```
library(dplyr)
library(magrittr)

rladies_global %>% filter(city == ‘Resistencia')
&& filter(city == ‘Corrientes')
```

RLadies
Resistencia - Corrientes

Manipulación de Datos con tidyr y dplyr

Viernes 7 de Diciembre de 2018

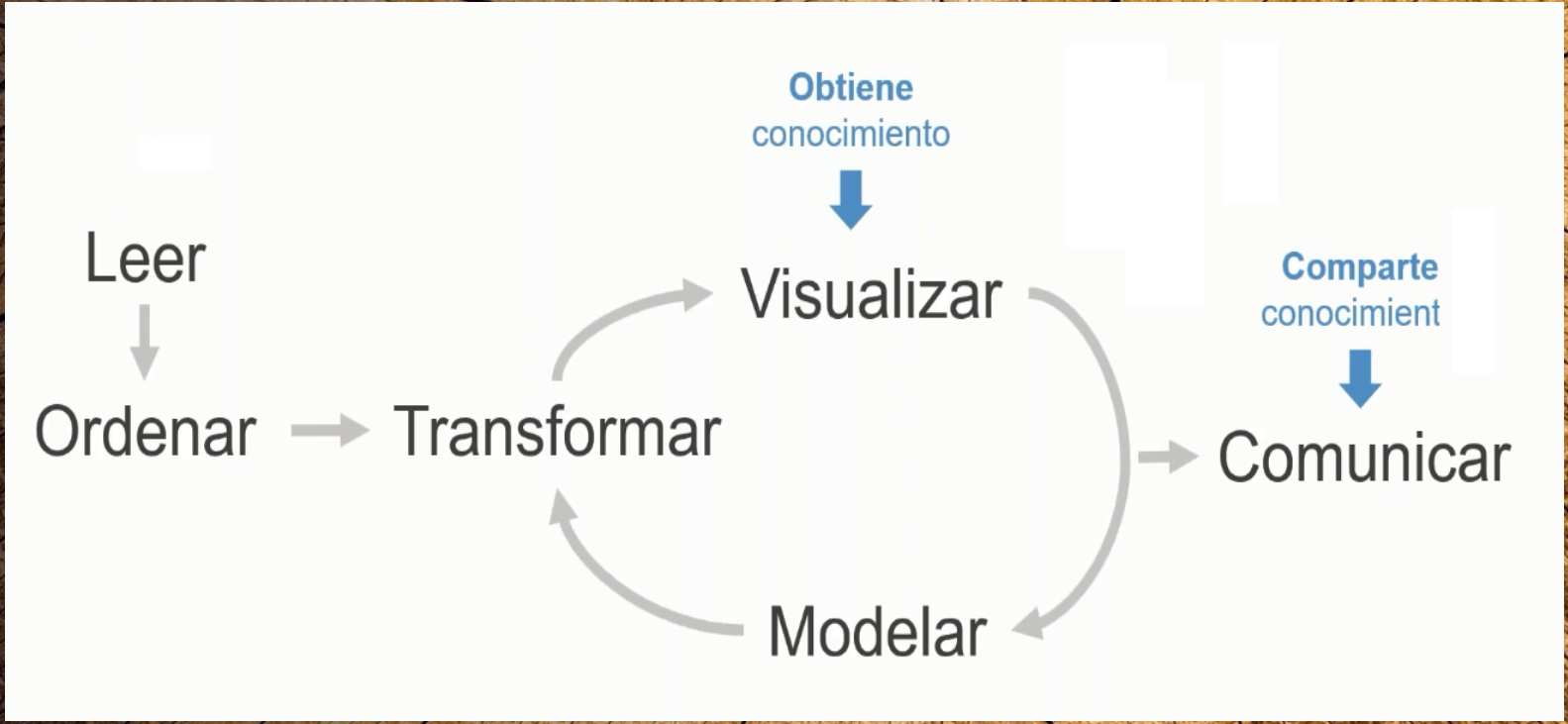
Hoy hablamos sobre…

PARTE 1: DATOS ORDENADOS CON TIDYR

PARTE 2: MANIPULACIÓN DE DATOS CON DPLYR

Operador %>%

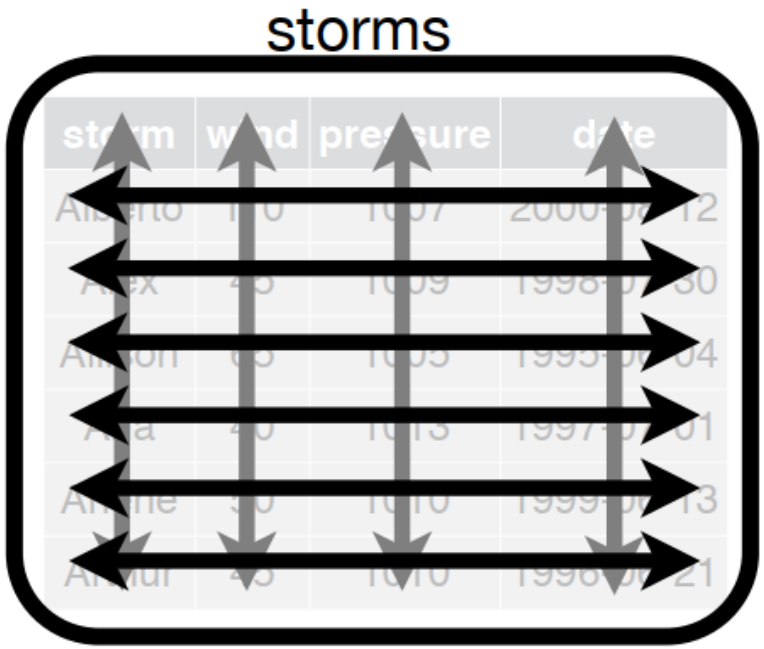
DATA SCIENCE WORKFLOW



PARTE 1

ORDENAR DATOS CON tidyr

Tidy data



1. Cada **variable** está en su **columna**
2. Cada **observación** está en una **fila**
3. Cada unidad de experimentación está en una tabla aparte

Las bases de datos para nuestro trabajo

install.packages(“devtools”)
devtools::install_github(“rstudio/EDAWR”)

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

cases

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

pollution

city	particle size	amount ($\mu g/m^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

Tidy data

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	[illegible]	1010	1996-06-21

- Storm name
- Wind Speed (mph)
- Air Pressure
- Date

cases

Country	[illegible]	[illegible]	[illegible]
FR	7000	[illegible]	7000
DE	5800	6000	6200
US	15000	[illegible]	13000

- Country
- Year
- Count

pollution

city	particle size	amount ($\mu g/m^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

- City
- Amount of large particles
- Amount of small particles

Tidy data

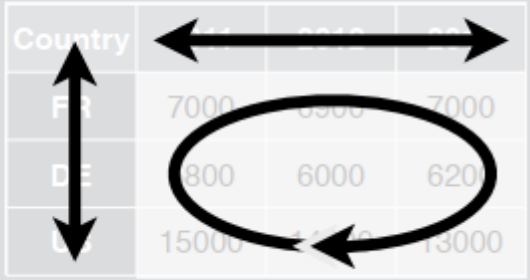
1. Instalamos el paquete tidyr
   install.packages("tidyr")
1. Cargamos la librería
   library(tidyr)
1. Funciones importantes: **gather()** y **spread()**
   ?gather
   ?spread

Nuestro turno de ordenar los datos

Ordenamos teniendo en cuenta 3 variables: ***country, year, n.***

cases

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000



Función gather()

Colapsar varias columnas en una sola columna

1. Una columna key que contiene los nombres de las columnas
2. Un valor que contiene los valores de las columnas.

gather (cases, year, n, 2:4)

número de columnas que colapsan

set de datos

columna key

columna values

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

gather()

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

gather (cases, year, n, 2:4)

número de columnas que colapsan

set de datos

columna key

columna values

Nuestro turno de ordenar los datos

Ordenamos teniendo en cuenta 3 variables: ***city, large, small.***

pollution

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	particle size	amount (μg/m³)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

Función spread()

Genera varias columnas a partir de dos columnas.

1. Un único valor en la columna *key* se convierte en una columna única.
2. Cada valor *value* se convierte en una fila en una nueva columna

spread (pollution, size, amount)

set de datos

nuevas columnas

nuevas filas

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

spread()

gather()

city	large	small
New York	23	14
London	22	16
Beijing	121	56

spread (pollution, size, amount)

nuevas filas

set de datos

nuevas columnas

Función separate()

Permite separar una columna en varias con un separador

```
separate(storms, date, c("year", "month", "day"), sep = "-")
```

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

storms2

storm	wind	pressure	year	month	day
Alberto	110	1007	2000	08	12
Alex	45	1009	1998	07	30
Allison	65	1005	1995	06	04
Ana	40	1013	1997	07	1
Arlene	50	1010	1999	06	13
Arthur	45	1010	1996	06	21

Función unite()

Permite unir columnas en una sola

```
unite(storms2, "date", year, month, day, sep = "-")
```

storms2

storm	wind	pressure	year	month	day
Alberto	110	1007	2000	08	12
Alex	45	1009	1998	07	30
Allison	65	1005	1995	06	04
Ana	40	1013	1997	07	1
Arlene	50	1010	1999	06	13
Arthur	45	1010	1996	06	21

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

Recapitulamos

* **gather()**: realizar observaciones a partir de variables
* **spread()**: realizar variables de observaciones
* Unir y separar columnas con **unite()** y **separate()**

PARTE 2

MANIPULAR DATOS CON dplyr

dplyr

- Paquete que nos ayuda a transformar datos tabulares.
- El paquete **dplyr** fue desarrollado por Hadley Wickham y es un versión optimizada de su paquete **plyr**.
- Proporciona una "gramática" (particularmente verbos) para la manipulación y operaciones con data frames.
- Instalación:

```
install.packages(“dplyr”)
library(dplyr)
install.packages(“nycflights13”)
library(nycflights13)
```

Extraer **variables** existentes: **select()**

Extraer **observaciones** existentes: **filter()**

Derivar nuevas **variables**: **mutate()**

Cambiar la unidad de análisis: **summarise()**

Organizar **filas** por variables: **arrange()**

Función **SELECT()**

select()

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

storm	pressure
Alberto	1007
Alex	1009
Allison	1005
Ana	1013
Arlene	1010
Arthur	1010

```
select(storms, storm, pressure)
```

Función **SELECT()**

select()

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

→

wind	pressure	date
110	1007	2000-08-12
45	1009	1998-07-30
65	1005	1995-06-04
40	1013	1997-07-01
50	1010	1999-06-13
45	1010	1996-06-21

```
select(storms, -storm)
```

Función **SELECT()**

select()

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

→

wind	pressure	date
110	1007	2000-08-12
45	1009	1998-07-30
65	1005	1995-06-04
40	1013	1997-07-01
50	1010	1999-06-13
45	1010	1996-06-21

```
select(storms, wind:date)
```

Función **FILTER()**

filter()

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

→

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Allison	65	1005	1995-06-04
Arlene	50	1010	1999-06-13

```
filter(storms, wind >= 50)
```

Función **FILTER()**

filter()

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

→

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Allison	65	1005	1995-06-04

```
filter(storms, wind >= 50,
  storm %in% c("Alberto", "Alex", "Allison"))
```

Función **MUTATE()**

mutate()

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

storm	wind	pressure	date	ratio
Alberto	110	1007	2000-08-12	9.15
Alex	45	1009	1998-07-30	22.42
Allison	65	1005	1995-06-04	15.46
Ana	40	1013	1997-07-01	25.32
Arlene	50	1010	1999-06-13	20.20
Arthur	45	1010	1996-06-21	22.44

```
mutate(storms, ratio = pressure / wind)
```

Función **MUTATE()**

mutate()

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

storm	wind	pressure	date	ratio	inverse
Alberto	110	1007	2000-08-12	9.15	0.11
Alex	45	1009	1998-07-30	22.42	0.04
Allison	65	1005	1995-06-04	15.46	0.06
Ana	40	1013	1997-07-01	25.32	0.04
Arlene	50	1010	1999-06-13	20.20	0.05
Arthur	45	1010	1996-06-21	22.44	0.04

```
mutate(storms, ratio = pressure / wind, inverse = ratio^-1)
```

Función **SUMMARISE()**

summarise()

city	particle size	amount (μg/m³)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

mean	sum	n
42	252	6

```
pollution %>% summarise(mean = mean(amount), sum = sum(amount), n = n())
```

Función **ARRANGE()**

arrange()

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

→

storm	wind	pressure	date
Ana	40	1013	1997-07-01
Alex	45	1009	1998-07-30
Arthur	45	1010	1996-06-21
Arlene	50	1010	1999-06-13
Allison	65	1005	1995-06-04
Alberto	110	1007	2000-08-12

```
arrange(storms, wind)
```

Función **ARRANGE()**

arrange()

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

storm	wind	pressure	date
Ana	40	1013	1997-07-01
Arthur	45	1010	1996-06-21
Alex	45	1009	1998-07-30
Arlene	50	1010	1999-06-13
Allison	65	1005	1995-06-04
Alberto	110	1007	2000-08-12

```
arrange(storms, wind, date)
```

Recapitulamos

* **select()**: extraer variables
* **filter()**: extraer observaciones
* **mutate()** : crear nuevas variables
* **summarize():** cambiar unidad de análisis
* **arrange():** organizar **filas** por variables

El operador pipe %>% nos permite conectar múltiples acciones en una única “pipeline”

Con %>% podemos reescribir los comandos anteriores

```
select(storms, storm, pressure)
```

```
storms %>% select(storm, pressure)
```

```
storms %>% filter(wind >= 50)
```

```
storms %>%
  filter(wind >= 50) %>%
  select(storm, pressure)
```

El operador pipe nos permite una sintaxis clara y entendible

```
select(storms, storm, pressure)
```

Empieza con un verbo

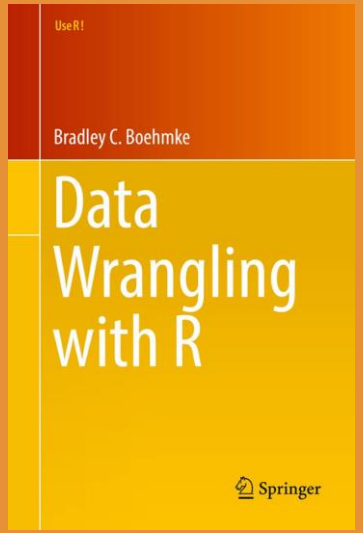
```
storms %>% select(storm, pressure)
```

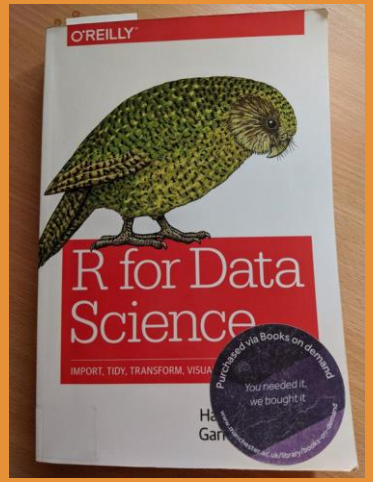
```
storms %>% filter(wind >= 50)
```

```
storms %>%
  filter(wind >= 50) %>%
  select(storm, pressure)
```

Empieza con un sustantivo (dataset) y luego la operación se indica con un verbo

Fuentes de Consulta

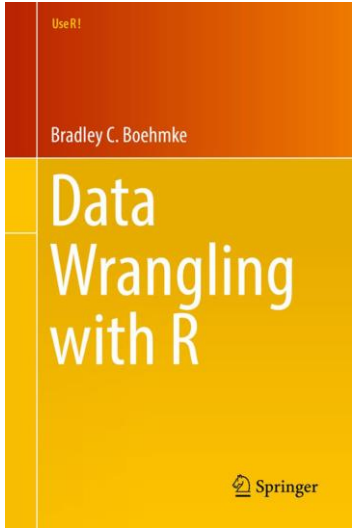


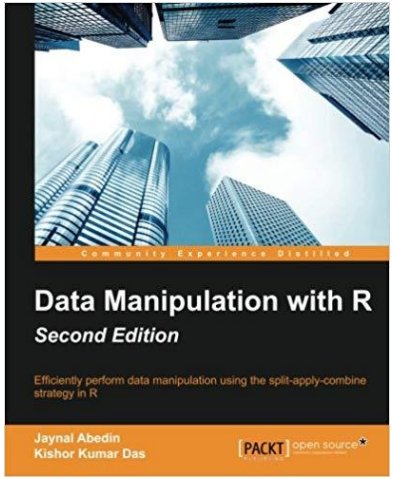


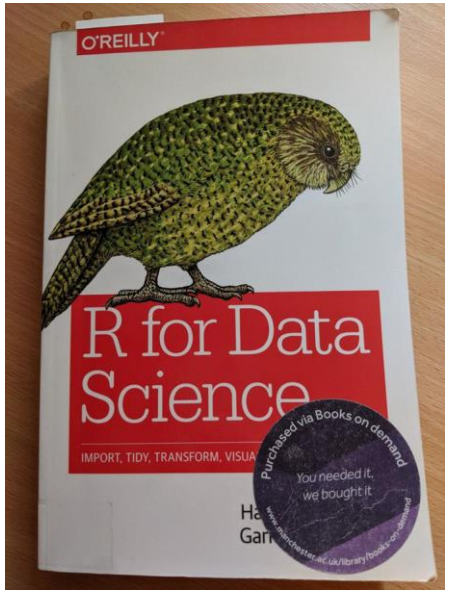
- Learning R (Github) http://bit.ly/2Aaq6d3
- R studio cheatsheets (dplyr, data.table)
- Documentación del CRAN
- Libros

Fuentes de consulta

- Learning R (Github) http://bit.ly/2Aaq6d3
- R studio cheatsheets (dplyr, data.table)
- Documentación del CRAN
- Libros





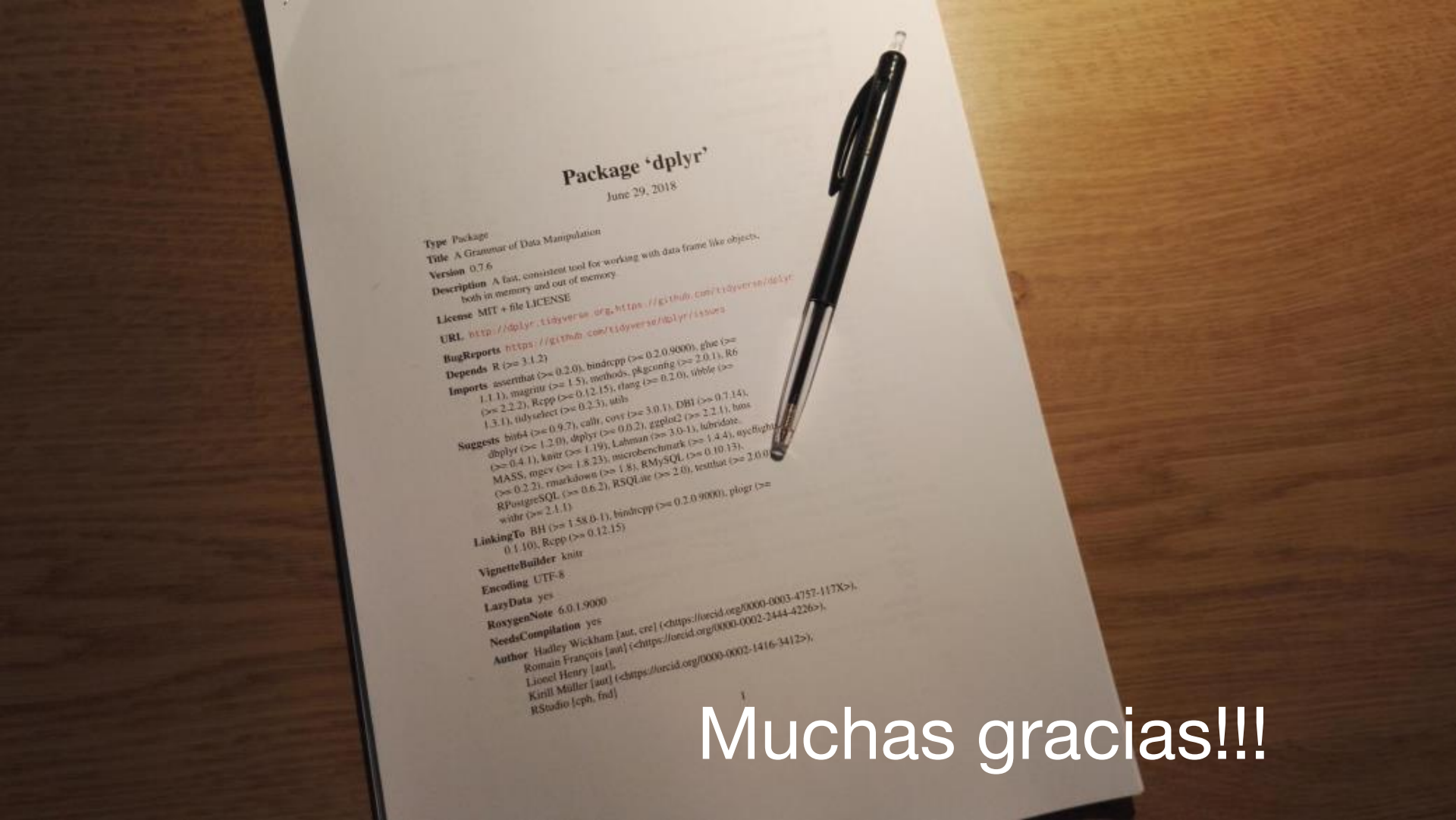


There’s no failure, only feedback

Gabriela de Queiroz
R-Ladies Founder

“If you're not having fun, you're not learning.”

Richard Feynman (1918 - 1988)
Nobel de Física -1965



Muchas gracias!!!

HASTA EL 2019!!!