ATP & WTA: analysis of tennis players by performance, sponsors, and social interaction

Marta Bernardelli 851178 Giulia Beccaria 851286 Alessia Perri 851374

Introduction

The idea for this project stems from the complaint of Romanian tennis player Sorana Cirstea where she points out how the world of women's tennis and the sponsors who invest in this environment, are tied to logics that go beyond sports performance and exclude meritocracy. Cirstea has made public some phrases allegedly uttered to her about the greater importance of her physical appearance over on-court performance by one of her former sponsors, specifically the clothing giant Adidas. In the past, many female tennis players have pointed out how it is an environment in which marketing needs often trump sporting merit and competition.
The goal of our work is to try to quantitatively test this phenomenon and see if there are any differences with men's tennis.
To do this we will combine the following information:

- sports performance data, such as the number of tournaments won, prize money and winning percentage
- data on sponsors: which brands collaborate with the tennis players?
- Google trends data, what is the most searched information related to a tennis player? Are they all tennis-related?
- number of followers and posts on instagram to get an estimate of the athlete's social engagement

We will extract these data for both males and females and combine them to see if matching them can show significant patterns in describing the phenomenon in question. The final goal of our work will be to build an infographic on Tableau that summarizes well what the data showed.

Data acquisition

Sponsor data

Data research related to sponsors was the starting point of our work. We started from the collection provided by the Toomanyrackets website where there is information regarding sponsorships for both the ATP and WTA players. Specifically, the following information can be found in the tables on the site:

- Name of the player
- Apparel sponsor

- Shoes sponsor
- Racket sponsor
- Other Sponsors

We performed the extraction of this data with the web scraping technique and ran it on Python thanks to the BeautifulSoup library. Using this technique, we extracted all the sponsor data of all the tennis players on the platform i.e. 103 male tennis players and 72 female tennis players. These will be the tennis players we will consider for our analysis, in fact we are going to extract from the other sources the information about the tennis players whose sponsors we know. This is what we obtain:

Name	Apparel	Shoes	Racket	OtherSponsor
Ajla Tomljanovic	Original Penguin	K-Swiss	Wilson	–
Alison Riske	K-Swiss	K-Swiss	Wilson	–
Alison Van Uytvanck	Fila	Fila	Head	Hubo, Signum Pro, AB Automotive, Wcup
Alize Cornet	Lotto	Lotto	Babolat	–
Anastasia Pavlyuchenkova	Lacoste	–	Wilson	–
...	...	...	...	...
Thanasi Kokkinakis	–	–	Yonex	–
Tim Van Rijthoven	Indian Maharadja	–	Yonex	–
Tommy Paul	New Balance	New Balance	Wilson	–
Ugo Humbert	Lacoste	Lacoste	Wilson	–
Vasek Pospisil	Ionik	Asics	Wilson	–

Performance data

As for the sports performance data, we decided to extract this data using the web scraping technique from the TennisExplorer website. It is one of the most comprehensive sources of data regarding tennis, in fact it contains for each tennis player a dedicated page. We selected the players of interest, and for each of them we extracted both biographical information and performance data:

- Name of the player
- Country of origin
- Height (cm)
- Weight (kg)
- Age
- Date of birth
- Current ranking (the current position in the ATP/WTA rankings, if the tennis player is no longer active it has null value)
- Higher ranking (the highest position the player has reached in the WTA/ATP rankings)
- Sex (woman or man)
- Plays (left if it is left-handed or right if it is not)
- Number of matches won

- Number of matches lost
- Titles won (number of tournaments won in singles)

We decided to extract only data related to singles tennis and not doubles tennis. We also performed web scraping for this data source with the help of Python's BeautifulSoap library. The result obtained is as follows:

	Name	Country	Height	Weight	Age	Birthday	Current rank	Higher rank	sex	plays	match win	match lose	titles
0	Tomljanovic Ajla - profile	Australia	180	67	29	(7. 5. 1993)	s: 35.	/ 33.	woman	right	\n387/	292\n	\nSummary:\n1/3\n-\n0/2\n-\n1/1\n-\n
1	Riske Alison - profile	USA	175	66	32	(3. 7. 1990)	s: 49.	/ 18.	woman	right	\n420/	341\n	\nTotal:\n3\n9\n
2	Van Uytvanck Alison - profile	Belgium	173	63	28	(26. 3. 1994)	s: 68.	/ 37.	woman	right	\n385/	260\n	\nTotal:\n7\n14\n
3	Cornet Alize - profile	France	173	60	33	(22. 1. 1990)	s: 34.	/ 11.	woman	right	\n555/	438\n	\nTotal:\n6\n3\n
4	Pavlyuchenkova Anastasia - profile	Russia	177	72	31	(3. 7. 1991)	s: 364	/ 11.	woman	right	\n465/	323\n	\nTotal:\n12\n5\n
...	...	...	...	...	...	...	...	...	...	...	...	...	...
99	Kokkinakis Thanasi - profile	Australia	193	84	26	(10. 4. 1996)	s: 159	/ 69.	man	right	\n251/	158\n	\nTotal:\n1\n5\n
100	Van Rijthoven Tim - profile	Netherlands	188	89	25	(24. 4. 1997)	s: 105	/ 101	man	right	\n239/	142\n	\nTotal:\n1\n9\n
101	Paul Tommy - profile	USA	185	82	25	(17. 5. 1997)	s: 35.	/ 28.	man	right	\n349/	213\n	\nTotal:\n1\n10\n
102	Humbert Ugo - profile	France	188	73	24	(26. 6. 1998)	s: 106	/ 25.	man	left	\n260/	173\n	\nTotal:\n3\n11\n
103	Pospisil Vasek - profile	Canada	193	88	32	(23. 6. 1990)	s: 99.	/ 25.	man	right	\n527/	372\n	\nTotal:\n0\n21\n

The performance data are up to date with the last tournament played, the Australian Open 2023.
Regarding sports performance, we also decided to extract data on career prize money, which is the total economic rewards that players received as a result of victories during the course of their careers. Regarding this data, the ATP and WTA sites periodically release a pdf file that you can download for free. For our work, we used the latest available version i.e. the one dated 2/01/2023.

Instagram data

Since not all tennis players have twitter accounts, in order to get information about their exposure on social media we decided to extract data about them from Instagram. Instagram was also chosen because of the numerous marketing campaigns that major tennis apparel sponsors carry out on the platform.
The main information of interest to us was as follows:

- number of posts on profile
- number of followers

To do this, however, it was essential to derive the nickname of the player's instagram account, as many of the times it cannot be traced back to it from the name. A good amount of nicknames could be retrieved from the database of Functional Tennis website and the rest of the nicknames we searched independently.

Once all the nicknames were found, to extract the information of interest we exploited Python's instaloader library, obtaining the following result:

simonahalep	302	1741993
sloanestephens	472	573014
sofia.kenin	349	205864
soranacirstea	663	159080
tamarazidansek997	146	13444
vera.zvonareva	164	14821
v.kudermetova	484	85184
vichka35	1199	699366
alexeipopyrin	216	48148
carlitosalcarazz	504	2139840
casperruud	392	429421
jenson_brooksby	126	36928
taylor_fritz	328	323910
adrianmannarino	218	20932
albertramosvinolas	302	25100
aledavidovichofficial	186	309034
alexdeminaur	440	259270
bublik	39	153090
alexzverev123	679	1832797
aljazbedene	303	16312
andyseppio	524	102400
andreyrublev	285	369758
andymurray	432	1910558
arthurrinder	198	18811
karatsev93	59	60673
benshelton	70	42537
benbonzi	180	46600
benpaire	1575	396581
bernard_tomic	189	4713
bornacoric	669	258638

where each instagram account searched is paired with the number of posts made and the number of users following the account.

Google Trends data

To see what words are most often used to describe players and the words most often associated with their names, we decided to use data from Google Trends. We preferred this data source to Twitter for the following two reasons:

- The tweets contain only the opinions of a limited part of the population, both in terms of age but also in terms of the fact that only people registered on this social network can express their opinions.
- The twitter api with elevated student access does not allow us to extract the full history of a given query (i.e., tweets written since 2006) but only a limited number of tweets written in the week prior to the extraction. This would have made little sense for our work and would have been very limiting in terms of being able to answer our research questions.

We then decided to extract with the help of the python Pytrends library, the related queries for each player considered with the following parameters:

- english as preferred language to extract the queries
- select the most frequent searches over the past 5 years

The information about each player is thus represented as a dictionary of pandas.DataFrames containing this two type of queries:

- **Top:** The most popular search queries. Scoring is on a relative scale where a value of 100 is the most commonly searched query, 50 is a query searched half as often, and a value of 0 is a query searched for less than 1% as often as the most popular query.
- **Rising:** Queries with the biggest increase in search frequency since the last time period. Results marked “Breakout” had a tremendous increase, probably because these queries are new and had few (if any) prior searches.

Here we can see the results of the extraction for a single player:

```
24                    hubert hurkacz   89950}}{'matteo
berrettini': {'top':                                    query
value
0                  berrettini tennis    100
1                             tennis     93
2            tennis matteo berrettini     92
3                                atp     68
4                   berrettini atp     67
5                         wimbledon     67
6          matteo berrettini wimbledon     64
7                matteo berrettini atp     60
8                  berrettini djokovic     58
9                             djokovic     56
10                              sinner     48
11                               nadal     38
12                            us open     37
13          matteo berrettini us open     37
14                     novak djokovic     27
15                      jannik sinner     26
16                            federer     26
17                         tomljanovic     26
18                            fognini     24
19  matteo berrettini ajla tomljanovic     23
20                    ajla tomljanovic     23
21               berrettini fidanzata     22
22         fidanzata matteo berrettini     22
23                   australian open     21
24                   berrettini oggi     21, 'rising':
```

```
24                    berrettini oggi     21, 'rising':
query   value
0                            wimbledon  406800
1                   berrettini djokovic  354350
2                              djokovic  345050
3                                sinner  292800
4                                 nadal  234900
5                              us open  227000
6                         jannik sinner  160600
7                               federer  158650
8                            tomljanovic  156150
9   matteo berrettini ajla tomljanovic  137750
10                     ajla tomljanovic  137550
11                  berrettini fidanzata  132150
12          fidanzata matteo berrettini  131900
13                       berrettini oggi  126800
14                          rafael nadal  120300
15                             tsitsipas  109600
16                                sonego  108850
17                        wimbledon 2021  106500
18                               musetti  106350
19                         roger federer   97800
20                               monfils   92200
21                               alcaraz   87150
22                                  ruud   82300
23                               hurkacz   77000
24                     alexander zverev   74250}}{'mikael
```

Google trends data reworking

To answer the main search question and interpret the results of the queries, we reprocessed the extracted results so that we could judge how much the words associated with the tennis player were related to tennis or other. To do this, we took all the words in the extracted queries and we divided them by player, thus obtaining a dictionary where the key was the player's name and the related value was a list containing all related words. Then we performed the following operations:

- Normalized the string by removing all accents and special signs
- We removed the player's name among the words related to him. For example, in the case of the list related to the key "matteo berrettini," we removed the words "matteo" and "berrettini" from the list. We did this for each player, because it would have been obvious and not very useful information
- Within each player-specific list we removed duplicate
- We have generally removed all English dictionary stopwords i.e., all words that do not give meaning such as "the", "and", "is" ect. To these types of words we have added numbers and years
- Translated some words into English

Putting everything into a pandas dataframe we got the following result.

name	sex	index	queries
ajla tomljanovic	woman	0	berrettini
ajla tomljanovic	woman	1	matteo
ajla tomljanovic	woman	2	kyrgios
ajla tomljanovic	woman	3	tennis
ajla tomljanovic	woman	4	nick
...	...	...	...
vasek pospisil	man	44	sofia
vasek pospisil	man	45	mackenzie
vasek pospisil	man	46	mcdonald
vasek pospisil	man	47	matteo
vasek pospisil	man	48	stefanos

At this point we constructed three bags of words in the following way:

- **ATP references:** This bag of words contains all the names of male tennis players in the dataset plus other references to male tennis that we found by exploring the words. It contains the names of ATP players
- **WTA references:** This bag of words contains all the names of female tennis players in the dataset plus other references to female tennis that we found by exploring the words. It contains the names of WTA players
- **Tennis references:** This bag of words contains words that refer to tennis generically such as "wimbledon," "tennis," "match," "atp," "wta," "rankings" etc.

In the bag of words creation phase, domain knowledge was crucial.
Then each word in the previous dataset was analyzed to see if it belonged to one of the three groups. If the word did not belong to any of the three groups, then it was classified as "others". In the following database pandas we can see the results of this operation:

	name	sex	index	queries	type
0	ajla tomljanovic	woman	0	berrettini	atpref
1	ajla tomljanovic	woman	1	matteo	atpref
2	ajla tomljanovic	woman	2	kyrgios	atpref
3	ajla tomljanovic	woman	3	tennis	tennisref
4	ajla tomljanovic	woman	4	nick	atpref
...	...	...	...	...	...
7225	vasek pospisil	man	44	sofia	wtaref
7226	vasek pospisil	man	45	mackenzie	atpref
7227	vasek pospisil	man	46	mcdonald	atpref
7228	vasek pospisil	man	47	matteo	atpref
7229	vasek pospisil	man	48	stefanos	atpref

In the analysis of the “other” words, we noticed some cases of contamination within related searches. To better understand what we are referring to, let us consider the following example:

	Name	Type	query
4950	dustin brown	top	bobby brown
4951	dustin brown	top	millie brown
4952	dustin brown	top	dustin brown tennis
4953	dustin brown	top	tennis
4954	dustin brown	top	brown tennis
4955	dustin brown	top	millie bobby brown
4956	dustin brown	top	stranger things dustin
4957	dustin brown	top	dustin stranger
4958	dustin brown	top	stranger things
4959	dustin brown	top	millie bobby brown stranger things
4960	dustin brown	top	nadal
4961	dustin brown	top	nadal dustin brown
4962	dustin brown	top	dustin brown atp
4963	dustin brown	top	atp

These are the words related to the tennis player Dustin Brown and we can see that it includes both tennis-related searches but also words referring to the world of the TV series Stranger Things.
The tennis player is in no way related to the TV series, but this contamination occurs due to the homonymy of the tennis player with a character from the TV series. These words are wrongly classified as 'other' because the tennis player is not really being talked about. For this reason, in order to improve the quality of our work, we have removed all these words so as not to have any kind of interference.

In addition to saving the data in the database format of pandas, we also decided to organize them in the form of json files. The latter is thus organized as follows:

```
diz["rafael nadal"]
```

```
{'atpref': ['federer',
  'djokovic',
  'roger',
  'novak',
  'carlos',
  'alcaraz',
  'berrettini',
  'medvedev',
  'tsitsipas',
  'daniil',
  'stefanos',
  'matteo',
  'jannik',
  'sinner'],
 'tennisref': ['tennis',
  'vs',
  'wimbledon',
  'roland',
  'garros',
  'atp',
  'french',
  'open',
  'australian',
  'us'],
 'other': ['news']}
```

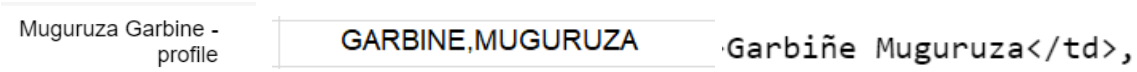
A collection of documents has been created where each one is dedicated to a player. On the left, we can see an example of the document on Rafael Nadal that we analyzed using Python. For each player, you can select the list of words associated with him divided by category. One can appreciate how, since there are no WTA references associated with Nadal, there is no entry for "wtaref".

We think this model is most useful when the focus of the work is on the individual player and the words associated with them. The dataframe pandas allows may be more intuitive when one wants to analyze the frequency of a word within a group of players.

Data quality and conflicts between data sources

In this part of the report, we will show all the conflicts between the different data sources we encountered during the integration phase. We will also illustrate the different conflict resolutions we decided to apply. The solutions chosen were made in relation to our research questions and with the aim of increasing the quality of our database.

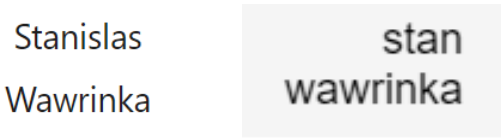
Name format problem



One of the main problems during the data integration phase is due to the different format of the names. Finding a single format for names was crucial both for the integration phase but also to have a good level of consistency. The picture shown above summarizes the main problems we tried to unify in creating a common name format for all three sources . We acted with the following directions:

- **The use of upper and lower case letters:** To overcome this problem, we decided to keep the name in lowercase in all extracted data sources. Also with regard to Google Trends data, queries related to the lowercase spelled name were searched.
- **First and Last Name vs. Last and First Name:** Since only the performance data had the form "Last Name and First Name" we decided to change it back to "First Name and Last Name" by building a special function on Python. We were particularly attentive to cases where the tennis player had 2 first names or 2 last names, even intervening individually.
- **special graphic signs such as the - or ~:** we decided to remove any particular graphic mark.

The final name format in the photo in light of these choices is therefore: garbine muguruza. In standardizing the names, we also encountered some problems from a semantic point of view like the one in this example:



These names refer to the same player but one represents the real name and the other an abbreviation of the same. In these cases we opted for the name format that

seemed to us to be the most commonly used to name the tennis player. The main reason for this choice is that the final name format was to be used to extract Google Trends data, and therefore an underutilized name format could have distorted the analysis by having few related queries found. This criterion was also applied in the case of some second names that could forge the analysis.

Missing values related to sponsor and completeness in general

As noted when collecting sponsor data, there are some missing data that are marked with a '-' symbol. We first changed this symbol to use a more suitable and recognisable one during the analysis phase. Then, we tried to quantify the presence of missing data by attribute and obtained the following result:

Attribute	Percentage of missing value
Name	0%
Apparel	7,42%
Shoes	11,42%
Racket	0%
Other Sponsor	81%

We can see that the majority of missing value comes from the attribute “Other Sponsor”. From the missing symbol, there are two assumptions one can make:

- the sponsor data was not collected for that particular tennis player
- the tennis player does not have a sponsor for that particular attribute

Aiming both to disambiguate this issue but also to increase the quality of our dataset, we tried to impute the missing data by obtaining additional information on it. To do this, we relied on information found on the internet and our knowledge of the domain. After this operation, the dataset still contains missing data, but in this case the missing data corresponds to cases where no information was found and therefore it is assumed that the athlete in the case has no sponsor for that particular characteristic. These are the percentage of missing values after our work:

Attribute	Percentage of missing value
Name	0%
Apparel	0%
Shoes	1,7%

Racket	0%
Other Sponsor	32%

We can therefore see how the percentages of missing values have been reduced considerably. The data related to other sponsors is still high but we think it is still reasonable to assume that not all athletes are linked to sponsors outside tennis.

Regarding the level of the object completeness, we are aware that the database we are building contains only a small part of all professional tennis players. Our work, however, was limited to considering those of which we had the information about the sponsors. While there are official databases with player performance data, there are no official databases with sponsor information. This has certainly limited our work but we still think that the most influential players in terms of sponsors are present in the database.

Inaccurate value related to stat value

When importing performance data, it was very important to put the data in the right context and be careful to assess the quality of each piece of information. Here are examples of some of the cautions we have encountered:

	Name	Country	Height	Weight	Age	Birthday	Current rank	Higher rank	sex	plays	match win	match lose	titles
0	Tomljanovic Ajla - profile	Australia	180	67	29	(7. 5. 1993)	s: 35.	/ 33.	woman	right	\n387/	292\n	\nSummary:\n1/3\n-\n0/2\n-\n1/1\n-\n
1	Riske Alison - profile	USA	175	66	32	(3. 7. 1990)	s: 49.	/ 18.	woman	right	\n420/	341\n	\nTotal:\n3\n9\n

In this case, for the tennis player Tomljanovic, the number of titles won was not reported correctly, but the information in the titles column refers to other characteristics of the athlete.

id	Name	Country	Height	Weight	Age	Birthday	Current rank	Higher rank	sex	plays	match win	match lose	titles
79	Shelton Ben - profile	USA	002)		nt	/Highest rank - singles: 89. / 89.	s: 283	/ 280	man	right	\n66/2	2\n5/	\nTotal:\n0\n4\n

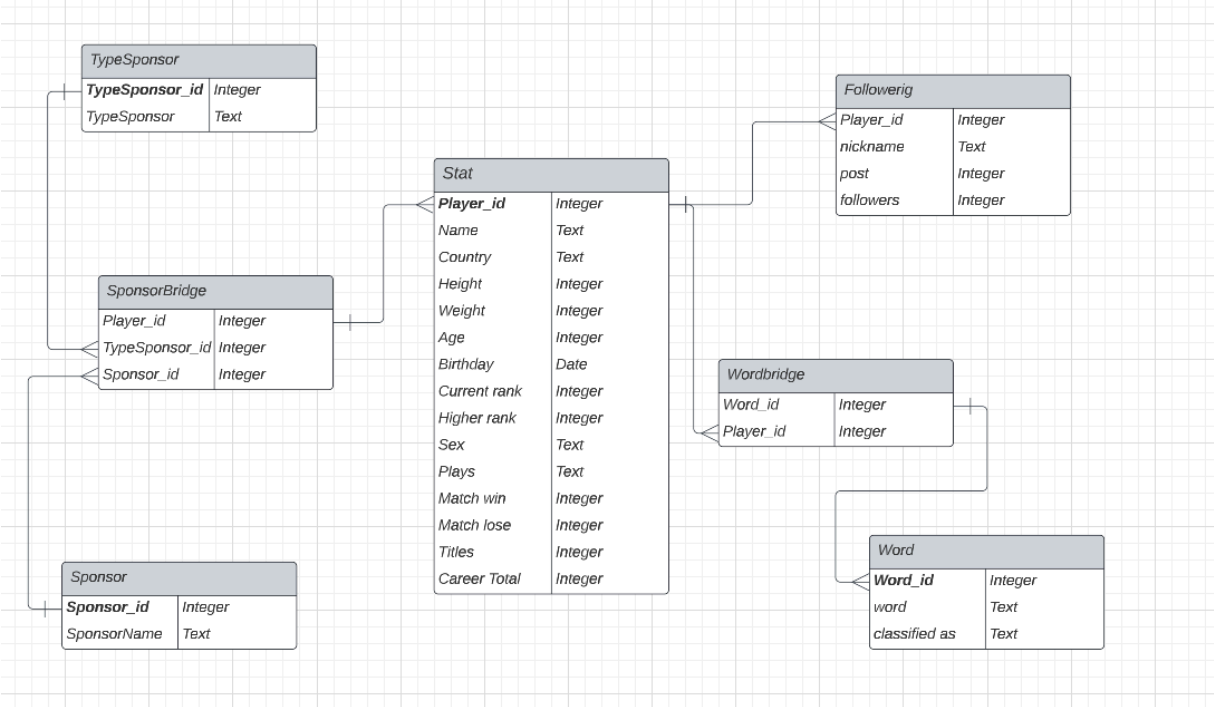
For this tennis player, however, there was no proper correspondence between cell instance and column name.
Sometimes, however, the error did not attack the syntactic sphere but the semantic one (e.g. Federer not winning tournaments in his career). In these cases, domain knowledge was essential to detect errors that would make the analysis of poor quality.

Data modeling

After due consideration, we decided to use a relational model for our integrated database. The choice fell on this model for the following reasons:

- Having decided to do the data analysis through the construction of an interactive tableau infographic, the tabular format was the one that best suited our needs.
- The other data models were better suited to highlighting parts of our work that were outside our scope. For example, a graph model was certainly better suited to highlighting the relationships between tennis players and the data from, for example, their head-to-head matches.
- Moreover, most of the data we had extracted was already in tabular format (e.g. the table of sponsors) and it would have been a stretch to switch to a different data model.

We therefore decided to build a relational database according to this schema:



The relational database consists of 7 tables organized as follows:

- **Stat:** it contains the personal and performance data of tennis players. Player_id is primary key of table
- **Followerig:** It contains data on players' instagram accounts. We assumed that a player could have a maximum of one Instagram account and that an account could only be associated with one player

- **Word**: It contains the list of words found in Google queries. Word_id is the primary key of the table and each word is classified as atpref, wtaref, tennisref or other. A word can only have one classification, e.g. a word cannot be considered tennisref and atpref at the same time. Finally, a word can be associated with several players and several players can be associated with one word, for this we need the **WordBridge** table
- **Wordbridge:** It is the table that allows the connection between players and the words associated with them
- **Sponsor:** This table contains the list of all named Sponsors and Sponsor_id is the primary key. The management of Sponsors is similar to the case of words, but with one difference, where previously a word was uniquely classified, here we have cases where the sponsor is not uniquely classified, for example Adidas who is both apparel sponsor and shoes sponsor. To handle this situation in addition to the **SponsorBridge** table we have created the **TypeSponsor** table
- **TypeSponsor:** It contains the four categories of sponsors considered: Apparel, Shoes, Racket and OtherSponsors. TypeSponsor_id is the primary key of the table
- **SponsorBridge:** it acts as a bridge between the two sponsor tables and the stat table

The type of scheme does not require too many joins and allows phenomena to be analyzed individually. For example, if one wants to analyze only the Instagram part and perhaps only needs that for other projects, one can do so without having to extract other data. Furthermore, although the use of this structure requires joins to analyze the information jointly, we think that the computational effort involved can be compensated for by a relatively small volume of data.
It can also be seen that the table of Sponsors has been redesigned to be part of a relational schema. This change was forced by the fact that the relational schema allows only one attribute value per cell, whereas in the previous representation we had more than one value. As far as word management was concerned, although the json format was valid, we decided to convert it to the relational database in order to have all the necessary information in one database.
Finally, although our choices were guided by our research questions, we still wanted to think about building a database that could also be useful for other purposes and be the starting point for other questions. We think, for example, that this type of representation may be the best for the application of machine learning models such as clustering and classification.

From the created database you can do any kind of extraction, here are two examples:

- selecting the number of 'OtherSponsors' for each tennis player

```
select Name,count(*)
from Sponsorbridge,stat,TypeSponsor
where Sponsorbridge.Player_id=stat.Player_id and TypeSponsor.TypeSponsor_id=Sponsorbridge.TypeSponsor_id
and TypeSponsor="OtherSponsor"
group by Name
order by count(*) desc;
```

Name	count(*)
roger federer	13
serena williams	12
naomi osaka	11
maria sakkari	9
petra kvitova	8
rafael nadal	7
kei nishikori	7
venus williams	6
milos raonic	6

- select Nike-sponsored tennis players with more than 100,000 followers but less than 5 titles won

```
select Name,followers
from Sponsorbridge,stat,followerig,Sponsortable
where Sponsorbridge.Player_id=stat.Player_id and stat.Player_id=followerig.Player_id
and Sponsortable.Sponsor_id=Sponsorbridge.Sponsor_id and Sponsortable.SponsorName="Nike"
and followers>100000 and titles<5
group by Name
order by followers desc;
```

Name	followers
emma raducanu	2542552
bianca andreescu	705075
paula badosa	677644
denis shapovalov	574317
frances tiafoe	442539
holger rune	345469
taylor fritz	323910
karen khachanov	279181
donna vekic	221918

Interpretation

As mentioned earlier, the objective of this work is to provide a Tableau infographic that could answer our research questions. We thought of an interactive infographic that would allow the user to navigate through the platform. In this part of the report we will illustrate the analyses made to the data with the help of the constructed

platform that can be found at this link. Of course, many other analyses are possible in addition to those in the infographics.

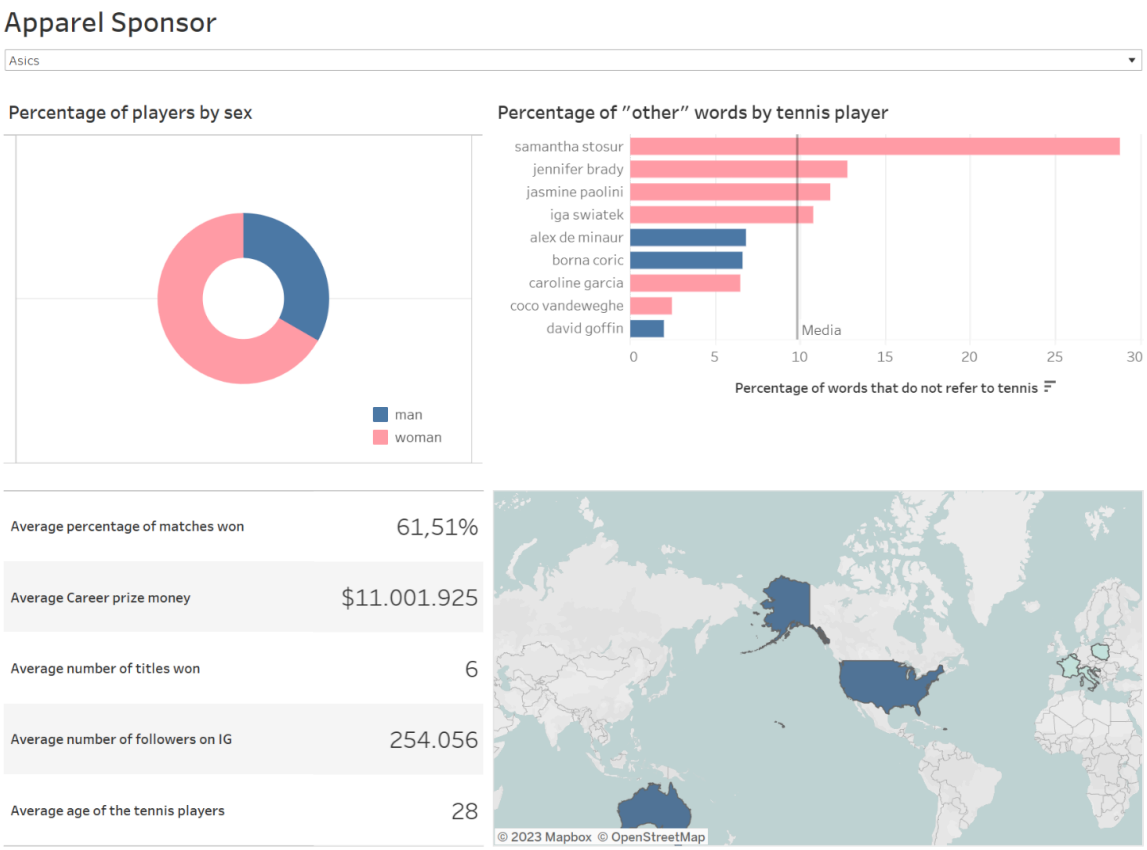
In particular, we decided to focus on four aspects based on our research questions to conduct our analysis:

- Apparel Sponsor
- Racket Sponsor
- Analysis of the related queries on Google
- Individual analysis of players

Apparel Sponsor and Racket Sponsor

To address these two matters, we built two dashboards with the same structure, one regarding apparel and one regarding rackets being sponsored, respectively.

Below is the dashboard obtained regarding apparel brands.



The dashboard consists of four charts. At the top the donut chart represents the percentage of women and men sponsoring the selected brand. Next to it one can see the names of the players involved and their respective percentages that relate to words that are not related to tennis. We decided to include the average vertically to

give a clearer idea of the phenomenon, in fact in this way it is easier to identify when the percentage deviates greatly from the average. We also inserted a table with the main statistics inherent to our research questions and always referring to the selected brand, so we calculated the average percentage of matches won, the average career prize money, the average number of titles won, the average number of followers on IG and the average age of the tennis players. Finally, we inserted a map displaying the origin of the players. In this case the color intensity is related to the number of players, the more the area has an intense color the higher the number of players related to the area.
The dashboard being interactive allows viewing charts related to a specific brand via the drop-down list at the top. In addition, by selecting a percentage of the pie chart the other charts adjust to the selection, or by selecting an area of interest in the map the dashboard updates focusing on the relevant location. In addition, hovering the cursor over each graph highlights the data and percentages corresponding to the portion of the graph queried.

Considering the data we have available, we can see that we have a total of 43 apparel brands that are divided into:

- 21 brands sponsored only by men
- 13 brands sponsored by both men and women of which 8 are sponsored more by men than women
- 9 brands sponsored only by women

While the 9 racket sponsors are divided into:

- 1 brand sponsored only by men
- 8 brands sponsored by both men and women of which 6 are sponsored more by men than women

From here it is evident how many brands consider only men as references, however, it is necessary to specify that some of them are sponsored by only one or two players (e.g. AMC, Artengo) and that is why there are no different proportions between men and women. This consideration also applies to brands sponsored only by women. While the most famous and sports related sponsors are sponsored by both men and women (such as Adidas, Lacoste, and Nike) and there are more players related to them.

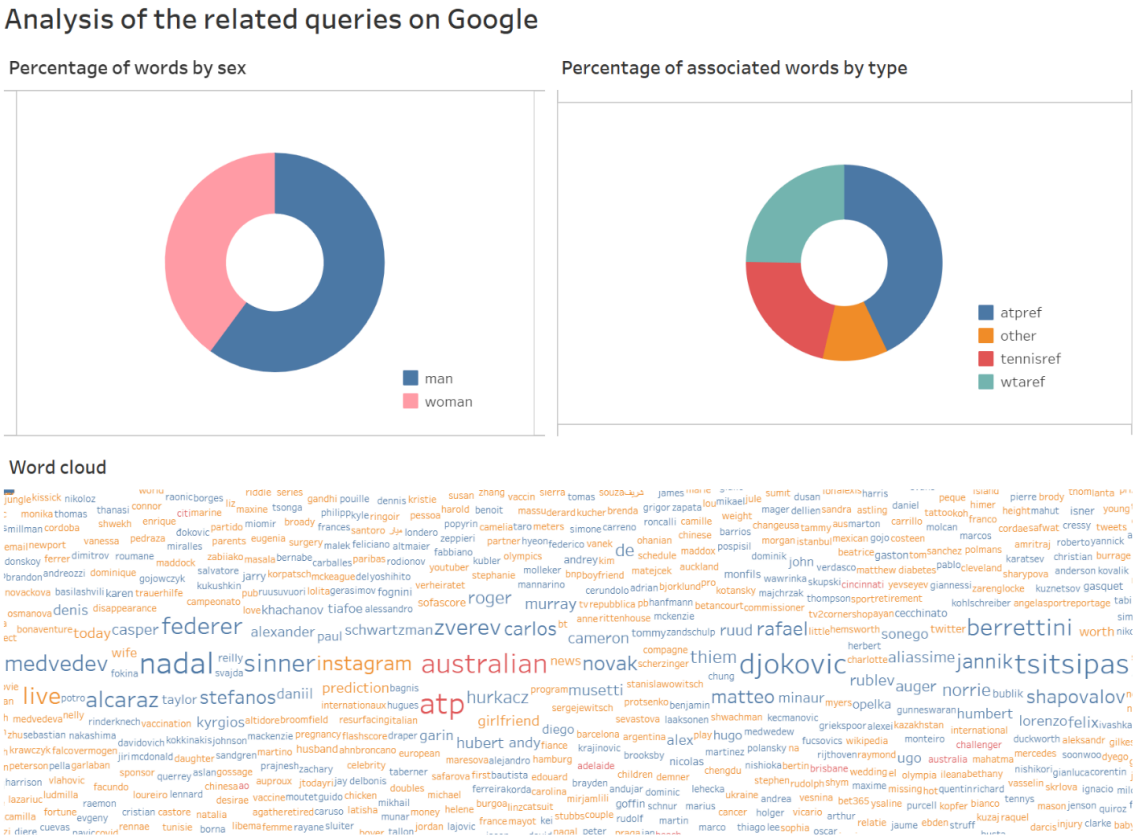
Brands that rely only on male or only on female figures have typically high percentages of matches won, on par with other brands. The average number of titles won is more stable and low; however, there are exceptions, such as the AMC brand sponsored only by Andy Murray that boasts very high values in all the statistics calculated in the table, or EleVen sponsored only by Venus Williams. The same reasoning fits the calculation of average career prize money and average number of followers taking into account the influence of the number of titles won.

We have noticed how sponsors tend to range widely across geographic areas and do not tend to be sponsored by players with a specific background. There are exceptions such as Joma, which is linked to six players, 4 of whom are from Spain.

Finally, some players are associated much more with words not related to tennis than others. For example, Shuai Peng has a very high percentage of words related to her disappearance and the report of harassment she suffered.

Analysis of the related queries on Google

In this third dashboard we displayed a general overview of the appearance of Google queries as follows:

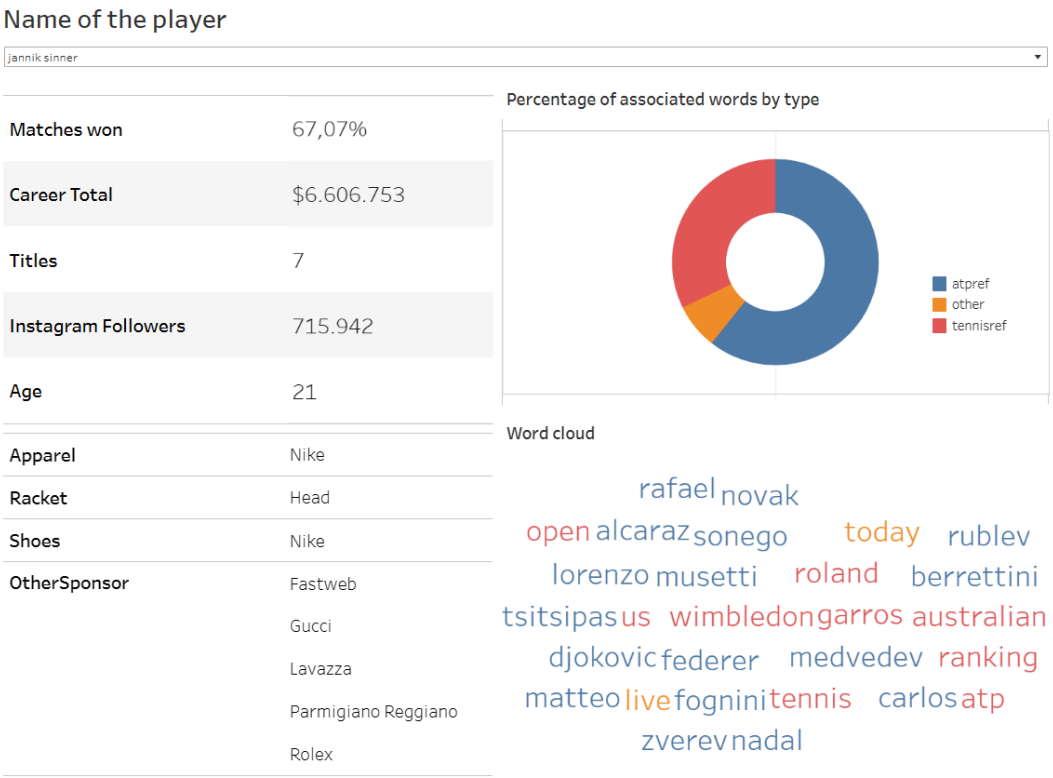


In the upper part of the dashboard we have represented the percentage of words associated with players divided for sex and the percentage of words divided by type by donut chart. In the lower part there is a word cloud with all words highlighted with different colors according to the type of word. In addition, the size of the word is proportional to the percentage of times it was searched.
Again the dashboard is interactive, in fact the pie charts allow filtering the words in the word cloud according to the selected category. In addition, the legend referring to the type of words also functions as a highlighter, so selecting the category of interest highlights the corresponding terms. By hovering the cursor over each portion of the graph, it is possible to observe the percentages obtained on the individual portion of the data.

According to the left donut chart, there are more words extracted from male players than from female players. This may be due to the fact that male players in the starting database are a little more than female players. Observing the dashboard without filtering the data, we can notice that according to the right donut chart, most of the terms are related to ATP (about 43%) followed by WTA terms (about 25%), tennis in general (about 21%), and finally other (about 10%). By selecting men, we can notice how the words related to ATP increase (about 65%) while the words related to WTA decrease (about 4%) and the other words (9% approximately). On the other hand, by selecting women from the donut chart, more noticeable changes occur, in fact, it can be seen that the words related to ATP are more than 8% and that the other words are about 13%. So, there are more references to ATP when filtering women than references to WTA when selecting men. Therefore, women are often associated with male players, even based on their personal relationships, and this happens more often than associating women with men. Even by selecting men and investigating the other category, it is possible to see that the percentage of times that Instagram was mentioned is about 3.5%, a much lower percentage than the same category associated with women, where the percentage related to the term Instagram increases up to about 10%.

Individual analysis of players

In the fourth and final dashboard we investigated the previous phenomenon from a different perspective, considering the individual player:



In this case we have included a drop-down list where the player can be selected, so updated graphs are displayed according to the choice made. We structured the dashboard by inserting a table with the main statistics referring to the player, thus the percentage of matches won, total prize money, number of titles won, number of followers on Instagram and age. Next to it is a pie chart with the percentage of words divided by type inherent to the player. In the second part we have included a second table with the apparel brand, racket brand, shoes brand, and the presence of any other brands. Finally there is a word cloud to represent the words related to Google queries again divided by type through colors. In this case we decided to remove duplicate terms to have a clearer view of specific terms.
The interactivity here is based on selecting the area of interest from the pie chart with subsequent filtering of the words in the word cloud.
This makes it more noticeable that some players are not associated with WTA or ATP depending on their gender, and sometimes the other section is absent. This, however, as also highlighted above, happens much more often with male players than with female players who are often searched in correlation with a man or with personal facts. In addition, it is possible to have a specific overview for each player regarding all the points addressed in the previously illustrated graphs.

Conclusion

In this project, we wanted to compare the world of men's and women's tennis and their relationship with sponsors to see if it was based on meritocracy or other factors. Although the topic is very broad, we think our work can partially answer the research questions and can give some food for thought on the situation of sponsors in the tennis world. However, our work has some limitations:

- **the difficulty in finding data on sponsors:** we only found data on 175 players who, despite being the best in world tennis, represent only a small part of all ATP and WTA professionals
- **lack of other significant variables:** Certainly, the phenomenon under investigation can be described by further elements, but for various reasons it was not possible to access them. An example of this is the Forbes data on the earnings of tennis players off the court or data on the period of sponsorship (how long has the athlete been sponsored by the brand?)

We do not exclude that in the future the work can be improved and we think it can be a starting point for further search queries such as clustering or more specific text mining techniques on the words of Google queries.