ATP & WTA:

Analysis of tennis players by performance, sponsors, and social interaction

Marta Bernardelli
Alessia Perri
Giulia Beccaria

INTRODUCTION

- **Problem statement:** Complaint by Romanian tennis player Sorana Cirstea regarding logics in women's tennis world that go beyond sports performance and exclude meritocracy.

- **Goal:** To quantitatively test the phenomenon of marketing needs trumping sporting merit in women's tennis.

- **Method:** Combine information on sports performance, sponsors, Google trends, and social engagement of both male and female tennis players.

- **Analysis:** Compare data to see if significant patterns can be found to describe the phenomenon.

- **Output:** Infographic on Tableau summarizing the data analysis

DATA ACQUISITION

01

DATA ACQUISITION

Sponsor data

103 male tennis players
72 female tennis players

- Name of the player
- Apparel sponsor
- Shoes sponsor
- Racket sponsor
- Other Sponsors

Web scraping from Toomanyrackets website
Python (BeautifulSoup library)

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	–

DATA ACQUISITION

Performance data

- Name of the player
- Country of origin
- Height (cm)
- Weight (kg)
- Age
- Date of birth
- Current ranking
- Higher ranking
- 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 (in singles)

	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

Web scraping from TennisExplorer website
Python (BeautifulSoup library)

DATA ACQUISITION

Instagram data

- number of posts on profile
- number of followers

API with Python Instaloader

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

DATA ACQUISITION

Google trends

- Why Google Trends?
- English as preferred language to extract the queries
- Select the most frequent searches over the past 5 years

```
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
```

API with Python Pytrends

DATA ACQUISITION

Google trends data reworking

- **Normalized** the strings by removing accents and special signs, removed the player's name, removed duplicates, and removed English dictionary stopwords.
- **Translated** some words into English and put the results into a pandas dataframe.
- Constructed three "bags of words" for **ATP references**, **WTA references**, and **tennis references**.
- **Analyzed** each word in the dataset to see if it belonged to one of the three groups or was classified as "others."
- Organized data in **pandas database** and **json files**.

DATA ACQUISITION

In the analysis of the “other” words there were some contamination within related searches.

Example: Dustin Brown

- tennis-related searches
- words referring to the world of the TV series Stranger Things

	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

First 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

Putting everything into a pandas dataframe we got the following result

Classification

	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

Three categories:

- ATP references
- WTA references
- Tennis references
- Others

JSON

```
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']}
```

DATA QUALITY AND CONFLICTS BETWEEN DATA SOURCES

02

Name format problem

- Different formats of names causing problems
- Aim to find a single format for names to improve consistency and integration

Garbiñe Muguruza</td>,

GARBINE,MUGURUZA

Muguruza Garbine - profile

Steps taken to unify names:

- Use of upper and lower case letters: names kept in lowercase in all data sources
- First and Last Name vs. Last and First Name: changed to "First Name and Last Name" format in performance data
- Removal of special graphic signs such as - or ~

Note: encountered semantic problems during standardization

Stanislas Wawrinka

stan wawrinka

Resulting name format: First Name and Last Name, lowercase, without special graphic signs. Example: **garbine muguruza**

Inaccurate value related to stat value

	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

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

Possible error:

- syntactic error
- semantic error

Missing values related to sponsor

Attribute	Percentage of missing value	Percentage of missing value after
Apparel	7,42%	0%
Shoes	11,42%	1,7%
Racket	0%	0%
Other Sponsor	81%	32%

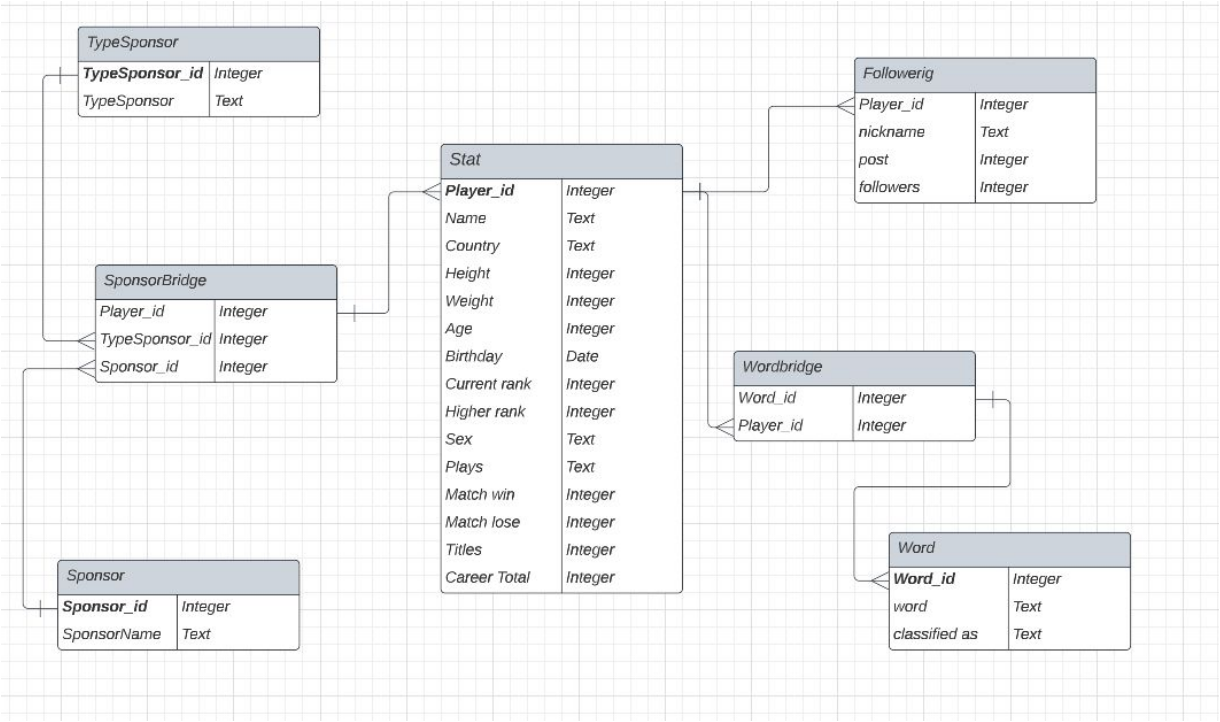
Missing values change:

- changed '-' symbol
- we relied on information found on the internet and our knowledge of the domain
- the percentages of missing values have been reduced considerably

DATA MODELLING

03

Relational Model



Reasons for the choice

- Tabular format was best suited for data analysis through interactive tableau infographic
- Other data models better suited for parts outside scope, such as graph model for highlighting relationships between tennis players
- Most of the extracted data was already in tabular format, switching would have been difficult

7 tables organized:

- Stat
- Followerig
- Word
- Wordbridge
- Sponsor
- TypeSponsor
- SponsorBridge

INTERPRETATION

Tableau

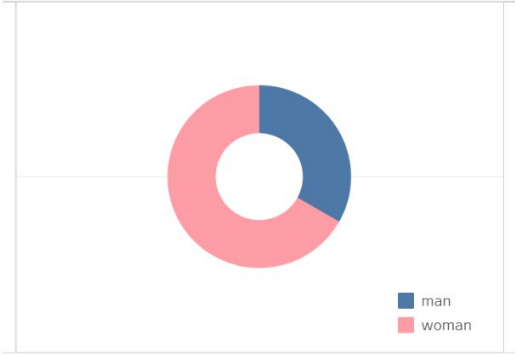
04

Apparel Sponsor and Racket Sponsor

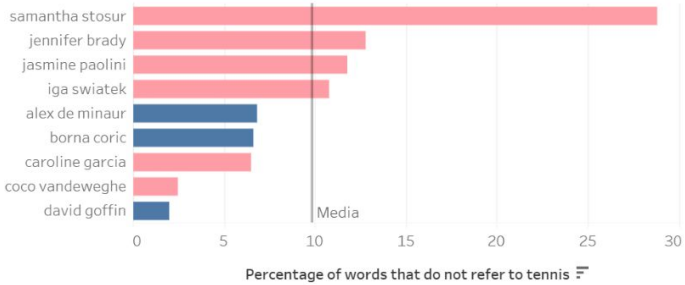
Apparel Sponsor

Asics

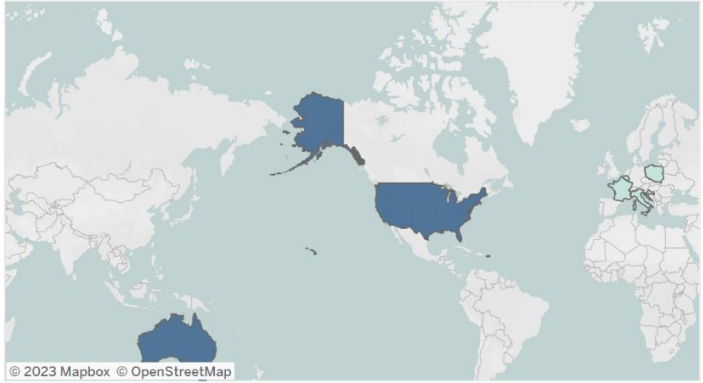
Percentage of players by sex



Percentage of "other" words by tennis player



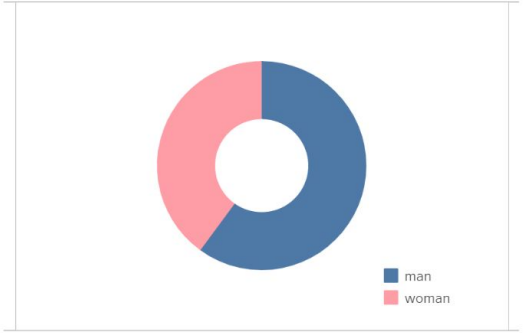
Average percentage of matches won	61,51%
Average Career prize money	$11.001.925
Average number of titles won	6
Average number of followers on IG	254.056
Average age of the tennis players	28



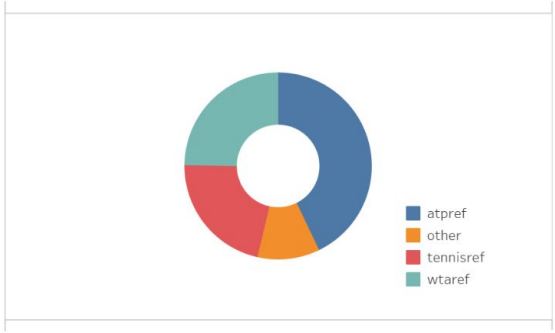
Analysis of the related queries on Google

Analysis of the related queries on Google

Percentage of words by sex



Percentage of associated words by type



Word cloud

Individual analysis of players

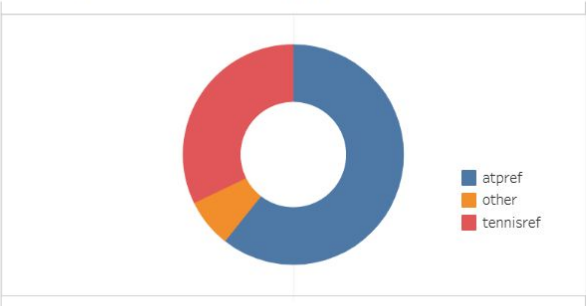
Name of the player

jannik sinner

Matches won	67,07%
Career Total	$6.606.753
Titles	7
Instagram Followers	715.942
Age	21

Apparel	Nike
Racket	Head
Shoes	Nike
OtherSponsor	Fastweb
	Gucci
	Lavazza
	Parmigiano Reggiano
	Rolex

Percentage of associated words by type



Word cloud

CONCLUSION

Goal: compare the world of men's and women's tennis and their relationship with sponsors, to determine whether it is based on merit or other factors

Limitation of the project:

- the difficulty in finding data on sponsors
- lack of other significant variables

Marta Bernardelli
Alessia Perri
Giulia Beccaria