How to conduct a research project

Lecture 16

Louis SIRUGUE

CPES 2 - Spring 2023

Welcome to the second semester of this course!

- Dedicated to an empirical **research project**:
  - **By pairs**, apply programming & econometric tools from S1 to your own research question
  - Find an example of what is expected here
- In Part I: **formal lectures**:
  - Today: The steps of the research process
  - The next two lectures: Refreshers from S1

Part 1: Guidelines and refreshers	
Lecture 1	How to conduct a research project
Lecture 2	Refresher: R Programming
Lecture 3	Refresher: Econometrics

- In Part II: **follow-ups** and **reports/presentations**

Part 2: Research project	
Lecture 4	***Presentation of your project***
Lecture 5-6	Follow-up: Data cleaning
Lecture 7	Follow-up: Descriptive statistics
Lecture 8	Follow-up: Visualizing the data
Lecture 9	Follow-up: Regression analysis
Lecture 10	***Midterm report feedback***
Lecture 11	Follow-up: Causality assessment
Lecture 12	Follow-up: Robustness
Lecture 13	Follow-up: Heterogeneity
Lecture 14	Follow-up: Last tips
Lecture 15	***Final presentation***

Today: How to conduct a research project

1. Preliminary steps
- 1.1. Research question
- 1.2. Finding data
- 1.3. Literature review

2. Data description
- 2.1. Data cleaning
- 2.2. Descriptive statistics
- 2.3. Data visualization

3. Analysis
- 3.1. Regression analysis
- 3.2. Robustness
- 3.3. Heterogeneity

4. Wrap up!

Today: How to conduct a research project

1. Preliminary steps

- 1.1. Research question
- 1.2. Finding data
- 1.3. Literature review

1. Preliminary steps

1.1. Research question

- The starting point of the research project is the **research question**
  - It is not easy to find a suitable research question, and not all questions are relevant
  - Here are some guidelines to help you in the process

- The question should lead to **explain rather than describe** a phenomenon
  - Do football teams win more home than away?
  - This basically calls to a descriptive statistic, not any explanation

- The question should be **specific enough**
  - What are the reasons why football teams win more often home than away?
  - You won't be able to cover all the determinants
  - Closed-ended questions recommended (Yes/No, to what extent, ...)

- It should be relatively **original** and **interesting to you!**
  - You're gonna spend the whole semester on that

1. Preliminary steps

1.1. Research question

- Here is an example of valid research question:

Do supporters help the home team win the match?

- This is the research question we will take as an example to see all the steps of the research process

- It is:
  - More about explanation than description
  - Specific enough, not too broad
  - Relatively original
  - Relatively interesting with respect to the sports literature

- And importantly there is **data available** to answer this question
  - There's no point having a good research question if you can't find data to answer it
  - Usually finding data comes after the idea of research question
  - But given the time constraint you should look for data while thinking about your research question

1. Preliminary steps

1.2. Finding data

- **Open** access online **data** is increasingly common

- Academic journals ask authors to share their data more and more systematically
  - The **academic literature** is a great source of data
  - Especially RCTs as they usually include many variables

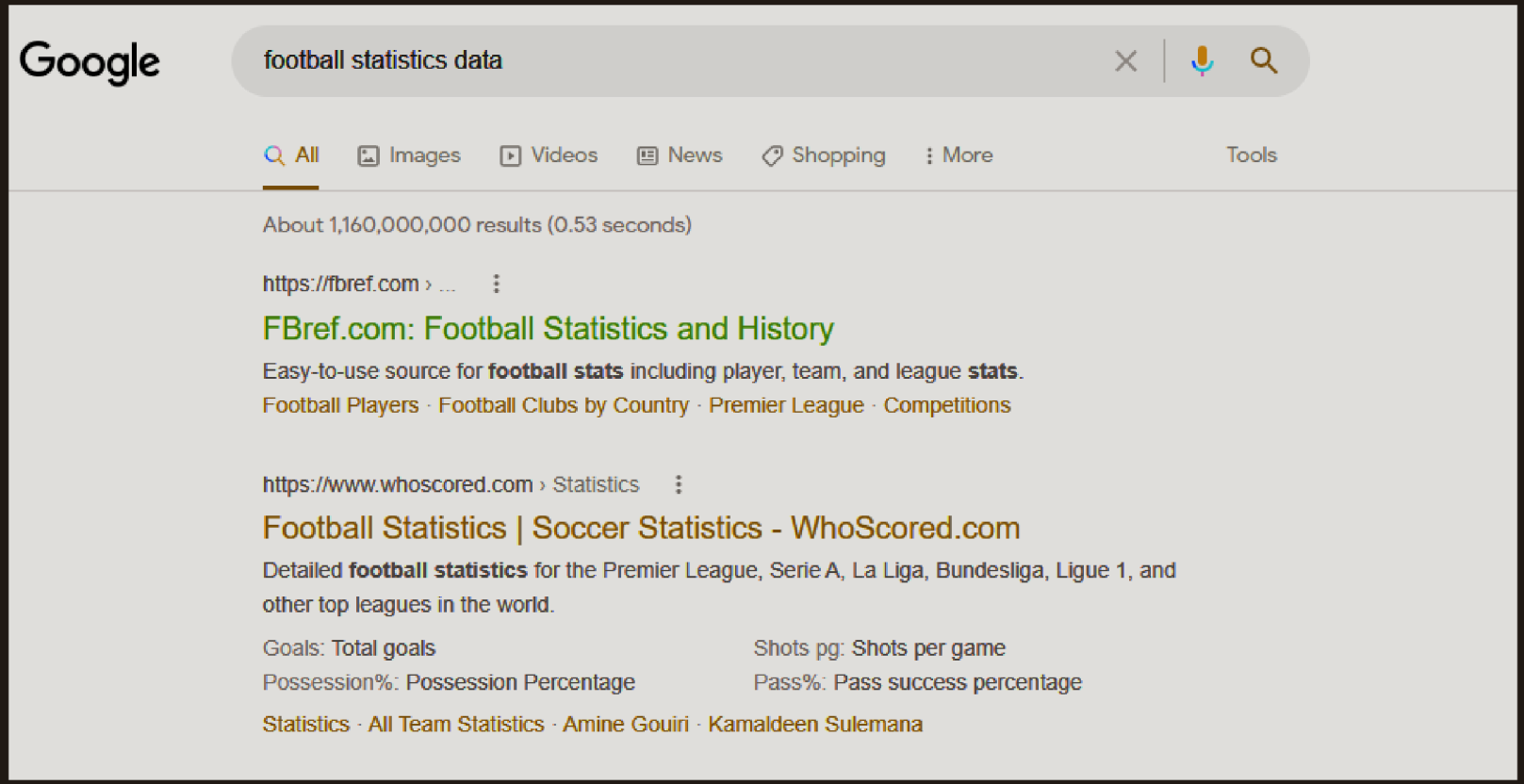
- At these two links you'll find an incredibly rich set of academic datasets
  - openICPSR
  - Harvard Dataverse
  - Browse the available datasets and check the corresponding articles, this may give you inspiration

- The American Economic Association also gathered a lot of data sources
  - Mostly from national statistical institutes
  - Here you may not find so much individual level data but rather local data
  - If relevant it is also possible to combine different data sources

1. Preliminary steps

1.2. Finding data

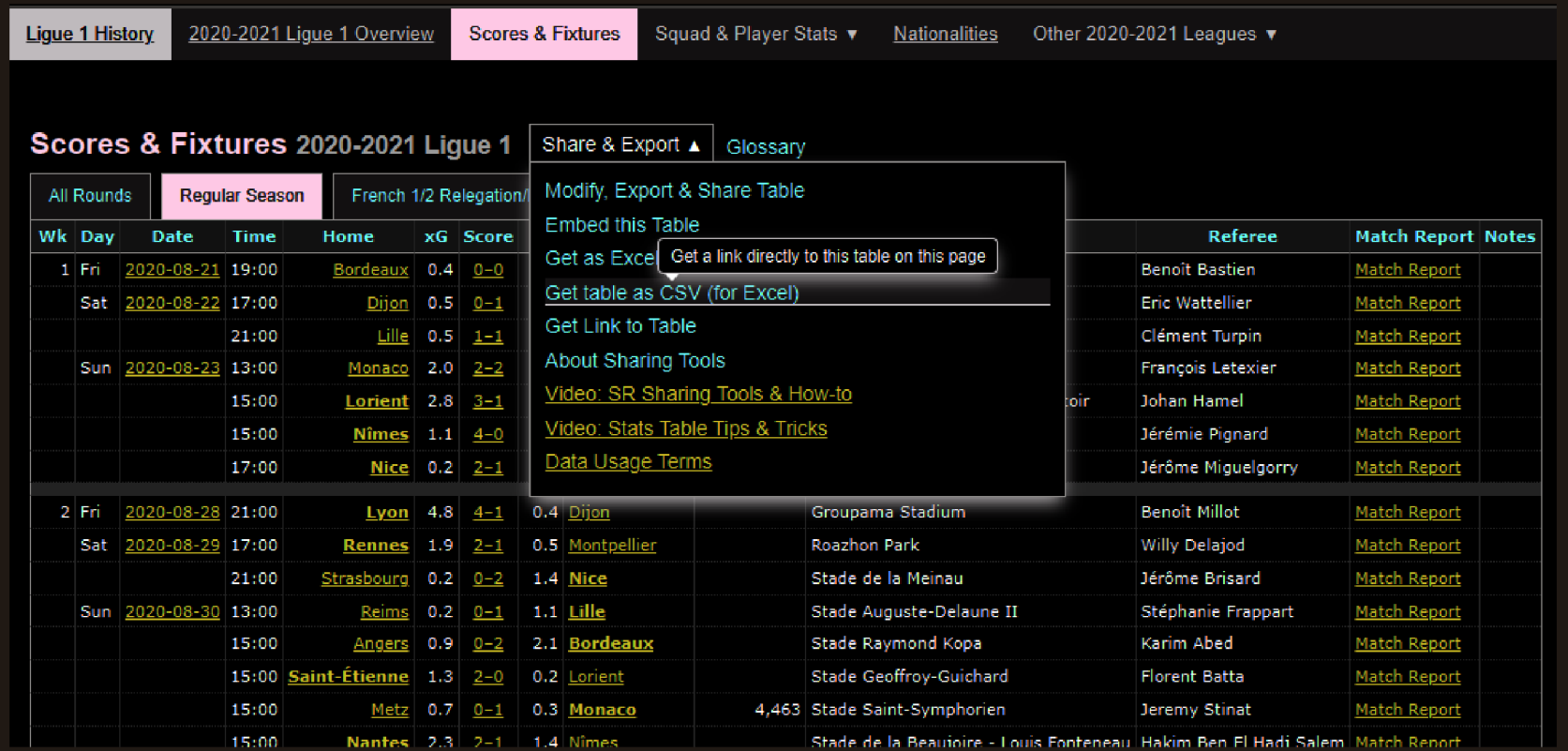
- Sometimes a simple Google search can be sufficient:



1. Preliminary steps

1.2. Finding data

- At fbref.com data on scores and attendance of football matches are available:



1. Preliminary steps

1.2. Finding data

- This data is appropriate for the exercise for two reasons

- It contains the **necessary variables** to study the research question
  - For each match the score and who played home and away
  - The number of people in the stadium

- And **additional variables** to use for robustness and heterogeneity analysis
  - The time in the day and day in the week of the match
  - The league/season (data available for several leagues/seasons)
  - (We'll come back to that point in a few slides)

- So we now have a valid research question and appropriate data to work with
  - But there is one last preliminary step
  - We need to know were the research idea stands with respect to the existing literature

1. Preliminary steps

1.3. Literature review

- A good research project should be relevant with respect to the academic literature on the issue

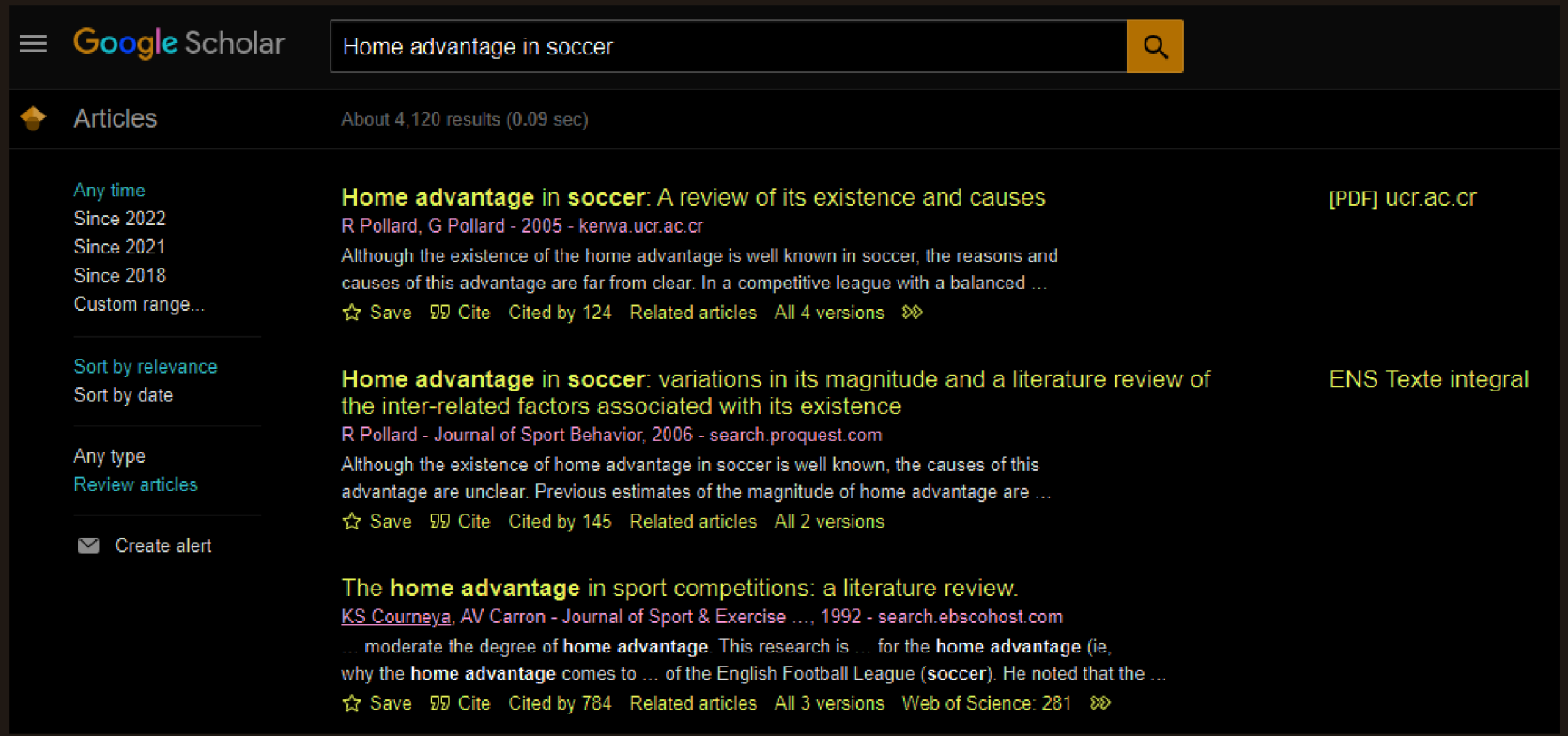
- You should find academic articles to get a sense of where your analysis will stand in the literature
  - What do we **already know** on the topic?
  - What **remains to be known**?
  - What is your **contribution** to the literature?

- You should refer to articles that are published in (peer reviewed) academic journals
  - You can find such articles on Google scholar
  - And via PSL explore

- The articles you should start by looking for are:
  - **Reviews/meta analyses** that will have a lot of references that may be relevant
  - Articles that are **as close as possible** to what you intend to do

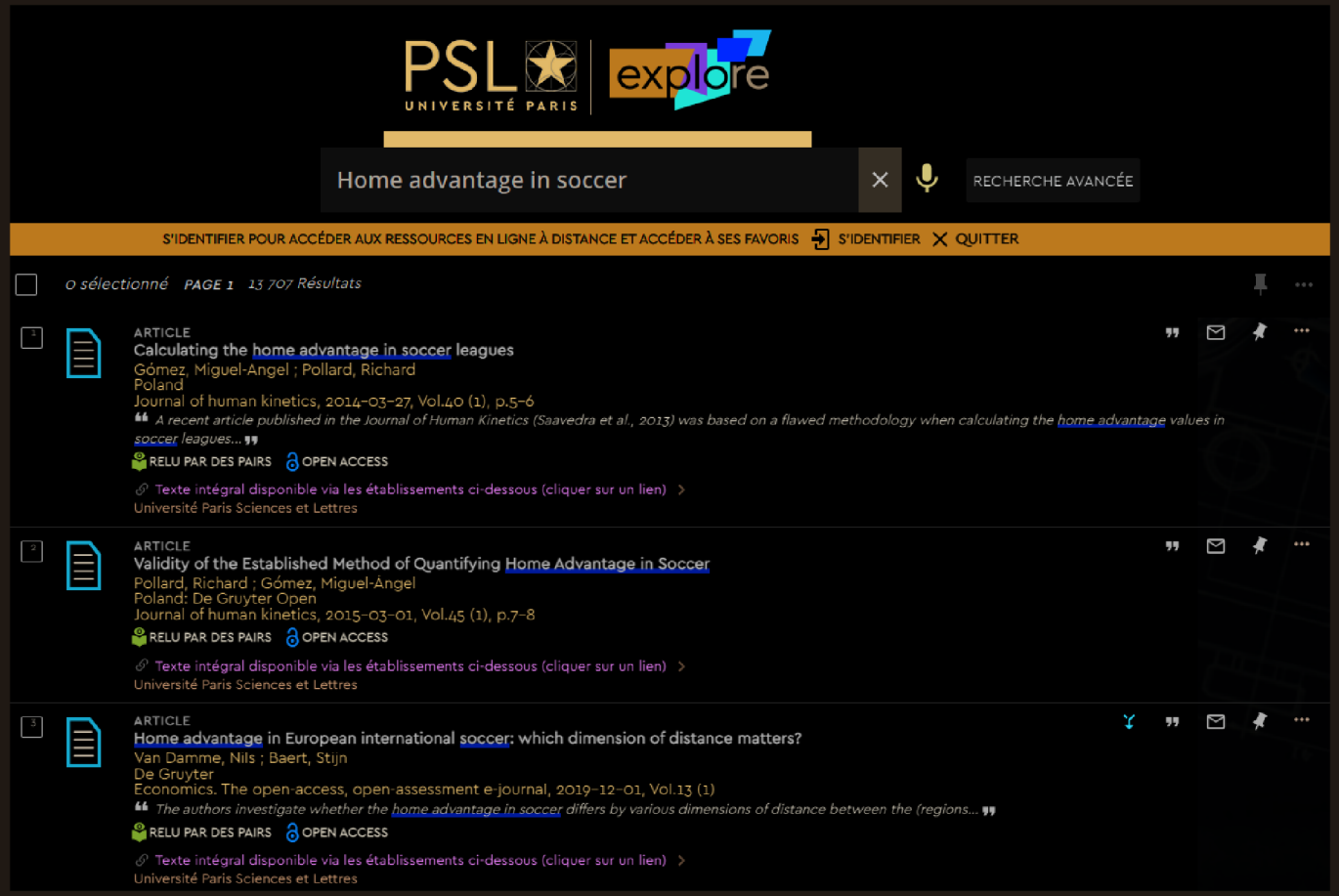
1. Preliminary steps

1.3. Literature review



1. Preliminary steps

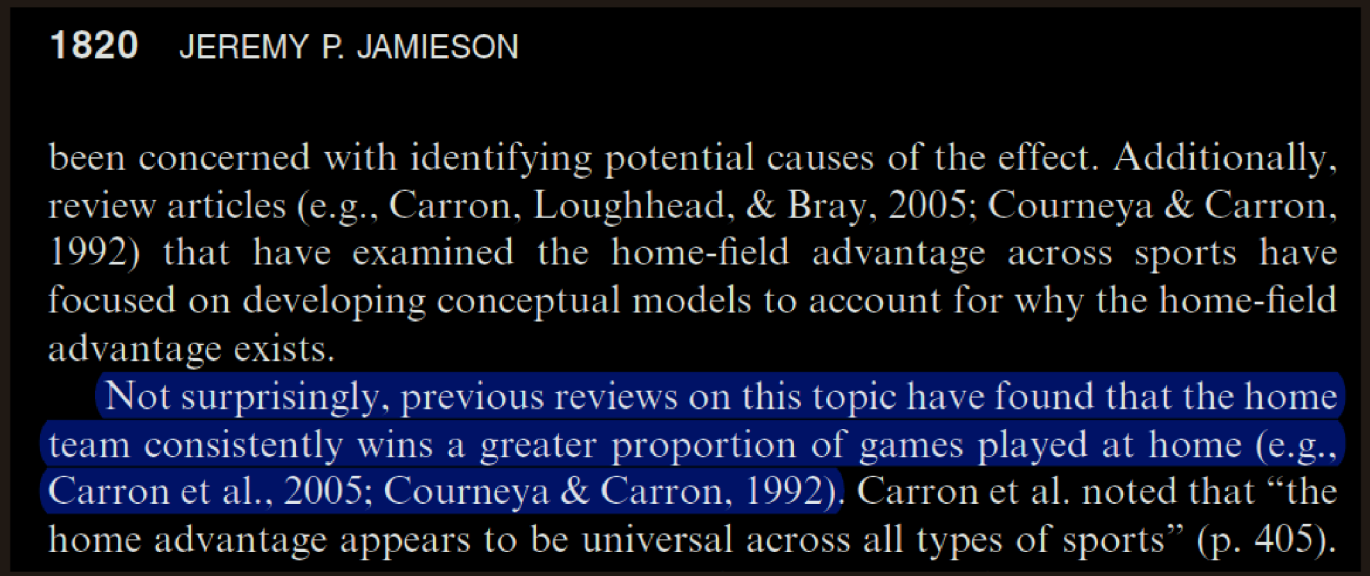
1.3. Literature review



1. Preliminary steps

1.3. Literature review

- Almost every academic article includes some review of the literature
  - You can go through it to find some inspiration and references



1820 JEREMY P. JAMIESON

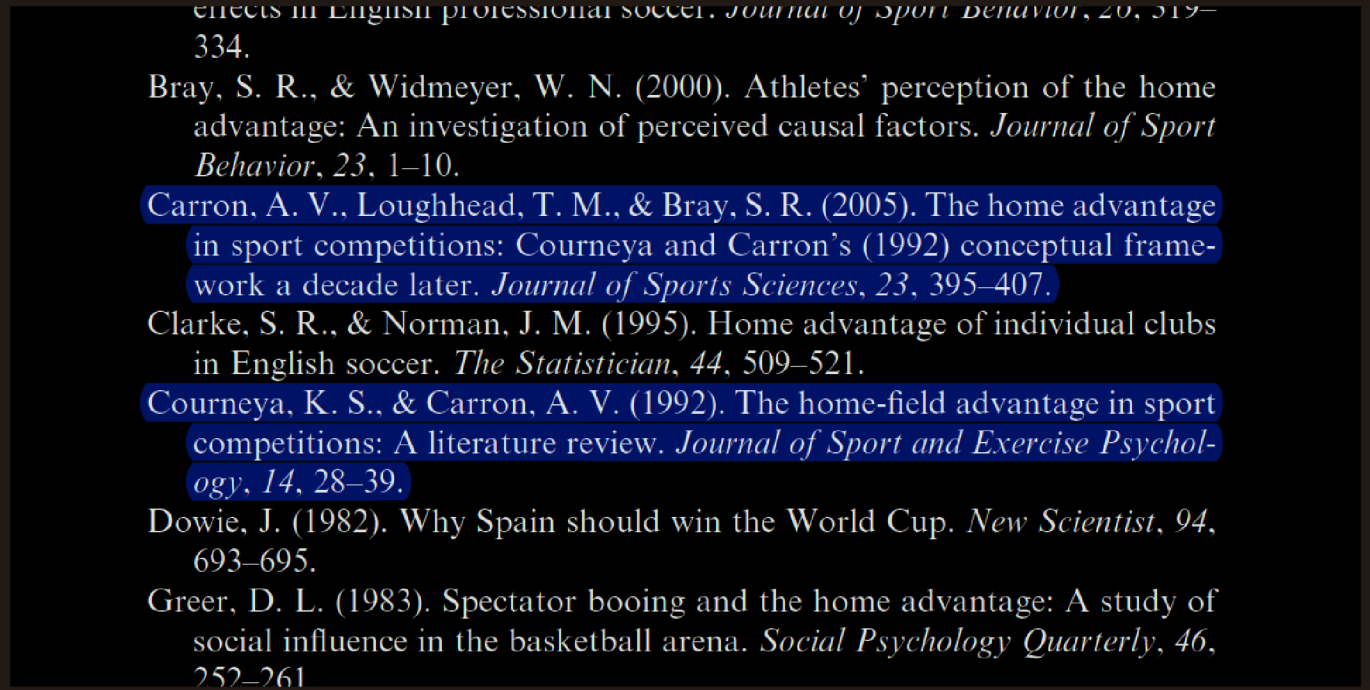
been concerned with identifying potential causes of the effect. Additionally, review articles (e.g., Carron, Loughhead, & Bray, 2005; Courneya & Carron, 1992) that have examined the home-field advantage across sports have focused on developing conceptual models to account for why the home-field advantage exists.

Not surprisingly, previous reviews on this topic have found that the home team consistently wins a greater proportion of games played at home (e.g., Carron et al., 2005; Courneya & Carron, 1992). Carron et al. noted that "the home advantage appears to be universal across all types of sports" (p. 405).

1. Preliminary steps

1.3. Literature review

- In an academic paper, every article mentioned in the text can be found in the *References* section at the end



effects in English professional soccer. *Journal of Sport Behavior*, 20, 319–334.

Bray, S. R., & Widmeyer, W. N. (2000). Athletes' perception of the home advantage: An investigation of perceived causal factors. *Journal of Sport Behavior*, *23*, 1–10.

Carron, A. V., Loughhead, T. M., & Bray, S. R. (2005). The home advantage in sport competitions: Courneya and Carron's (1992) conceptual framework a decade later. *Journal of Sports Sciences*, *23*, 395–407.

Clarke, S. R., & Norman, J. M. (1995). Home advantage of individual clubs in English soccer. *The Statistician*, *44*, 509–521.

Courneya, K. S., & Carron, A. V. (1992). The home-field advantage in sport competitions: A literature review. *Journal of Sport and Exercise Psychology*, *14*, 28–39.

Dowie, J. (1982). Why Spain should win the World Cup. *New Scientist*, *94*, 693–695.

Greer, D. L. (1983). Spectator booing and the home advantage: A study of social influence in the basketball arena. *Social Psychology Quarterly*, *46*, 252–261

➜ Citing articles this way is also something you will have to do

1. Preliminary steps

1.3. Literature review

- The way you should refer to academic articles in the text is codified:

Referring to an academic article in-text

	One author	Two authors	More authors
Within the sentence	Smith (2012) showed that ...	Smith and Watson (2012) showed that ...	Smith et al. (2012) showed that ...
Outside the sentence	It has been shown that ... (Smith, 2012)	It has been shown that ... (Smith and Watson, 2012)	It has been shown that ... (Smith et al., 2012)

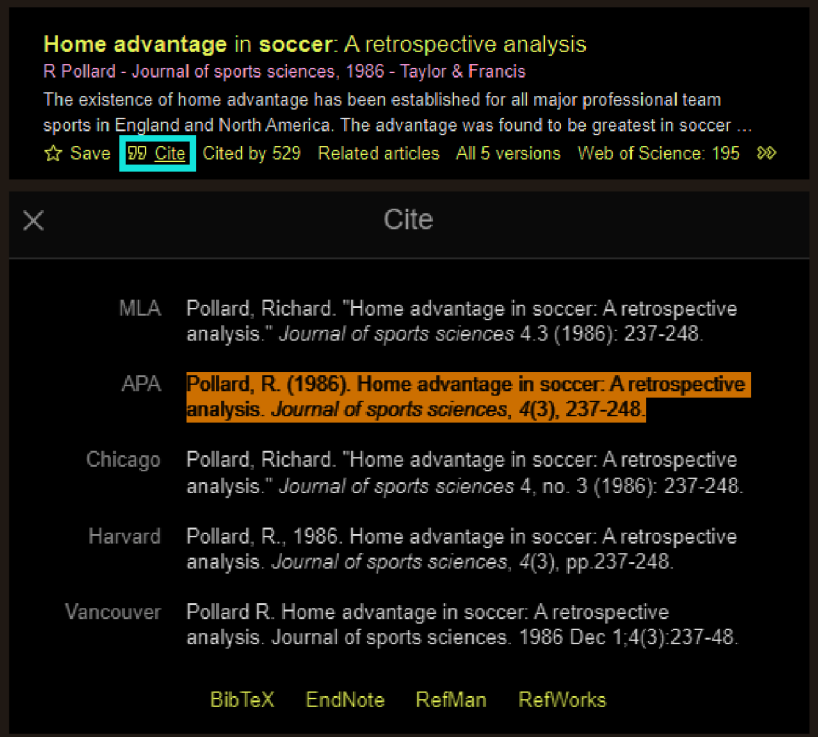
- Conventionally (in Economics) authors are listed by alphabetical order of surname
- The reference of every article you cite should be added by alphabetical order in a Reference section at the end
  - How to write the reference in this last section is also codified
  - Take a look at the research project example available here to see what it should look like

1. Preliminary steps

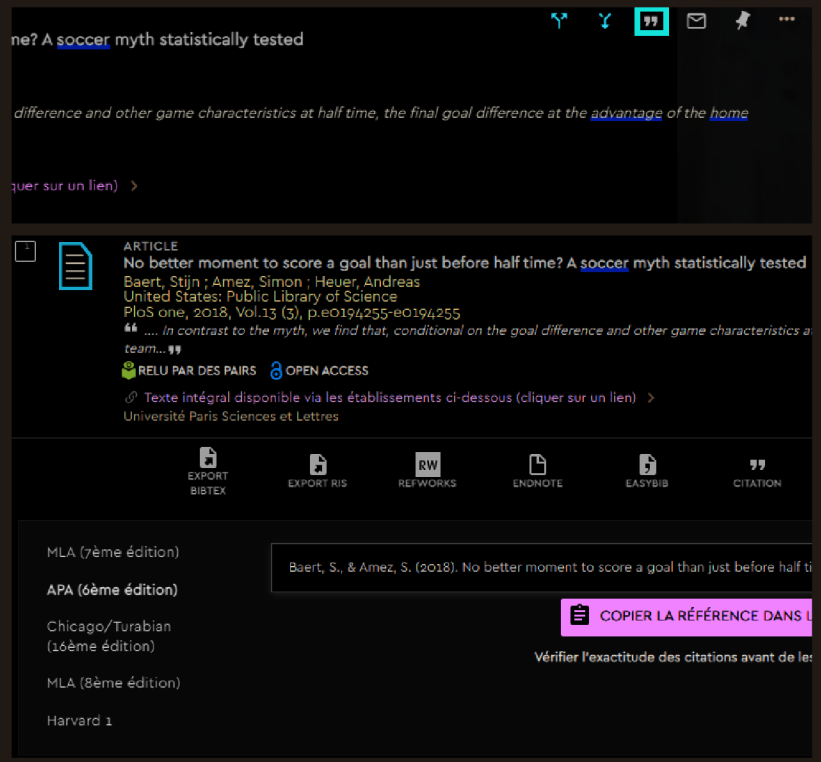
1.3. Literature review

- To find the proper reference of an article, click on the Cite button and copy-paste it in your *References* section

On Google scholar



On PSL explore



Overview

1. Preliminary steps ✓
- 1.1. Research question
- 1.2. Finding data
- 1.3. Literature review

2. Data description
- 2.1. Data cleaning
- 2.2. Descriptive statistics
- 2.3. Data visualization

3. Analysis
- 3.1. Regression analysis
- 3.2. Robustness
- 3.3. Heterogeneity

4. Wrap up!

Overview

1. Preliminary steps ✓
- 1.1. Research question
- 1.2. Finding data
- 1.3. Literature review

2. Data description
- 2.1. Data cleaning
- 2.2. Descriptive statistics
- 2.3. Data visualization

2. Data description

2.1. Data cleaning

- The **first** thing to do with the **data** is to **clean** it
  - You should open the data and take a close look at it to understand what's inside

```
library(tidyverse)
data_match <- read.csv("data_match.csv")
dim(data_match)
```

```
## [1] 4845   16
```

- The data contains 4845 rows and 16 variables
  - Let's see what these variables are

```
names(data_match)
```

```
##  [1] "Wk"          "Day"         "Date"        "Time"        "Home"
##  [6] "xG"          "Score"       "xG.1"        "Away"        "Attendance"
## [11] "Venue"       "Referee"     "Match.Report" "Notes"      "League"
## [16] "Season"
```

2. Data description

2.1. Data cleaning

```
str(data_match)
```

```
## 'data.frame':    4845 obs. of  16 variables:
##  $ Wk          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Day         : chr  "Fri" "Sat" "Sat" "Sat" ...
##  $ Date        : chr  "2018-08-10" "2018-08-11" "2018-08-11" "2018-08-11" ...
##  $ Time        : chr  "20:45" "17:00" "20:00" "20:00" ...
##  $ Home        : chr  "Marseille" "Nantes" "Montpellier" "Lille" ...
##  $ xG          : num  2.8 1.6 2 1.5 2.5 1 1.3 1 0.2 2.8 ...
##  $ Score       : chr  "4-0" "1-3" "1-2" "3-1" ...
##  $ xG.1        : num  0.3 2.2 2 0.5 1.8 1.9 0.5 0.5 1.7 0.2 ...
##  $ Away        : chr  "Toulouse" "Monaco" "Dijon" "Rennes" ...
##  $ Attendance  : int  60756 32760 12765 25708 9534 26006 21421 48263 23079 47289 ...
##  $ Venue       : chr  "Orange Vélodrome" "Stade de la Beaujoire - Louis Fonteneau" "Stade de la Mosson"
##  $ Referee     : chr  "Ruddy Buquet" "Jérôme Brisard" "Florent Batta" "Willy Delajod" ...
##  $ Match.Report: chr  "Match Report" "Match Report" "Match Report" "Match Report" ...
##  $ Notes       : chr  NA NA NA NA ...
##  $ League      : chr  "Ligue 1" "Ligue 1" "Ligue 1" "Ligue 1" ...
##  $ Season      : chr  "2018-2019" "2018-2019" "2018-2019" "2018-2019" ...
```

2. Data description

2.1. Data cleaning

- The dataset contains the following 16 variables:
  - **Wk:** Season week when the match took place
  - **Day:** Week day when the match took place
  - **Date:** Date of the match
  - **Time:** Time of the match
  - **Home:** Team that played home
  - **xG:** Expected number of goals for home team
  - **Score:** Score of the match
  - **xG.1:** Expected number of goals for away team
  - **Away:** Team that played away
  - **Attendance:** Number of supporters in the stadium
  - **Venue:** Name of the stadium where the match took place
  - **Referee:** Name of the referee
  - **Match.Report:** Link to an online report of the match
  - **Notes:** Miscellaneous information on the match
  - **League:** Name of the league
  - **Season:** Season from 2018-2019 to 2020-2021

2. Data description

2.1. Data cleaning

- We can keep only the relevant variables and look at the first rows of the data

```
data_match <- data_match %>%
  select(Day, Date, Time, Home, Score, Away, Attendance, League, Season)

kable(head(data_match, n = 5), caption = "Outlook of the data:")
```

Outlook of the data:

Day	Date	Time	Home	Score	Away	Attendance	League	Season
Fri	2018-08-10	20:45	Marseille	4-0	Toulouse	60756	Ligue 1	2018-2019
Sat	2018-08-11	17:00	Nantes	1-3	Monaco	32760	Ligue 1	2018-2019
Sat	2018-08-11	20:00	Montpellier	1-2	Dijon	12765	Ligue 1	2018-2019
Sat	2018-08-11	20:00	Lille	3-1	Rennes	25708	Ligue 1	2018-2019
Sat	2018-08-11	20:00	Angers	3-4	Nîmes	9534	Ligue 1	2018-2019

2. Data description

2.1. Data cleaning

- Data cleaning involves:
  - **Recoding variables** in a practical way (it may imply **creating new variables**)
  - **Removing observations** that are not relevant, if any, typically **missing values**
  - Potentially **joining** data, and **pivoting** variables from wide to long or conversely

Day	Date	Time	Home	Score	Away	Attendance	League	Season
Fri	2018-08-10	20:45	Marseille	4-0	Toulouse	60756	Ligue 1	2018-2019

- Here is what we can do already:
  - Divide the score variable into two variables, for home and away
  - Create a variable indicating who won

- Some datasets are cleaner than others, but there's always some data cleaning to do

2. Data description

2.1. Data cleaning

- Recoding variables

```
data_match <- data_match %>%

  # Separate the home and away score into 2 variables
  separate(Score, c("Home", "Away"), "-") %>%

  # Convert these variables as numeric
  mutate(Home = as.numeric(Home),
         Away = as.numeric(Away),

         # Generate a variable for the outcome of the match depending on who scored the most
         Winner = case_when(Home > Away ~ "Home",
                            Home == Away ~ "Draw",
                            Home < Away ~ "Away"))
```

- Let's take a look at the cleaned data
  - (Pay attention to rows 11, 22, ...)

2. Data description

2.1. Data cleaning

Show 7 entries Search:

	Day	Date	Time	Attendance	Home	Away	League	Season	Winner
1	Fri	2018-08-10	20:45	60756	4	0	Ligue 1	2018-2019	Home
2	Sat	2018-08-11	17:00	32760	1	3	Ligue 1	2018-2019	Away
3	Sat	2018-08-11	20:00	12765	1	2	Ligue 1	2018-2019	Away
4	Sat	2018-08-11	20:00	25708	3	1	Ligue 1	2018-2019	Home
5	Sat	2018-08-11	20:00	9534	3	4	Ligue 1	2018-2019	Away
6	Sat	2018-08-11	20:00	26006	2	1	Ligue 1	2018-2019	Home
7	Sat	2018-08-11	20:00	21421	0	1	Ligue 1	2018-2019	Away

Showing 1 to 7 of 4,845 entries Previous 1 2 3 4 5 ... 693 Next

2. Data description

2.1. Data cleaning

- **Between each week** of competition there is a **empty line** with missing values
  - These rows are not actual observations so we should **delete** them

```
data_match <- data_match %>% filter(!is.na(Home))
```

- But we still need to **check** for actual **missing values**

```
data_match %>% summarise_all(~sum(is.na(.)))
```

```
##   Day Date Time Attendance Home Away League Season Winner
## 1   0    0    0       1670    0    0      0      0      0
```

- There is no missing value except for the **Attendance** variable that has **many NAs**
  - This is suspicious, we should **investigate** more

2. Data description

2.1. Data cleaning

- We must check the distribution of the variable:

```
summary(data_match$Attendance)
```

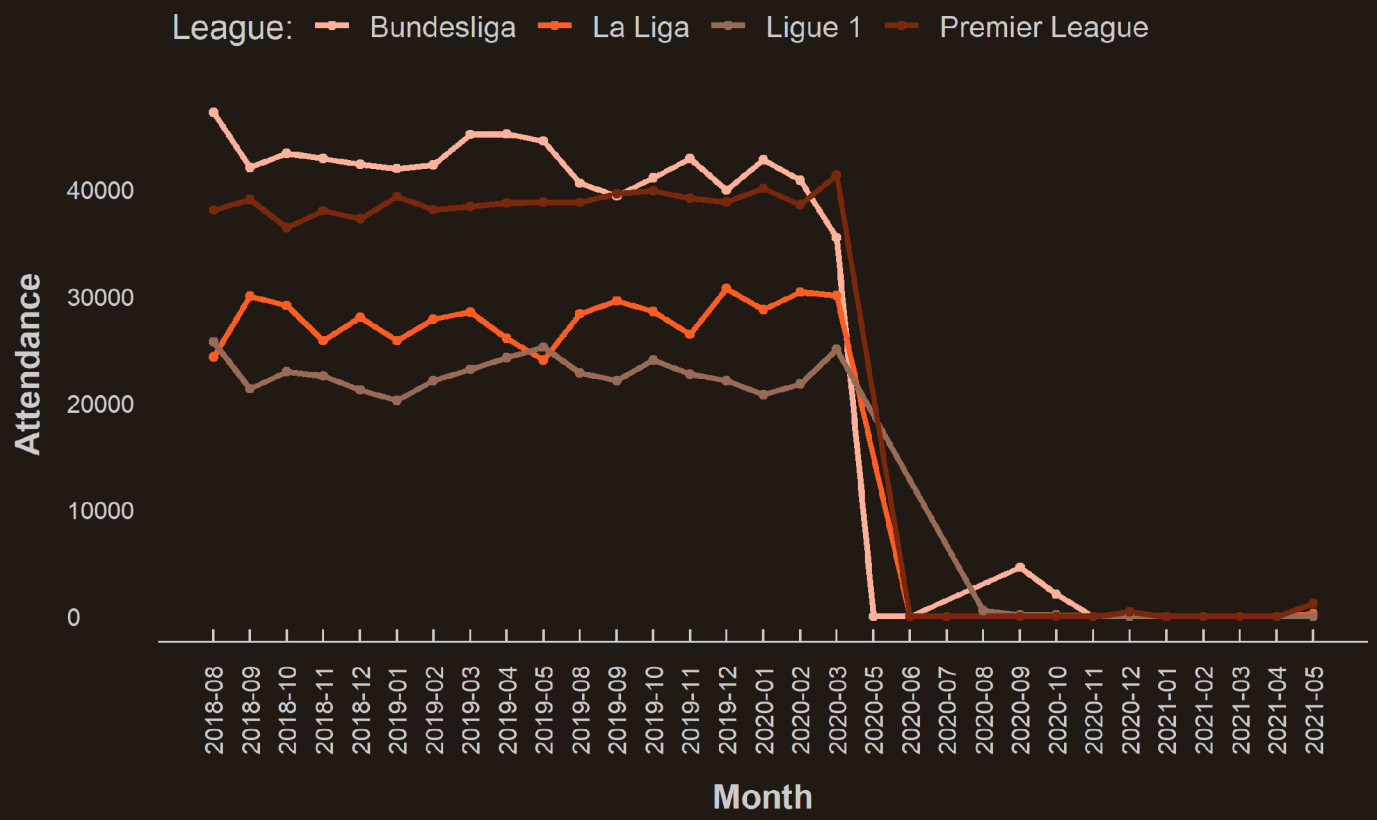
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      13   16158   27717   31790   45014   93426    1670
```

- Except for NAs, the distribution seems fine
  - But the number of spectators per match starts at 13 while the COVID-19 pandemic prevented many matches from having any attendance
  - These **NAs** for attendance may actually **mean 0 attendance**
  - This is particularly plausible given that there is no other variable with missing values in the data

➜ *To check this hypothesis we can plot the evolution of the monthly attendance,* ***replacing NAs by 0s***
(code in the research project example)

2. Data description

2.1. Data cleaning



- This graph confirms the hypothesis
  - There is a drop to 0 attendance (the NAs) due to the pandemic right after March 2020
  - Missing values for the Attendance variable should indeed be recoded as 0

```
data_match <- data_match %>%
  mutate(Attendance =
    ifelse(is.na(Attendance), 0,
           Attendance))
```

2. Data description

2.2. Descriptive statistics

- Now that the data is clean, we should describe it with **relevant statistics**
  - For **categorical variables**: Number of observations per category
  - For **continuous variables**: Summarizing the distribution

```
data_match %>%
  group_by(Winner) %>%
  summarise(N = n(), Pct = 100 * (n() / nrow(.))) %>%
  kable(., "Distribution of match outcomes")
```

Distribution of match outcomes

Winner	N	Pct
Away	1343	31.70
Draw	1067	25.18
Home	1827	43.12

2. Data description

2.2. Descriptive statistics

- The number of observations per season/league is also interesting to know:
  - *(code in the research project example)*

Number of matches:

	2018-2019	2019-2020	2020-2021	Total
Bundesliga	306	306	306	**918**
La Liga	380	380	380	**1140**
Ligue 1	380	279	380	**1039**
Premier League	380	380	380	**1140**
Total	**1446**	**1345**	**1446**	**4237**

- The distribution of the main continuous variables can also be summarized by season/league
  - *(code in the research project example)*

Season 2018-2019

	Min	Q1	Median	Mean	Q3	Max
Attendance						
Bundesliga	19205	29230.50	40911.0	43453.18	52500.00	81365
La Liga	3592	12074.50	19367.5	27118.68	39587.75	93265
Ligue 1	0	12795.75	17577.5	22807.27	27378.50	64696
Premier League	9980	25034.75	31948.0	38181.29	53282.75	81332
Goals away						
Bundesliga	0	0.00	1.0	1.39	2.00	6
La Liga	0	0.00	1.0	1.13	2.00	6
Ligue 1	0	0.00	1.0	1.09	2.00	5
Premier League	0	0.00	1.0	1.25	2.00	6
Goals home						
Bundesliga	0	1.00	2.0	1.79	3.00	8
La Liga	0	1.00	1.0	1.45	2.00	8
Ligue 1	0	1.00	1.0	1.47	2.00	9
Premier League	0	1.00	1.0	1.57	2.00	6

Season 2019-2020

	Min	Q1	Median	Mean	Q3	Max
Attendance						
Bundesliga	0	0.0	27062.5	29783.37	49025.0	81365
La Liga	0	0.0	16001.5	20694.99	33583.5	93426
Ligue 1	0	12418.0	15814.0	22427.67	29440.5	65421
Premier League	0	10346.5	30534.0	29796.04	45594.5	73737
Goals away						
Bundesliga	0	1.0	1.0	1.55	2.0	6
La Liga	0	0.0	1.0	1.04	2.0	5
Ligue 1	0	0.0	1.0	1.03	2.0	5
Premier League	0	0.0	1.0	1.21	2.0	9
Goals home						
Bundesliga	0	1.0	1.0	1.66	2.0	8
La Liga	0	1.0	1.0	1.44	2.0	6
Ligue 1	0	1.0	1.0	1.49	2.0	6
Premier League	0	1.0	1.0	1.52	2.0	8

Season 2020-2021

	Min	Q1	Median	Mean	Q3	Max
Attendance						
Bundesliga	0	0	0	503.57	0	11500
La Liga	0	0	0	33.54	0	4800
Ligue 1	0	0	0	46.90	0	5000
Premier League	0	0	0	224.22	0	10000
Goals away						
Bundesliga	0	0	1	1.36	2	5
La Liga	0	0	1	1.14	2	6
Ligue 1	0	1	1	1.36	2	5
Premier League	0	0	1	1.34	2	7
Goals home						
Bundesliga	0	1	1	1.68	2	8
La Liga	0	0	1	1.37	2	6
Ligue 1	0	0	1	1.40	2	6
Premier League	0	0	1	1.35	2	9

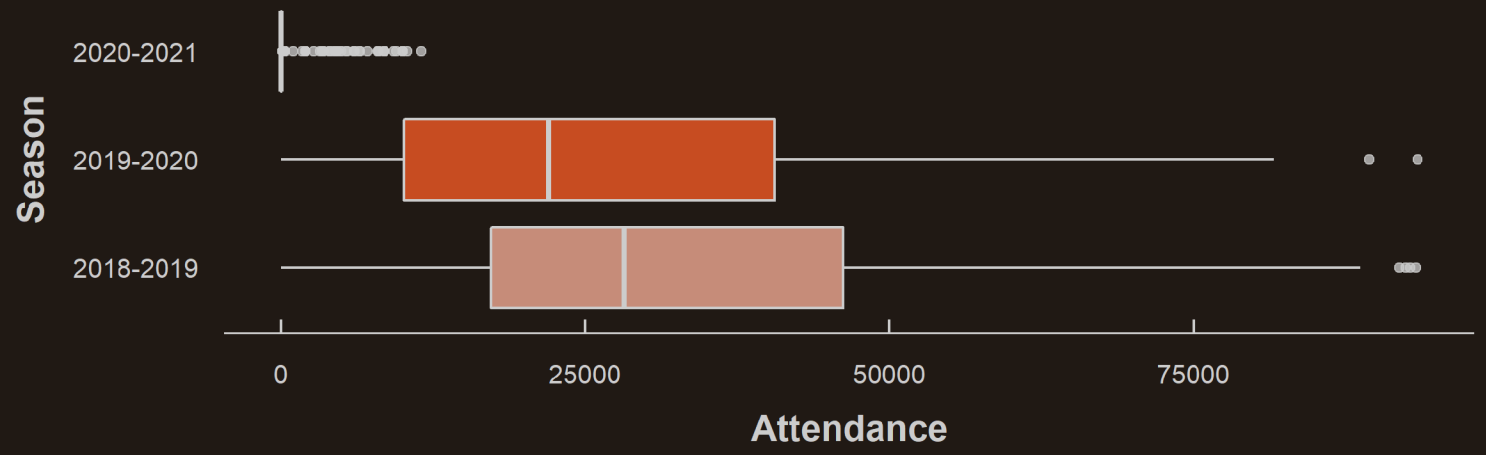
2. Data description

2.3. Data visualization

- The last step before the analysis is to **visualize the data**
  - The idea is also to describe the data, but **with relevant graphs**

➜ Attendance:

```
ggplot(data_match, aes(x = Season, y = Attendance, fill = Season)) +
  geom_boxplot(show.legend = F, alpha = .75) + coord_flip()
```

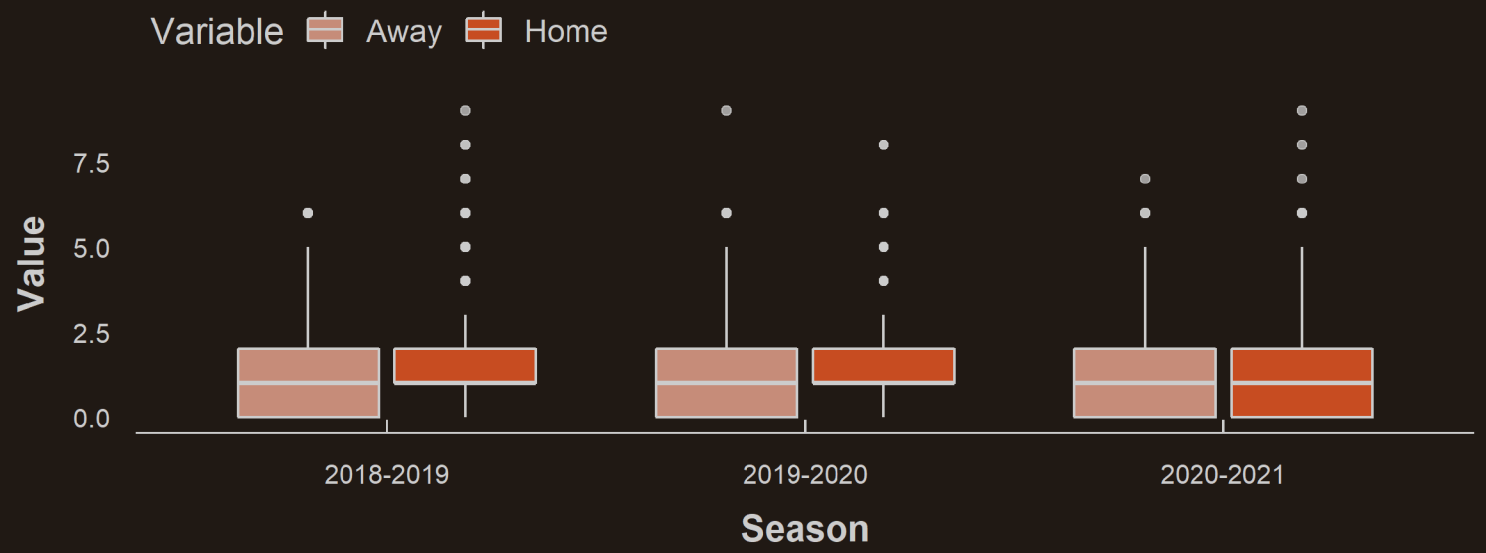


2. Data description

2.3. Data visualization

➜ Goals home vs. away:

```
data_match %>%
  pivot_longer(c(Home, Away), names_to = "Variable", values_to = "Value") %>%
  ggplot(., aes(x = Season, y = Value, fill = Variable)) + geom_boxplot(alpha = .75)
```

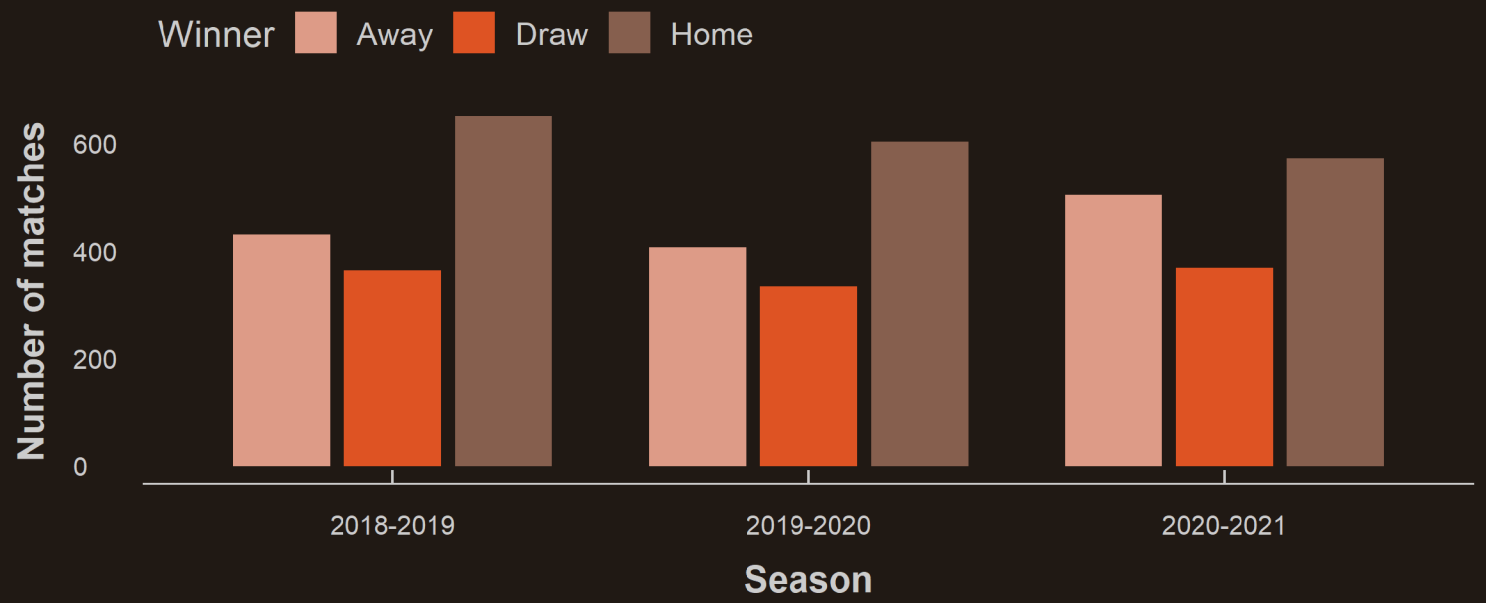


2. Data description

2.3. Data visualization

➜ Winner:

```
ggplot(data_match, aes(x = Season, fill = Winner)) + ylab("Number of matches") +
  geom_bar(stat = "count", position = position_dodge(width = .8), width = .7, alpha = .85)
```



Overview

1. Preliminary steps ✓
- 1.1. Research question
- 1.2. Finding data
- 1.3. Literature review

2. Data description ✓
- 2.1. Data cleaning
- 2.2. Descriptive statistics
- 2.3. Data visualization

3. Analysis
- 3.1. Regression analysis
- 3.2. Robustness
- 3.3. Heterogeneity

4. Wrap up!

Overview

1. Preliminary steps ✓
- 1.1. Research question
- 1.2. Finding data
- 1.3. Literature review

2. Data description ✓
- 2.1. Data cleaning
- 2.2. Descriptive statistics
- 2.3. Data visualization

3. Analysis
- 3.1. Regression analysis
- 3.2. Robustness
- 3.3. Heterogeneity

3. Analysis

3.1. Regression analysis

- The first step of the regression analysis is to write down properly the **equation** you want to estimate:

$$1\{Winner_m = \text{Home}\} = \alpha + \beta \times 1\{Public_m = \text{Yes}\} + \varepsilon_m$$

- Where for a given match m:
  - $1\{Winner_m = \text{Home}\}$ takes the value 1 if the winning team is that playing home and 0 otherwise
  - $1\{Public_m = \text{Yes}\}$ takes the value 1 if there is public in the stadium and and 0 otherwise

- The two variables of interest should be coded properly for the regression:

```
data_match <- data_match %>%
  mutate(Winner_home = ifelse(Winner == "Home", 1, 0),
         Public = ifelse(Attendance > 0, "Public", "No public"))
```

3. Analysis

3.1. Regression analysis

- Then the regression should be properly reported:

```
stargazer(lm(Winner_home~Public, data_match),
          dep.var.labels = c("Home win"),
          keep.stat = c("n", "adj.rsq"),
          type = "text")
```

- And the coefficient of interest properly interpreted:

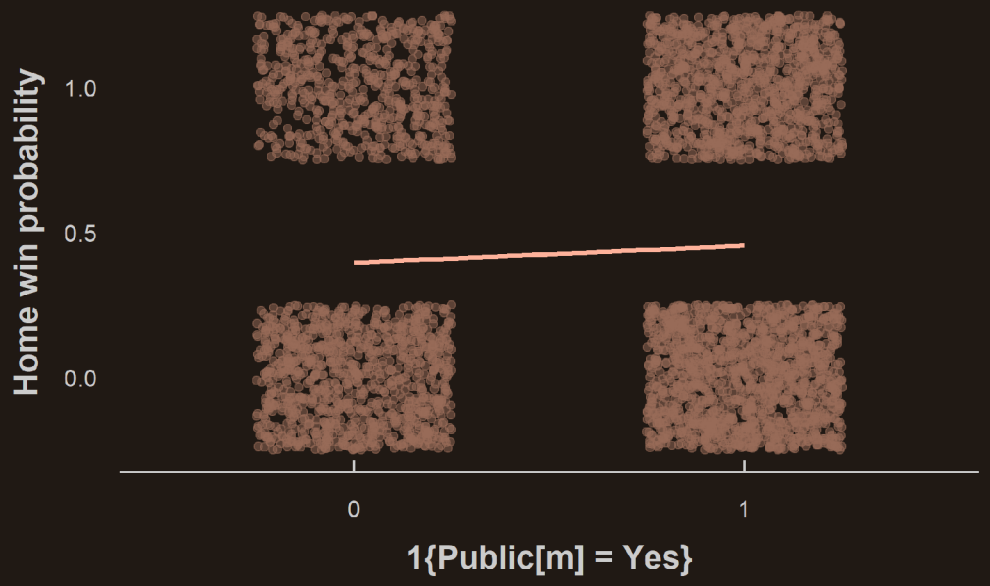
The presence of supporters in the audience increases by 5.9 percentage points on average the probability for the home team to win the match relative to loose or draw, everything else equal. The coefficient is statistically significantly different from 0 at the 1% significance level.

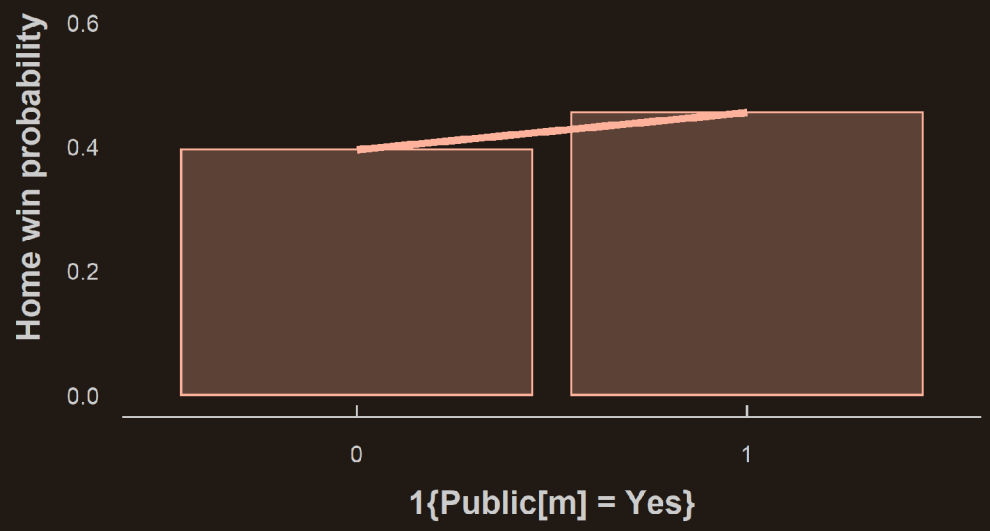
```
##
## ========================================
##                   Dependent variable:
##               -------------------------
##                         Home win
## ----------------------------------------
## PublicPublic           0.059***
##                         (0.016)
##
## Constant               0.395***
##                         (0.012)
##
## ----------------------------------------
## Observations             4,237
## Adjusted R2              0.003
## ========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

3. Analysis

3.1. Regression analysis

- It is also a good practice to provide a visual representation of the relationship you estimate
  - In this case it is not simple because both variables are binary
  - But geom_jitter() allows to add noise in the location of each data point around the 4 possible coordinates
  - *(code in the research project example)*





3. Analysis

3.1. Regression analysis

- It is also crucial to discuss whether or not the effect is **causal**
  - Self-selection issue?
  - Omitted variable bias?
  - Under which assumptions the effect would be causal?

- **Self-selection** issue
  - Teams that play home or away cannot self-select into whether there is public or not
  - Variations in x are fully driven by decision teams have no control over

- **Omitted variable** bias
  - There may be other variables correlated with both x and y that drive this relationship
  - When no public (i.e., pandemic), the trip to the stadium may be less tiring because there is less congestion on the roads due to remote working, or for any other reason

Thus, this result can be considered as causal only if there was no change concomitant to the attendance restrictions that could have a differentiated impact on the teams that play home and away

3. Analysis

3.2. Robustness

- Assessing the robustness of the result consists in progressively **adding control variables** in the regression
  - If the result is robust this should **not affect too much the magnitude** of the coefficient
  - If the result is robust the coefficient should **remain statistically significant**

- We can control for the day and time of the match
  - These factors could be linked to the mechanism related to transport
  - Even though it cannot rule out this mechanism, it can suggest whether or not time and day is a channel
  - We can also control for the league in case the effect is driven by differences across leagues

```
stargazer(lm(Winner_home ~ Public, data_match),
          lm(Winner_home ~ Public + League, data_match),
          lm(Winner_home ~ Public + League + Time, data_match),
          lm(Winner_home ~ Public + League + Time + Day, data_match),
          dep.var.labels = c("Home win"), type = "text", keep.stat = c("n", "adj.rsq"))
```

```
##
## ================================================================
##                                Dependent variable:
##                            ------------------------------------
##                                       Home win
## ----------------------------------------------------------------
## PublicPublic              0.059*** 0.060*** 0.064*** 0.066***
##                           (0.016)  (0.016)  (0.017)  (0.017)
##
## LeagueBundesliga                    0.003    -0.013   -0.020
##                                    (0.022)  (0.060)  (0.062)
##
## LeagueLa Liga                       0.019    -0.010   -0.007
##                                    (0.021)  (0.032)  (0.032)
##
## LeaguePremier League                0.014    -0.505   -0.460
##                                    (0.021)  (0.500)  (0.504)
##
## Time12:00 (13:00)                            0.490    0.449
##                                             (0.502)  (0.505)
##
## Time12:30 (13:30)                            0.409    0.339
##                                             (0.498)  (0.500)
##
## Time12:45                                    -0.440   -0.465
##                                             (0.501)  (0.501)
##
## Time13:00                                    -0.037   -0.055
##                                             (0.085)  (0.087)
```

- As control variables are included:
  - The **magnitude** of the coefficient does not vary much
  - The statistical **significance** does not change either

→ *So the result is robust to controlling for these characteristics*

3. Analysis

3.2. Robustness

- Note that robustness is not necessarily about including controls
  - It can be about **excluding/including** some observations (e.g., outliers)
  - About **changing the definition** of one or several variables, etc.

- For instance, the independent variable of the regression could be defined in two ways:
  - So far: Probability of winning relative to loosing or draw
  - Alternative: Probability of winning relative to loosing only, omitting draws

```
data_match <- data_match %>%
  mutate(Winner_home2 = ifelse(Winner != "Draw", Winner_home, NA))

stargazer(lm(Winner_home ~ Public, data_match),
          lm(Winner_home2 ~ Public, data_match),
          keep.stat = c("n", "adj.rsq"), model.numbers = FALSE,
          dep.var.labels = c("Home win vs. Home loss", "Home win vs. Home loss/Draw"))
```

3. Analysis

3.2. Robustness

```
##
## ================================================================
##                                Dependent variable:
##                    ------------------------------------------------
##                    Home win vs. Home loss Home win vs. Home loss/Draw
## ----------------------------------------------------------------
## PublicPublic              0.059***                  0.078***
##                           (0.016)                   (0.018)
##
## Constant                  0.395***                  0.529***
##                           (0.012)                   (0.014)
##
## ----------------------------------------------------------------
## Observations               4,237                     3,170
## Adjusted R2                0.003                     0.006
## ================================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

- Coefficients cannot be compared directly because they are mechanically inflated by the omission of the possibility of draw
  - But the ratio of the effect of public in the stadium on the probability to win, relative to the probability to win when there is no public, is very similar in the two cases (≈0.15)
  - And both statistically significantly different from 0 at the 99% confidence level

➜ Also robust

3. Analysis

3.3. Heterogeneity

- The last step of the analysis is to investigate the potential heterogeneity of the results
  - **Homogenous** effect: The coefficient is more or less **the same** for everybody
  - **Heterogenous** effects: The coefficient **varies a lot** depending on individual (/match) characteristics
  - It can be according to sex, education, income, or here league for instance

- While **robustness** consisted in controlling for variables
  - Estimating the relationship **net of the effect** of other (potentially confounding) **variables**

- **Heterogeneity** consists in interacting x with a third variable
  - By how much the relationship between x and y **varies depending on** the value of a **third variable**

```
stargazer(lm(Winner_home ~ Public, data_match),
          lm(Winner_home ~ Public + League, data_match),
          lm(Winner_home ~ Public + League + Public * League, data_match),
          dep.var.labels = c("Home win"), type = "text", keep.stat = c("n", "adj.rsq"))
```

```
##
## ================================================================
##                                      Dependent variable:
##                                  ---------------------------
##                                             Home win
## ----------------------------------------------------------------
## PublicPublic                       0.059***  0.060***   0.075**
##                                     (0.016)   (0.016)   (0.032)
##
## LeagueBundesliga                                0.003     0.024
##                                                (0.022)   (0.037)
##
## LeagueLa Liga                                   0.019     0.035
##                                                (0.021)   (0.034)
##
## LeaguePremier League                            0.014     0.016
##                                                (0.021)   (0.034)
##
## PublicPublic:LeagueBundesliga                             -0.034
##                                                           (0.046)
##
## PublicPublic:LeagueLa Liga                                -0.026
##                                                           (0.044)
##
## PublicPublic:LeaguePremier League                         -0.002
##                                                           (0.044)
##
## Constant                           0.395***  0.385***  0.376***
##                                     (0.012)   (0.018)   (0.025)
```

- Point estimates are:
  - Ligue 1: 7.5pp
  - Bundesliga: 7.5-3.4=4.1pp
  - La Liga: 7.5-2.6=4.9pp
  - Premier League: 7.5-0.2=7.3pp

- But none of the coefficients associated with interaction terms are statistically significantly different from 0
  - It's sufficiently likely that these variations across groups are just random noise for us not being able to conclude that there is heterogeneity across leagues, at least none we can detect

Overview

1. Preliminary steps ✓
- 1.1. Research question
- 1.2. Finding data
- 1.3. Literature review

2. Data description ✓
- 2.1. Data cleaning
- 2.2. Descriptive statistics
- 2.3. Data visualization

3. Analysis ✓
- 3.1. Regression analysis
- 3.2. Robustness
- 3.3. Heterogeneity

4. Wrap up!

Wrap up!

Preliminary steps

- It all starts with a good **research question**:
  - More about **explanation** than description
  - **Specific** enough, not too broad (Yes/No question, to what extent, ...)
  - Relatively **original and interesting** to you!

- That can be studied with **data**:
  - openICPSR, Harvard Dataverse, American Economic Association, ...
  - With the **necessary variables** to study the research question
  - And **additional variables** to use for robustness and heterogeneity analysis

- And that is relevant with respect to the academic **literature** on the issue:
  - What do we **already know** on the topic?
  - What **remains to be known**?
  - What is your **contribution** to the literature?

➜ That's what you have to do for the next 3 weeks!

Wrap up!

Preliminary steps

- During **lecture 4** you'll have to do at **5 to 10-minute presentation** with slides in which you should:
  - Present and motivate your **research question**
  - Present your **data** (source, main variables description)
  - Present a short review of the **related literature**
- You can come up with **your own** research question or take one **from an existing article**
  - When you have an idea send it by e-mail with the data to be sure it's fine and not taken already
- It will be **graded** (detailed grading scheme here):
  - 25% of the grade on this presentation
  - 30% of the grade on the midterm report
  - 45% of the grade on the final research project/presentation

Please go through the example to get familiar with what is expected from you

- All the following steps of the research process will be subject to weekly 10mn follow-ups by group

Wrap up!

Data description

- After opening and eyeballing the data, the first thing to do is **data cleaning**
  - **Recoding variables** in a practical way (it may imply **creating new variables**)
  - **Removing observations** that are not relevant, if any, typically **missing values**
  - Potentially **joining**, and **pivoting** variables from wide to long or conversely

```
mutate() %>% filter() %>% select()
```

- It should then be summarized with relevant **descriptive statistics**
  - For **categorical variables:** Number of observations per category
  - For **continuous variables:** Summarizing the distribution

```
summarise(N = n()) // summary(variable)
```

- And the last step of the data description is **data visualization** (+/- same thing but with graphs)

```
ggplot(., aes()) +
```

Wrap up!

Analysis

- The analysis should be carried out as follows
  - **Write down the equation** to estimate
  - Estimate it and **interpret properly the coefficient(s)** of interest
  - **Represent graphically** the estimated relationship

$$1\{Winner_m = \text{Home}\} = \alpha + \beta \times 1\{Public_m = \text{Yes}\} + \varepsilon_m$$

```
stargazer(lm(Winner_home~Public, data_match))
```

- And it should be followed by these three steps
  - A discussion on the **causality** of the estimated effect: OVB, selection, ...
  - A **robustness** assessment: Include control variables, omit groups, ...
  - A **heterogeneity** analysis: Interact with third variable(s)

- At some point, add the introduction (with literature review), conclusion, and references sections

Wrap up!

Some important remarks

- Your final document should be an html file produced with **R Markdown**
  - It should be well formatted (stargazer, kable, LaTeX, inline code, ...)
  - It can be written in English or in French

- It should be **reproducible**
  - The R Markdown should knit without error
  - Every data modification should be in the code, it should produce the html document from the raw data

- **When you send an email** related to a coding issue
  - Send your .Rmd and the data in attachment, do not copy-paste your code in the mail nor send screenshots
  - You should first have viewed your data at each step to see where the problem comes from
  - And copy-pasted your error message with keywords on Google to try to understand the problem

- Beware of **technical issues**
  - Knit your .Rmd regularly to check it works
  - Save your files regularly and on multiples devices/on your mailbox