I had to manually enter the country names against these codes while creating a geographic datatype for country.

3. For the visualization title **Analysis of Player win % in Grandslams vs ATP events,** I had to make a lot of calculated fields to ensure that my calculation and analysis is being portrayed correctly and accurately. I had to make used of the match scores dataset but still had to create many additional fields like **UNIQUE PLAYER ID** which was needed to identify who is the player I am identifying in my analysis. To get a count of the total number of matches played by a player, another filed titled, **Total Matches Played** was created which had the summation of the number of matches won and lost by a player which was not directly mentioned in the dataset**.** This was followed with a calculated field that talks about the **No of matches** won by a player and finally the overall **win % .**

   If one sees carefully , since my analysis was only for the Big 4 of tennis, it was easy for me to manually add their player ids in the calculated fields, but for 200 players, this would be impossible for me and cause a lot of manual labor.

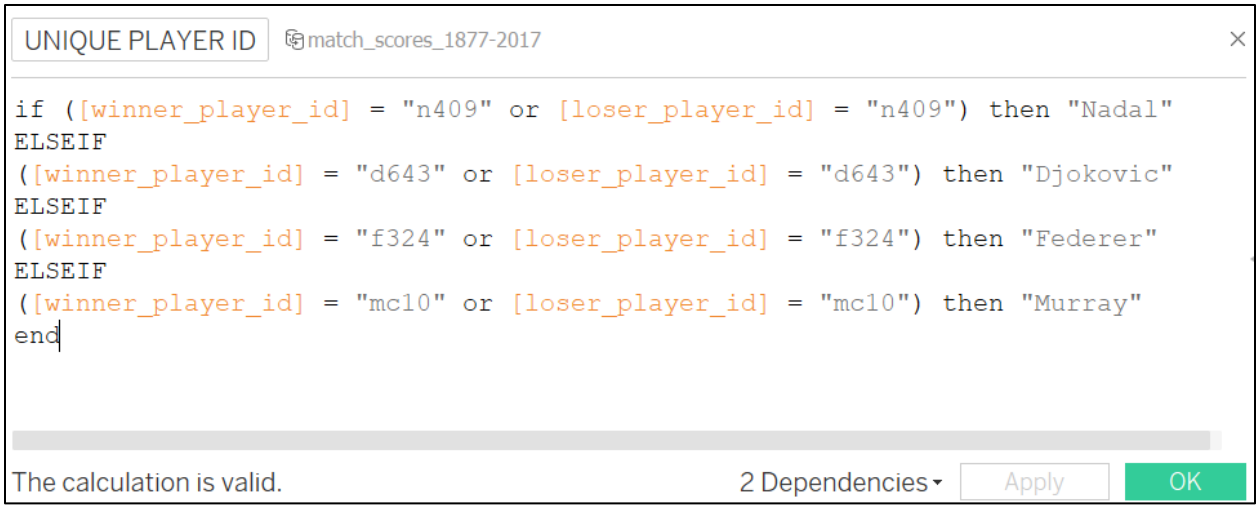

Figure 3

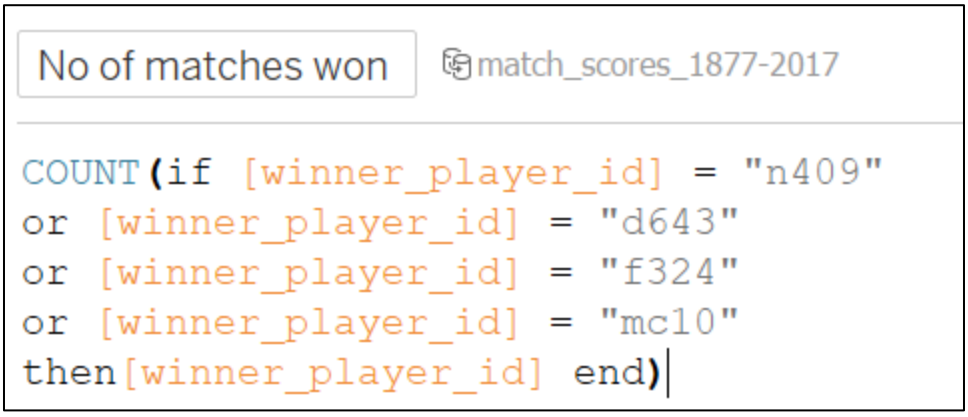

Figure 4

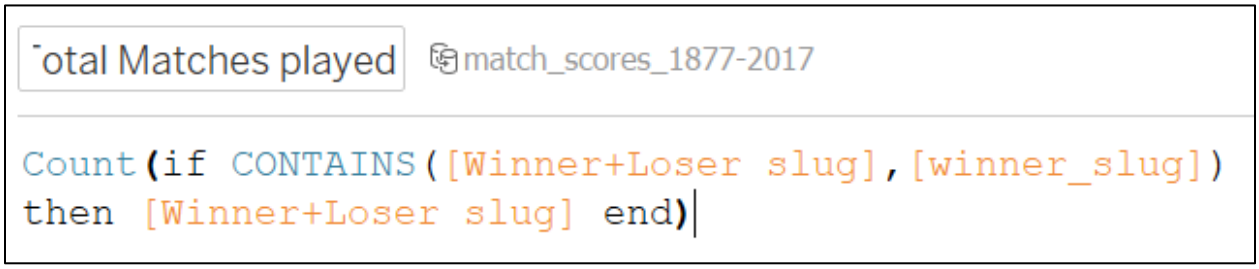

Figure 5

4. I was making use of multiple data files for my analysis and it was very important to come up with a primary key for linking the data. **Match Id** and **Tournament ID** were the important fields in this scenario and the problem was that not all the csv files had the same format of reporting the these. Some had a format of xxx-xxx-xxxx while another one had xxx-xxx.

Standardizing the same was a big task and I had to use tableau's formulae capabilities to match the primary keys.

5. One big challenge for me was that I felt I was restricted to use many creative chart types. Most of my analysis were best represented using bar charts. Other charts seemed do misrepresent a lot of information being conveyed. To get around the monotonous views, I tried making use of some geography maps and dual axis charts having a bar and a line. For one correlation, the scatter plot seemed to be doing justice and thus I made use of that as well.

Dashboard Screenshots

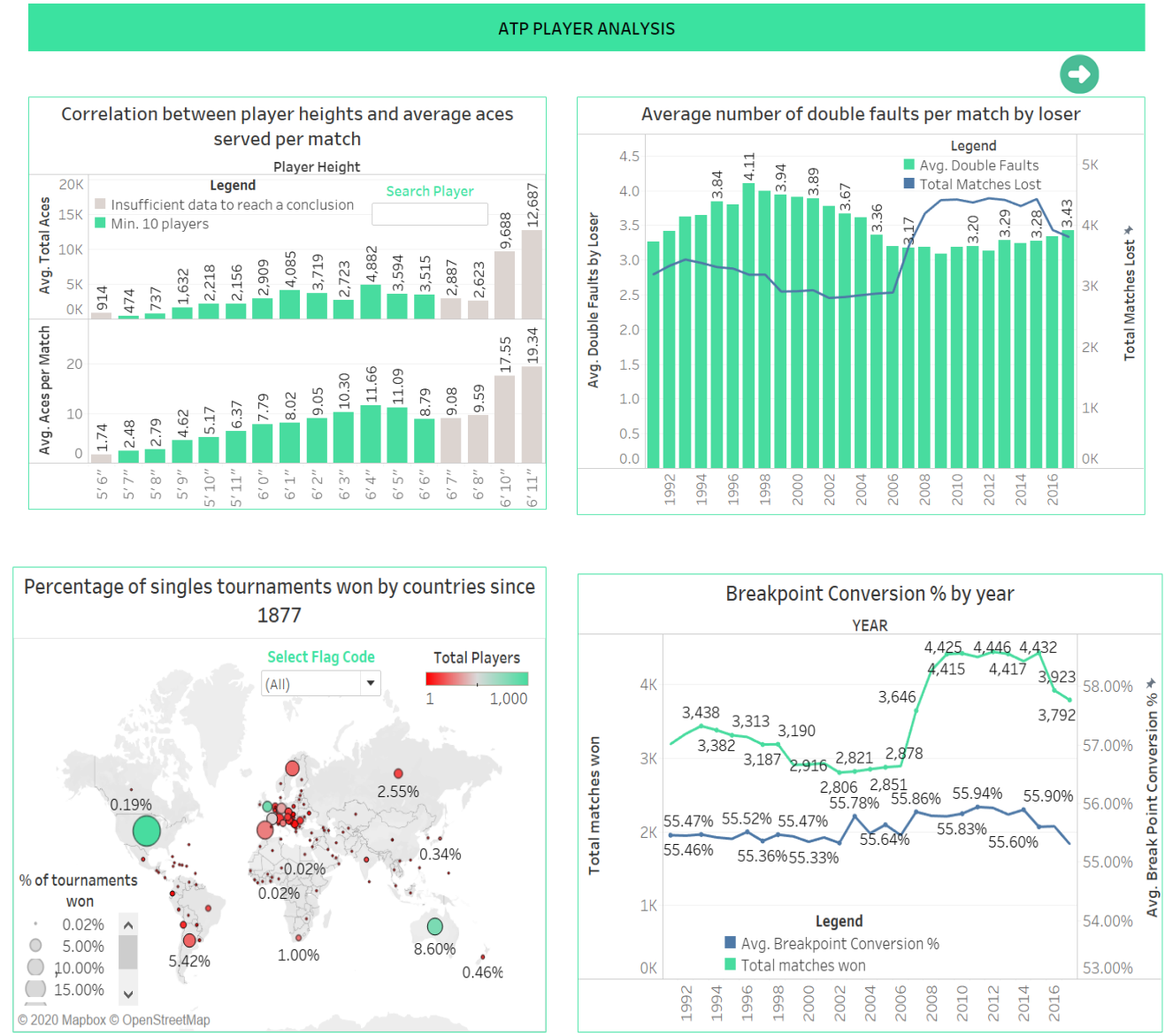

Figure 6: Player Analysis Dashboard

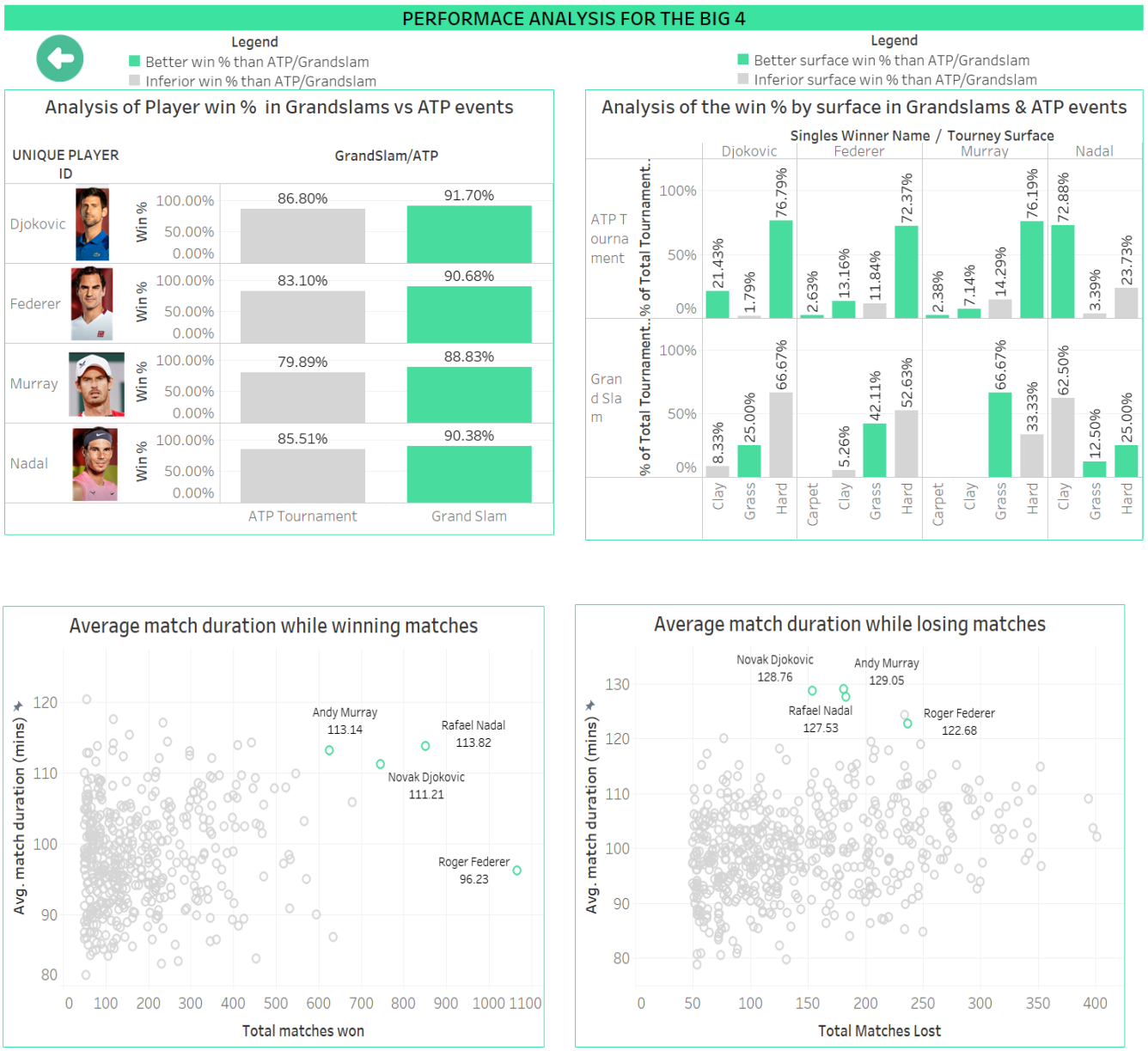

Figure 7: BIG 4 Analysis Dashboard

References for images used in the dashboard

- https://www.atptour.com/en/players/andy-murray/mc10/overview
- https://www.atptour.com/en/players/novak-djokovic/d643/overview
- https://www.atptour.com/en/players/rafael-nadal/n409/overview
- https://www.atptour.com/en/players/roger-federer/f324/overview

- https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.pinterest.com%2Fpin%2F391250286363285006%2F&psig=AOvVaw0Gb7572luhbmji1rzSb_BD&ust=1608144332983000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCJDjn9PS0O0CFQAAAAAdAAAAABAJ
- https://www.google.com/url?sa=i&url=http%3A%2F%2Fwww.icons101.com%2Ficon%2Fid_67317%2Fsetid_2086%2FFlat_Round_by_Roundicons%2FArrowRight&psig=AOvVaw3fNgD-fynl4gkUl-WanguH&ust=1608144365099000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCMCZteDS0O0CFQAAAAAdAAAAABAJ