Football Findings

This work was made in order to find useful insights into football.

We hope to find external and internal factors that could affect results in general, such as time, League, playing as Local, etc.

That could be useful for making more informed decisions in bets.

The data used in this work was scraped from https://www.espn.com/soccer/ (https://www.espn.com/soccer/) and organized by JOSEPH MOHR. It includes data for over 2300 matches and information about the top 6 European Leagues.

Importing libraries

For this work, we are going to use different cleaning and manipulation tools and functions from different libraries. Let's import them at the beginning.

```
library(ggplot2)
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(readr)
library(skimr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Importing data

For this case study, we are going to use two tables.

Importing the first table

This table includes data about all the matches of the most competitive 6 Leagues in Europe from 2000 to 2021.

```
matches <- read.csv("matches.csv")
```

Importing the second table

This table has data about the number of games played, won and lost matches, number of points of each team, etc. For the seasons from 2000 to 2021.

```
summary <- read.csv("all_tables.csv")
```

Analyzing the distribution of the first table.

We are going to start working with the table called “matches”.

```
 skim_without_charts(matches)
```

```
## Warning: There was 1 warning in `dplyr::summarize()`.
## ℹ In argument: `dplyr::across(tidyselect::any_of(variable_names),
##   mangled_skimmers$funs)`.
## ℹ In group 0: .
## Caused by warning:
## ! There were 150 warnings in `dplyr::summarize()`.
## The first warning was:
## ℹ In argument: `dplyr::across(tidyselect::any_of(variable_names),
##   mangled_skimmers$funs)`.
## Caused by warning in `sorted_count()`:
## ! Variable contains value(s) of "" that have been converted to "empty".
## ℹ Run ¤]8;;ide:run:dplyr::last_dplyr_warnings()☎dplyr::last_dplyr_warnings()¤]8;;☎ to see the 149 remaining w
arnings.
```

Data summary

Name	matches
Number of rows	25724
Number of columns	210

_____

Column type frequency:	
factor	159
numeric	51
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
home	0	1	FALSE	204	Man: 400, Tot: 400, Eve: 399, Ars: 398
away	0	1	FALSE	203	Ars: 400, Liv: 400, Che: 399, Eve: 398
date	0	1	FALSE	1525	Sat: 82, Sat: 80, Sat: 77, Sat: 75
time..utc.	0	1	FALSE	75	15:: 3084, 14:: 2675, 13:: 2368, 14:: 1977
attendance	0	1	FALSE	13284	emp: 4320, 30,: 194, 15,: 191, 20,: 186
venue	0	1	FALSE	306	emp: 1519, Old: 381, Goo: 380, Anf: 379
league	0	1	FALSE	75	201: 383, 200: 381, 200: 381, 200: 380
part_of_competition	0	1	FALSE	18	na: 21755, Re: 1512, 20: 306, 20: 306
game_status	0	1	FALSE	4	FT: 25710, Aba: 12, AET: 1, FT-: 1
shootout	0	1	FALSE	2	Fal: 25723, Tru: 1
home_possessionPct	0	1	FALSE	74	emp: 2031, 50%: 884, 51%: 879, 55%: 875
away_possessionPct	0	1	FALSE	74	emp: 2031, 50%: 884, 49%: 879, 45%: 875
home_shotsSummary	0	1	FALSE	435	emp: 2031, 0 (: 394, 11 : 385, 12 : 378
away_shotsSummary	0	1	FALSE	341	emp: 2031, 10 : 521, 9 (: 513, 8 (: 478
home_goal_minutes	0	1	FALSE	10494	emp: 5791, 56’: 100, 67’: 98, 59’: 93

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
home_goal_scorers	0	1	FALSE	14440	emp: 5791, Ari: 25, Dar: 22, Jer: 22
away_goal_minutes	0	1	FALSE	7749	emp: 8307, 82’: 117, 81’: 116, 50’: 115
away_goal_scorers	0	1	FALSE	11475	emp: 8307, Cri: 25, Iag: 24, Pie: 23
home_starting_1_num	0	1	FALSE	105	1.0: 10382, 1: 3991, 13.: 3244, 25.: 1038
home_starting_1	0	1	FALSE	792	Man: 231, Pep: 226, Dav: 222, Pet: 220
home_starting_2_num	0	1	FALSE	119	5.0: 2183, 4.0: 2119, 6.0: 1766, 3.0: 1439
home_starting_2	0	1	FALSE	2011	Mat: 164, Iñi: 159, Die: 157, emp: 143
home_starting_3_num	0	1	FALSE	113	5.0: 2351, 4.0: 2045, 3.0: 1965, 6.0: 1529
home_starting_3	0	1	FALSE	2124	Rya: 146, Gar: 145, Ger: 144, emp: 143
home_starting_4_num	0	1	FALSE	113	3.0: 3454, 5.0: 1244, 6.0: 1011, 4.0: 980
home_starting_4	0	1	FALSE	2837	Lei: 159, emp: 143, Mar: 141, Fil: 139
home_starting_5_num	0	1	FALSE	126	2.0: 3325, 3.0: 1068, 22.: 916, 5.0: 912
home_starting_5	0	1	FALSE	2995	Jua: 189, Dan: 153, Kyl: 147, emp: 143
home_starting_6_num	0	1	FALSE	124	8.0: 1615, 6.0: 1066, 10.: 992, 14.: 927
home_starting_6	0	1	FALSE	3555	emp: 143, Ser: 141, Bru: 94, Raú: 89
home_starting_7_num	0	1	FALSE	122	8.0: 2488, 6.0: 1377, 10.: 1101, 14.: 1077
home_starting_7	0	1	FALSE	3057	emp: 143, And: 92, Fra: 86, Mic: 86
home_starting_8_num	0	1	FALSE	124	8.0: 1790, 10.: 1230, 7.0: 1075, 6.0: 1041
home_starting_8	0	1	FALSE	3716	emp: 143, Jua: 108, Luk: 101, Xav: 95
home_starting_9_num	0	1	FALSE	115	9.0: 2473, 10.: 1719, 7.0: 1496, 11.: 1305
home_starting_9	0	1	FALSE	3586	emp: 165, Rob: 126, Kar: 120, Har: 107
home_starting_10_num	0	1	FALSE	120	10.: 2371, 9.0: 2198, 11.: 1977, 7.0: 1774

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
home_starting_10	0	1	FALSE	3222	emp: 175, Cri: 171, Ant: 115, Son: 105
home_starting_11_num	0	1	FALSE	125	9.0: 2266, 10.: 1451, 11.: 1437, 7.0: 1378
home_starting_11	0	1	FALSE	2984	emp: 179, Lio: 147, Nic: 109, Mar: 93
home_bench_1_num	0	1	FALSE	145	1.0: 1583, 13.: 1384, 1: 966, 12.: 808
home_bench_1	0	1	FALSE	4615	emp: 142, Ste: 138, Boa: 108, Stu: 96
home_bench_1_minute	0	1	FALSE	6	na: 25578, emp: 142, 18': 1, 46': 1
home_bench_2_num	0	1	FALSE	146	1.0: 805, 13.: 784, 22.: 696, 1: 606
home_bench_2	0	1	FALSE	6550	emp: 143, Chr: 54, Adr: 50, Tob: 47
home_bench_2_minute	0	1	FALSE	12	na: 25570, emp: 143, 77': 2, 45': 1
home_bench_3_num	0	1	FALSE	147	13.: 650, 1.0: 596, 16.: 546, 21.: 537
home_bench_3	0	1	FALSE	7504	emp: 144, Dan: 48, Raf: 47, Joe: 44
home_bench_3_minute	0	1	FALSE	91	na: 24123, emp: 144, 45': 124, 77': 48
home_bench_4_num	0	1	FALSE	151	13.: 713, 19.: 673, 20.: 669, 15.: 660
home_bench_4	0	1	FALSE	7687	emp: 188, Cos: 37, Dav: 33, Raf: 31
home_bench_4_minute	0	1	FALSE	97	na: 23055, emp: 188, 45': 144, 76': 87
home_bench_5_num	0	1	FALSE	150	emp: 1114, 20.: 752, 8.0: 745, 19.: 699
home_bench_5	0	1	FALSE	7725	emp: 1114, Adr: 27, Jer: 27, Sho: 27
home_bench_5_minute	0	1	FALSE	105	na: 9578, 45': 1203, emp: 1114, 82': 468
away_starting_1_num	0	1	FALSE	107	1.0: 10330, 1: 3977, 13.: 3241, 25.: 1041
away_starting_1	0	1	FALSE	823	Man: 234, Pep: 228, Pet: 222, Dav: 211
away_starting_2_num	0	1	FALSE	113	5.0: 2176, 4.0: 2125, 6.0: 1785, 2.0: 1403
away_starting_2	0	1	FALSE	2067	Mat: 167, Iñi: 159, Die: 157, Ser: 147

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
away_starting_3_num	0	1	FALSE	117	5.0: 2348, 4.0: 2004, 3.0: 1968, 6.0: 1553
away_starting_3	0	1	FALSE	2186	Ger: 147, emp: 142, Gar: 139, Jos: 137
away_starting_4_num	0	1	FALSE	117	3.0: 2825, 3: 1182, 5.0: 962, 6.0: 838
away_starting_4	0	1	FALSE	2901	Lei: 152, Mar: 147, Nac: 147, emp: 144
away_starting_5_num	0	1	FALSE	123	2.0: 3294, 3.0: 1098, 5.0: 918, 22.: 830
away_starting_5	0	1	FALSE	3136	Jua: 183, Kyl: 157, emp: 150, Dan: 150
away_starting_6_num	0	1	FALSE	126	8.0: 1734, 6.0: 1195, 10.: 1122, 14.: 1034
away_starting_6	0	1	FALSE	3686	emp: 153, Ser: 151, Gar: 75, Cas: 74
away_starting_7_num	0	1	FALSE	125	8.0: 2353, 6.0: 1408, 14.: 1105, 10.: 992
away_starting_7	0	1	FALSE	3195	emp: 155, Fra: 90, Luc: 89, Gar: 88
away_starting_8_num	0	1	FALSE	123	8.0: 1752, 10.: 1197, 6.0: 1062, 7.0: 1059
away_starting_8	0	1	FALSE	3900	emp: 158, Luk: 102, Xav: 94, Jua: 87
away_starting_9_num	0	1	FALSE	125	9.0: 2404, 10.: 1576, 7.0: 1472, 11.: 1260
away_starting_9	0	1	FALSE	3796	emp: 181, Kar: 123, Rob: 119, Lui: 108
away_starting_10_num	0	1	FALSE	122	10.: 2375, 9.0: 2062, 11.: 1948, 7.0: 1720
away_starting_10	0	1	FALSE	3370	emp: 188, Cri: 148, Ant: 120, Son: 92
away_starting_11_num	0	1	FALSE	125	9.0: 2679, 10.: 1902, 11.: 1733, 7.0: 1698
away_starting_11	0	1	FALSE	3105	emp: 196, Lio: 141, Nic: 97, Way: 95
away_bench_1_num	0	1	FALSE	146	1.0: 1613, 13.: 1385, 1: 895, 12.: 780
away_bench_1	0	1	FALSE	4678	emp: 154, Ste: 142, Boa: 108, Jer: 99
away_bench_1_minute	0	1	FALSE	8	na: 25563, emp: 154, 75': 2, 60': 1
away_bench_2_num	0	1	FALSE	146	13.: 1208, 1.0: 1073, 22.: 780, 5.0: 724

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
away_bench_2	0	1	FALSE	6676	emp: 156, Adr: 57, Chr: 51, Ser: 46
away_bench_2_minute	0	1	FALSE	14	na: 25554, emp: 156, 66’: 2, 84’: 2
away_bench_3_num	0	1	FALSE	147	13.: 1014, 1.0: 810, 21.: 716, 15.: 690
away_bench_3	0	1	FALSE	7600	emp: 167, Raf: 51, Dan: 42, Iag: 41
away_bench_3_minute	0	1	FALSE	95	na: 24093, emp: 167, 45’: 102, 46’: 61
away_bench_4_num	0	1	FALSE	154	19.: 702, 13.: 691, 15.: 648, 20.: 647
away_bench_4	0	1	FALSE	7834	emp: 216, Cos: 35, Raf: 35, Joe: 31
away_bench_4_minute	0	1	FALSE	98	na: 22902, emp: 216, 45’: 167, 76’: 80
away_bench_5_num	0	1	FALSE	148	emp: 1257, 9.0: 779, 20.: 759, 8.0: 759
away_bench_5	0	1	FALSE	7786	emp: 1257, Jer: 32, Jam: 31, Raf: 28
away_bench_5_minute	0	1	FALSE	105	na: 9532, 45’: 1282, emp: 1257, 78’: 447
home_bench_6_num	0	1	FALSE	153	emp: 4096, 7.0: 800, 9.0: 791, 11.: 771
home_bench_6	0	1	FALSE	6743	emp: 4096, Adr: 30, Cla: 28, Raú: 25
home_bench_6_minute	0	1	FALSE	102	na: 6093, emp: 4096, 45’: 968, 79’: 473
away_bench_6_num	0	1	FALSE	142	emp: 4190, 9.0: 957, 7.0: 894, 11.: 869
away_bench_6	0	1	FALSE	6773	emp: 4190, Oli: 31, Ádá: 28, Adr: 26
away_bench_6_minute	0	1	FALSE	105	na: 5964, emp: 4190, 45’: 1060, 75’: 480
away_starting_12	0	1	FALSE	473	emp: 24797, Lui: 13, Jos: 12, Ped: 11
home_starting_12	0	1	FALSE	416	emp: 24857, Laf: 13, Lui: 13, Lui: 12
home_formation	0	1	FALSE	24	emp: 7155, 4-2: 5920, 4-4: 3593, 4-3: 2760
away_formation	0	1	FALSE	24	emp: 7170, 4-2: 5765, 4-4: 3224, 4-3: 2725
home_bench_7_num	0	1	FALSE	136	emp: 4492, 9.0: 1275, 11.: 1129, 7.0: 1052

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
home_bench_7	0	1	FALSE	6183	emp: 4492, Ped: 35, Mik: 33, Iva: 32
home_bench_7_minute	0	1	FALSE	108	emp: 4492, na: 3806, 45’: 1011, 77’: 542
away_bench_7_num	0	1	FALSE	129	emp: 4660, 9.0: 1223, 11.: 1124, 7.0: 1040
away_bench_7	0	1	FALSE	6306	emp: 4660, Cla: 35, Jav: 33, Adr: 32
away_bench_7_minute	0	1	FALSE	106	emp: 4660, na: 3669, 45’: 1182, 78’: 546
home_bench_8	0	1	FALSE	3463	emp: 18228, Giu: 17, Van: 16, Jos: 15
home_bench_8_minute	0	1	FALSE	102	emp: 18228, na: 2092, 45’: 430, 82’: 182
away_bench_8_num	0	1	FALSE	101	emp: 18243, 9.0: 338, 11.: 332, 7.0: 310
away_bench_8	0	1	FALSE	3462	emp: 18243, Jos: 17, Fed: 16, Ani: 13
away_bench_8_minute	0	1	FALSE	98	emp: 18243, na: 1959, 45’: 482, 78’: 175
away_bench_9	0	1	FALSE	2976	emp: 19134, Ada: 14, Ama: 14, Mar: 13
away_bench_9_minute	0	1	FALSE	98	emp: 19134, na: 1079, 45’: 406, 71’: 191
away_starting_13	0	1	FALSE	194	emp: 25383, Álv: 9, Ari: 7, Lui: 7
home_bench_9	0	1	FALSE	3007	emp: 19031, Ama: 18, Fed: 16, Ale: 13
home_bench_9_minute	0	1	FALSE	96	emp: 19031, na: 1142, 45’: 385, 78’: 183
home_starting_13	0	1	FALSE	175	emp: 25417, Jos: 9, Pab: 8, Ped: 7
home_starting_14	0	1	FALSE	97	emp: 25600, Jua: 3, Lui: 3, Pab: 3
home_bench_10	0	1	FALSE	1774	emp: 21927, Gré: 16, Cri: 12, Zak: 11
home_bench_10_minute	0	1	FALSE	94	emp: 21927, 45’: 272, na: 135, 74’: 130
home_bench_11	0	1	FALSE	1373	emp: 22708, Gor: 15, Cri: 14, Pat: 14
home_bench_11_minute	0	1	FALSE	89	emp: 22708, 45’: 191, 74’: 98, 84’: 98
away_bench_10	0	1	FALSE	1752	emp: 21950, Gor: 16, Pio: 14, And: 11

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
away_bench_10_minute	0	1	FALSE	94	emp: 21950, 45’: 299, na: 146, 72’: 114
away_starting_14	0	1	FALSE	114	emp: 25569, Jes: 4, Die: 3, Jos: 3
League	0	1	FALSE	6	Eng: 7979, Spa: 6817, Ger: 6124, Ita: 2270
home_starting_15	0	1	FALSE	43	emp: 25674, Ped: 4, Edu: 2, Fre: 2
away_starting_15	0	1	FALSE	63	emp: 25654, Álv: 3, Lui: 3, Alb: 2
away_starting_16	0	1	FALSE	31	emp: 25691, Jos: 2, Jua: 2, Val: 2
away_starting_17	0	1	FALSE	15	emp: 25710, Alb: 1, Ber: 1, Giu: 1
away_bench_11	0	1	FALSE	1319	emp: 22851, Pat: 20, Kho: 15, Ark: 13
away_bench_11_minute	0	1	FALSE	84	emp: 22851, 45’: 170, 78’: 94, 80’: 94
away_bench_12	0	1	FALSE	917	emp: 23771, Gor: 16, Joa: 14, Sam: 14
away_bench_12_minute	0	1	FALSE	80	emp: 23771, 45’: 135, 77’: 65, 78’: 62
away_bench_13	0	1	FALSE	48	emp: 25675, Jua: 2, Mat: 2, Adr: 1
away_bench_13_minute	0	1	FALSE	34	emp: 25675, 69’: 5, 45’: 4, 64’: 4
home_starting_16	0	1	FALSE	18	emp: 25706, Die: 2, Adr: 1, Asi: 1
home_bench_12	0	1	FALSE	974	emp: 23628, Joa: 16, Fed: 14, Ric: 14
home_bench_12_minute	0	1	FALSE	87	emp: 23628, 45’: 132, 68’: 73, 73’: 70
home_starting_17	0	1	FALSE	8	emp: 25717, Fab: 1, Fra: 1, Jua: 1
home_bench_13	0	1	FALSE	42	emp: 25683, Alb: 1, Ale: 1, Ant: 1
home_bench_13_minute	0	1	FALSE	28	emp: 25683, 45’: 3, 64’: 3, 73’: 3
away_starting_18	0	1	FALSE	8	emp: 25717, Die: 1, Die: 1, Jan: 1
away_starting_19	0	1	FALSE	8	emp: 25717, Ebi: 1, Iva: 1, Lau: 1
away_bench_14	0	1	FALSE	6	emp: 25719, Ale: 1, Dav: 1, Del: 1

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
away_bench_14_minute	0	1	FALSE	6	emp: 25719, 55’: 1, 64’: 1, 65’: 1
home_starting_18	0	1	FALSE	5	emp: 25720, Dav: 1, Hal: 1, Lui: 1
home_starting_19	0	1	FALSE	4	emp: 25721, Hal: 1, Lui: 1, Mar: 1
home_bench_14	0	1	FALSE	5	emp: 25720, Cri: 1, Iñi: 1, Jua: 1
home_bench_14_minute	0	1	FALSE	5	emp: 25720, 63’: 1, 68’: 1, 82’: 1
away_starting_20	0	1	FALSE	6	emp: 25719, Cac: 1, Dav: 1, Mar: 1
away_starting_21	0	1	FALSE	5	emp: 25720, Dan: 1, Ebi: 1, Luk: 1
home_starting_20	0	1	FALSE	3	emp: 25722, Bou: 1, Mar: 1
home_starting_21	0	1	FALSE	3	emp: 25722, Bou: 1, Mar: 1
away_starting_22	0	1	FALSE	3	emp: 25722, Luk: 1, Rad: 1
home_bench_15	0	1	FALSE	2	emp: 25723, Ian: 1
home_bench_15_minute	0	1	FALSE	2	emp: 25723, 56’: 1

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
id	0	1.00	383131.25	161629.48	18123	252476.75	396775.5	522549.25	614925
year	0	1.00	2013.01	5.95	2001	2008.00	2014.0	2018.00	2021
home_score	0	1.00	1.57	1.32	0	1.00	1.0	2.00	10
away_score	0	1.00	1.20	1.17	0	0.00	1.0	2.00	13
home_foulsCommitted	2031	0.92	12.97	5.11	0	10.00	13.0	16.00	37
away_foulsCommitted	2031	0.92	13.31	5.23	0	10.00	13.0	17.00	36
home_yellowCards	2031	0.92	1.81	1.34	0	1.00	2.0	3.00	8

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
away_yellowCards	2031	0.92	2.09	1.39	0	1.00	2.0	3.00	9
home_redCards	2031	0.92	0.09	0.31	0	0.00	0.0	0.00	3
away_redCards	2031	0.92	0.12	0.35	0	0.00	0.0	0.00	4
home_offsides	2031	0.92	2.34	1.95	0	1.00	2.0	3.00	14
away_offsides	2031	0.92	2.16	1.87	0	1.00	2.0	3.00	17
home_wonCorners	2031	0.92	5.54	3.13	0	3.00	5.0	7.00	20
away_wonCorners	2031	0.92	4.40	2.73	0	2.00	4.0	6.00	20
home_saves	2031	0.92	2.83	2.05	0	1.00	3.0	4.00	18
away_saves	2031	0.92	3.48	2.37	0	2.00	3.0	5.00	18
away_starting_12_num	24797	0.04	15.12	7.96	2	9.00	14.0	19.00	91
home_starting_12_num	24857	0.03	14.97	7.02	2	10.00	14.0	19.00	69
home_bench_8_num	18228	0.29	22.82	18.81	1	11.00	19.0	28.00	99
away_bench_9_num	19134	0.26	22.43	18.79	1	10.00	19.0	27.00	99
away_starting_13_num	25383	0.01	14.32	6.47	2	9.00	12.0	19.00	43
home_bench_9_num	19031	0.26	22.85	19.58	1	10.00	19.0	27.00	99
home_starting_13_num	25417	0.01	13.95	6.29	2	9.00	12.0	18.00	40
home_starting_14_num	25600	0.00	14.93	6.14	2	10.00	14.0	19.00	33
home_bench_10_num	21927	0.15	22.60	20.73	1	10.00	18.0	26.00	99
home_bench_11_num	22708	0.12	23.27	21.49	1	10.00	17.0	27.00	99
away_bench_10_num	21950	0.15	22.55	20.84	1	9.00	17.0	26.00	99
away_starting_14_num	25569	0.01	14.72	6.29	5	9.00	14.0	19.00	33
home_starting_15_num	25674	0.00	14.74	6.75	5	9.25	13.0	21.00	33

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
away_starting_15_num	25654	0.00	14.74	6.89	2	9.00	14.5	20.00	35
away_starting_16_num	25691	0.00	16.97	8.48	7	10.00	16.0	22.00	38
away_starting_17_num	25710	0.00	17.14	6.01	7	12.00	18.5	21.75	27
away_bench_11_num	22851	0.11	23.28	21.49	1	10.00	17.0	27.00	99
away_bench_12_num	23771	0.08	22.75	21.47	1	10.00	17.0	25.00	99
away_bench_13_num	25675	0.00	20.33	21.03	5	9.00	14.0	22.00	99
home_starting_16_num	25706	0.00	14.61	6.28	6	9.25	12.0	18.75	27
home_bench_12_num	23628	0.08	23.27	21.79	1	10.00	17.0	27.00	99
home_starting_17_num	25717	0.00	20.00	9.20	7	14.50	18.0	26.00	34
home_bench_13_num	25683	0.00	22.22	19.88	2	9.00	20.0	25.00	90
away_starting_18_num	25717	0.00	16.00	8.89	7	9.00	14.0	22.00	29
away_starting_19_num	25717	0.00	15.00	6.27	8	9.50	14.0	20.50	23
away_bench_14_num	25719	0.00	13.80	6.87	7	9.00	11.0	19.00	23
home_starting_18_num	25720	0.00	16.75	6.85	7	15.25	18.5	20.00	23
home_starting_19_num	25721	0.00	15.67	4.93	10	14.00	18.0	18.50	19
home_bench_14_num	25720	0.00	11.50	3.42	8	9.50	11.0	13.00	16
away_starting_20_num	25719	0.00	16.80	9.78	8	11.00	14.0	18.00	33
away_starting_21_num	25720	0.00	21.00	12.38	10	13.00	18.0	26.00	38
home_starting_20_num	25722	0.00	13.50	6.36	9	11.25	13.5	15.75	18
home_starting_21_num	25722	0.00	13.50	6.36	9	11.25	13.5	15.75	18
away_starting_22_num	25722	0.00	15.00	7.07	10	12.50	15.0	17.50	20
home_bench_15_num	25723	0.00	24.00	NA	24	24.00	24.0	24.00	24

Comparing match results with time.

In this section, we analyze the relationship between the total of goals per match and the time it was taking place.

Changing to time format

The first step is changing the format of the column time because it is recognized as a string. We need R to recognize it as a time format.

For this, we are going to add a new column to the data frame called “time_converted”.

```
matches <- mutate(matches, time_converted=hm(time..utc.))
```

Calculating total goals per match

This data frame has data about the Local and Visit scores but, it does not have the total goals per match. For this, we add a new column where we calculate the total.

```
matches <- mutate(matches, total_goals = home_score + away_score)
```

Removing matches with more than 10 goals

Matches with more than 10 goals are extremely rare and not representative, which is why they can cause bias, so it has been decided to remove them.

```
matches <- filter(matches, total_goals <=10)
League_insights <-summarize(group_by(matches,League),count=n(), total_goals_mean=mean(total_goals))
View(League_insights)
```

Sun time

Our first analysis will be how the sun can affect the number of goals in a match.

Differentiating between sun time and non-sun time

First, we will create a new column called sun time, defining sun time as before 17:00 hours.

```
matches <- mutate(matches, sun_time = time_converted <= hm('17:00'))
```

Creating a new table with insights

This table will calculate some insights about the total goals in sun time and non-sun time

```
sun_time_sum <-summarize(group_by(matches,sun_time), goal_sum = sum(total_goals), count=n(),
                         goal_mean=mean(total_goals))
knitr::kable(sun_time_sum, format = "markdown")
```

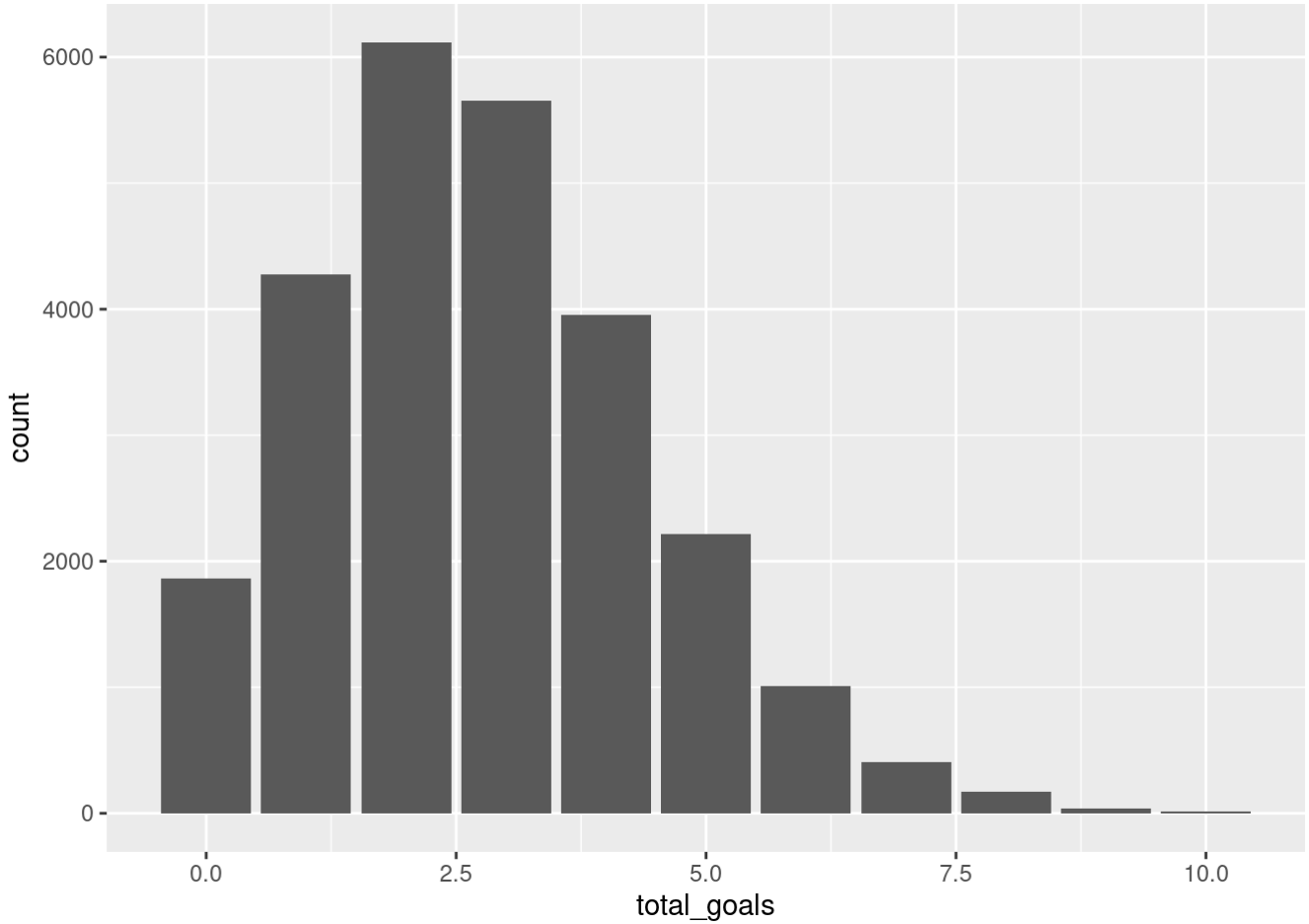
sun_time	goal_sum	count	goal_mean
FALSE	23578	8584	2.746738
TRUE	47535	17135	2.774147

We can not see much difference in the average total of goals.

Visualizing distributions of goals in football.

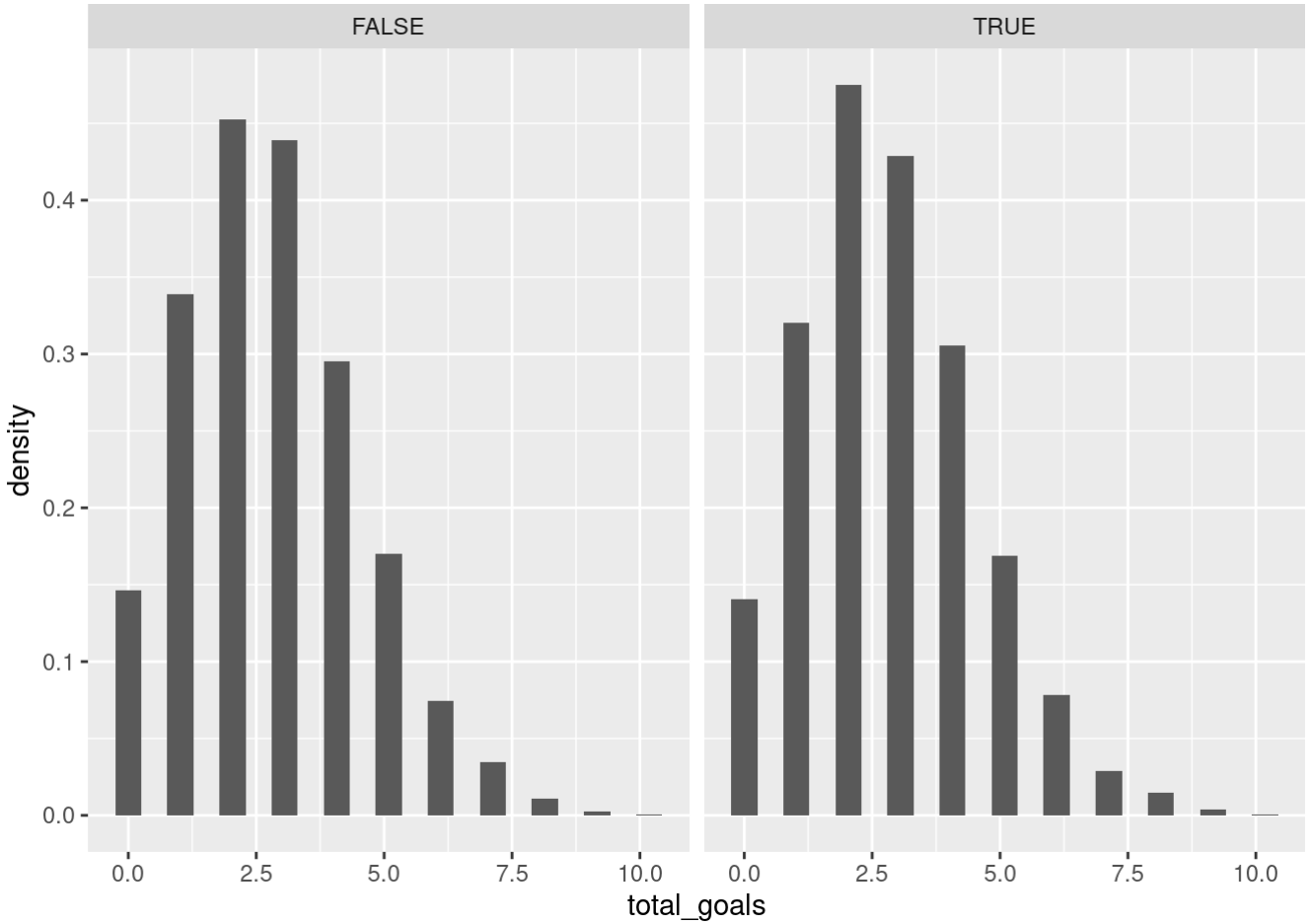
The table did not give us too much information. Therefore, we will compare the distribution of total goals depending on the time, but first, let's see how the totals of goals are distributed in general.

```
ggplot(data=matches)+
  geom_bar(mapping=aes(x=total_goals))
```



Visualizing if the distribution of goals depends on whether or not there is sunlight.

```
ggplot(data=matches)+
  geom_histogram(mapping=aes(x=total_goals, y= after_stat(density)),binwidth=0.509, inherit.aes=TRUE, bins=10)+
  facet_wrap(~sun_time)
```



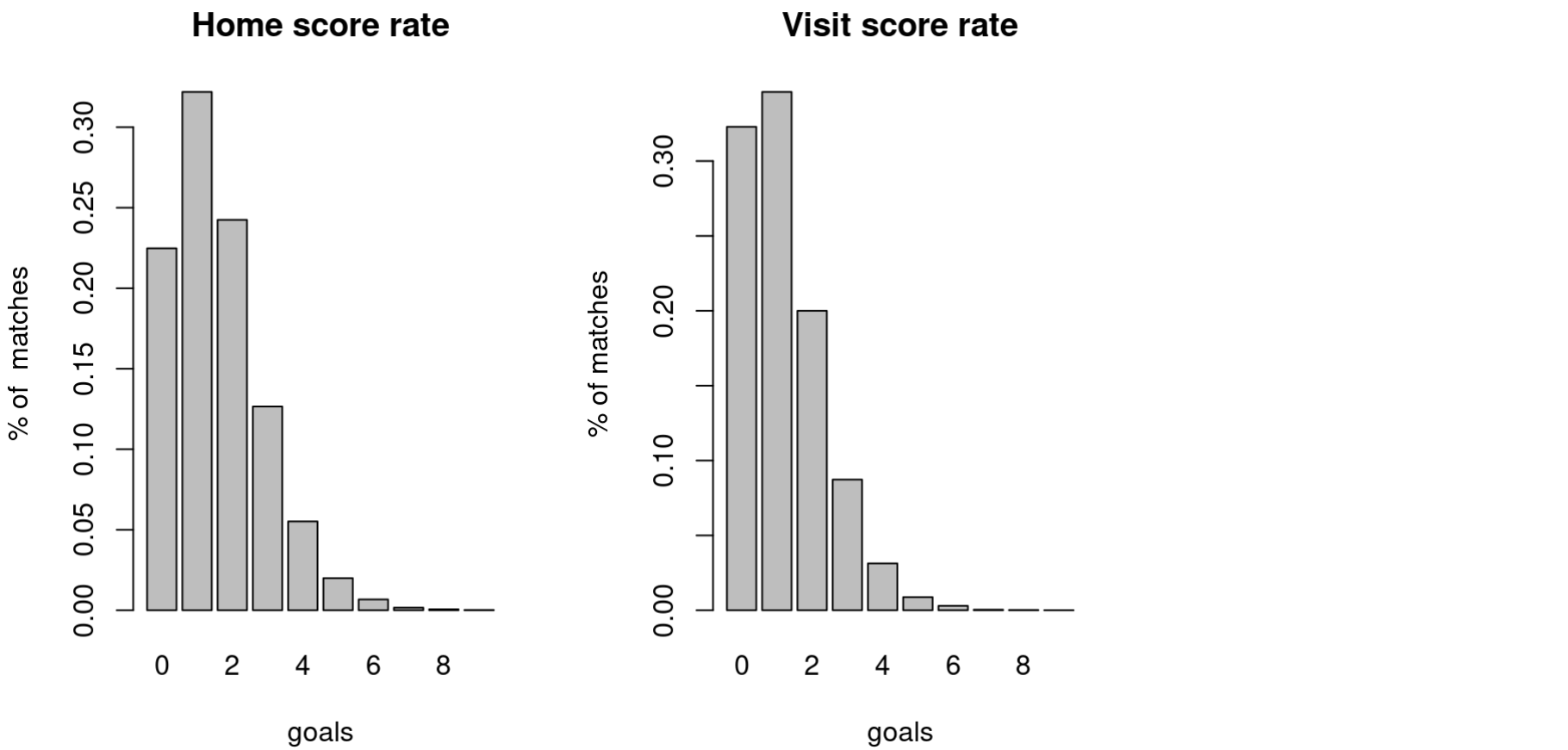
We can see the distributions are very similar but slightly different. That can be due to the different sample sizes.

Local and Visit Scores

In the summary of the first chart, we can see that the average local and visit scores were different.

The average home score is 1.57 goals meanwhile the visit one is 1.20 goals. Let’s see how scores are distributed depending on this factor. ### Visualizing Home score and Visit Score rate

```
par(mfrow = c(1, 2))
barplot(prop.table(table(matches$home_score)), main= "Home score rate", ylab= "% of  matches", xlab = "goals")
barplot(prop.table(table(matches$away_score)), main= "Visit score rate", ylab= "% of matches", xlab= "goals")
```




```
par(mfrow = c(1, 1))
```

We can appreciate a clear difference, being way more usual with a 0 score for visits in general.

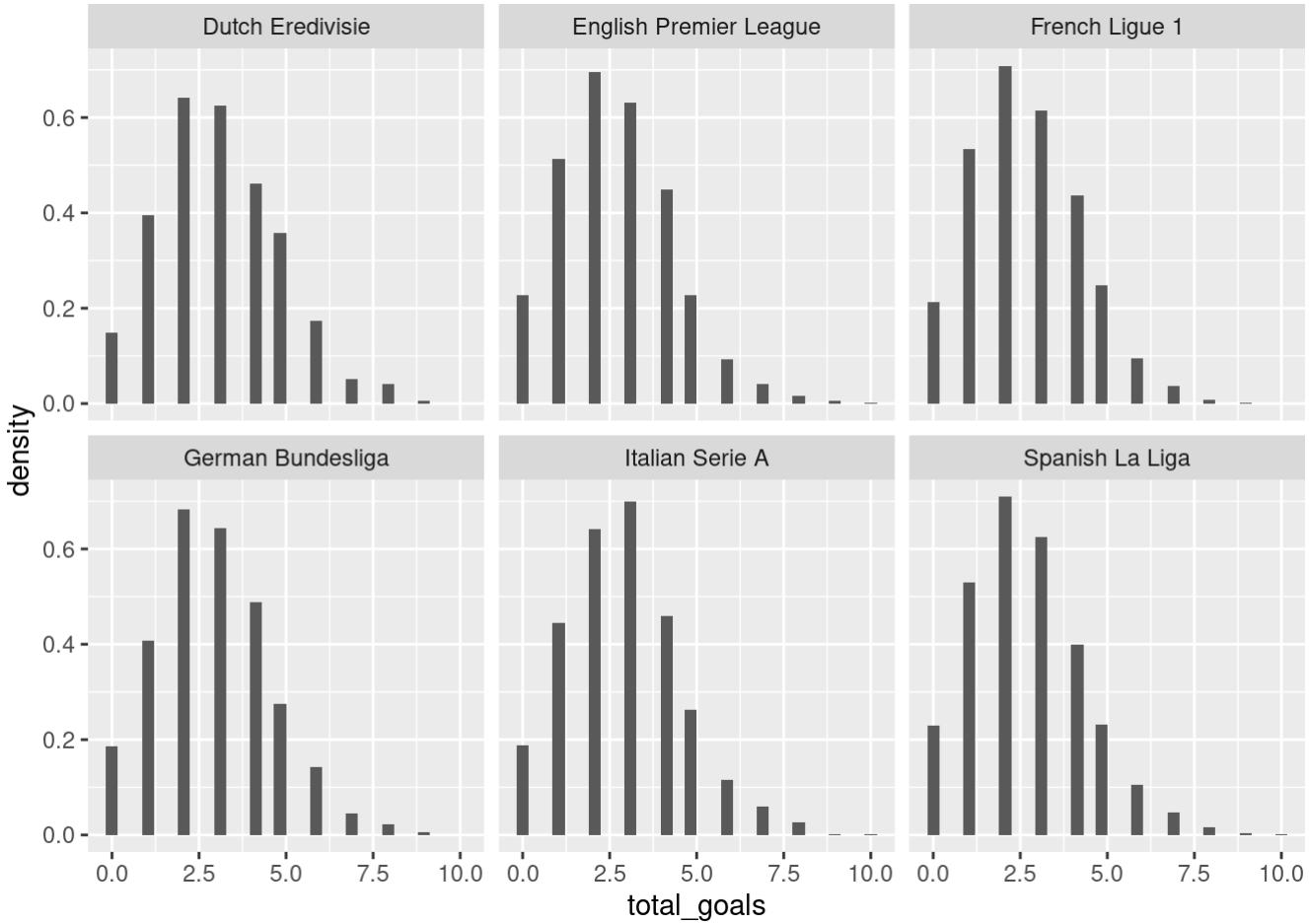
League comparation total goals

Let’s make a visual that shows the distribution of the total goals for each League.

Total goals rate in each European League (2000-2021)

```
ggplot(data = matches)+
  geom_histogram(mapping = aes(x = total_goals, y= after_stat(density)))+
  facet_wrap(~League)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can see some differences between them.

Sample sizes and the average total goals per match for each League

```
League <- arrange(summarize(group_by(matches,League), count = n(), total_goal_mean = mean(total_goals)), total_go
al_mean)
knitr::kable(League, format = "markdown")
```

League	count	total_goal_mean
French Ligue 1	1400	2.657143

League	count	total_goal_mean
Spanish La Liga	6816	2.666227
English Premier League	7978	2.677864
Italian Serie A	2270	2.872247
German Bundesliga	6122	2.913100
Dutch Eredivisie	1133	3.090909

In both the visual and table, we can appreciate that dutch Eredivisie has fewer matches with 0 or 1 goal than the other Leagues, plus the average total goals is higher. But the sample size is considerably smaller, so we can not take these results as conclusive, but it may be a good starting point for further investigations.

Analyzing the second table.

Now we will use the table “summary” to get insights. First let’s analyze its distribution.

```
skim_without_charts(summary)
```

Data summary

Name	summary
Number of rows	2484
Number of columns	12
_____	
Column type frequency:	
factor	2
numeric	10
_____	

Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Team	0	1	FALSE	227	MON: 38, LEV: 36, BOL: 29, BAR: 25
League	0	1	FALSE	6	Spa: 440, Ita: 432, Eng: 420, Fre: 400

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Place	0	1	10.15	5.59	1	5	10	15	20
GP	0	1	36.17	2.81	24	34	38	38	38
W	0	1	13.47	5.58	1	10	12	17	33
D	0	1	9.23	3.00	0	7	9	11	20
L	0	1	13.47	5.20	0	10	14	17	29
GF	0	1	48.22	17.59	0	38	46	57	121
GA	0	1	48.22	14.99	0	40	49	57	98
GD	0	1	0.00	24.07	-69	-16	-3	12	89
P	0	1	49.57	16.05	9	39	46	59	102
Year	0	1	2010.77	6.21	2000	2005	2011	2016	2021

Champions insights

Let’s create a new chart that gives us insights into the champions of all the Leagues.

First, we have to filter the table to only give us information about teams in the 1st position. Let’s remove the data for the years 2000 and 2019 since they include incomplete data.

Filtering champions

```
champions <- arrange(filter(summary, Place==1, Year>2000, Year != 2019), P, GP)
```

Given that we can create another table with the most important insights.

Summary stats of eache league

This table will show how different the stats of the champions of each League are within the period from 2000 to 2021.

```
insights_champions <- summarize(group_by(champions,League),count=n(), total_points_mean=mean(P), game_played_mean
=mean(GP),
                                Win_mean=mean(W), Lost_mean=mean(L), Draw_mean=mean(D), max(P), min(P), number_ch
ampions=n_distinct(Team))
knitr::kable(insights_champions, format = "markdown")
```

League	count	total_points_mean	game_played_mean	Win_mean	Lost_mean	Draw_mean	max(P)	min(P)	number_champions
Dutch Eredivisie	20	80.55000	34.00000	24.90000	3.250000	5.850000	88	71	5
English Premier League	20	88.85000	37.95000	27.45000	4.000000	6.500000	100	80	5
French Ligue 1	19	83.26316	37.94737	24.94737	4.578947	8.421053	96	68	7
German Bundesliga	20	77.95000	34.00000	23.95000	3.950000	6.100000	91	69	5
Italian Serie A	20	85.90000	37.35000	26.05000	3.550000	7.750000	102	71	3
Spanish La Liga	20	87.90000	37.95000	27.00000	4.050000	6.900000	100	75	4

Most winning teams in Europe

To see which teams have more titles, we can create a visualization. But first let's filter our table "champions" ton only contain info about teams with 3 or more titles.

```
insights_champions_team <- summarize(group_by(champions, Team, League),count=n(), total_points_mean=mean(P), game
_played_mean=mean(GP),
                                     Win_mean=mean(W), Lost_mean=mean(L), Draw_mean=mean(D), max(P), min(P))
```

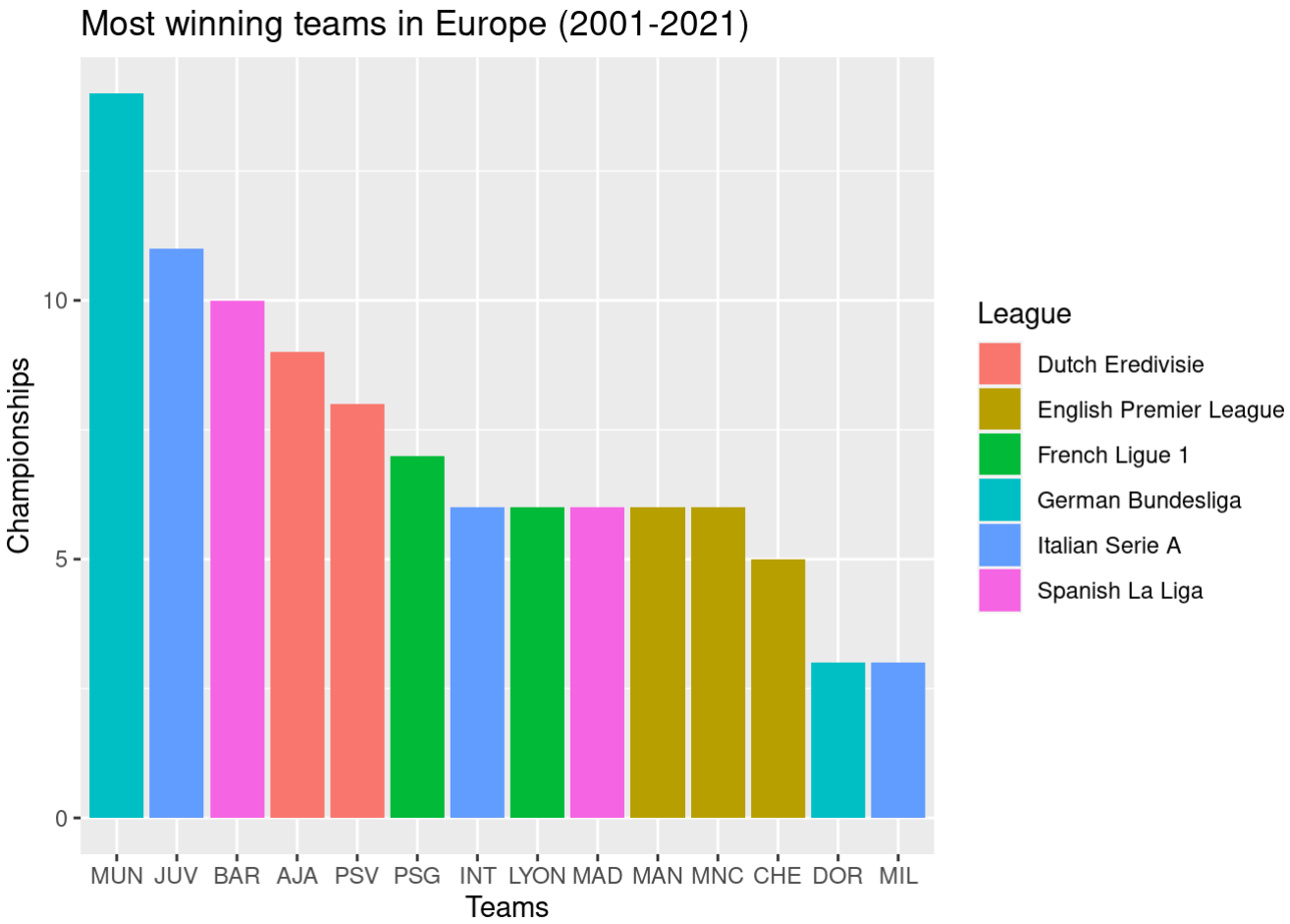
```
## `summarise()` has grouped output by 'Team'. You can override using the
## `.groups` argument.
```

```
insights_champions_winners <- filter(insights_champions_team, count >= 3)
```

Now let's create a viz showing this.

Showing the most winning teams.

```
ggplot(data=insights_champions_winners)+
  geom_bar(mapping=aes(x = reorder(Team, -count), y = count, fill = League), stat = "identity", position="dodge")
+
  labs(
    x = "Teams",
    y = "Championships",
    title = "Most winning teams in Europe (2001-2021)"
    )
```

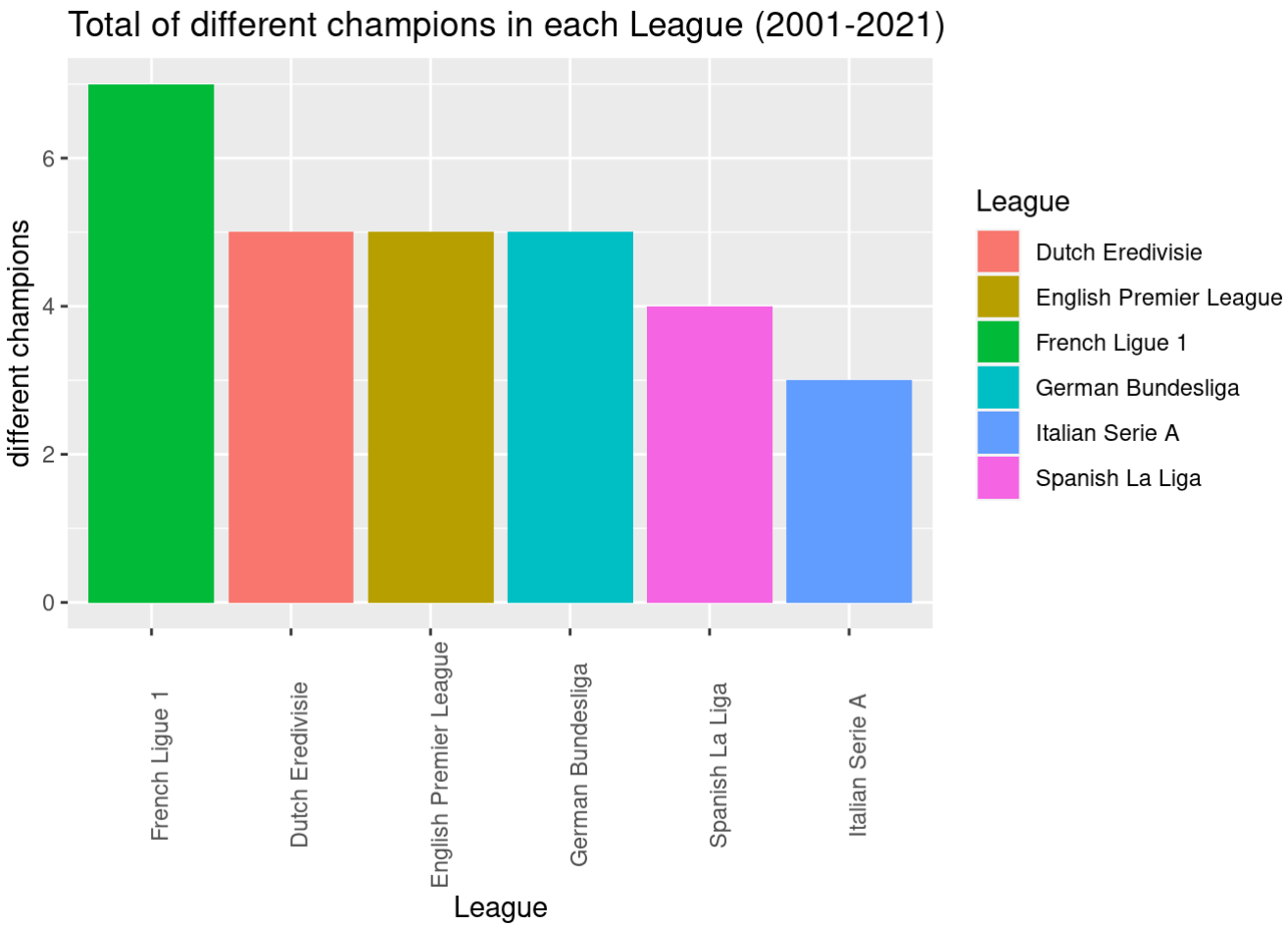


We can appreciate that Bayern Munich is the most winning team in Europe from 2000 to 2021.

Most dominated Leagues

Let’s create a Viz which tells us how many different teams have won their League in the period (2000-2021)

```
ggplot(data=insights_champions)+
  geom_bar(mapping=aes(x = reorder(League, -number_champions), y = number_champions, fill = League), stat = "iden
tity", position="dodge")+
  labs(
    x = "League",
    y = "different champions",
    title = "Total of different champions in each League (2001-2021)"
  )+
  theme(axis.text.x = element_text(angle = 90))
```



We can see that Italian Serie A is the most-dominated League, having only three different champions in more than 20 years.

Rate of championships

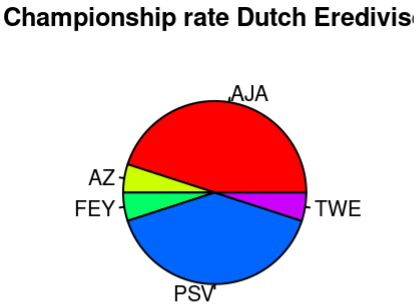
First, let's create a table for each League containing all the champions.

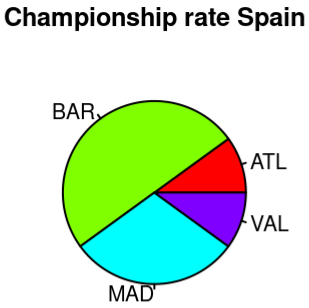
```
Dutch <- (filter(insights_champions_team, League == "Dutch Eredivisie"))
Spanish <- (filter(insights_champions_team, League == "Spanish La Liga"))
English <- (filter(insights_champions_team, League == "English Premier League"))
German <- (filter(insights_champions_team, League == "German Bundesliga"))
Italian <- (filter(insights_champions_team, League == "Italian Serie A"))
French <- (filter(insights_champions_team, League == "French Ligue 1"))
```

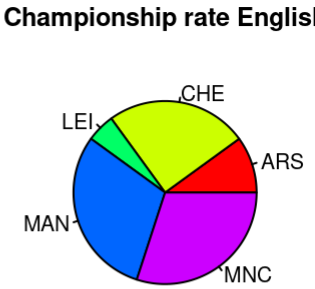
Creating pie charts to see the rates.

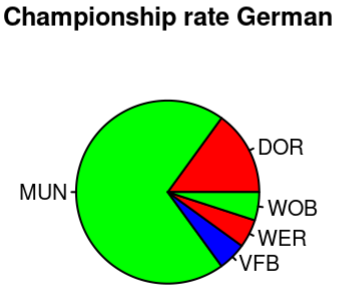
These pie charts are showing the rate of all the championships in each League (2000-2021)

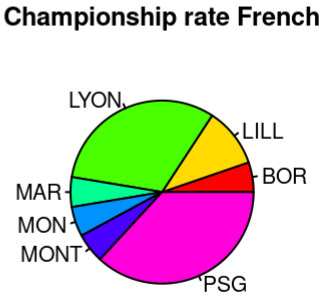
```
par(mfrow = c(2, 3))
pie(Dutch$count,Dutch$Team, main= "Championship rate Dutch Eredivise", col = rainbow(length(Dutch$Team)))
pie(Spanish$count,Spanish$Team, main= "Championship rate Spain", col = rainbow(length(Spanish$Team)))
pie(English$count,English$Team, main= "Championship rate English", col = rainbow(length(English$Team)))
pie(German$count,German$Team, main= "Championship rate German", col = rainbow(length(Italian$Team)))
pie(Italian$count,Italian$Team, main= "Championship rate Italian", col = rainbow(length(French$Team)))
pie(French$count,French$Team, main= "Championship rate French", col = rainbow(length(French$Team)))
```












```
par(mfrow = c(1, 1))
```

Difficulty

Let's see how difficult is to become a champion in each League. For this, we will measure it with the win rate (win games/ total games). In general, the higher the necessary win rate is, the harder to get a championship.

For that, we have to create a new column in the table "champions" calculating the "win_rate".

```
champions <- mutate(champions, win_rate= W/GP)
```

Taking the data from the previous table let's create a new one with the main insights about the win rate for each League.

Win Rate

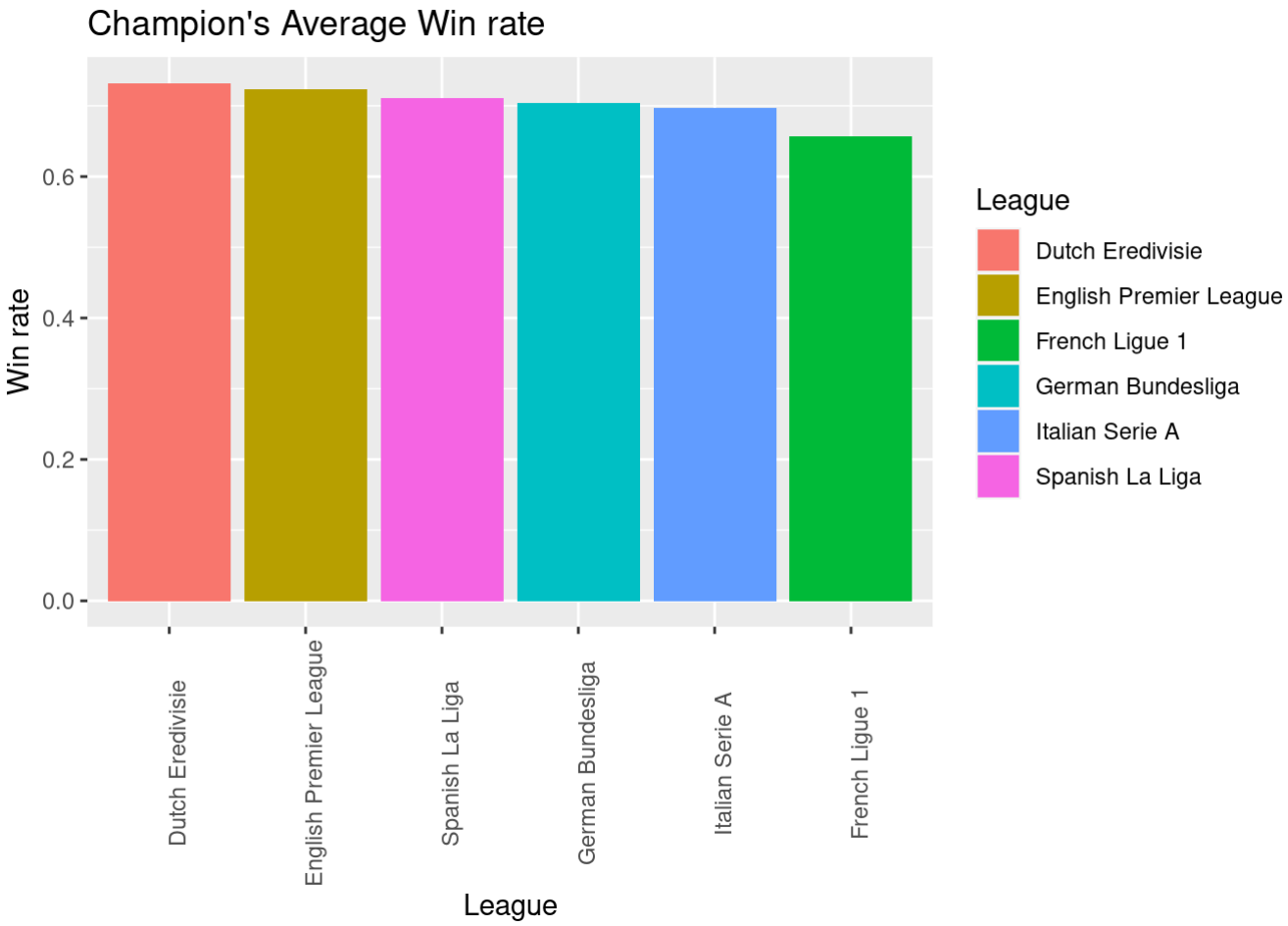
```
insights_wr <- arrange(summarize(group_by(champions, League), WR_mean=mean(win_rate), WR_max=max(win_rate), WR_mi
n=min(win_rate)), WR_mean)

knitr::kable(insights_wr, format = "markdown")
```

League	WR_mean	WR_max	WR_min
French Ligue 1	0.6574455	0.7894737	0.5000000
Italian Serie A	0.6966321	0.8684211	0.5882353
German Bundesliga	0.7044118	0.8529412	0.5882353
Spanish La Liga	0.7114509	0.8421053	0.5526316
English Premier League	0.7233642	0.8421053	0.6052632
Dutch Eredivisie	0.7323529	0.8529412	0.5882353

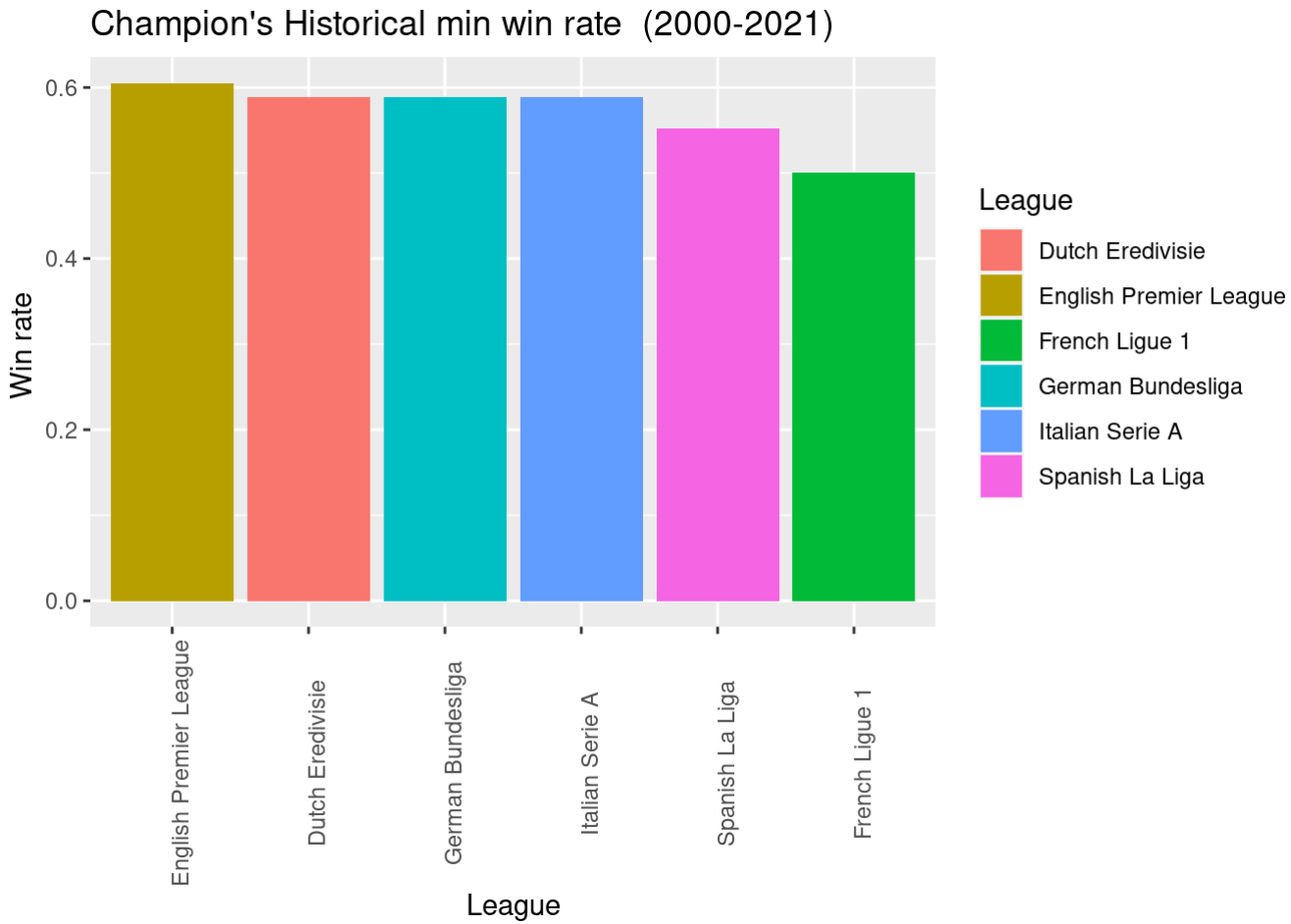
Let's see the difference in the average champion's win rate.

```
ggplot(data=insights_wr)+
  geom_bar(mapping=aes(x = reorder(League, -WR_mean), y = WR_mean, fill = League), stat = "identity", position="d
odge")+
  labs(
    x = "League",
    y = " Win rate ",
    title = "Champion's Average Win rate "
  )+
  theme(axis.text.x = element_text(angle = 90))
```



Let's see the historical minimum Win Rate needed to become a champion in each League.

```
ggplot(data=insights_wr)+
  geom_bar(mapping=aes(x = reorder(League, -WR_min), y = WR_min, fill = League), stat = "identity", position="dod
ge")+
  labs(
    x = "League",
    y = "Win rate ",
    title = "Champion's Historical min win rate  (2000-2021)"
  )+
  theme(axis.text.x = element_text(angle = 90))
```



Conclusions

We can conclude that external factors such as sunlight may not affect the matches as it might be thought, at least in Europe, where the temperatures are not as extreme as in other continents. That could be a reason why differences in scores are not visible in the matches that take place in this Region. Another explanation could be that temperatures and weather affect both teams equally, so the performance changes in defensive and offensive players, so the amount of goals scored tends to be the same.

The difference in scores between locals and visitants is much more visible. It is more usual that the Local team scores more goals than the visitant.

We can see a lot of differences between the total of goals and championships distribution within Leagues. For example, Italian Serie A is the League with fewer different champions having only 3, but the most dominated League is the German Bundesliga, being highly dominated by Bayern Munich.