

# R desde Cero

Yanina Bellini Saibene

@yabellini

bellini.yanina@inta.gob.ar



# Antes de arrancar



un super repasito de la 1ra clase



# ¡Ayuda!

- Para ayuda con R la fuente principal es Google
  - Agregando “R tidy” a cualquier pregunta (mejor en inglés)
  - Incluyendo mensajes de error (mejor en inglés)
  - Los foros (StackOverflow, Rstudio Community)
  - Las hojas de referencia
  - La comunidad R



# CalculadoRa

- R también es una calculadora

```
> 1 + 2
```

- Respuesta

```
[1] 3
```

- Si quiero guardar la respuesta

```
> resultado <- 1 + 2
```

- Para ver la respuesta

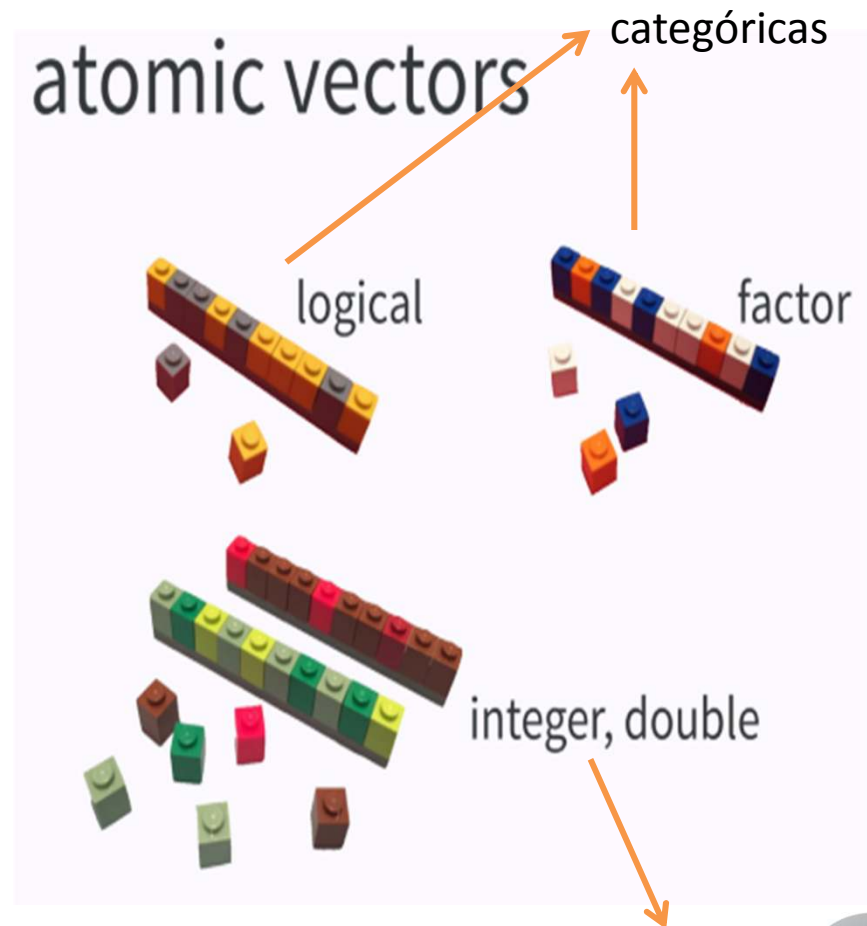
```
> resultado
```

**Variable:** un solo dato.

**Vector:** varios datos.

Condición especial:

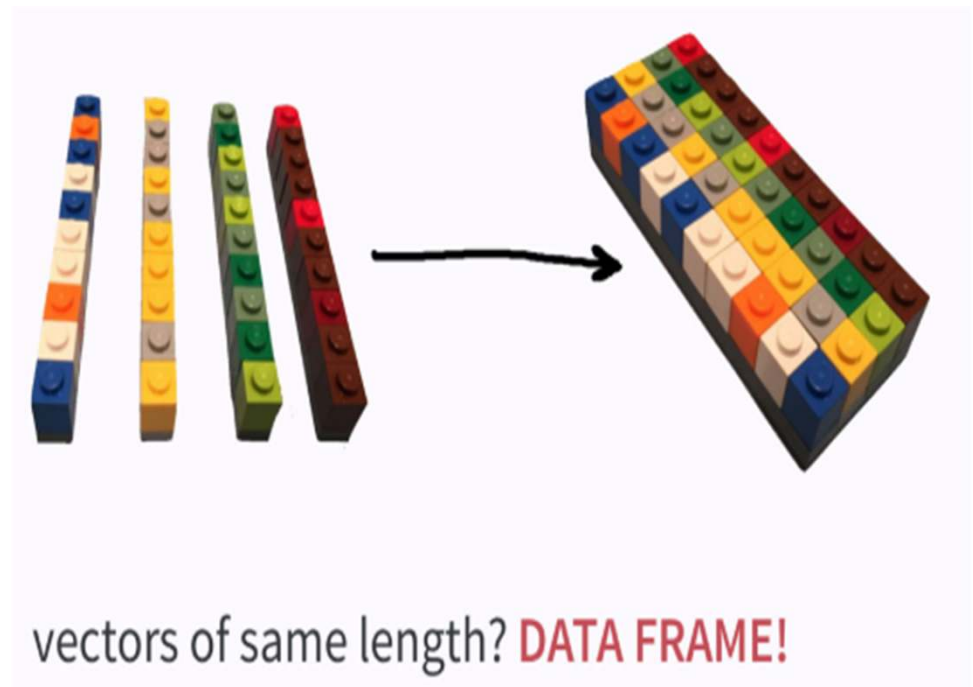
- Todo en el vector debe ser el mismo tipo de datos (double, integer, complex, logical o character)



# **data.frame:** matriz de datos, varias filas y columnas

las columnas de datos que cargamos en los data frames son todos vectores

**Por eso R hace que todo en una columna sea el mismo tipo de datos básicos.**



mmmmm.....y como es eso?



Entremos a Rstudio....



# Analicemos la salida:

Tibble: es un tipo de data.frame

Cantidad de filas que tiene el tibble

Cantidad de columnas

columnas/variables

```
> vuelos
# A tibble: 336,776 x 19
  año   mes   día horario_salida salida_programa~ atraso_salida horario_llegada
  <int> <int> <int>         <int>         <int>         <dbl>         <int>
1  2013     1     1           517           515             2           830
2  2013     1     1           533           529             4           850
3  2013     1     1           542           540             2           923
4  2013     1     1           544           545            -1          1004
5  2013     1     1           554           600            -6           812
6  2013     1     1           554           558            -4           740
7  2013     1     1           555           600            -5           913
8  2013     1     1           557           600            -3           709
9  2013     1     1           557           600            -3           838
10 2013     1     1           558           600            -2           753
# ... with 336,766 more rows, and 5 more variables: tiempo_vuelo <dbl>, distancia <
```

10 primeras filas/casos

Tipo de dato



# Analicemos la salida:

columnas/variables

filtro

	anio	mes	dia	horario_salida	salida_programada	atraso_salida	horario_llegada	llegada_programada	atraso_llegada
1	2013	1	1	517	515	2	830	819	11
2	2013	1	1	533	529	4	850	830	20
3	2013	1	1	542	540	2	923	850	73
4	2013	1	1	544	545	-1	1004	1022	-18
5	2013	1	1	554	600	-6	812	837	-25
6	2013	1	1	554	558	-4	740	728	12
7	2013	1	1	555	600	-5	913	854	59
8	2013	1	1	557	600	-3	709	723	-14
9	2013	1	1	557	600	-3	838	846	-8

Showing 1 to 10 of 336,776 entries

Cantidad de filas

## Analicemos la salida:

## Cantidad de columnas

Cantidad de filas

## Tipo de dato

```
observations: 336,776
variables: 19
```

336,776

[illegible]

columnas/variables



# Analicemos la salida:

```
> kable(aerolineas)
```

codigo_carrier	nombre
9E	Endeavor Air Inc.
AA	American Airlines Inc.
AS	Alaska Airlines Inc.
B6	JetBlue Airways
DL	Delta Air Lines Inc.
EV	ExpressJet Airlines Inc.
F9	Frontier Airlines Inc.
FL	AirTran Airways Corporation
HA	Hawaiian Airlines Inc.
MQ	Envoy Air
OO	Skywest Airlines Inc.
UA	United Air Lines Inc.
US	US Airways Inc.
VX	Virgin America
WN	Southwest Airlines Co.
YV	Mesa Airlines Inc.

columns/variables

valores

¿A qué hace referencia UNA fila en este conjunto de datos de vuelos?

- A. Datos de una aerolínea
- B. Datos de un vuelo.
- C. Datos de un aeropuerto.
- D. Datos de vuelos múltiples.

¿Cuáles son algunos ejemplos en este conjunto de datos de variables **categóricas**?

¿Qué las hace diferentes a las variables **cuantitativas**?

¿Qué ejemplos de variables **cuantitativas** encontramos en vuelos?



# Importando y explorando datos propios







Foto: gentilha Mauro Lepore



# Mate break



# Datos ordenados y datos limpios

El 80% del tiempo del **análisis de datos** se **utiliza** en el proceso de **limpieza y preparación** de los datos.

Esta tarea se realiza varias veces durante el análisis de los datos

Datos ordenados (Tidy Data): **estructuración** de conjuntos de datos para **facilitar el análisis**.



# Principios de Tidy Data

CULTIVAR	Días a floración	Altura (cm)	Vuelco (%)	Densidad (pl/ha)	Humedad de grano	Rendimiento de granos (kg/ha)	Aceite (%)
ACA 203 CL	85	181	0	48554	6.1	2719	43.6
ACA 861	85	166	0	47521	6.1	2319	51.8
ACA 869	87	189	3	45455	6.0	2300	54.0

Observación

Variable ó Atributo

1. Cada **variable** es una **columna**.
2. Cada **observación** es una **fila**.
3. Cada **tipo de unidad de observación** forma una **tabla**.



# Síntomas comunes de datos desordenados

- Los encabezados de columna son valores, no nombres de variables.
- Múltiples variables se almacenan en una columna.
- Las variables se almacenan tanto en filas como en columnas.
- Múltiples tipos de unidades de observación se almacenan en la misma tabla.
- Una sola unidad de observación se almacena en varias tablas.

# Síntomas comunes de datos desordenados

**Los encabezados de columna son valores, no nombres de variables**

DEPART	CABECERA	SUP_JURI	AVENA	CEBADA	CENTENO	TRIGO	GIRASOL	MAIZ
CHICALCO	LA PASTORIL	9117	0	0	0	0	0	0
LIMAY MAHUIDA	LIMAY MAHUIDA	9985	0	0	0	0	0	0
CHALILEO	SANTA ISABEL	8917	0	0	0	0	0	0
HUCAL	BERNASCONI	6047	3363	182	219	12606	1289	226
LOVENTUE	VICTORICA	9235	208	39	77	1256	603	337
RANCUL	RANCUL	4933	2679	413	1614	12135	33910	14696




DEPART	CABECERA	SUP_JURI	CULTIVO	SUPERFICIE
CHICALCO	LA PASTORIL	9117	AVENA	0
CHICALCO	LA PASTORIL	9117	CEBADA	0
CHICALCO	LA PASTORIL	9117	CENTENO	0
CHICALCO	LA PASTORIL	9117	TRIGO	0
LOVENTUE	VICTORICA	9235	GIRASOL	0
CHICALCO	LA PASTORIL	9117	MAIZ	0

# Síntomas comunes de datos desordenados

**Múltiples variables se almacenan en una columna**

CULTIVAR	Días a floración	Días a madurez	Altura (cm)	Vuelco (%)
ACA 203 CL (ACA)	85	122	181	0
ACA 861 (ACA)	85	122	166	0
Aguara 6 (ADVANTA)	85	126	174	0
CACIQUE 312 CL (EL CENCERRO)	90	127	161	1
KWS 480 CL (KWS)	90	127	164	1
LG 56.78 CLP (LIMAGRAIN)	88	127	188	4



CULTIVAR	EMPRESA	Días a floración	Días a madurez	Altura (cm)	Vuelco (%)
ACA 203 CL	ACA	85	122	181	0
ACA 861	ACA	85	122	166	0
Aguara 6	ADVANTA	85	126	174	0
CACIQUE 312 CL	CENCERRO	90	127	161	1
KWS 480 CL	KWS	90	127	164	1
LG 56.78 CLP	LIMAGRAIN	88	127	188	4

# Síntomas comunes de datos desordenados

**Las variables se almacenan tanto en filas como en columnas.**

id	año	mes	elemento	1	2	3	4	5	6	7	8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—

id	fecha	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

# Síntomas comunes de datos desordenados

**Múltiples tipos de unidades de observación se almacenan en la misma tabla.**

ID	PROVIN	CAP_PROV	DEPART	CABECERA	SUP_JURI	AVENA	CEBADA	CENTENO
42063	LA PAMPA	SANTA ROSA	CHICALCO	LA PASTORIL	9117	0	0	0
42091	LA PAMPA	SANTA ROSA	LIMAY MAHUIDA	LIMAY MAHUIDA	9985	0	0	0
42049	LA PAMPA	SANTA ROSA	CHALILEO	SANTA ISABEL	8917	0	0	0
42077	LA PAMPA	SANTA ROSA	HUCAL	BERNASCONI	6047	3363	182	219



**Los datos de provincia y capital de la provincia se repite por cada departamento**

**Tabla Provincias**

**Tabla Departamentos**

**Tabla CultivosXDeptos**



# Síntomas comunes de datos desordenados

**Múltiples tipos de unidades de observación se almacenan en la misma tabla.**

ID	PROVIN	DEPART	CABECERA	SUP_JURI	CULTIVO	SUPERFICIE		CEBADA	CENTENO
42063	LA PAMPA	SA	CHICALCO	LA PASTORIL	9117	AVENA	0	0	0
42091	LA PAMPA	SA	CHICALCO	LA PASTORIL	9117	CEBADA	0	0	0
42049	LA PAMPA	SA	CHICALCO	LA PASTORIL	9117	CENTENO	0	0	0
42077	LA PAMPA	SA	CHICALCO	LA PASTORIL	9117	TRIGO	0	0	0
		SA	LOVENTUE	VICTORICA	9235	GIRASOL	0	182	219
		SA	CHICALCO	LA PASTORIL	9117	MAIZ	0		

**Tabla Provincias**

**Tabla Departamentos**

**Tabla CultivosXDeptos**

**Los datos de provincia y capital de la provincia se repite por cada departamento**



# Síntomas comunes de datos desordenados

Múltiples tipos de unidades de observación se almacenan en la misma tabla.

**Tabla Provincias**

PROVIN	CAP_PROV
LA PAMPA	SANTA ROSA

**Tabla Departamentos**

ID	PROVIN	DEPART	CABECERA	SUP_JURI
42063	LA PAMPA	CHICALCO	LA PASTORIL	9117
42091	LA PAMPA	LIMAY MAHUIDA	LIMAY MAHUIDA	9985
42049	LA PAMPA	CHALILEO	SANTA ISABEL	8917
42077	LA PAMPA	HUCAL	BERNASCONI	6047

**Tabla CultivosXDeptos**

DEPART	CULTIVO	SUPERFICIE
42077	AVENA	3363
42077	CEBADA	182
42077	CENTENO	219



# Tidy Data

- Cuando se recolectan datos por primera vez, siempre es mejor pensar una estructura ordenada desde el inicio
- Cuando nos envían datos ya registrados, debemos analizar su estructura y generar una que sea ordenada
- La estructura ordenada hará la tarea de manejo de datos mucho más sencilla.





# Ordenemos datos juntos

- ¿Esta tabla está ordenada (Tidy)?

ID del envío	3651	3655	3662	3663
Título	Telemetría L	Evaluación d	Desarrollo d	Caminos Rur
Resumen	Organizacio	La evapotran	Existe una br	En un país tar
Primer nombre (Autor 1)	Pablo	Mónica	Santiago	diego
Segundo Nombre (Autor 1)	Guillermo			gabriel
Apellidos (Autor 1)	Di Nanno	Bocco	Lombardo	giordano
País (Autor 1)	AR	AR	UY	AR
Filiación (Autor 1)	INTA - Instit	Facultad de C	Instituto Plai	26884654
Correo electrónico (Autor 1)	pablo.dinani	mbocco@gm	slombardo@d	dgiordano@t
URL (Autor 1)			http://www.planagropecu	
Resumen biográfico (Autor 1)	Investigador en tecnologí	Agrónomo	Director de C	
Primer nombre (Autor 2)		Miguel	Federico	Maria
Segundo Nombre (Autor 2)				Beatriz
Apellidos (Autor 2)		Nolasco	Arias	Rodolfo
País (Autor 2)		AR	UY	AR
Filiación (Autor 2)		Facultad de C	Instituto Plan Agropecuari	
Correo electrónico (Autor 2)		mnolasconq	farias@plan	miriambrodu
URL (Autor 2)				
Resumen biográfico (Autor 2)			Desarrollado	Directora de
Primer nombre (Autor 3)		Silvina		Griselda
Segundo Nombre (Autor 3)				
Apellidos (Autor 3)		Sayago		Galeano



Foto: gentiliza Mauro Lepore



# Mate break



# Seguimos ordenando: Proyectos

La vida de muchos proyectos comienza como notas aleatorias, algún código, luego un manuscrito, y eventualmente **todo está mezclado**.



# Carpeta: Proyecto Analisis De Datos

Informe1.docx

Informe2.pdf

Grafico1.jpg

Nuevografico1.jpg

Informeconcorrecciones.docx

Informe\_final2-docx

Informemasfinal.docx

Tabla1.xlsx

Resultados.xlsx

Informe\_ultimodeverdad.docx

Informe\_listoparaentregar.docx

ProcesoData.R

Resultados\_final.xlsx

Seminario.pptx

PresentacionDirector.pptx

Para\_leer\_urgente

Datosnuevos.xlsx



# Hay muchas razones de porqué debemos **siempre** evitar esto

- Es realmente difícil saber cuál versión de tus datos es la original y cuál es la modificada;
- Es muy complicado porque se mezclan archivos con varias extensiones juntas;
- Probablemente te lleve mucho tiempo encontrar lo que necesitas, y relacionar las figuras correctas con el código exacto que ha sido utilizado para generarlas.
- Imaginate 3, 6 o 12 meses después de haberlo hecho.



# Un buen diseño del proyecto **hará tu vida más fácil**

- Ayudará a garantizar la integridad de tus datos;
- Hace que sea más simple compartir tu código con alguien más (un compañero de trabajo, colaborador o supervisor);
- Permite relacionar fácilmente tu código con las partes y versiones de tu informe;
- Hace que sea más fácil retomar un proyecto después de un descanso.





A ver...vamor a ordenarnos un  
poquitito



Entremos a RStudio

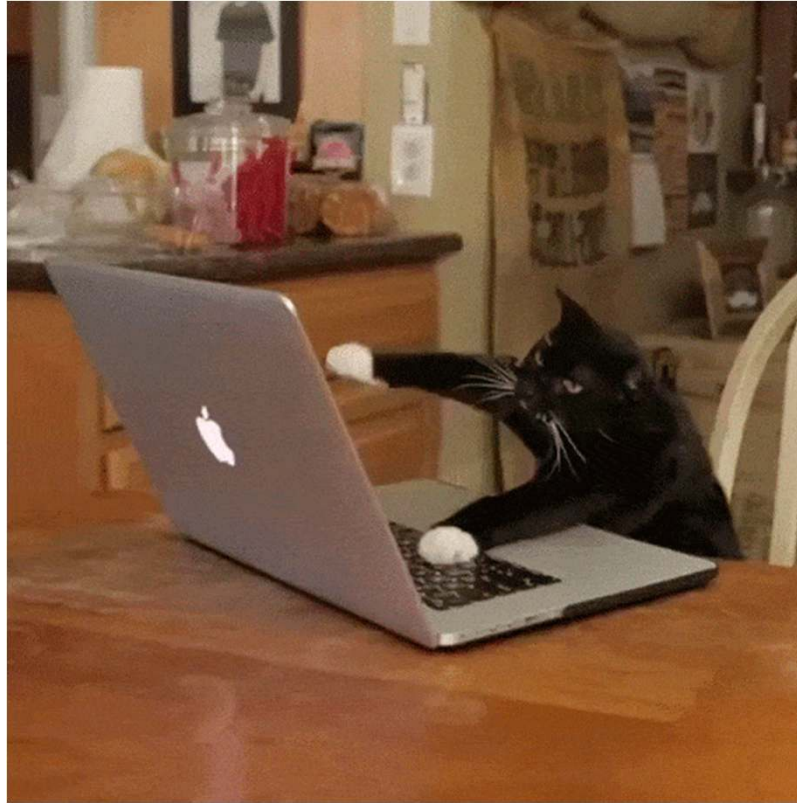
# Recomendaciones para la organización de proyectos

- Coloque cada proyecto en su propio directorio, el cual lleva el nombre del proyecto.
- Coloque documentos de texto asociados con proyecto en el directorio **doc**.
- Coloque los datos sin procesar y los metadatos en el directorio **data**, y archivos generados durante la limpieza y análisis en el directorio **resultados**.
- Coloque los scripts fuente del proyecto y los programas en el directorio **codigo**, y programas traídos de otra parte o compilados localmente en el directorio **bin**.
- Nombre todos archivos de tal manera que reflejen su contenido o función.





# ¿Preguntas, comentarios?



## Eso es todo por hoy

# Fuentes/Bibliografía de esta clase:

- <https://swcarpentry.github.io/r-novice-gapminder-es/>
- <https://moderndive.com/index.html>
- <https://flor14.github.io/Fundamentos de R/>
- <https://vita.had.co.nz/papers/tidy-data.pdf>