


# R introduction, scraping and text analysis

---

Toni Rodon

February 10, 2022

Universitat de Barcelona

 [www.tonirodon.cat](http://www.tonirodon.cat)

 [@tonirodon](https://twitter.com/tonirodon)

# Hello everyone

- Assistant Professor

# Hello everyone

- Assistant Professor
- Research Fellow at the LSE

# Hello everyone

- Assistant Professor
- Research Fellow at the LSE
- Political Behaviour

# Hello everyone

- Assistant Professor
- Research Fellow at the LSE
- Political Behaviour
- Comparative Politics

# Hello everyone

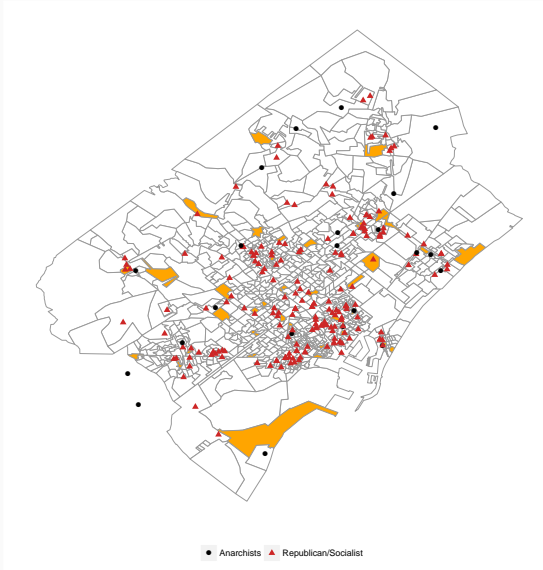
- Assistant Professor
- Research Fellow at the LSE
- Political Behaviour
- Comparative Politics
- Historical Political Economy

# Hello everyone

- Assistant Professor
- Research Fellow at the LSE
- Political Behaviour
- Comparative Politics
- Historical Political Economy
- And lots of maps!

# Digitization

Amat, Francesc; Boix, Carles; Muoz, Jordi; Rodon, Toni (forthcoming) From Political Mobilization to Electoral Participation: Turnout in Barcelona in the 1930s, *The Journal of Politics*.





# Who are you?

- Name / surname
- What do you do?
- Usual data analysis software
- Have you ever used R? If yes (or, if no), why?

## Today

The objective of this course is to learn how to use R.

We will also learn how to **analyze**, **gather** and **work** with social science data (in R).

We will also learn how to not be afraid of using R.

After Tuesday's session, most likely you will feel overwhelmed. This is perfectly fine!

- R is a powerful environment and programming language for statistical computing and graphics.

# History of R

- R is a powerful environment and programming language for statistical computing and graphics.
- R was first developed by Robert Gentleman and Ross Ihaka (U of Auckland, NZ) during the 1990s.

# History of R

- R is a powerful environment and programming language for statistical computing and graphics.
- R was first developed by Robert Gentleman and Ross Ihaka (U of Auckland, NZ) during the 1990s.
- The language used by R is a “dialect” of the S statistical programming language.

# History of R

- R is a powerful environment and programming language for statistical computing and graphics.
- R was first developed by Robert Gentleman and Ross Ihaka (U of Auckland, NZ) during the 1990s.
- The language used by R is a “dialect” of the S statistical programming language.
- R version 1.0.0 is released.

# History of R

- R is a powerful environment and programming language for statistical computing and graphics.
- R was first developed by Robert Gentleman and Ross Ihaka (U of Auckland, NZ) during the 1990s.
- The language used by R is a “dialect” of the S statistical programming language.
- R version 1.0.0 is released.
- CRAN package repository features 18,859 available packages.

- The R system is divided into 2 conceptual parts



- The R system is divided into 2 conceptual parts
  - The “base” R system that you download from CRAN

- The R system is divided into 2 conceptual parts
  - The “base” R system that you download from CRAN
  - Everything else

- Hundreds (thousands!) of tutorials online.

- Hundreds (thousands!) of tutorials online.
  - An introduction to R <https://cran.r-project.org/doc/manuals/r-release/R-intro.html>

- Hundreds (thousands!) of tutorials online.
  - An introduction to R <https://cran.r-project.org/doc/manuals/r-release/R-intro.html>
  - R for Data Science <https://r4ds.had.co.nz/>

- Hundreds (thousands!) of tutorials online.
  - An introduction to R <https://cran.r-project.org/doc/manuals/r-release/R-intro.html>
  - R for Data Science <https://r4ds.had.co.nz/>
  - Quantitative Social Science  
<https://press.princeton.edu/books/hardcover/9780691167039/quantitative-social-science>

- Hundreds (thousands!) of tutorials online.
  - An introduction to R <https://cran.r-project.org/doc/manuals/r-release/R-intro.html>
  - R for Data Science <https://r4ds.had.co.nz/>
  - Quantitative Social Science  
<https://press.princeton.edu/books/hardcover/9780691167039/quantitative-social-science>
  - Quantitative politics with R <http://qpplr.com/>

- Hundreds (thousands!) of tutorials online.
  - An introduction to R <https://cran.r-project.org/doc/manuals/r-release/R-intro.html>
  - R for Data Science <https://r4ds.had.co.nz/>
  - Quantitative Social Science  
<https://press.princeton.edu/books/hardcover/9780691167039/quantitative-social-science>
  - Quantitative politics with R <http://qpplr.com/>
  - Geocomputation with R <https://bookdown.org/robinlovelace/geocompr/intro.html>



- Hundreds (thousands!) of tutorials online.
  - An introduction to R <https://cran.r-project.org/doc/manuals/r-release/R-intro.html>
  - R for Data Science <https://r4ds.had.co.nz/>
  - Quantitative Social Science  
<https://press.princeton.edu/books/hardcover/9780691167039/quantitative-social-science>
  - Quantitative politics with R <http://qpplr.com/>
  - Geocomputation with R <https://bookdown.org/robinlovelace/geocompr/intro.html>
  - A Business Analyst's Introduction to Business Analytics  
<https://www.causact.com/>

- Hundreds (thousands!) of tutorials online.
  - An introduction to R <https://cran.r-project.org/doc/manuals/r-release/R-intro.html>
  - R for Data Science <https://r4ds.had.co.nz/>
  - Quantitative Social Science  
<https://press.princeton.edu/books/hardcover/9780691167039/quantitative-social-science>
  - Quantitative politics with R <http://qpplr.com/>
  - Geocomputation with R <https://bookdown.org/robinlovelace/geocompr/intro.html>
  - A Business Analyst's Introduction to Business Analytics  
<https://www.causact.com/>
  - Big book of R <https://www.bigbookofr.com/index.html>

## Why use R?

- Data handling, wrangling, and storage

# Why use R?

- Data handling, wrangling, and storage
- Wide array of statistical methods and graphical techniques available

# Why use R?

- Data handling, wrangling, and storage
- Wide array of statistical methods and graphical techniques available
- Easy to install on any platform and use (and its free!)

# Why use R?

- Data handling, wrangling, and storage
- Wide array of statistical methods and graphical techniques available
- Easy to install on any platform and use (and its free!)
- Open source with a large and growing community of peers

# Why use R?

- Data handling, wrangling, and storage
- Wide array of statistical methods and graphical techniques available
- Easy to install on any platform and use (and its free!)
- Open source with a large and growing community of peers
- An enthusiastic community (Stackoverflow, R-help mailing list)

# Why use R?

- Data handling, wrangling, and storage
- Wide array of statistical methods and graphical techniques available
- Easy to install on any platform and use (and its free!)
- Open source with a large and growing community of peers
- An enthusiastic community (Stackoverflow, R-help mailing list)
- Used by New York Times, Facebook, Google, Twitter...



# Why use R?

- Data handling, wrangling, and storage
- Wide array of statistical methods and graphical techniques available
- Easy to install on any platform and use (and its free!)
- Open source with a large and growing community of peers
- An enthusiastic community (Stackoverflow, R-help mailing list)
- Used by New York Times, Facebook, Google, Twitter...
- Data analysts are in high demand!

## Why did I decide to use R

- GIS analysis in Stata was almost non-existent

## Why did I decide to use R

- GIS analysis in Stata was almost non-existent
- Stata graphs are (very) ugly

# Why did I decide to use R

- GIS analysis in Stata was almost non-existent
- Stata graphs are (very) ugly
- Most top universities mainly use R (or Python)

# Why did I decide to use R

- GIS analysis in Stata was almost non-existent
- Stata graphs are (very) ugly
- Most top universities mainly use R (or Python)
- Using R made me think about what I was doing

# Why did I decide to use R

- GIS analysis in Stata was almost non-existent
- Stata graphs are (very) ugly
- Most top universities mainly use R (or Python)
- Using R made me think about what I was doing
- R had functions Stata did not have—scraping, text analysis...

# Why did I decide to use R

- GIS analysis in Stata was almost non-existent
- Stata graphs are (very) ugly
- Most top universities mainly use R (or Python)
- Using R made me think about what I was doing
- R had functions Stata did not have—scraping, text analysis...
- It was very annoying to waste my time (paying) or searching for a license key

## Why did I decide to use R

- I found it very weird to be using a graph scheme in Stata to copy R graphs



## Why did I decide to use R

- I found it very weird to be using a graph scheme in Stata to copy R graphs
- I don't know anyone transitioning from R to Stata

## Why did I decide to use R

- I found it very weird to be using a graph scheme in Stata to copy R graphs
- I don't know anyone transitioning from R to Stata
- It pays off!

## Things you can do in R

- Automating your weekly reports (automating tasks)

# Things you can do in R

- Automating your weekly reports (automating tasks)
- Analyzing data (modeling) or creating your on data model (in a fairly easy way)

# Things you can do in R

- Automating your weekly reports (automating tasks)
- Analyzing data (modeling) or creating your on data model (in a fairly easy way)
- Creating nicely formatted documents (communicating results)

# Things you can do in R

- Automating your weekly reports (automating tasks)
- Analyzing data (modeling) or creating your on data model (in a fairly easy way)
- Creating nicely formatted documents (communicating results)
- Create interactive maps and export them as a web page

# Things you can do in R

- Automating your weekly reports (automating tasks)
- Analyzing data (modeling) or creating your on data model (in a fairly easy way)
- Creating nicely formatted documents (communicating results)
- Create interactive maps and export them as a web page
- Analyze your whatsapp messages ([https://cran.r-project.org/web/packages/rwhatsapp/vignettes/Text\\_Analysis\\_using\\_WhatsApp\\_data.html](https://cran.r-project.org/web/packages/rwhatsapp/vignettes/Text_Analysis_using_WhatsApp_data.html) or Telegram messages (<https://cran.r-project.org/web/packages/telegram/README.html>))

# Things you can do in R

- Automating your weekly reports (automating tasks)
- Analyzing data (modeling) or creating your on data model (in a fairly easy way)
- Creating nicely formatted documents (communicating results)
- Create interactive maps and export them as a web page
- Analyze your whatsapp messages ([https://cran.r-project.org/web/packages/rwhatsapp/vignettes/Text\\_Analysis\\_using\\_WhatsApp\\_data.html](https://cran.r-project.org/web/packages/rwhatsapp/vignettes/Text_Analysis_using_WhatsApp_data.html) or Telegram messages (<https://cran.r-project.org/web/packages/telegram/README.html>))
- Analyze your instagram account (<https://github.com/pablobarbera/instaR>)



# Things you can do in R

- Automating your weekly reports (automating tasks)
- Analyzing data (modeling) or creating your on data model (in a fairly easy way)
- Creating nicely formatted documents (communicating results)
- Create interactive maps and export them as a web page
- Analyze your whatsapp messages ([https://cran.r-project.org/web/packages/rwhatsapp/vignettes/Text\\_Analysis\\_using\\_WhatsApp\\_data.html](https://cran.r-project.org/web/packages/rwhatsapp/vignettes/Text_Analysis_using_WhatsApp_data.html) or Telegram messages (<https://cran.r-project.org/web/packages/telegram/README.html>))
- Analyze your instagram account (<https://github.com/pablobarbera/instaR>)
- Machine learning

# Things you can do in R

- Automating your weekly reports (automating tasks)
- Analyzing data (modeling) or creating your on data model (in a fairly easy way)
- Creating nicely formatted documents (communicating results)
- Create interactive maps and export them as a web page
- Analyze your whatsapp messages ([https://cran.r-project.org/web/packages/rwhatsapp/vignettes/Text\\_Analysis\\_using\\_WhatsApp\\_data.html](https://cran.r-project.org/web/packages/rwhatsapp/vignettes/Text_Analysis_using_WhatsApp_data.html) or Telegram messages (<https://cran.r-project.org/web/packages/telegram/README.html>))
- Analyze your instagram account (<https://github.com/pablobarbera/instaR>)
- Machine learning
- Create 3-D objects (<https://www.tylermw.com/3d-ggplots-with-rayshader/>)

## Things you can do in R (II)

- Extract text from images (<https://cran.r-project.org/web/packages/tesseract/vignettes/intro.html> or [https://ropensci.org/tutorials/tabulizer\\_tutorial/](https://ropensci.org/tutorials/tabulizer_tutorial/))

## Things you can do in R (II)

- Extract text from images (<https://cran.r-project.org/web/packages/tesseract/vignettes/intro.html> or [https://ropensci.org/tutorials/tabulizer\\_tutorial/](https://ropensci.org/tutorials/tabulizer_tutorial/))
- Many new libraries almost every day ([https://cran.r-project.org/web/packages/available\\_packages\\_by\\_date.html](https://cran.r-project.org/web/packages/available_packages_by_date.html))

## Things you can do in R (II)

- Extract text from images (<https://cran.r-project.org/web/packages/tesseract/vignettes/intro.html> or [https://ropensci.org/tutorials/tabulizer\\_tutorial/](https://ropensci.org/tutorials/tabulizer_tutorial/))
- Many new libraries almost every day ([https://cran.r-project.org/web/packages/available\\_packages\\_by\\_date.html](https://cran.r-project.org/web/packages/available_packages_by_date.html))
- It facilitates reproducibility

## Things you can do in R (II)

- Extract text from images (<https://cran.r-project.org/web/packages/tesseract/vignettes/intro.html> or [https://ropensci.org/tutorials/tabulizer\\_tutorial/](https://ropensci.org/tutorials/tabulizer_tutorial/))
- Many new libraries almost every day ([https://cran.r-project.org/web/packages/available\\_packages\\_by\\_date.html](https://cran.r-project.org/web/packages/available_packages_by_date.html))
- It facilitates reproducibility
- It facilitates replicability

## Things you can do in R (II)

- Extract text from images (<https://cran.r-project.org/web/packages/tesseract/vignettes/intro.html> or [https://ropensci.org/tutorials/tabulizer\\_tutorial/](https://ropensci.org/tutorials/tabulizer_tutorial/))
- Many new libraries almost every day ([https://cran.r-project.org/web/packages/available\\_packages\\_by\\_date.html](https://cran.r-project.org/web/packages/available_packages_by_date.html))
- It facilitates reproducibility
- It facilitates replicability
- Using R you can even order a pizza!

# Disadvantages

- R is not perfect



# Disadvantages

- R is not perfect
- Steeper learning curve than other languages/software

# Disadvantages

- R is not perfect
- Steeper learning curve than other languages/software
- R is not always the best tool for everything

# Disadvantages

- R is not perfect
- Steeper learning curve than other languages/software
- R is not always the best tool for everything
  - Digitize old maps

# Disadvantages

- R is not perfect
- Steeper learning curve than other languages/software
- R is not always the best tool for everything
  - Digitize old maps
  - Some tools for text analysis

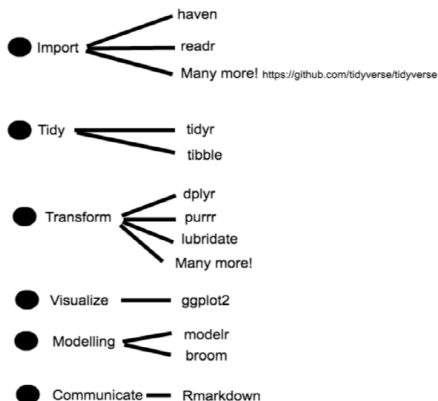
# Disadvantages

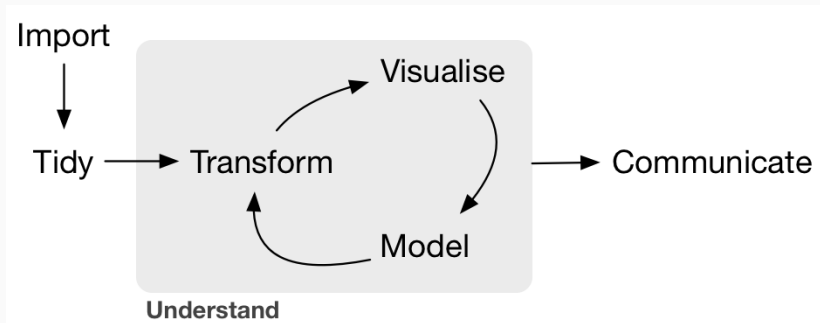
- R is not perfect
- Steeper learning curve than other languages/software
- R is not always the best tool for everything
  - Digitize old maps
  - Some tools for text analysis
  - etc.

# Disadvantages

- R is not perfect
- Steeper learning curve than other languages/software
- R is not always the best tool for everything
  - Digitize old maps
  - Some tools for text analysis
  - etc.
- R works for small/medium sized data (me: up to 8 million observations)

# Things you can do in R (III)







### **Hadley Wickham**

The bad news is that when ever you learn a new skill youre going to suck. It's going to be frustrating. The good news is that is typical and happens to everyone and it is only temporary. You cant go from knowing nothing to becoming an expert without going through a period of great frustration and great suckiness.

**Kosuke Imai**

One can learn data analysis only by doing, not by reading.

- Do not use the console, write scripts instead

- Do not use the console, write scripts instead
- Don't be lazy—don't use the menu!

- Do not use the console, write scripts instead
- Don't be lazy—don't use the menu!
- Think before you code

- Do not use the console, write scripts instead
- Don't be lazy—don't use the menu!
- Think before you code
- Code is a medium of communication

- Do not use the console, write scripts instead
- Don't be lazy—don't use the menu!
- Think before you code
- Code is a medium of communication  
Between you and the computer

- Do not use the console, write scripts instead
- Don't be lazy—don't use the menu!
- Think before you code
- Code is a medium of communication

Between you and the computer

Between you and other people



- Do not use the console, write scripts instead
- Don't be lazy—don't use the menu!
- Think before you code
- Code is a medium of communication

Between you and the computer

Between you and other people

Between you and your future you

- Do not use the console, write scripts instead
- Don't be lazy—don't use the menu!
- Think before you code
- Code is a medium of communication
  - Between you and the computer
  - Between you and other people
  - Between you and your future you
- Comment your code

- Do not use the console, write scripts instead
- Don't be lazy—don't use the menu!
- Think before you code
- Code is a medium of communication
  - Between you and the computer
  - Between you and other people
  - Between you and your future you
- Comment your code
- Code should (must!) work from the beginning to the end

# Future self



- R is an object-based language

# Creating objects

- R is an object-based language
- Everything has a name

# Creating objects

- R is an object-based language
- Everything has a name
- Everything is an object

# Creating objects

- R is an object-based language
- Everything has a name
- Everything is an object
- Every object has a class



# Creating objects

- R is an object-based language
- Everything has a name
- Everything is an object
- Every object has a class
- There is no agreement about how to name things. You will likely see variables named `snake_case` or `SnakeCase` or `snakecase`

# Creating objects

- R is an object-based language
- Everything has a name
- Everything is an object
- Every object has a class
- There is no agreement about how to name things. You will likely see variables named `snake_case` or `SnakeCase` or `snakecase`
- There is consensus, however, that variable names can't have blank spaces

## Other important points

- R is case-sensitive: `Toni_Rodon` is not the same variable or object than `toni_rodon` or even than `toni_Rodon`

## Other important points

- R is case-sensitive: `Toni_Rodon` is not the same variable or object than `toni_rodon` or even than `toni_Rodon`
- An object can only be of one class (not hybrid objects)

## Other important points

- R is case-sensitive: `Toni_Rodon` is not the same variable or object than `toni_rodon` or even than `toni_Rodon`
- An object can only be of one class (not hybrid objects)
- It is possible to work with multiple objects at the same time (multiple “datasets”)

## Other important points

- R is case-sensitive: `Toni_Rodon` is not the same variable or object than `toni_rodon` or even than `toni_Rodon`
- An object can only be of one class (not hybrid objects)
- It is possible to work with multiple objects at the same time (multiple “datasets”)
- This means we often need to “call” the object (i.e. when using a variable in a dataset)

## Examples of R objects

- character string (i.e. words)

## Examples of R objects

- character string (i.e. words)
- number



## Examples of R objects

- character string (i.e. words)
- number
- vector

## Examples of R objects

- character string (i.e. words)
- number
- vector
- matrix

## Examples of R objects

- character string (i.e. words)
- number
- vector
- matrix
- list

## Examples of R objects

- character string (i.e. words)
- number
- vector
- matrix
- list
- data frame

- NA: not available, missing (`is.na`)

## Special values

- NA: not available, missing (`is.na`)
- undefined (`is.null`)

## Special values

- NA: not available, missing (`is.na`)
- undefined (`is.null`)
- TRUE: logical TRUE (`isTRUE`)

## Special values

- NA: not available, missing (`is.na`)
- undefined (`is.null`)
- TRUE: logical TRUE (`isTRUE`)
- FALSE: logical FALSE (`!isTRUE`)



- R stores spreadsheets like data in a data frame

- R stores spreadsheets like data in a data frame
- These are really collections of list vectors of the same length

- R stores spreadsheets like data in a data frame
- These are really collections of list vectors of the same length
- Tip: Create data frames whenever you can

- On its own, R can't do all that much

## Packages or libraries

- On its own, R can't do all that much
- To really make use of R's capabilities, we need packages

- On its own, R can't do all that much
- To really make use of R's capabilities, we need packages
- A package bundles together code, data, documentation, and tests

- On its own, R can't do all that much
- To really make use of R's capabilities, we need packages
- A package bundles together code, data, documentation, and tests
- We install packages from two sources:

- On its own, R can't do all that much
- To really make use of R's capabilities, we need packages
- A package bundles together code, data, documentation, and tests
- We install packages from two sources:
  - The Comprehensive R Archive Network (CRAN)



- On its own, R can't do all that much
- To really make use of R's capabilities, we need packages
- A package bundles together code, data, documentation, and tests
- We install packages from two sources:
  - The Comprehensive R Archive Network (CRAN)
  - Github

## New information on libraries?

- Twitter.

## New information on libraries?

- Twitter.
- rblogger <https://www.r-bloggers.com/>

## New information on libraries?

- Twitter.
- rlogger <https://www.r-bloggers.com/>
- ropensci <https://ropensci.org/>

## New information on libraries?

- Twitter.
- rlogger <https://www.r-bloggers.com/>
- ropensci <https://ropensci.org/>
- github [github.com/](https://github.com/)

## New information on libraries?

- Twitter.
- rlogger <https://www.r-bloggers.com/>
- ropensci <https://ropensci.org/>
- github [github.com/](https://github.com/)
- #tidytuesday

- Basics

# Today

- Basics
- Data wrangling



# Today

- Basics
- Data wrangling
- Plotting

# Today

- Basics
- Data wrangling
- Plotting
- Univariate analysis

- Basics
- Data wrangling
- Plotting
- Univariate analysis
- Bivariate analysis

- Basics
- Data wrangling
- Plotting
- Univariate analysis
- Bivariate analysis
- Multivariate analysis

- Please ask for a breack if you feel your brain is melting.

- Please ask for a breack if you feel your brain is melting.
- I will show some code.

- Please ask for a breack if you feel your brain is melting.
- I will show some code.
- You will try to reproduce it—and replicate it.

- Please ask for a breack if you feel your brain is melting.
- I will show some code.
- You will try to reproduce it—and replicate it.
- Exercises - and we will go over them together.



**R**

Let's go for it!

- [https://github.com/hbctraining/Intro-to-R/blob/master/lessons/01\\_introR-R-and-RStudio.md](https://github.com/hbctraining/Intro-to-R/blob/master/lessons/01_introR-R-and-RStudio.md)


# R introduction, scraping and text analysis

---

Toni Rodon

February 10, 2022

Universitat de Barcelona

 [www.tonirodon.cat](http://www.tonirodon.cat)

 [@tonirodon](https://twitter.com/tonirodon)