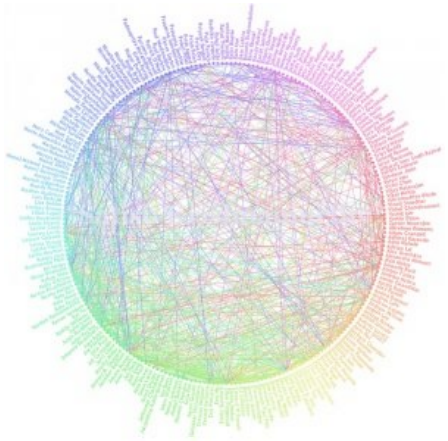
HAW Hamburg
Lars Ullrich

UEFA Euro 2016

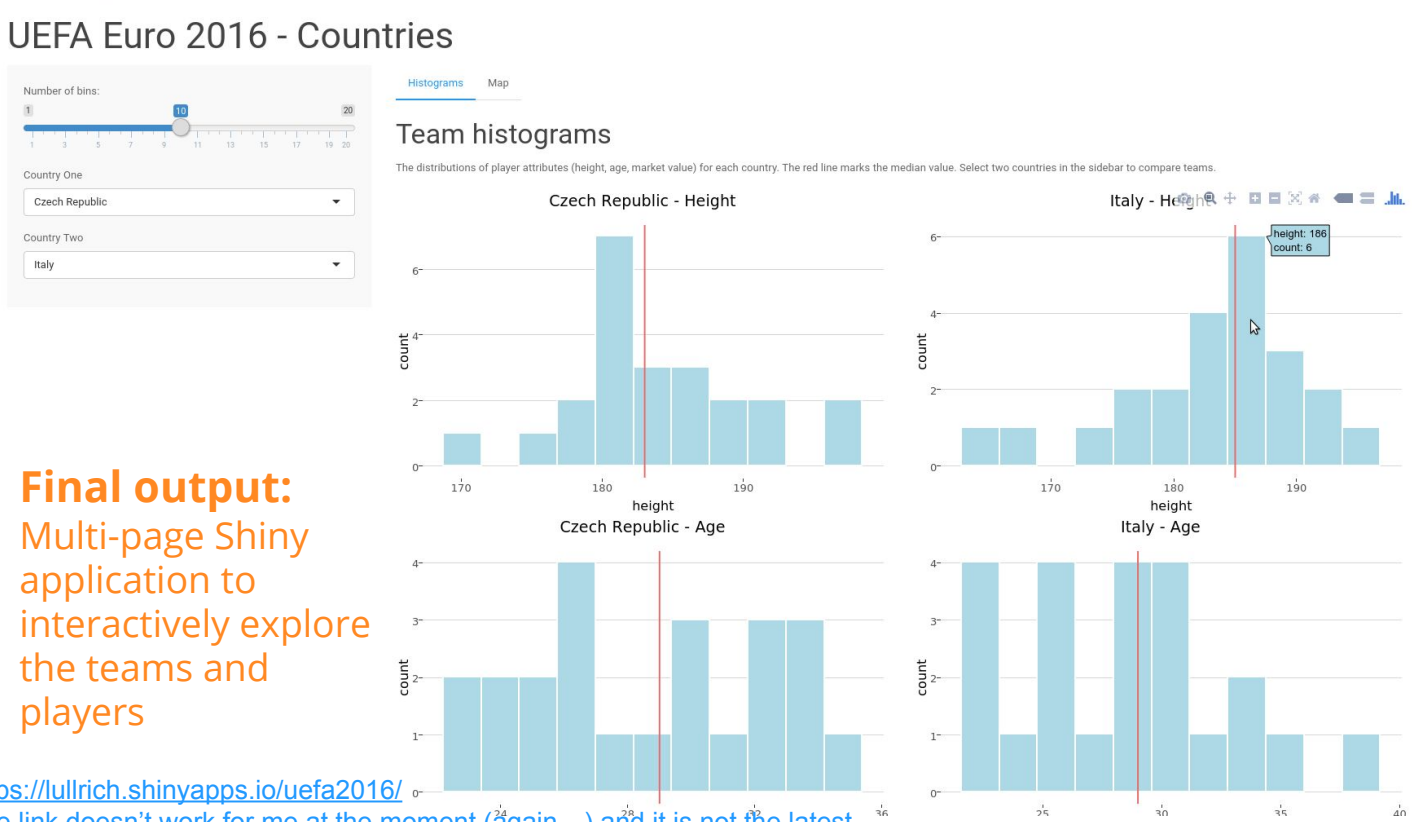


Agenda

- Main Research Question
  - What are the characteristics of the teams competing in UEFA Euro 2016?
  - Can the data gathering and presentation be automated for future events? (Feasibility)?
- Sub Research Questions (per country or per position)
  - What is the distribution of player height per team?
  - What is the distribution of player age per team?
  - What is the distribution of player market value per team?
  - How do they compare on these characteristics?
  - Where are the clubs located the players play for (geolocation of home ground)?
- Conslusion

Main Research Question (Final Output)

"What are the characteristics of the teams competing in UEFA Euro 2016?"



Final output:
Multi-page Shiny application to interactively explore the teams and players

https://lullrich.shinyapps.io/uefa2016/
The link doesn't work for me at the moment (again…) and it is not the latest version. To start the app locally, see main.R (the main entry point)

Motivation:

- Interested in Football and intrigued by the idea to quantify national teams
- Interested in web development and data visualization

Sub Research Question (Final Output)

"Where are the clubs located the players play for?"

Number of bins:

1 10 20

1 3 5 7 9 11 13 15 17 19 20

Country One

Czech Republic

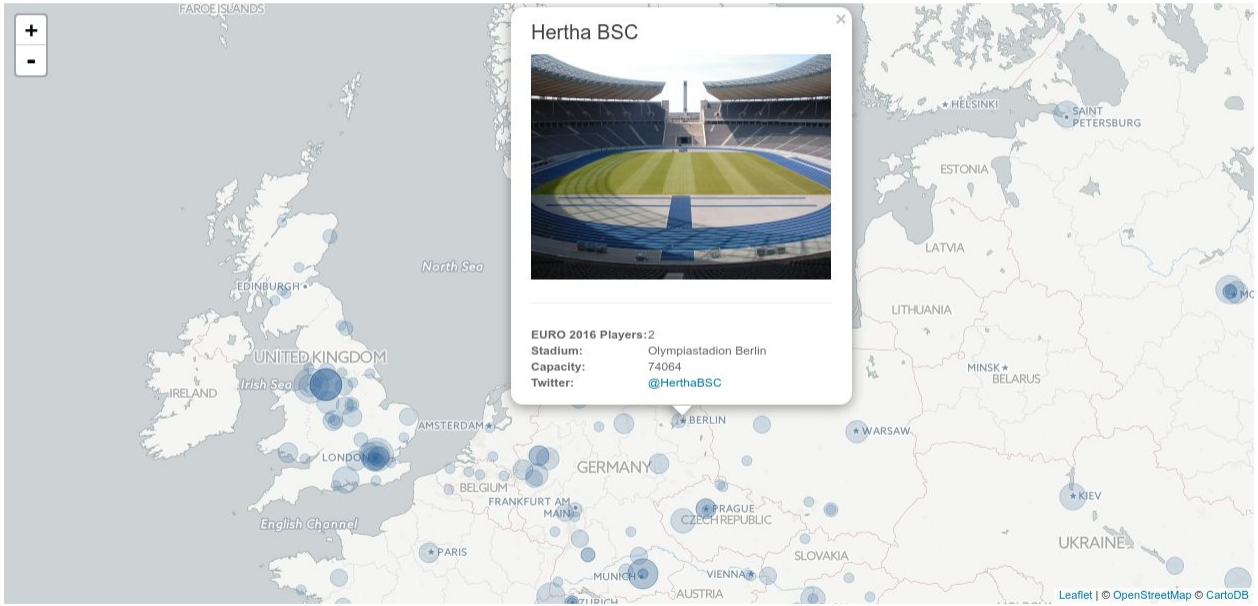
Country Two

Italy

Histograms | Map

Map of player clubs

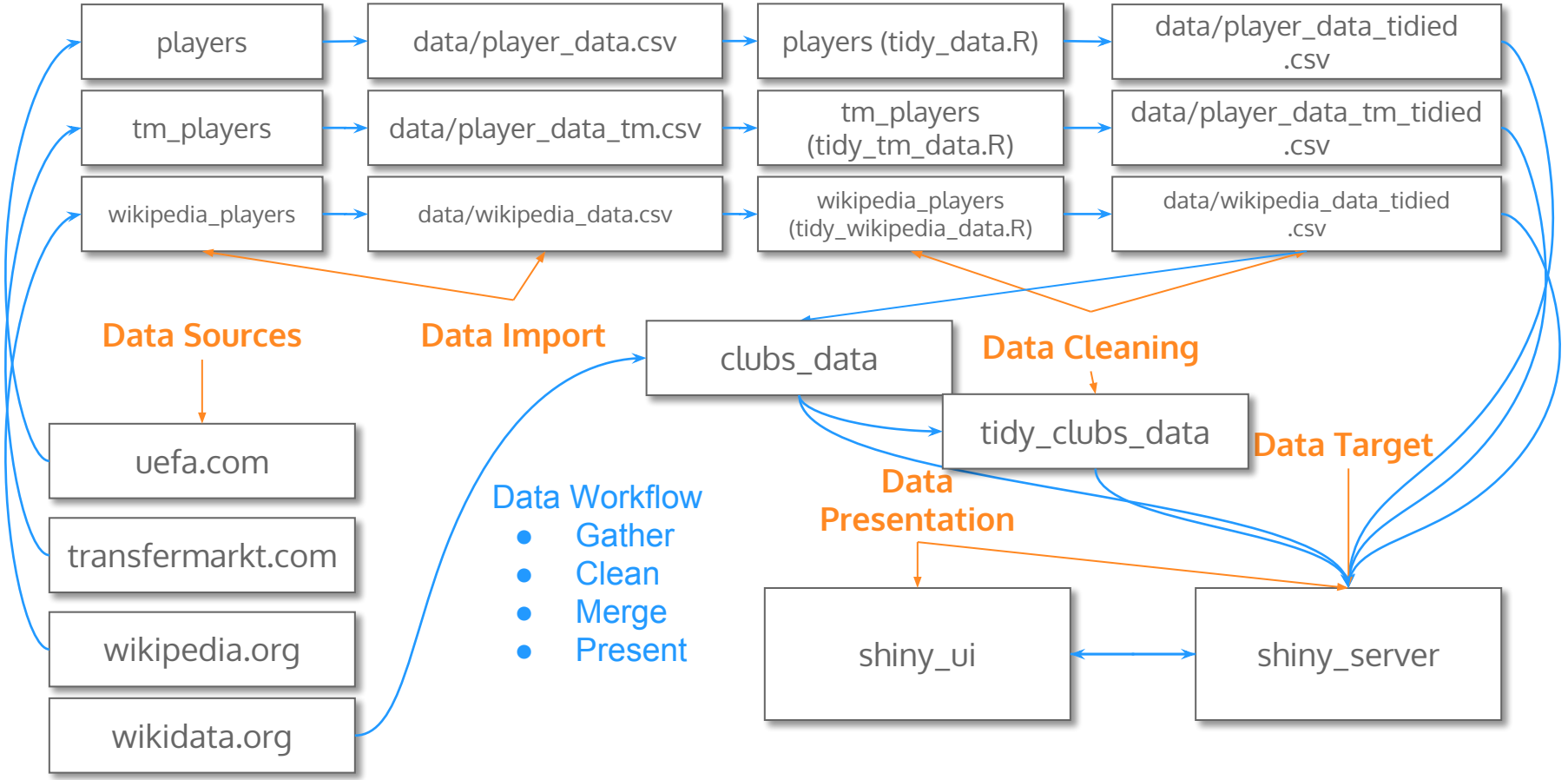
A map of every club that has players participating in EURO 2016. The size of the marker is equivalent to the number of players the club sent to France. Click on a circle to see information about the club.



https://lullrich.shinyapps.io/uefa2016/

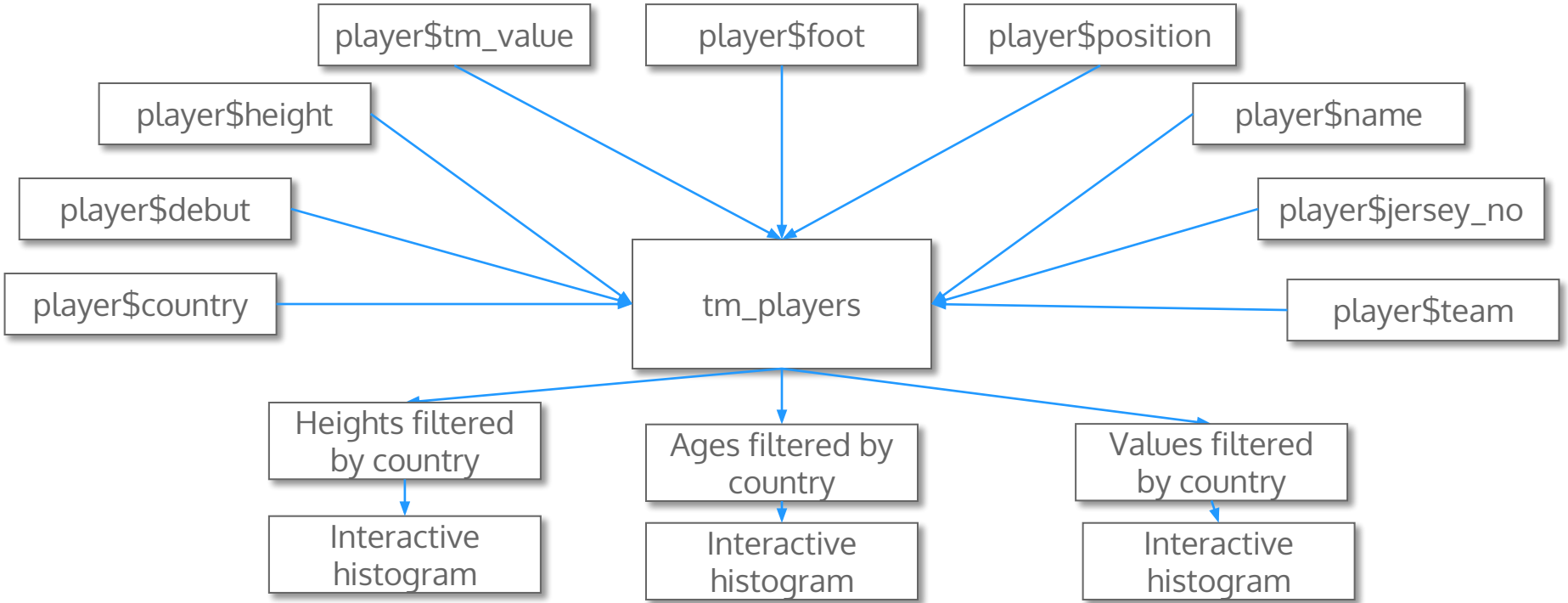
Status Update

Data Pipeline



Status Update

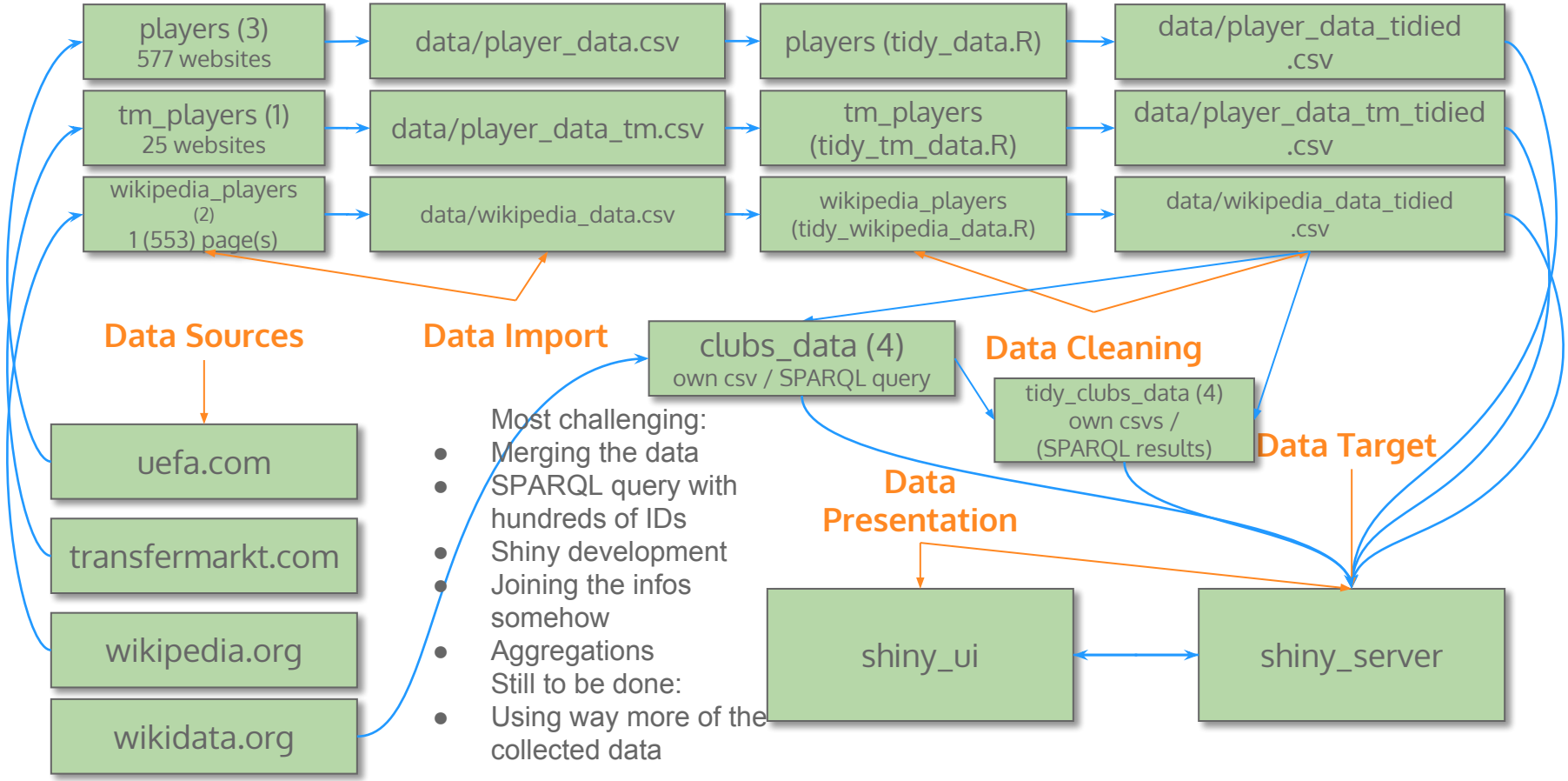
Data Elements (Sample) - data source transfermarkt.com



- A custom function gathers the data from each team details page on transfermarkt.com and constructs the player list.
- From this list a data.frame is created
- The list of data.frames is merged into one data.table

Status Update
Data Workflow

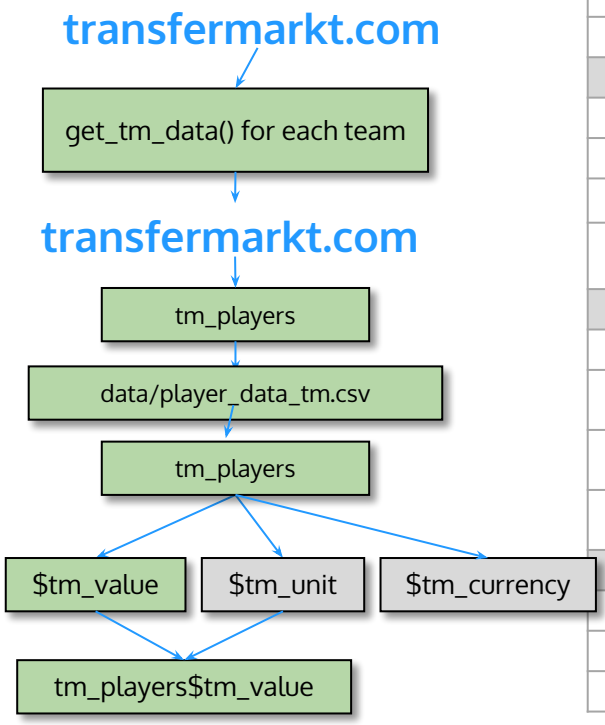
done started doubtful



Agenda

- Main Research Question
- **Sub Research Questions**
- Conclusion

OVERVIEW



What is the distribution of player value per team?		
Data Output		
	Name	tm_player_data$tm_value
	Type	numeric
Data Input 1		
	name	No name. These values are just piped through to get to the next step.
	Data Format	URL for each country in HTML table
	Data type	chr
	Data Source (DS)	http://www.transfermarkt.com/europameisterschaft-2016/startseite/pokalwettbewerb/EM16
Data Input 2		
	name	tm_players
	Data type	Data.table
	Data Category	Internet
	Data Source (DS)	http://www.transfermarkt.com/germany/kader/verein/3262/saison_id/2016 (sample)
Data Input 3		
	Data Format	R Output
	Data name	tm_players$tm_value
	Data type	Chr

Continued

Internal Flow		
	Activity 1	Get the URL for each team
	Activity 2	Construct the needed URL s from the scraped URLs (so they point to the squad details view)
	Activity 3	Visit each URL from step 2 and extract the information from HTML elements within a horribly formatted table
	Activity 4	Save the data
	Activity 5	Split the tm_value column into tm_value, tm_unit and tm_currency
	Activity 6	format tm_value according to international number formatting (decimal point) and turn it into a numeric value
	Activity 7	Multiply the value according to tm_value to arrive at the correct result (Th times 1000, Mill times 1000000)
Functions		
	custom function	Get_tm_data, month_to_num
	contains:	Unique, list, data.frame(base); mutate, extract_numeric (dplyr); unite, separate (tidyr); str_replace(stringr); read_html, html_node, html_text, html_attr (rvest); now, ymd (lubridate); fread, data.table (data.table)

Data Input Samples mentioned in RQ OVERVIEW slide before

Data Input		
	Data Format	**HTML TABLE/HTML TAG**
	Data Category	Internet
	Data Source	http://www.transfermarkt.com/europameisterschaft-2016/startseite/pokalwettbewerb/EM16
	Data element in DS	CSS-Selector: "td.links.no-border-links.hauptlink > a.vereinprofil_tooltip", href
	Data quality	Ok, neeed to alter the URLs to get what i was after
	Selection reason	Gateway to the market value data I needed

Data Input		
	Data Format	**HTML**
	Data Category	Internet
	Data Source	http://www.transfermarkt.com/germany/kader/verein/3262/saison_id/2016 (sample)
	Data element in DS	xpath: "//*[@id='yw1']/table/tbody/tr/td[10]"
	Data quality	poor, horribly nested table
	Selection reason	Best way to get market values

Data Input		
	Data Format	**CSV**
	Data Category	R Output
	Data Source	“data/tm_player_data.csv”
	Data element in DS	$tm_value
	Data quality	poor, values needed quite a bit reformatting
	Selection reason	-

Data Output		
	Data Format	**R OUTPUT**
	Data name	tm_player_data$tm_value
	Data type	Data table: numeric

INPUT

What is the distribution of player value per team?

transfermarkt.com

get_tm_data() for each team

PARTICIPATING TEAMS OF THE EURO 2016

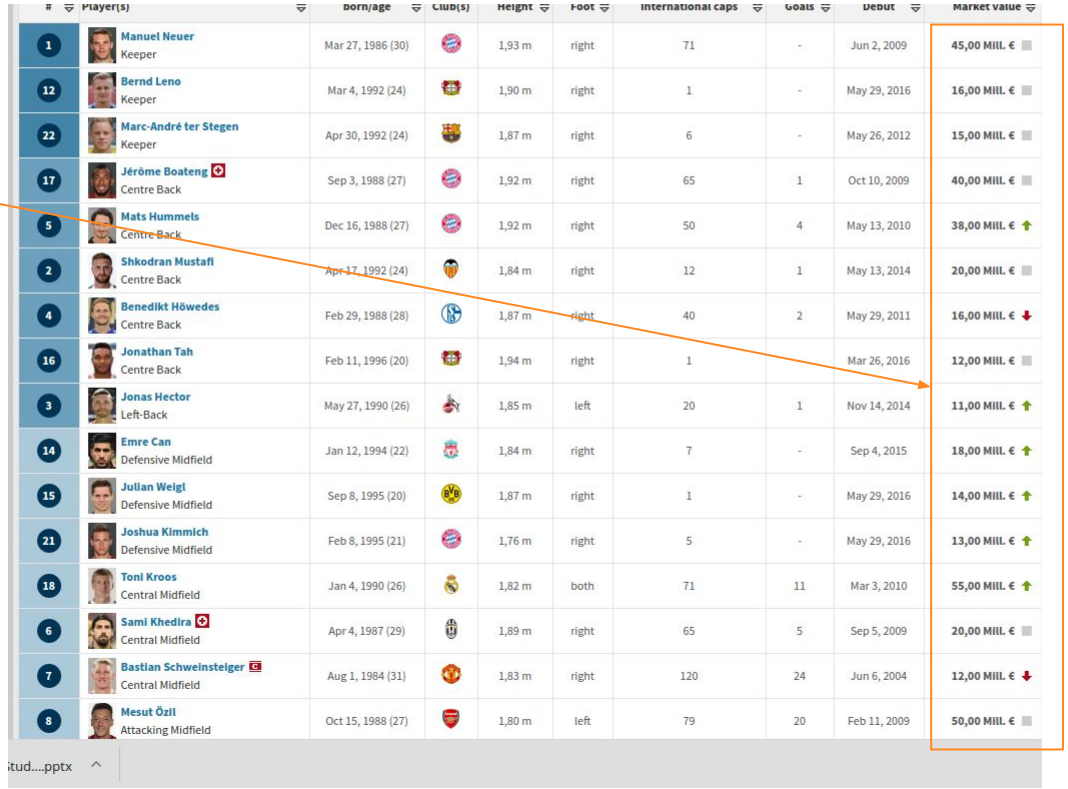
Club	Squad	ø-Age	Legionäre	Market value	ø-Market value
Spain	23	28,2	10	597,00 Mill. €	25,96 Mill. €
Germany	23	25,9	9	585,00 Mill. €	25,43 Mill. €
France	23	27,9	20	510,50 Mill. €	22,20 Mill. €
England	23	26,0	0	477,00 Mill. €	20,74 Mill. €
Belgium	23	26,6	22	472,00 Mill. €	20,52 Mill. €
Portugal	23	28,5	16	364,00 Mill. €	15,83 Mill. €
Croatia	22	26,5	18	296,75 Mill. €	13,49 Mill. €
Italy	23	29,0	6	276,50 Mill. €	12,02 Mill. €
Poland	23	27,7	14	194,45 Mill. €	8,45 Mill. €
Turkey	23	27,1	7	180,50 Mill. €	7,85 Mill. €
Wales	23	27,5	20	180,10 Mill. €	7,83 Mill. €
Switzerland	22	25,9	20	172,85 Mill. €	7,86 Mill. €
Austria	22	27,7	20	126,45 Mill. €	5,75 Mill. €
Ukraine	23	28,0	7	121,90 Mill. €	5,30 Mill. €
Russia	23	29,2	2	117,10 Mill. €	5,09 Mill. €
Republic of Ireland	23	30,0	23	87,30 Mill. €	3,80 Mill. €
Slovakia	24	29,2	21	87,10 Mill. €	3,63 Mill. €
Sweden	19	26,8	17	69,35 Mill. €	3,65 Mill. €
Romania	23	28,9	16	54,60 Mill. €	2,37 Mill. €
Czech Republic	18	27,9	11	49,70 Mill. €	2,76 Mill. €

INPUT

What is the distribution of player value per team?

transfermarkt.com

tm_players



#	Player(s)	Born/age	Club(s)	Height	Foot	International caps	Goals	Debut	Market value
1	Manuel Neuer Keeper	Mar 27, 1986 (30)		1,93 m	right	71	-	Jun 2, 2009	45,00 Mill. €
12	Bernd Leno Keeper	Mar 4, 1992 (24)		1,90 m	right	1	-	May 29, 2016	16,00 Mill. €
22	Marc-André ter Stegen Keeper	Apr 30, 1992 (24)		1,87 m	right	6	-	May 26, 2012	15,00 Mill. €
17	Jérôme Boateng Centre Back	Sep 3, 1988 (27)		1,92 m	right	65	1	Oct 10, 2009	40,00 Mill. €
5	Mats Hummels Centre Back	Dec 16, 1988 (27)		1,92 m	right	50	4	May 13, 2010	38,00 Mill. €
2	Shkodran Mustafi Centre Back	Apr 17, 1992 (24)		1,84 m	right	12	1	May 13, 2014	20,00 Mill. €
4	Benedikt Höwedes Centre Back	Feb 29, 1988 (28)		1,87 m	right	40	2	May 29, 2011	16,00 Mill. €
16	Jonathan Tah Centre Back	Feb 11, 1996 (20)		1,94 m	right	1		Mar 26, 2016	12,00 Mill. €
3	Jonas Hector Left-Back	May 27, 1990 (26)		1,85 m	left	20	1	Nov 14, 2014	11,00 Mill. €
14	Emre Can Defensive Midfield	Jan 12, 1994 (22)		1,84 m	right	7	-	Sep 4, 2015	18,00 Mill. €
15	Julian Weigl Defensive Midfield	Sep 8, 1995 (20)		1,87 m	right	1	-	May 29, 2016	14,00 Mill. €
21	Joshua Kimmich Defensive Midfield	Feb 8, 1995 (21)		1,76 m	right	5	-	May 29, 2016	13,00 Mill. €
18	Toni Kroos Central Midfield	Jan 4, 1990 (26)		1,82 m	both	71	11	Mar 3, 2010	55,00 Mill. €
6	Sami Khedira Central Midfield	Apr 4, 1987 (29)		1,89 m	right	65	5	Sep 5, 2009	20,00 Mill. €
7	Bastian Schweinsteiger Central Midfield	Aug 1, 1984 (31)		1,83 m	right	120	24	Jun 6, 2004	12,00 Mill. €
8	Mesut Özil Attacking Midfield	Oct 15, 1988 (27)		1,80 m	left	79	20	Feb 11, 2009	50,00 Mill. €

INPUT

What is the distribution of player value per team?

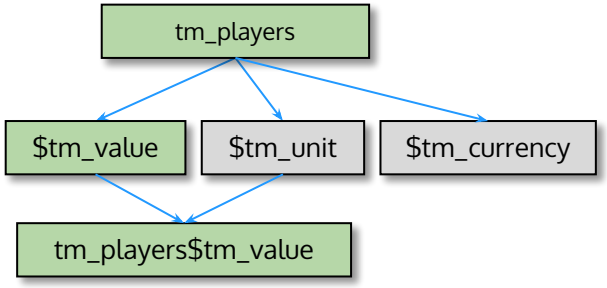
transfermarkt.com

tm_players

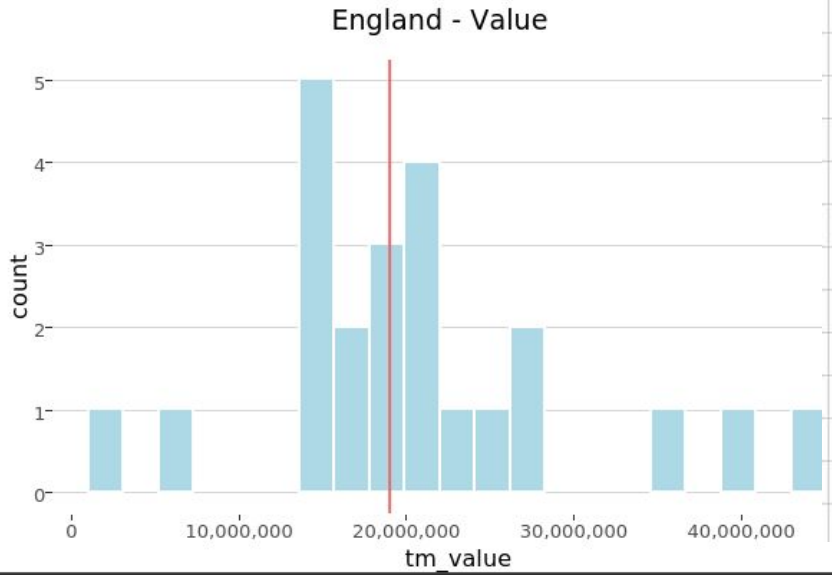
	jersey_no	name	tm_id	tm_position	birthday	tm_team_id	tm_team	foot	tm_value	height	debut	country
1	13	David de Gea	59377	Keeper	Nov 7, 1990 (25)	985	Manchester United	right	40,00 Mill. €	1,89 m	Jun 8, 2014	Spain
2	23	Sergio Rico	207302	Keeper	Sep 1, 1993 (22)	368	Sevilla FC	right	15,00 Mill. €	1,94 m	Jun 1, 2016	Spain
3	1	Iker Casillas	3979	Keeper	May 20, 1981 (35)	720	FC Porto	left	3,00 Mill. €	1,82 m	Jun 3, 2000	Spain
4	3	Gerard Piqué	18944	Centre Back	Feb 2, 1987 (29)	131	FC Barcelona	right	40,00 Mill. €	1,93 m	Feb 11, 2009	Spain
5	15	Sergio Ramos	25557	Centre Back	Mar 30, 1986 (30)	418	Real Madrid	right	40,00 Mill. €	1,83 m	Mar 26, 2005	Spain
6	4	Marc Bartra	99922	Centre Back	Jan 15, 1991 (25)	131	FC Barcelona	right	8,00 Mill. €	1,83 m	Nov 16, 2013	Spain
7	18	Jordi Alba	69751	Left-Back	Mar 21, 1989 (27)	131	FC Barcelona	left	35,00 Mill. €	1,70 m	Oct 11, 2011	Spain
8	2	César Azpilicueta	57500	Right-Back	Aug 28, 1989 (26)	631	Chelsea FC	right	20,00 Mill. €	1,78 m	Feb 6, 2013	Spain
9	12	Héctor Bellerín	191217	Right-Back	Mar 19, 1995 (21)	11	Arsenal FC	right	18,00 Mill. €	1,78 m	May 29, 2016	Spain
10	16	Juanfran	16635	Right-Back	Jan 9, 1985 (31)	13	Atlético Madrid	right	16,00 Mill. €	1,81 m	May 26, 2012	Spain
11	5	Sergio Busquets	65230	Defensive Midfield	Jul 16, 1988 (27)	131	FC Barcelona	right	60,00 Mill. €	1,89 m	Apr 1, 2009	Spain
12	17	Mikel San José	52469	Defensive Midfield	May 30, 1989 (27)	621	Athletic Bilbao	right	10,00 Mill. €	1,86 m	Sep 4, 2014	Spain
13	19	Bruno Soriano	44412	Defensive Midfield	Jun 12, 1984 (32)	1050	Villarreal CF	left	8,00 Mill. €	1,84 m	Aug 11, 2010	Spain
14	10	Cesc Fàbregas	8806	Central Midfield	May 4, 1987 (29)	631	Chelsea FC	right	45,00 Mill. €	1,75 m	Mar 1, 2006	Spain
15	6	Andrés Iniesta	7600	Central Midfield	May 11, 1984 (32)	131	FC Barcelona	right	35,00 Mill. €	1,71 m	May 27, 2006	Spain
16	14	Thiago	60444	Central Midfield	Apr 11, 1991 (25)	27	Bayern Munich	right	25,00 Mill. €	1,74 m	Aug 10, 2011	Spain
17	21	David Silva	35518	Attacking Midfield	Jan 8, 1986 (30)	281	Manchester City	left	33,00 Mill. €	1,70 m	Nov 15, 2006	Spain
18	8	Koke	74229	Left Wing	Jan 8, 1992 (24)	13	Atlético Madrid	right	60,00 Mill. €	1,76 m	Aug 14, 2013	Spain
19	11	Pedro	65278	Right Wing	Jul 28, 1987 (28)	631	Chelsea FC	both	23,00 Mill. €	1,67 m	May 29, 2010	Spain

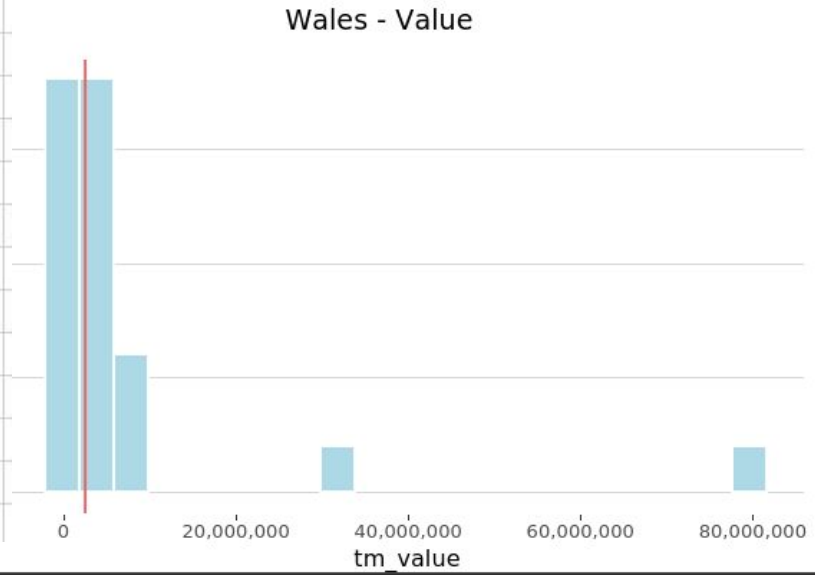
Output

What is the distribution of player value per team?



tm_value
40000000
15000000
3000000
40000000
40000000
8000000
35000000
20000000
18000000
16000000
60000000
10000000
8000000
45000000
35000000
25000000
33000000
60000000
23000000



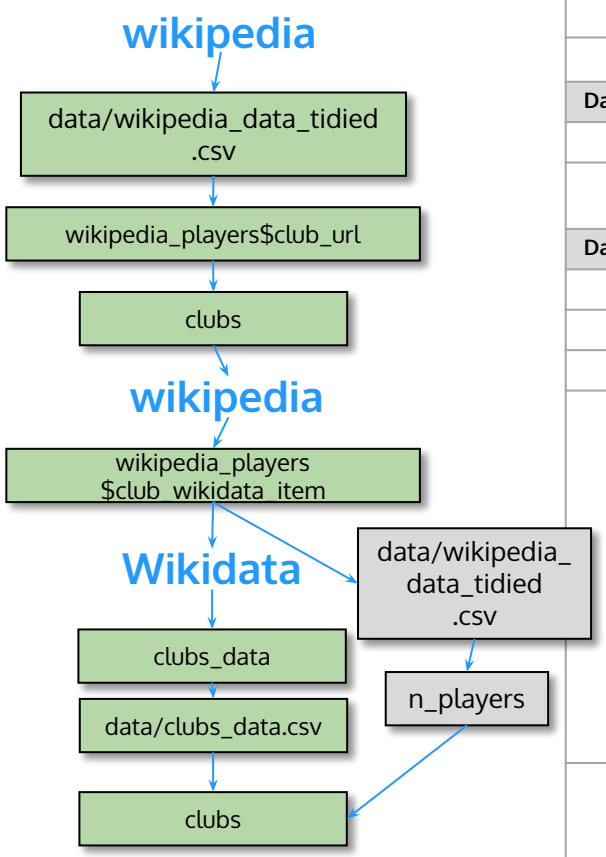


Additional Notes

Where are the clubs located the players play for (geolocation of home ground)? + additional club information

- **Challenges**
  - Html was not nice to scrape
  - Splitting, formatting and multiplying the value
- **Extra Work** (Functions that were not covered in class)
  - Data.table (and everything belonging to that package including the syntax for data manipulation)
  - Also rvest, lubridate, dplyr, tidyr
  - Custom functions
  - Shiny for presentation

OVERVIEW



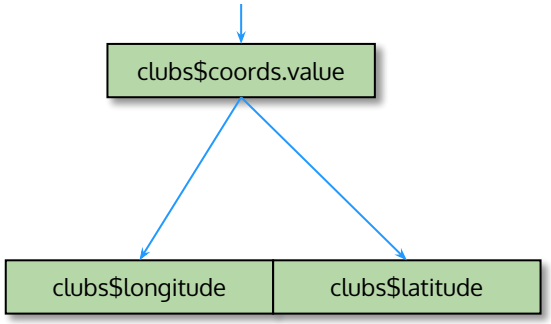
Where are the clubs located the players play for (geolocation of home ground)?

Data Output		
	Name	clubs_data$latitude,clubs_data$longitude
	Type	numeric
Data Input 1 (from collected data stored in local csv)		
	name	wikipedia_players$club_url (Input for the collected data = Wikipedia)
	type	Data.table element (for original Input = html tag)
Data Input 2 (collected from the web / computed)		
	name	clubs
	type	Chr-vector of urls from visiting every club page on Wikipedia using the distinct URLs in wikipedia_players and applying a custom function to extract the Wikidata link.
Data Input 3 (collected from web)		
	Data Category	Linked Data
	Data Source (DS)	Wikidata
	URL	https://query.wikidata.org/sparql?query={SPARQL} (SPARQL endpoint)

	Location of data element in DS	query <- sprintf("SELECT ?team ?teamLabel ?venueLabel ?capacity ?coords ?team_twitter ?image WHERE {
		VALUES (?team) {%s}.
		?team wdt:P641 wd:Q2736.
		?team wdt:P115 ?venue.
		?venue wdt:P625 ?coords.
		?venue wdt:P1083 ?capacity.
		OPTIONAL{
		?team wdt:P2002 ?team_twitter.
		}
		OPTIONAL{
		?team wdt:P18 ?image.
		}
		
		SERVICE wikibase:label{bd:serviceParam wikibase:language 'en'.}}", formatted_ids)
		
		The VALUES keywords makes it possible to pre-filter a query. In this case by a list of 218 possible entity IDs for ?team. The final query can be found in “data/query.txt”.

	Data quality	(+) most clubs are found (+) most have coords

Continued



Internal Flow		
	Activity 1	Read csv
	Activity 2	• Get the unique club_url values • Visit every URL and get the wikidata link -> add to the data.table
	Activity 3	Construct a query with a list of item IDs to insert into VALUES
	Activity 4	Query data
	Activity 5	Parse the returned json
	Activity 6	Seperate the coords.value into two new columns longitude and latitude by splitting the string and extracting the numeric values
Functions		
	custom function	get_wikidata_link
	contains:	unique, sprintf (base); seperate (dplyr);str_c, str_detect, str_extract (stringr); read_html, html_node, html_attr (rvest); now (lubridate); fread, data.table (data.table); GET, content (httr); fromJSON (jsonlite)

Data Input Samples mentioned in RQ OVERVIEW slide before

Data Input		
	Data Format	**HTML TABLE**
	Data Category	Internet
	Data Source	https://en.wikipedia. org/wiki/UEFA_Euro_2016_squads
	Data element in DS	Every table with class “table.sortable.wikitable”
	Data quality	Fairly good
	Selection reason	Possibility to get to the wikidata links for every player and team
Data Input		
	Data Format	**SPARQL**
	Data Category	Semantic Web
	Data Source	wikidata.org
	Data element in DS	See query above
	Data quality	good
	Selection reason	Easiest to get all locations

Data Input		
	Data Format	**HTML TAG**
	Data Category	Internet
	Data Source	https://en.wikipedia. org/wiki/UEFA_Euro_2016_squads
	Data element in DS	<td style="border:0"><img alt=" England" src="//upload.wikimedia. org/wikipedia/en/thumb/b/be/Flag_of_England. svg/23px-Flag_of_England.svg.png" width="23" height="14" class="thumbborder" srcset="//upload. wikimedia. org/wikipedia/en/thumb/b/be/Flag_of_England. svg/35px-Flag_of_England.svg.png 1.5x, //upload. wikimedia. org/wikipedia/en/thumb/b/be/Flag_of_England. svg/46px-Flag_of_England.svg.png 2x" data-file-width="800" data-file-height="480"> Tottenham Hotspur</td> //td[7]/a/@href in every table
	Data quality	good
	Selection reason	
	Data Format	**R OUTPUT**
	Data name	clubs$longitude, clubs$latitude
	Data type	Data table: both numeric

INPUT

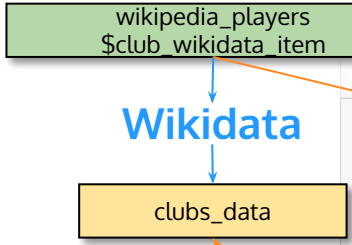
Where are the clubs located the players play for (geolocation of home ground)? + additional club information

wikipedia_players$clubs_url

#	Pos.	Player	Date of birth (age)	Caps	Goals	Club
1	GK	Hugo Lloris *(captain)*	26 December 1986 (aged 29)	75	0	Tottenham Hotspur
2	DF	Christophe Jallet	31 October 1983 (aged 32)	11	1	Lyon
3	DF	Patrice Evra	15 May 1981 (aged 35)	73	0	Juventus
4	DF	Adil Rami	27 December 1985 (aged 30)	28	1	Sevilla
5	MF	N'Golo Kanté	29 March 1991 (aged 25)	4	1	Leicester City
6	MF	Yohan Cabaye	14 January 1986 (aged 30)	46	4	Crystal Palace
7	FW	Antoine Griezmann	21 March 1991 (aged 25)	27	7	Atlético Madrid
8	MF	Dimitri Payet	29 March 1987 (aged 29)	19	3	West Ham United
9	FW	Olivier Giroud	30 September 1986 (aged 29)	49	17	Arsenal
10	FW	André-Pierre Gignac	5 December 1985 (aged 30)	27	7	UANL
11	FW	Anthony Martial	5 December 1995 (aged 20)	9	0	Manchester United
12	MF	Morgan Schneiderlin	8 November 1989 (aged 26)	15	0	Manchester United
13	DF	Eliaquim Mangala	13 February 1991 (aged 25)	7	0	Manchester City
14	MF	Blaise Matuidi	9 April 1987 (aged 29)	44	8	Paris Saint-Germain
15	MF	Paul Pogba	15 March 1993 (aged 23)	31	5	Juventus
16	GK	Steve Mandanda	28 March 1985 (aged 31)	22	0	Marseille
17	DF	Lucas Digne	20 July 1993 (aged 22)	13	0	Roma
18	MF	Moussa Sissoko	16 August 1989 (aged 26)	38	1	Newcastle United
19	DF	Bacary Sagna	14 February 1983 (aged 33)	57	0	Manchester City
20	FW	Kingsley Coman	13 June 1996 (aged 19)	5	1	Bayern Munich
21	DF	Laurent Koscielny	10 September 1985 (aged 30)	29	1	Arsenal
22	DF	Samuel Umtiti	14 November 1993 (aged 22)	0	0	Lyon
23	GK	Benoît Costil	3 July 1987 (aged 28)	0	0	Rennes

INPUT

Where are the clubs located the players play for (geolocation of home ground)? + additional club information



Wikidata Query Service — Examples — Prefixes — Tools — Help

```
SELECT ?team ?teamLabel ?venueLabel ?capacity ?coords ?team_twitter WHERE {
  VALUES (?team) {(wd:Q18741) (wd:Q704) (wd:Q1422) (wd:Q10329) (wd:Q19481) (wd:Q19467) (wd:Q8701) (wd:Q18747) (wd:Q9617) (wd:Q849823) (wd:Q18656) (wd:Q50602) (wd:Q483020) (wd:Q132885) (wd:Q2739) (wd:Q18716) (wd:Q1
  ?team wdt:P641 wd:Q2736.
  ?team wdt:P115 ?venue.
  ?venue wdt:P625 ?coords.
  ?venue wdt:P1083 ?capacity.
  OPTIONAL{
    ?team wdt:P2002 ?team_twitter.
    }
  SERVICE wikibase:label{bd:serviceParam wikibase:language 'en'.}}
```

Press *[CTRL-SPACE]* to activate auto completion. Data last updated: 00:24:26 CEST, 22 Jun 2016

▶ Run | Clear | 215 Results in 828 ms | Display | Download | Search

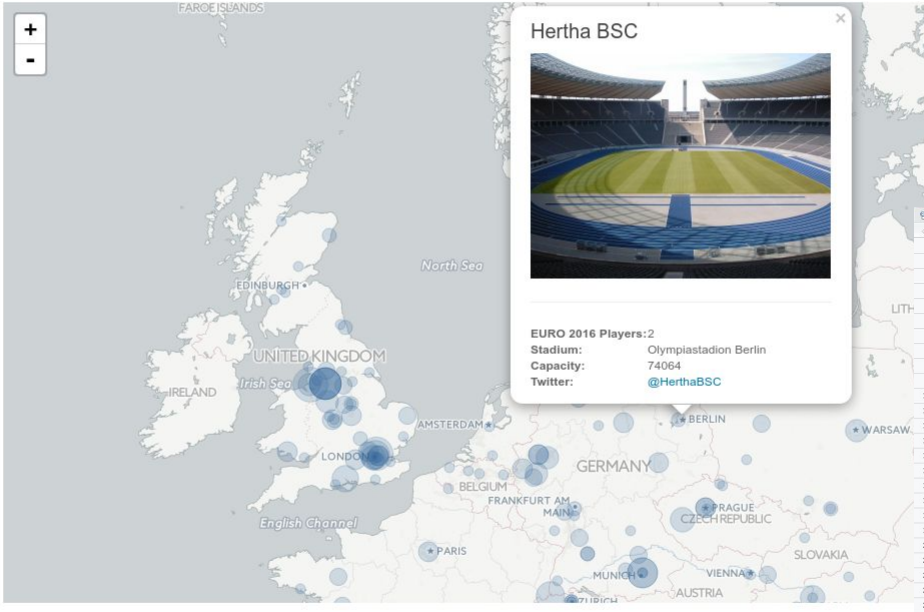
team	teamLabel	venueLabel	capacity	coords	team_twitter
wd:Q631	Football Club Internazionale Milano	Stadio Giuseppe Meazza	26000	Point(9.124 45.47808)	inter
wd:Q704	Olympique Lyonnais	Parc Olympique lyonnais	59186	Point(4.979825 45.764327777)	
wd:Q1422	Juventus F.C.	Juventus Stadium	41800	Point(7.641247 45.109594)	juventusfces
wd:Q1457	Unione Calcio Sampdoria	Stadio Luigi Ferraris	36703	Point(8.952428 44.416431)	sampdoria
wd:Q1543	A.C. Milan	Stadio Giuseppe Meazza	26000	Point(9.124 45.47808)	acmilan
wd:Q1893	Bologna F.C. 1909	Stadio Renato Dall'Ara	38279	Point(11.309722222 44.492222222)	
wd:Q1900	Cagliari Calcio	Stadio Is Arenas	16500	Point(9.181944 39.230833)	
wd:Q2052	ACF Fiorentina	Stadio Artemio Franchi	47282	Point(11.282258 43.780822)	violachannel
wd:Q2074	Genoa Cricket and Football Club	Stadio Luigi Ferraris	36703	Point(8.952428 44.416431)	GenoaCFC
wd:Q2609	S.S. Lazio	Stadio Olimpico	72698	Point(12.454714 41.933872)	officialsslazio
wd:Q2641	Napoli	Stadio San Paolo	76824	Point(14.193008 40.827967)	
wd:Q2674	U.S. Città di Palermo	Stadio Renzo Barbera	36349	Point(13.342275 38.152767)	palermocalcioit

Output

Where are the clubs located the players play for (geolocation of home ground)? + additional club information

Map of player clubs

A map of every club that has players participating in EURO 2016. The size of the marker is equivalent to the number of players the club sent to France. Click on a circle to see information about the club.



get_club_data.R × clubs_data ×

Filter

	team.value	capacity.value	coords.value	team_twitter.value	teamLabel.value	venueLabel.value	club_wikidata_id
1	http://www.wikidata.org/entity/Q631	26000	Point(9.124 45.47808)	inter	Football Club Internazionale Milano	Stadio Giuseppe Meazza	Q631
2	http://www.wikidata.org/entity/Q704	59186	Point(4.979825 45.764327777)	NA	Olympique Lyonnais	Parc Olympique lyonnais	Q704
3	http://www.wikidata.org/entity/Q1422	41800	Point(7.641247 45.109594)	juventusfces	Juventus F.C.	Juventus Stadium	Q1422
4	http://www.wikidata.org/entity/Q1457	36703	Point(8.952428 44.416431)	sampdoria	Unione Calcio Sampdoria	Stadio Luigi Ferraris	Q1457
5	http://www.wikidata.org/entity/Q1543	26000	Point(9.124 45.47808)	acmilan	A.C. Milan	Stadio Giuseppe Meazza	Q1543
6	http://www.wikidata.org/entity/Q1893	38279	Point(11.309722222 44.492222222)	NA	Bologna F.C. 1909	Stadio Renato Dall'Ara	Q1893
7	http://www.wikidata.org/entity/Q1900	16500	Point(9.181944 39.230833)	NA	Cagliari Calcio	Stadio Is Arenas	Q1900
8	http://www.wikidata.org/entity/Q2052	47282	Point(11.282258 43.780822)	violachannel	ACF Fiorentina	Stadio Artemio Franchi	Q2052
9	http://www.wikidata.org/entity/Q2074	36703	Point(8.952428 44.416431)	GenoaCFC	Genoa Cricket and Football Club	Stadio Luigi Ferraris	Q2074
10	http://www.wikidata.org/entity/Q2609	72698	Point(12.454714 41.933872)	officialsslazio	S.S. Lazio	Stadio Olimpico	Q2609
11	http://www.wikidata.org/entity/Q2641	76824	Point(14.193008 40.827967)	NA	Napoli	Stadio San Paolo	Q2641
12	http://www.wikidata.org/entity/Q2674	36349	Point(13.342275 38.152767)	palermocalcioit	U.S. Città di Palermo	Stadio Renzo Barbera	Q2674
13	http://www.wikidata.org/entity/Q2710	24400	Point(14.22944444 42.45527778)	PescaraCalcio	Delfino Pescara 1936	Stadio Adriatico	Q2710
14	http://www.wikidata.org/entity/Q2714	17477	Point(-0.401606 51.649928)	NA	Watford Football Club	Vicarage Road	Q2714
15	http://www.wikidata.org/entity/Q2739	72698	Point(12.454714 41.933872)	officialasroma	A.S. Roma	Stadio Olimpico	Q2739
16	http://www.wikidata.org/entity/Q2768	28140	Point(7.65 45.041667)	torinofc_1906	Torino Football Club	Stadio Olimpico di Torino	Q2768
17	http://www.wikidata.org/entity/Q2798	30642	Point(13.200136 46.081603)	Udinese_1896	Mońitos clup	Stadio Friuli	Q2798
18	http://www.wikidata.org/entity/Q4512	60441	Point(9.231944 48.792222)	NA	Q4512	Mercedes-Benz Arena	Q4512
19	http://www.wikidata.org/entity/Q5794	40157	Point(-2.966388888 53.438888888)	Everton	Everton F.C.	Goodison Park	Q5794
20	http://www.wikidata.org/entity/Q6664	23860	Point(12.261944 44.140556)	NA	A.C. Cesena	Stadio Dino Manuzzi	Q6664
21	http://www.wikidata.org/entity/Q6703	19847	Point(10.955 43.726389)	NA	Empoli F.C.	Stadio Carlo Castellani	Q6703
22	http://www.wikidata.org/entity/Q7156	99354	Point(2.1225 41.380833)	fcbarcelona_es	FC Barcelona	Camp Nou	Q7156
23	http://www.wikidata.org/entity/Q8466	49780	Point(7.775833 49.434444)	rote_teufel	1. FC Kaiserslautern	Fritz-Walter-Stadion	Q8466
24	http://www.wikidata.org/entity/Q8603	21151	Point(10.923333 44.653611)	SassuoloUS	U.S. Sassuolo Calcio	Stadio Alberto Braglia	Q8603
25	http://www.wikidata.org/entity/Q8682	85454	Point(-3.68835 40.45306)	Realmadrid	Real Madrid Club de Fútbol	Estadio Santiago Bernabéu	Q8682
26	http://www.wikidata.org/entity/Q8687	83332	Point(-2.949369 43.264205)	NA	Athletic Club	San Mamés Stadium	Q8687
27	http://www.wikidata.org/entity/Q8701	54960	Point(-3.720556 40.401667)	Atleti	Atlético Madrid	Vicente Calderón Stadium	Q8701
28	http://www.wikidata.org/entity/Q8749	32	Point(-8.740328 42.211861)	rccelta_oficial	Celta Vigo	Balaídos	Q8749
29	http://www.wikidata.org/entity/Q8857	30044	Point(-4.426422 36.734092)	MalagaCF	Málaga CF	La Rosaleda Stadium	Q8857
30	http://www.wikidata.org/entity/Q9616	41800	Point(-0.191111111 51.481666666)	chelseafc	Chelsea F.C.	Stamford Bridge stadium	Q9616
31	http://www.wikidata.org/entity/Q9617	60338	Point(-0.108611111 51.555)	Arsenal	Arsenal F.C	Emirates Stadium	Q9617
32	http://www.wikidata.org/entity/Q9653	30311	Point(1.144830555 52.055061111)	NA	Ipswich Town F.C.	Portman Road	Q9653
33	http://www.wikidata.org/entity/Q10300	14708	Point(-3.658961 40.391944)	NA	Rayo Vallecano	Campo de Fútbol de Vallecas	Q10300
34	http://www.wikidata.org/entity/Q10333	53000	Point(-0.358333 39.474722)	valenciacf	Valencia CF	Mestalla Stadium	Q10333
35	http://www.wikidata.org/entity/Q10499	21822	Point(-4.764642 37.872294)	cordobacfsad	Córdoba CF	Estadio Nuevo Arcángel	Q10499
36	http://www.wikidata.org/entity/Q12297	22000	Point(-0.103611 39.944167)	NA	Villarreal Club de Fútbol	Estadio El Madrigal	Q12297

Showing 1 to 36 of 215 entries

Additional Notes

Where are the clubs located the players play for (geolocation of home ground)? + additional club information

- **Challenges**
  - Querying SPARQL endpoints with multiple possible values for a variable (finding out about and using the VALUES keyword
  - Getting the wikidata item for every team from a list of wikipedia pages in a data.table
  - Merging the data
- **Extra Work** (Functions that were not covered in class)
  - Data.table (and everything belonging to that package including the syntax for data manipulation)
  - Also rvest, lubridate, dplyr, tidyr, jsonlite
  - Custom functions
  - Additional SPARQL keyword VALUES
  - Shiny & leaflet for presentation

Agenda

- Main Research Question
- Sub Research Questions
- **Conclusion**

Conclusion

- **How good was the data that could be used as input?**
  - Data quality overall was ok and data was plentiful
  - Data had to be extensively cleaned
- **How good is the data that could be computed?**
  - The data for the map is very good
  - The data for the histograms is ok, quality of the computation depends on user input for bins
- **How much more work would need to be put into this project to come to a reliable answer?**
  - There is not a reliable answer to my rather vague question. The sub questions have been answered rather reliably. The data seems reliable.
  - Too much time was spent gathering data and figuring out shiny and time in class was almost always spent helping others (and in preparation of tutorials) which I loved to do,

Conclusion

- **Outlook: How could this research questions further be extended?**
  - There was way more data gathered than i ended up using, so there is a solid foundation for developing the shiny app further.
  - It remains to be seen, if some of the code for scraping the information will be re-usable come the next European or World Cup.
  - If the layout of the “special” pages stays mostly the same (which is doubtful), it would hopefully be easy to gather the information again with minimal effort
  - Not only the shiny app itself, but also just the map can be extended, either with information already available or by altering the SPARQL query to include further geographical information

Additional notes

- **Overall Problems**
  - I focused too much on data gathering and cleaning
  - There is much unused potential in the data
  - I wasted much time, trying to unify all the data from my different sources and couldn't get it to work because a common key was missing and couldn't be constructed artificially by me
  - I've written much more code than this that ultimately had to be discarded
  - Shiny development is quite finicky and it isn't always clear right away what went wrong especially when deploying to shinyapp.io
  - I started out commenting my code quite extensively, but as time went on and frustration grew, I stopped commenting as much. But there is still much additional information in the files themselves and I also tried my best to write clear and concise code

Requirements

CHECK	REQUIREMENT		Files
yes	**SPARQL Query**		data_gathering/get_club_data.R
yes	**Parse HTML table**		data_gathering/get_wikipedia_data.R
yes	**Parse HTML tags**		All files in “data_gathering”
	Data Frame/Data Table	Select some columns	All files in “data_cleaning”; server.R
		Rename columns	data_cleaning/tidy_club_data.R a.o.
		Filter rows	All files in “data_cleaning”; server.R
		Combine two data frames (e.g., with merger)	data_cleaning/tidy_club_data.R
unclear	**Connect two web pages**		Not sure. data_cleaning/tidy_club_data.R
yes	**User defined function**		setup.R
	Get one URL from one website and use it to call another website		All files in “data_gathering”
	Manipulate Strings		All files in “data_cleaning”

Overview Applied Functions (for every Rscript.file)

setup.R

	package:UDFs	package:rvest	package:stringr	package:utils	package:xml2
1	as.list	html_attr	str_extract	install.packages	read_html
2	c	html_node	str_extract_all	*NA*	*NA*
3	data.frame	html_nodes	*NA*	*NA*	*NA*
4	library	html_text	*NA*	*NA*	*NA*
5	list	*NA*	*NA*	*NA*	*NA*
6	names	*NA*	*NA*	*NA*	*NA*
7	options	*NA*	*NA*	*NA*	*NA*
8	require	*NA*	*NA*	*NA*	*NA*
9	sapply	*NA*	*NA*	*NA*	*NA*
10	tolower	*NA*	*NA*	*NA*	*NA*
11	unique	*NA*	*NA*	*NA*	*NA*

Overview Applied Functions (for every Rscript.file)

server.R

	package:UDFs	package:base	package:data.table	package:ggplot2	package:ggthemes	package:leaflet	package:plotly	package:shiny	package:stats	package:stringr
1	select_data1	library	fread	aes	theme_hc	addCircleMarkers	ggplotly	reactive	median	str_c
2	select_data2	options	*NA*	geom_histogram	*NA*	addProviderTiles	renderPlotly	shinyServer	*NA*	*NA*
3	*NA*	return	*NA*	geom_vline	*NA*	leaflet	*NA*	*NA*	*NA*	*NA*
4	*NA*	sqrt	*NA*	ggplot	*NA*	renderLeaflet	*NA*	*NA*	*NA*	*NA*
5	*NA*	unique	*NA*	ggtitle	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
6	*NA*	*NA*	*NA*	scale_x_continuous	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*

Overview Applied Functions (for every Rscript.file)

ui.R

	package:UDFs	package:data.table	package:leaflet	package:plotly	package:shiny
1	library	fread	leafletOutput	plotlyOutput	column
2	options	*NA*	*NA*	*NA*	fluidPage
3	order	*NA*	*NA*	*NA*	fluidRow
4	*NA*	*NA*	*NA*	*NA*	h3
5	*NA*	*NA*	*NA*	*NA*	mainPanel
6	*NA*	*NA*	*NA*	*NA*	p
7	*NA*	*NA*	*NA*	*NA*	selectInput
8	*NA*	*NA*	*NA*	*NA*	shinyUI
9	*NA*	*NA*	*NA*	*NA*	sidebarLayout
10	*NA*	*NA*	*NA*	*NA*	sidebarPanel
11	*NA*	*NA*	*NA*	*NA*	sliderInput
12	*NA*	*NA*	*NA*	*NA*	tabPanel
13	*NA*	*NA*	*NA*	*NA*	tabsetPanel
14	*NA*	*NA*	*NA*	*NA*	titlePanel

Overview Applied Functions (for every Rscript.file)

data_gathering/get_data.R

	package:UDFs	package:data.table	package:lubridate	package:rvest	package:stringr	package:utils	package:xml2
1	colnames	rbindlist	now	html_attr	str_replace_all	write.csv	read_html
2	lapply	*NA*	*NA*	html_nodes	str_to_lower	*NA*	*NA*
3	sapply	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
4	unique	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
5	unlist	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*

Overview Applied Functions (for every Rscript.file)

data_gathering/get_tm_data.R

	package:UDFs	package:data.table	package:lubridate	package:rvest	package:stringr	package:utils	package:xml2
1	lapply	rbindlist	now	html_attr	str_c	write.csv	read_html
2	sapply	*NA*	*NA*	html_nodes	str_replace	*NA*	*NA*
3	unique	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*

Overview Applied Functions (for every Rscript.file)

data_gathering/get_wikipedia_data.R

	package:UDFs	package:base	package:data.table	package:lubridate	package:rvest	package:utils	package:xml2
1	get_wikidata_link	lapply	rbindlist	now	html_attr	write.csv	read_html
2	*NA*	list	*NA*	*NA*	html_nodes	*NA*	*NA*
3	*NA*	rep	*NA*	*NA*	html_table	*NA*	*NA*
4	*NA*	unique	*NA*	*NA*	html_text	*NA*	*NA*
5	*NA*	unlist	*NA*	*NA*	*NA*	*NA*	*NA*

Overview Applied Functions (for every Rscript.file)

data_gathering/get_club_data.R

	package:UDFs	package:data.table	package:httr	package:jsonlite	package:lubridate	package:stringr	package:utils
1	colnames	data.table	content	fromJSON	now	str_c	write.csv
2	list	fread	GET	*NA*	*NA*	str_detect	*NA*
3	sapply	*NA*	*NA*	*NA*	*NA*	str_extract	*NA*
4	sprintf	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
5	unique	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
6	write	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*

Overview Applied Functions (for every Rscript.file)

data_cleaning/tidy_data.R

	package:UDFs	package:base	package:countrycode	package:data.table	package:dplyr	package:lubridate	package:stringr	package:tidyr	package:utils
1	contains	as.numeric	countrycode	fread	mutate	dmy	str_replace_all	extract_numeric	write.csv
2	*NA*	c	*NA*	set	select	*NA*	*NA*	separate	*NA*
3	*NA*	colnames	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*

Overview Applied Functions (for every Rscript.file)

data_cleaning/tidy_tm_data.R

	package:UDFs	package:base	package:data.table	package:dplyr	package:lubridate	package:stringr	package:tidyr	package:utils
1	month_to_num	c	fread	mutate	ymd	str_replace	extract_numeric	write.csv
2	*NA*	*NA*	*NA*	*NA*	*NA*	str_replace_all	separate	*NA*
3	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	unite	*NA*

Overview Applied Functions (for every Rscript.file)

data_cleaning/tidy_wikipedia_data.R

	package:UDFs	package:data.table	package:lubridate	package:stringr	package:tidyr	package:utils
1	colnames	fread	ymd	str_extract	extract_numeric	write.csv
2	tolower	setnames	*NA*	str_replace	*NA*	*NA*
3	trimws	*NA*	*NA*	str_replace_all	*NA*	*NA*

Overview Applied Functions (for every Rscript.file)

data_cleaning/tidy_club_data.R

	package:UDFs	package:data.table	package:stringr	package:tidyr	package:utils
1	c	fread	str_c	extract_numeric	write.csv
2	is.na	setkey	str_extract	separate	*NA*
3	list	*NA*	*NA*	*NA*	*NA*
4	merge	*NA*	*NA*	*NA*	*NA*