Hello

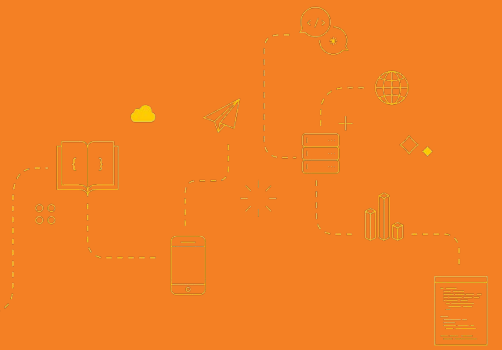
Stack Overflow <3 Data Science

Introducción a R para desarrolladores

@g3rv4

Agenda

- Qué es data science
- Por qué R?
- Ejemplo: "Qué lenguajes se usan de noche?"
- Funciones útiles
- Demo! "Qué tags dan más reputación en Stack Overflow en español?"
- Arrancando con data science



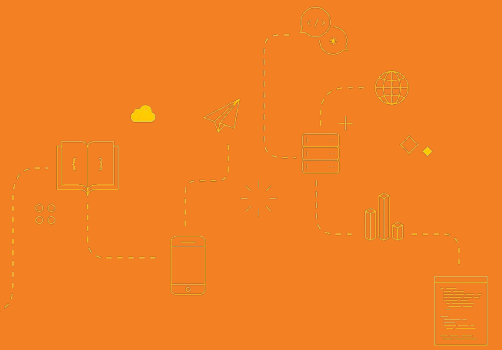
Qué es data science? (o ciencia de datos)

"*The field of data science is emerging at the intersection of the fields of **social science** and **statistics**, **information and computer science**, and **design***"

datascience@**berkeley**

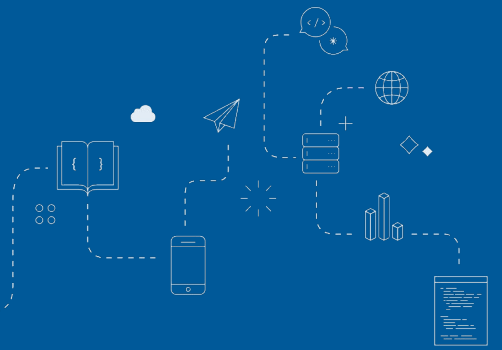
Data science = ciencias sociales + estadística + computación + diseño

Gervasio traduciendo a datascience@**berkeley**



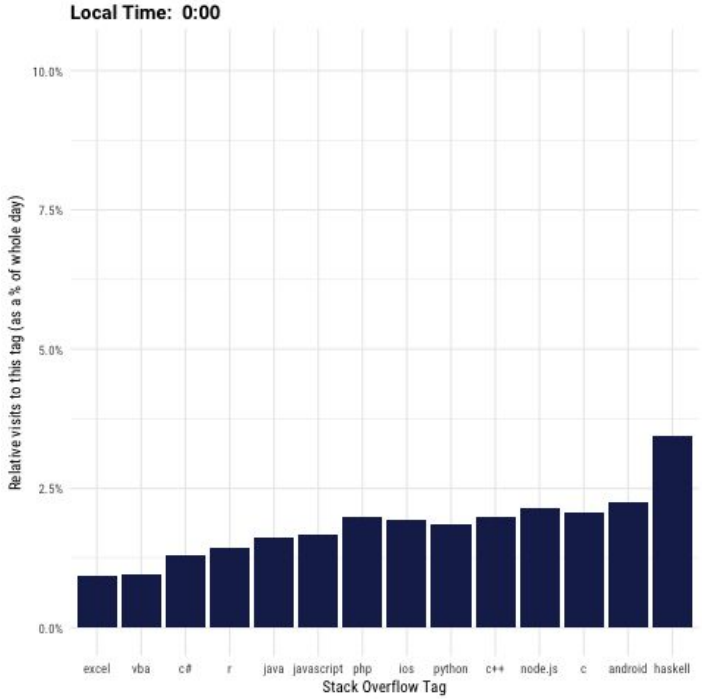
Por qué R?

- Curva de aprendizaje accesible
- Excelente comunidad
- Es la herramienta que usan los estadísticos
  - Nos permite jugar a todos al mismo juego
- Está en la vanguardia del conocimiento
- Hace cosas difíciles de forma fácil
  - Entrenar modelos de ML es muy sencillo… hasta yo lo puedo hacer



Ejemplo: Qué lenguajes se usan de noche?

Un análisis de @drob



gmc.uy/lenguajes-en-la-noche

gmc.uy/tabs-espacios

gmc.uy/ds-en-produccion

stackoverflow

Funciones útiles

Las vamos a usar en la demo

Las básicas

- Las "tablas" con datos se llaman dataframes
- Asignamos con <-
  - `a <- 1` asigna 1 a la variable a
- Usamos pipes! (pero… el pipe es…`%>%`)
- `x %>% f()` es lo mismo que `f(x)`
- `x %>% f(1)` es lo mismo que `f(x, 1)`
- Permite encadenar operaciones
- Usamos tidyverse para filtrar, agrupar, modificar dataframes de forma sencilla y rápida

filter

Nombre	Edad	Pais
Luciano	29	PY
Romina	15	UY
Martín	35	AR
Daniela	22	UY

%>% filter(Edad > 17)

Nombre	Edad	Pais
Luciano	29	PY
Martín	35	AR
Daniela	22	UY

mutate

Nombre	Edad	Pais
Luciano	29	PY
Romina	15	UY
Martín	35	AR
Daniela	22	UY

%>% mutate(EsMayor = Edad > 17)

Nombre	Edad	Pais	EsMayor
Luciano	29	PY	TRUE
Romina	15	UY	FALSE
Martín	35	AR	TRUE
Daniela	22	UY	TRUE

group_by y summarize

Nombre	Edad	Pais	EsMayor
Luciano	29	PY	TRUE
Romina	15	UY	FALSE
Martín	35	AR	TRUE
Daniela	22	UY	TRUE

```
%>%
  group_by(Pais) %>%
  summarize(EdadPromedio = mean(Edad),
                Mayores = sum(EsMayor))
```

Pais	EdadPromedio	Mayores
AR	35.0	1
PY	29.0	1
UY	18.5	1

Inner_join (también tiene left_join y anti_join)

usuarios <-

Nombre	Edad	Pais	EsMayor
Luciano	29	PY	TRUE
Romina	15	UY	FALSE
Martín	35	AR	TRUE
Daniela	22	UY	TRUE

paises <-

Codigo	Nombre
PY	Paraguay
UY	Uruguay
AR	Argentina

```
usuarios %>%
  inner_join(paises, by = c("Pais"="Codigo"),
             suffix = c(".usuario", ".pais"))
```

Nombre.usuario	Edad	Pais	EsMayor	Nombre.pais
Luciano	29	PY	TRUE	Paraguay
Romina	15	UY	FALSE	Uruguay
Martín	35	AR	TRUE	Argentina
Daniela	22	UY	TRUE	Uruguay

tidytext y unnest_tokens

Nombre	Edad	Pais	Gustos
Luciano	29	PY	Dulce de leche, Chocolate, Vainilla
Romina	15	UY	Frutilla, Limón
Martín	35	AR	Durazno, Chocolate amargo, Dulce de leche
Daniela	22	UY	Menta, Chocolate, Limón

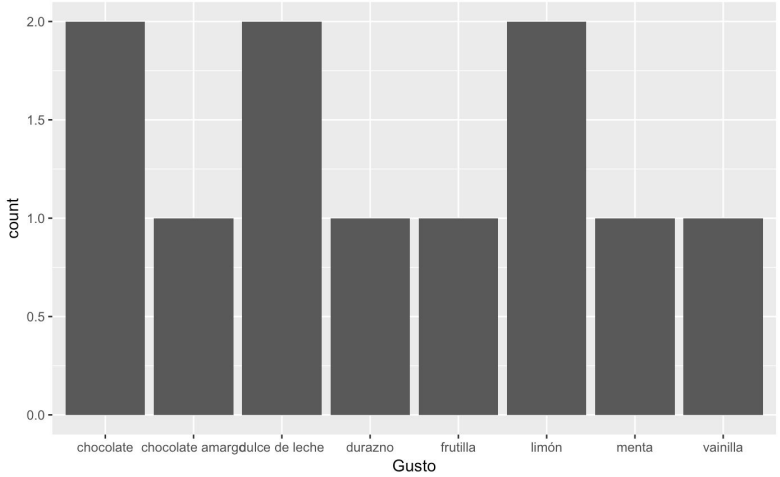
```
%>%
  unnest_tokens(Gusto, Gustos, token='regex', pattern=', ')
```

Nombre	Edad	Pais	Gusto
Daniela	22	UY	menta
Daniela	22	UY	chocolate
Daniela	22	UY	limón
Luciano	29	PY	dulce de leche
Luciano	29	PY	chocolate
Luciano	29	PY	vainilla
Martín	35	AR	durazno
Martín	35	AR	chocolate amargo
Martín	35	AR	dulce de leche
Romina	15	UY	frutilla
Romina	15	UY	limón

ggplot2

Nombre	Edad	Pais	Gusto
Daniela	22	UY	menta
Daniela	22	UY	chocolate
Daniela	22	UY	limón
Luciano	29	PY	dulce de leche
Luciano	29	PY	chocolate
Luciano	29	PY	vainilla
Martín	35	AR	durazno
Martín	35	AR	chocolate amargo
Martín	35	AR	dulce de leche
Romina	15	UY	frutilla
Romina	15	UY	limón

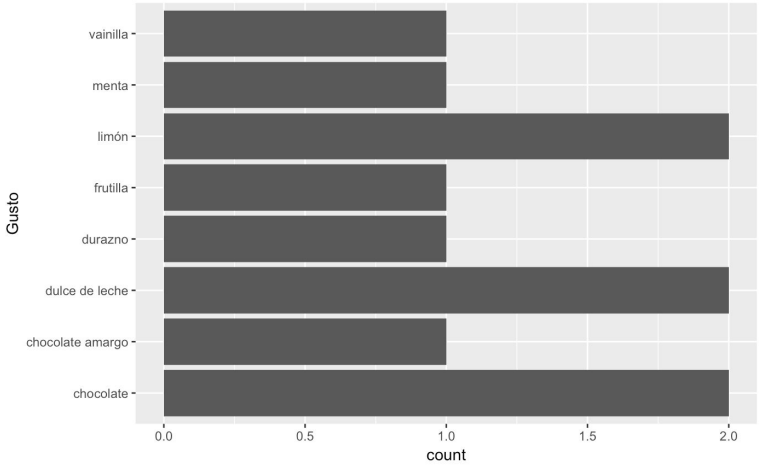
```
%>%
  ggplot(aes(Gusto)) +
  geom_histogram(stat="count")
```



ggplot2

Nombre	Edad	Pais	Gusto
Daniela	22	UY	menta
Daniela	22	UY	chocolate
Daniela	22	UY	limón
Luciano	29	PY	dulce de leche
Luciano	29	PY	chocolate
Luciano	29	PY	vainilla
Martín	35	AR	durazno
Martín	35	AR	chocolate amargo
Martín	35	AR	dulce de leche
Romina	15	UY	frutilla
Romina	15	UY	limón

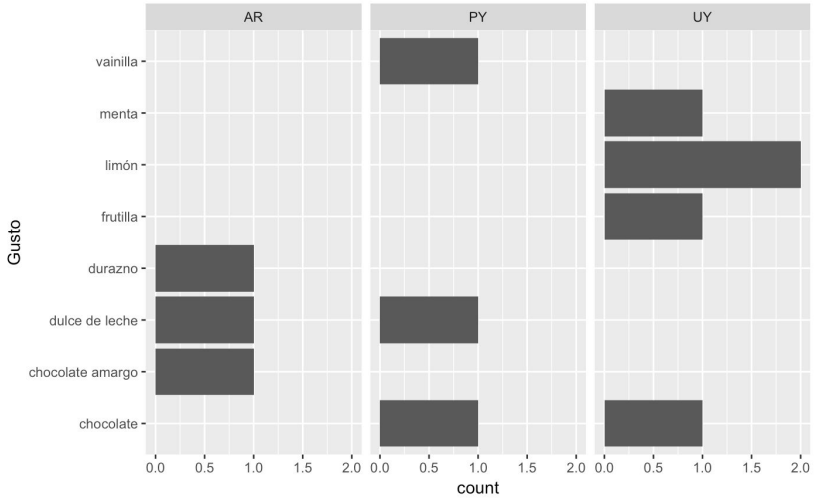
```
%>%
  ggplot(aes(Gusto)) +
  geom_histogram(stat="count") +
  coord_flip()
```



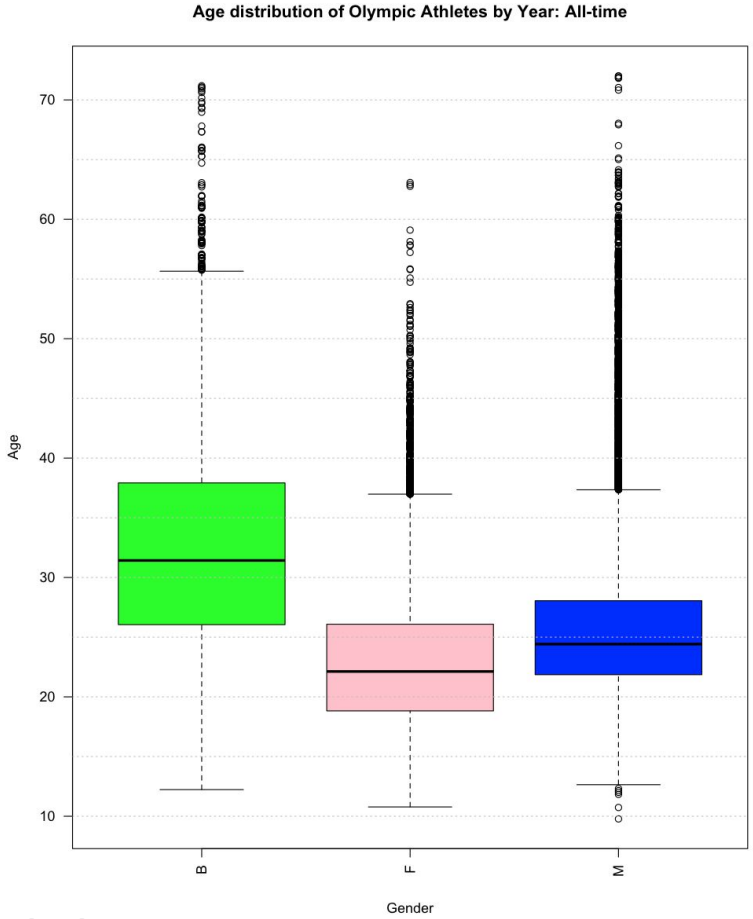
ggplot2

Nombre	Edad	Pais	Gusto
Daniela	22	UY	menta
Daniela	22	UY	chocolate
Daniela	22	UY	limón
Luciano	29	PY	dulce de leche
Luciano	29	PY	chocolate
Luciano	29	PY	vainilla
Martín	35	AR	durazno
Martín	35	AR	chocolate amargo
Martín	35	AR	dulce de leche
Romina	15	UY	frutilla
Romina	15	UY	limón

```
%>%
  ggplot(aes(Gusto)) +
  geom_histogram(stat="count") +
  coord_flip() +
  facet_wrap(~Pais)
```



Boxplots



Crédito: statsinthewild.com

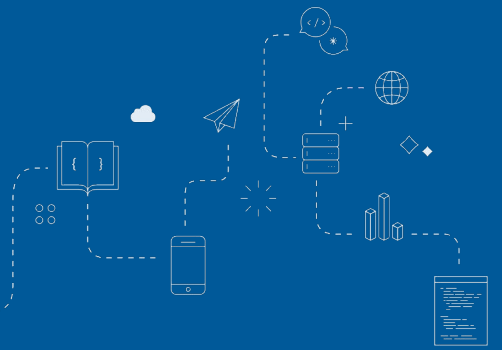
En la conferencia expliqué cómo interpretar un boxplot, pero tenía animaciones…

Así que te recomiendo que googlees cómo entender un boxplot si te quedan dudas… porque por más de que intenté, el PDF no es amigo de las animaciones.

Demo

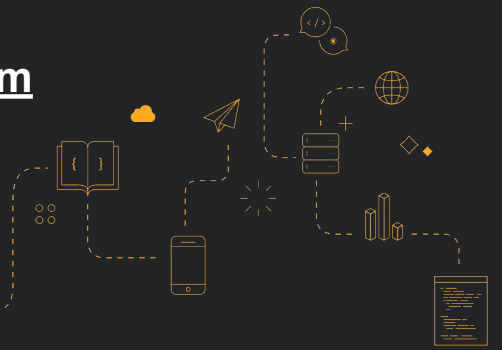
Qué tags de Stack Overflow en español dan más reputación al responder?

- Bajé el dump de Stack Overflow en español de archive.org
  - **gmc.uy/data-dump**
- Los convertí a CSV
- Vamos a usar sólo Posts y Votes
- Buena suerte, futuro yo!



Arrancando con data science

- Bajá RStudio (es open source)
- Podés bajar el código de la demo en **gmc.uy/nerdearla**
  - Ahí, además de la demo, hago un modelo de ML para predecir el puntaje de una respuesta
  - Preguntá :) en Stack Overflow, en otras comunidades o a mí :)
- Está disponible la encuesta a desarrolladores de este año (con más de 64.000 encuestad@s) para analizar en **gmc.uy/encuesta**
  - Incluye datos de salarios!
- Hay muchos datasets abiertos interesantes, mirá **kaggle.com**
- JUGÁ! Y si en tu trabajo no se puede… {{ chivo }}
  - Visitá stackoverflow.com/jobs !!!



Thanks!

@g3rv4