

DATA SCIENCE, R Y GITHUB

Taller Manos a la Data

Arturo Chian

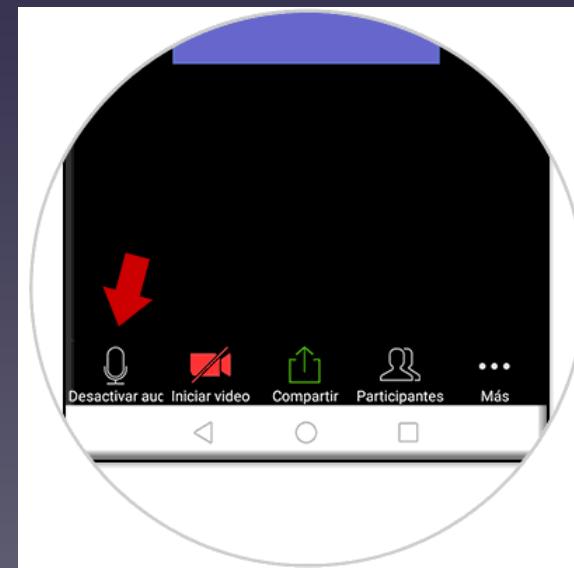


Un presentación BEST <http://besteamperu.org/>

Por favor silenciar los micrófonos



¿Cómo hacerlo?



Agenda del día de hoy (0-10)

- ➡ Motivarnos: Sin motivación, no aprenderemos! + web
- ➡ Qué es Manos a la data
- ➡ Qué es Data Science.
- ➡ Qué es R y RStudio.
- ➡ Qué es Markdown y Rmarkdown.
- ➡ Qué es Git y Github.
- ➡ Trabajar con nuestra primera base de datos de Manos a la Data!

¡Bienvenidos a Manos a la data!



¿Qué es manos a la data?

- ➡ Un proyecto opensource liderado por BEST!
- ➡ Se busca impulsar la Ciencia de Datos en Perú para todas las carreras!
- ➡ Se busca ayudar en la enseñanza de DS con ejemplos concretos por semana de nuestra realidad con diversas temáticas y tipos de bases de datos (PDF, excels, txt, csv, mapas, etc).
- ➡ Se busca crear un ecosistema opensource de proyectos de Data Science en Perú.
- ➡ Siéntase libres de compartir las Bases de Datos semanales para practicar o enseñar.
- ➡ ¡Recuerden que también es importante ofrecer créditos!

¿Qué queremos de ti?

- ➡ Difundan la ciencia de datos en Perú, no existen barreras!
- ➡ Queremos que crezcan!
- ➡ Aprender + aportar = crecer
- ➡ Compartan 1 base de datos pública de tu interés en nuestro Github!

Motivación



¿Por qué R? ¿Qué puede hacer R? (10-20)

- ☒ Es la lingua franca de estadística.
- ☒ Una comunidad científica muy activa.
- ☒ Esta diapositiva está en R.
- ☒ Varios libros están escritos con R.
- ☒ Este curso y su web está en R.
- ☒ Se pueden crear blogs en R.
- ☒ Se pueden crear productos en R. (Se explicará más adelante!)

Creación de libros abiertos

The screenshot shows a digital textbook interface for 'Forecasting: Principles and Practice' by Rob J Hyndman and George Athanasopoulos. The left sidebar contains a table of contents with sections like Preface, 1 Getting started, 2 Time series graphics, etc. The main content area displays the book's title, authors, and publisher information. Below the title, the 'Preface' section is visible, followed by a welcome message and a description of the textbook's purpose.

Forecasting: Principles and Practice

Rob J Hyndman and George Athanasopoulos
Monash University, Australia

Preface

Welcome to our online textbook on forecasting.

This textbook is intended to provide a comprehensive introduction to forecasting

Rob J Hyndman
George Athanasopoulos

FORECASTING

Fuente: *Forecasting: Principles and Practice*

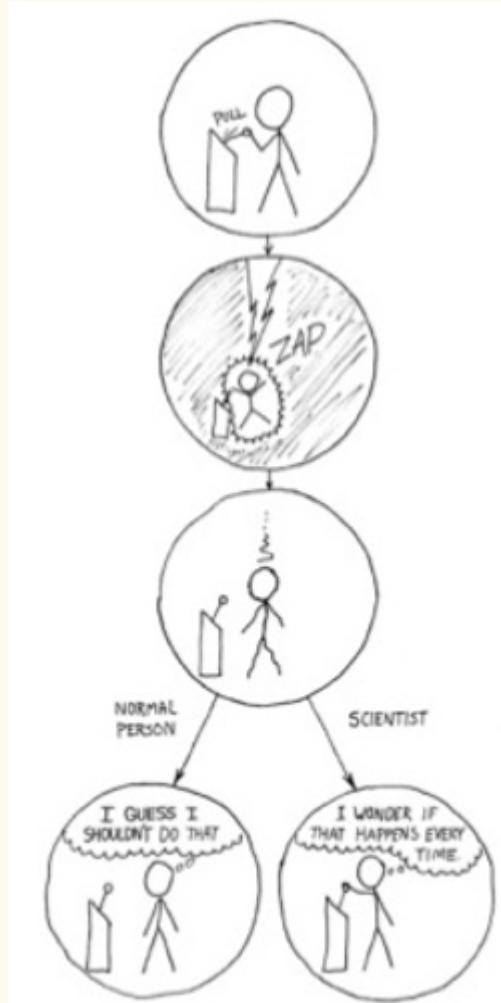
10 / 53

Nuevas formas de hacer tesis

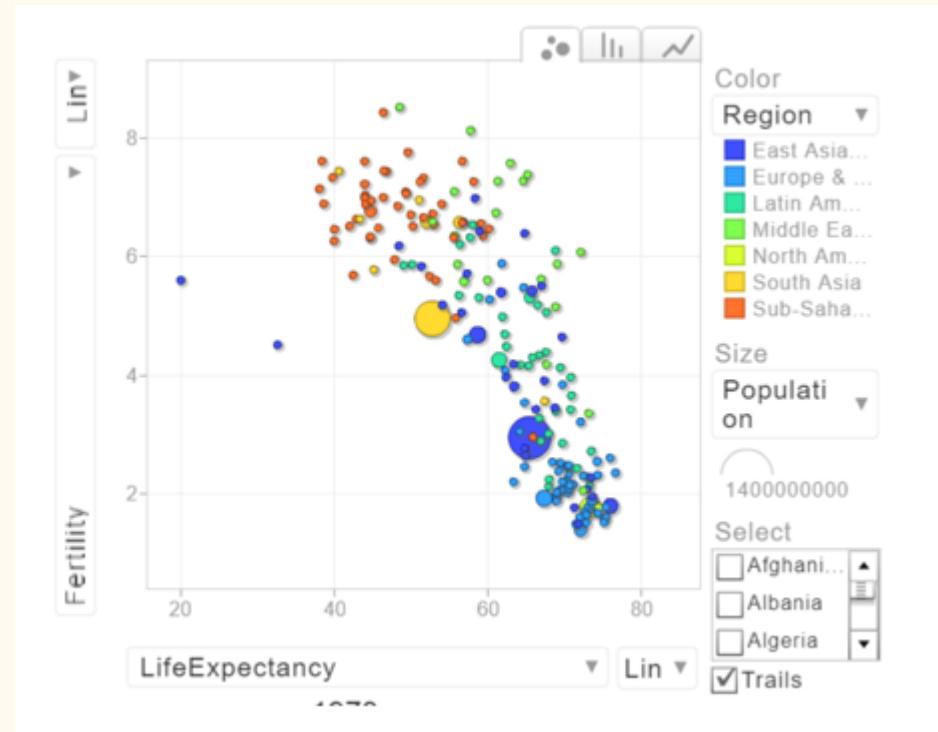
The screenshot shows a portion of a README.md file. At the top, there's a header with a file icon and the text "README.md". Below this, the title "thesisdown" is displayed in a large, bold, black font. A horizontal line follows the title. The text "This project was inspired by the [bookdown](#) package and is an updated version of my Senior Thesis template in the [reedtemplates](#) package [here](#)." is present. Another horizontal line follows. The text "Currently, the PDF and gitbook versions are fully-functional. The word and epub versions are developmental, have no templates behind them, and are essentially calls to the appropriate functions in bookdown." is displayed. A third horizontal line follows. The text "If you are new to working with `bookdown / rmarkdown`, please read over the documentation available in the `gitbook` template at <https://thesisdown.netlify.com/>. This is also available below at http://ismayc.github.io/thesisdown_book." is shown. A fourth horizontal line follows. The text "The current output for the four versions is here:" is followed by a bulleted list:

- [PDF](#) (Generating LaTeX file is available [here](#) with other files at in the `book` directory.)
- [Word](#)
- [ePub](#)
- [gitbook](#)

Promueve la investigación reproducible



Gráficos Dinámicos



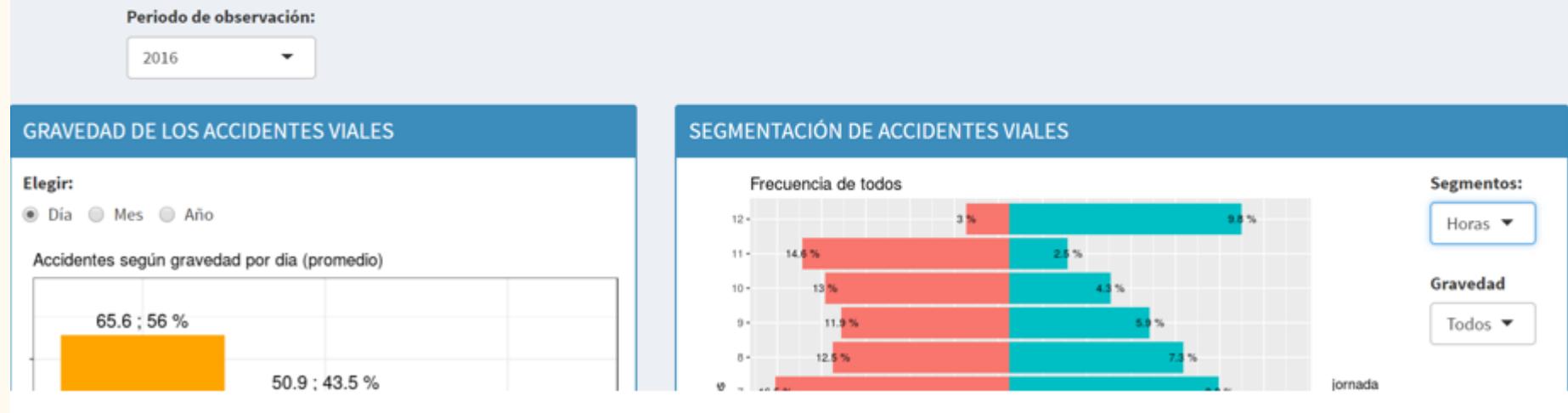
Desarrollo de aplicaciones fácil de realizar

Pasa-Segura-Medellín

Aplicación web sobre accidentalidad vial en Medellín usando Ciencia de Datos

Vers. 3 (2018-1). Años de observación: 2014 - 2017_1 (ene-jul 2017). Grupo de investigación IDINNOV investigacion@idinnov.com

1. DESCRIPCIÓN DE VARIABLES DE ACCIDENTALIDAD PARA PERIODOS 2014 - 2017_1



Fuente: Pasa Segura Medellin

Aplicable a todas las ciencias

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes 

Published: April 3, 2014 • <https://doi.org/10.1371/journal.pcbi.1003531>

Sum of P

Article	Authors	Metrics	Comments	Media Coverage
				

Abstract

Author Summary

Introduction

Methods

Results/Discussion

Supporting Information

Acknowledgments

Fuente: paquete phyloseq Github

References

Abstract

Current practice in the normalization of microbiome count data is inefficient in the statistical sense. For apparently historical reasons, the common approach is either to use simple proportions (which does not address heteroscedasticity) or to use *rarefying* of counts, even though both of these approaches are inappropriate for detection of differentially abundant species. Well-established statistical theory is available that simultaneously accounts for library size differences and biological variability using an appropriate mixture model. Moreover, specific implementations for DNA sequencing read count data (based on a Negative Binomial model for instance) are already available in RNA-Seq focused R packages such as edgeR and DESeq. Here we summarize the supporting statistical theory and use simulations and empirical data to demonstrate substantial improvements provided by a relevant mixture model framework over simple proportions or rarefying. We show how both proportions and rarefied counts result in a high rate of false positives in tests for species that are differentially abundant across

¿Qué es Ciencia de Datos? (20-30)

“

For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. ...All in all I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

Tukey (1962). *The future of Data Analysis, The Annals of Mathematical Statistics*

¿Qué es Ciencia de Datos? (20-30)

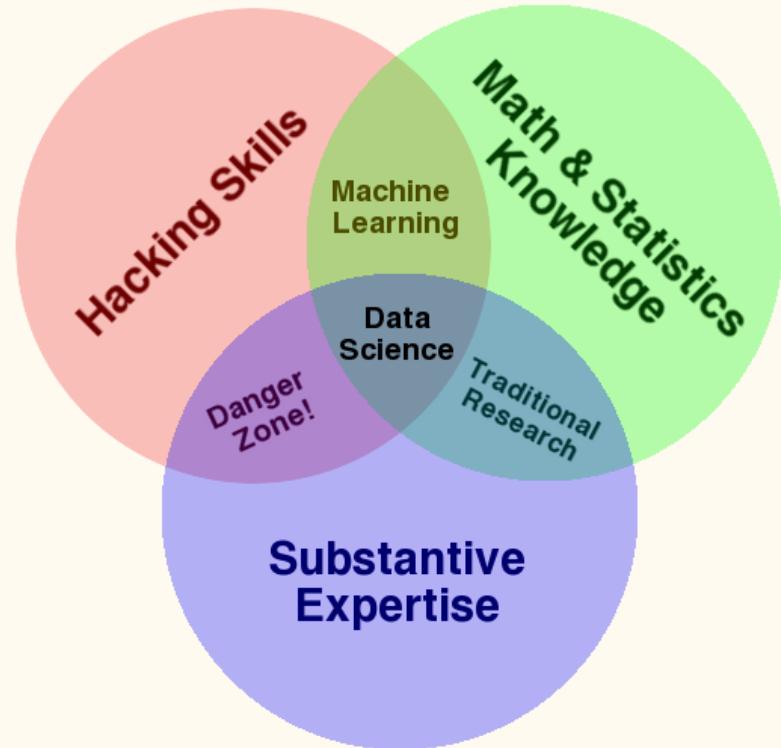
‘

Hace 50 años, John Tukey llamó a una reforma académica en estadística, a través de uno de los más importantes papers de esa época, llamado “The Future of Data Analysis”, donde señalaba la necesidad futura de una ciencia cuyo interés sea aprender de la data o análisis de datos. Hace unos 20 a 10 años, John Chamber, Jeff Wu, Bill Cleveland y Leo Breiman, dieron una serie de argumentos, de forma independiente sobre expandir los límites de la estadística teórica: Chambers enfatizaba la importancia de la preparación de datos, más que el modelaje estadístico; Breiman, prefería enfatizar la predicción antes que la inferencia; y Cleveland y Wu sugerían llamar a este nuevo campo Data Science por su estrecha relación a la data.

Arturo Chian (2018). A propósito de los 25 años de R y 50 años de Data Science (Parte 1), Blog de Behavioral Economics & Data Science Team

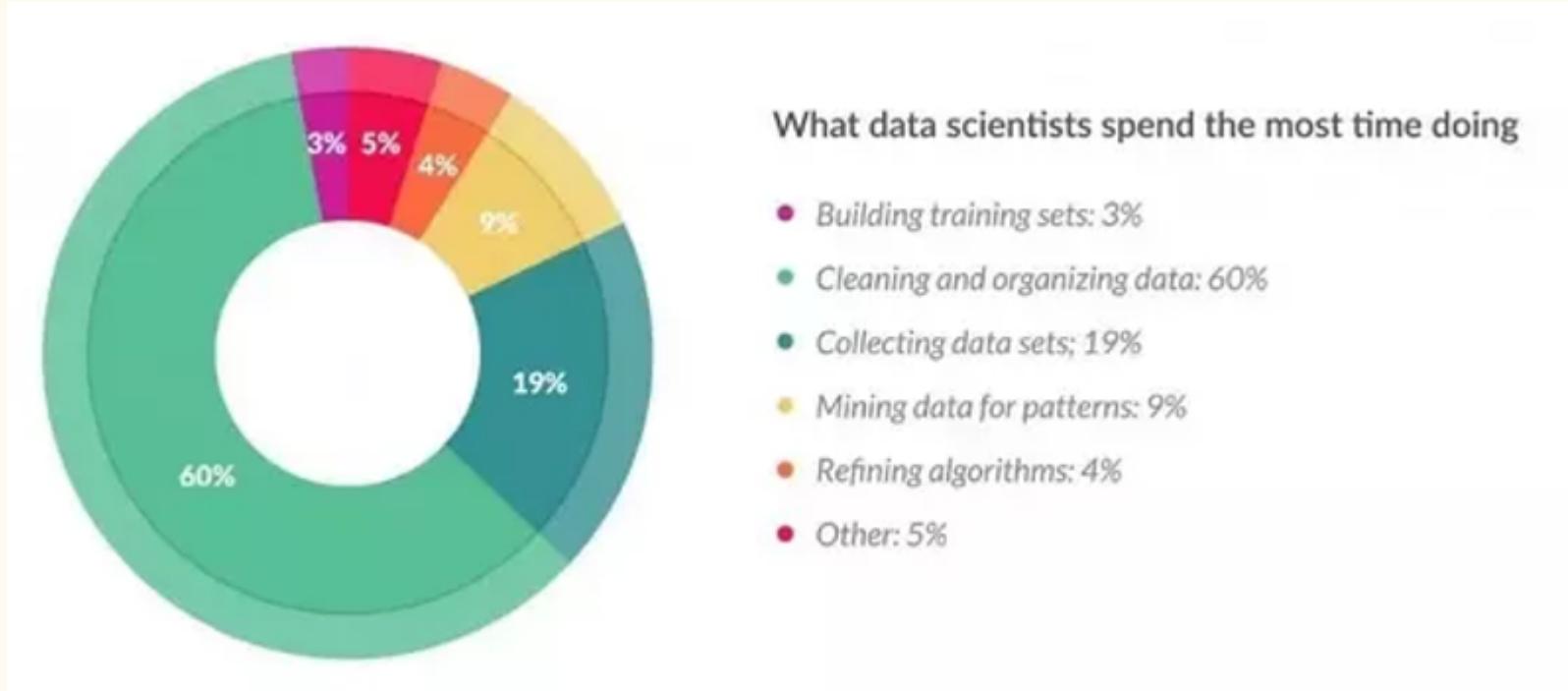
Definiendo Data Science

Diagrama de Venn - Drew Conway



1. **Hacking Skills:** Capacidad de resolver problemas programando.
2. **Math & Statistics knowledge:** Aplicar de forma correcta estadística.
3. **Conocimiento de experto:** Comprender la data en su campo de investigación (economía, biología, psicología, derecho, etc).

¿Qué hace un Data Scientist en el día a día?



¿Cómo muchos hemos aprendido a usar R? (30-45)



¡Hora de aprender!



Diferenciando entre R y RStudio

R Studio

Es uno de los más IDE más usados.
Facilita al usuario para programar R y
otros lenguajes.



R

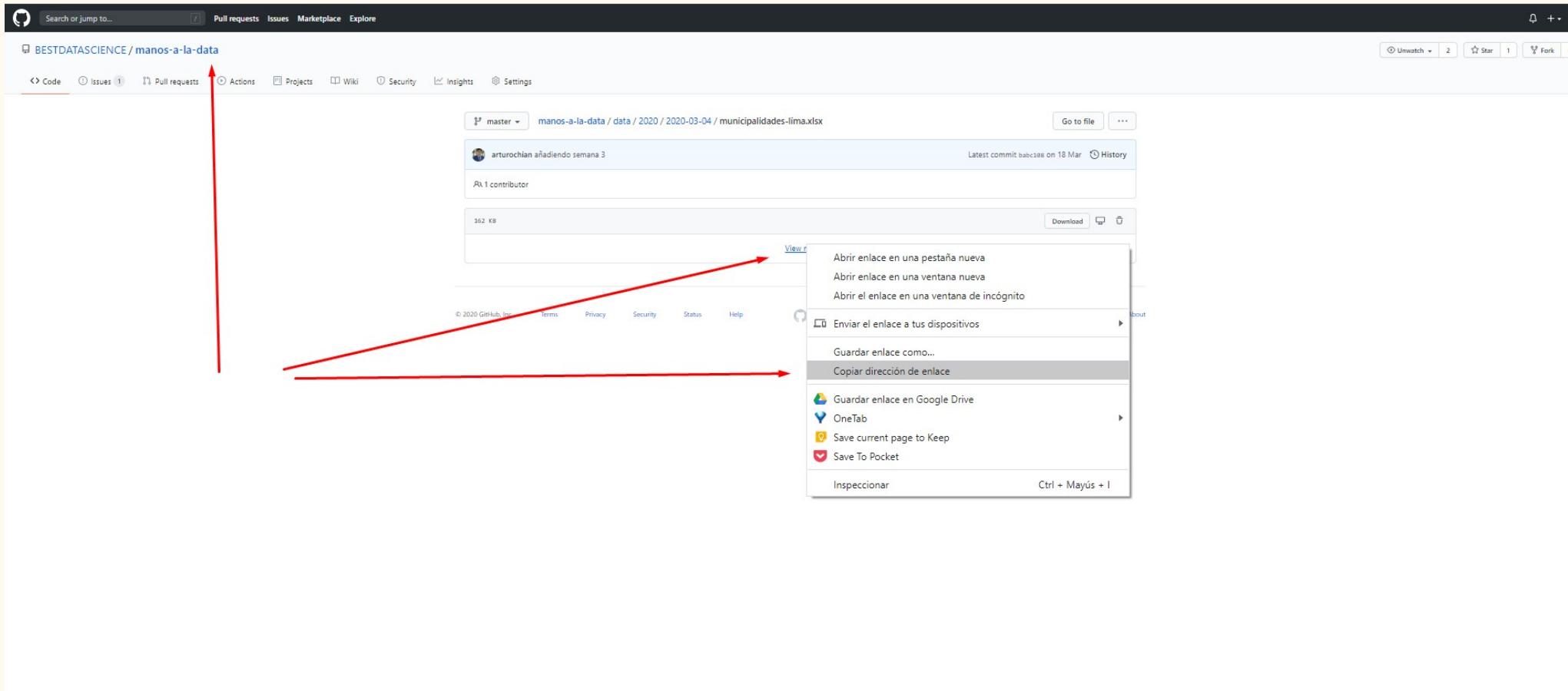
Es el software estadístico por
excelencia. ***La lingua franca*** en
estadística.



Práctica 1: Leer un archivo de Manos a la data

1. Entra a la [semana 1 de Manos a la Data](#)
2. Identificar el archivo csv de la [semana 1](#)
3. Instala el paquete tidyverse. Escribe `install.packages("tidyverse")`. **Importante
respetar las comillas y mayúsculas y minúsculas en general al instalar paquetes.**
4. Crea un R Script usando la función `readr::read_csv()` (paquete readr, función `read_csv`)
5. Utiliza el operador "`<-`" y ponle el nombre `municipalidades`.
6. Utiliza la función `View()` para ver la data.

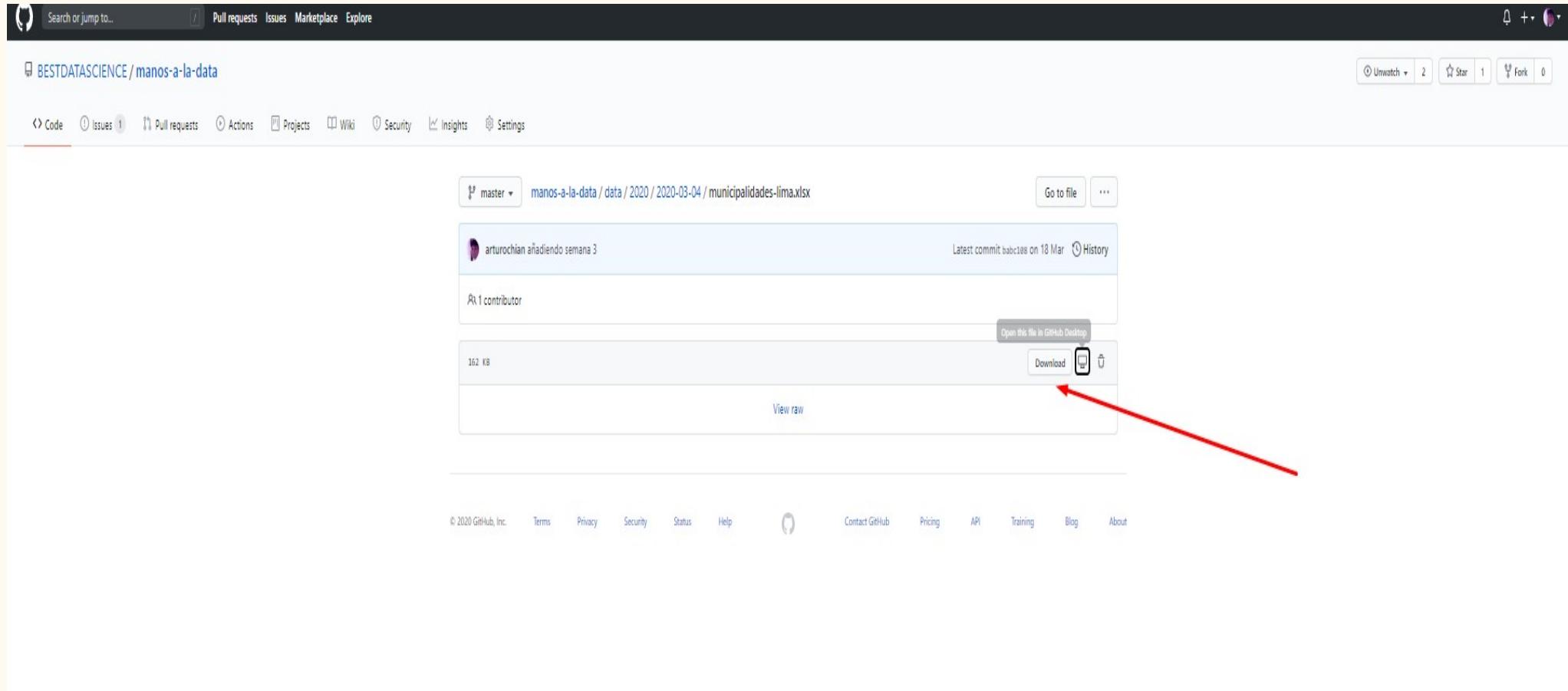
Práctica 1: Leer un archivo de Manos a la data (web)



Práctica 1: Leer un archivo de Manos a la data

```
install.packages("tidyverse") #esto solo se hace 1 vez
library(tidyverse) # carga el paquete
municipalidades <- readr::read_csv('https://raw.githubusercontent.com/BESTDAT/
View(municipalidades) # veamos cómo está cargada la data!
```

Práctica 1: Leer un archivo de Manos a la data (descargado)

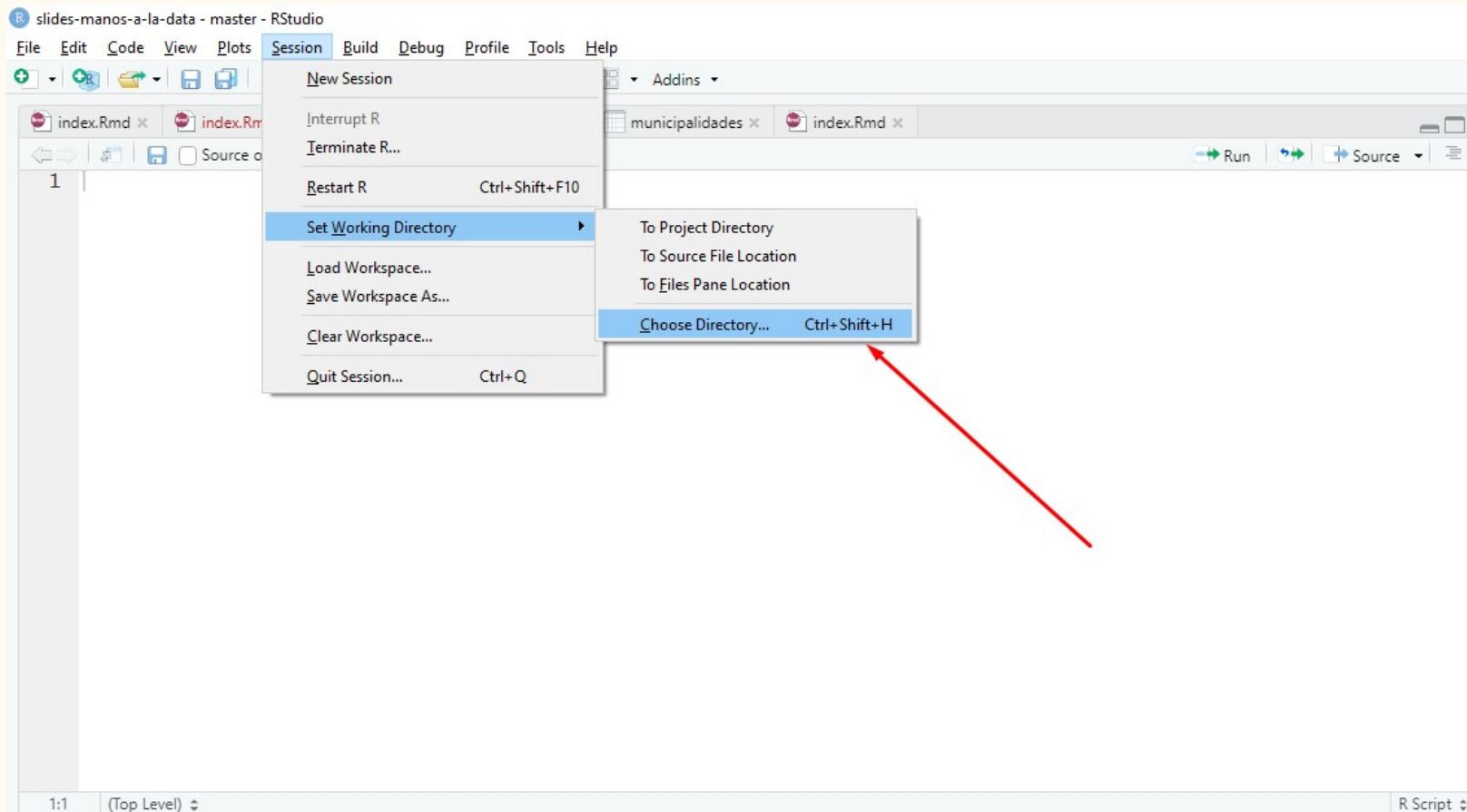


Práctica 1: Leer un archivo de Manos a la data (descargado)

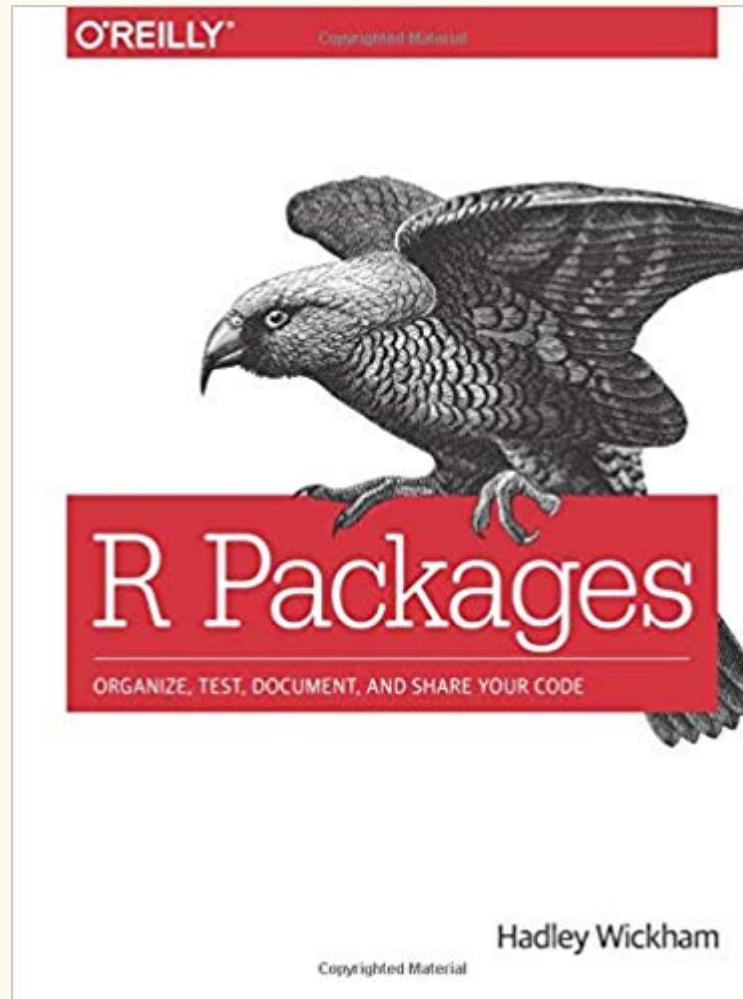
```
library(tidyverse) # carga el paquete  
#  
setwd("~/clases best/semana 1/") # solo si has descargado el archivo  
municipalidades <- readr::read_csv("municipalidades.csv")  
  
View(municipalidades) # veamos cómo está cargada la data!
```

Práctica 1: Leer un archivo de Manos a la data (descargado)

Recuerda que tu directorio de trabajo (working directory), es donde vas a trabajar y guardar tus archivos. Puedes usar esta opción si estás iniciando en R.



Conceptos: Paquetes



Conceptos: Paquetes

‘

Packages are the fundamental units of reproducible R code. They include reusable R functions, the documentation that describes how to use them, and sample data.

Hadley (2015). R packages

Conceptos: Paquetes

‘

In R, the fundamental unit of shareable code is the package. A package bundles together code, data, documentation, and tests, and is easy to share with others. As of January 2015, there were over 6,000 packages available on the Comprehensive R Archive Network, or CRAN, the public clearing house for R packages. This huge variety of packages is one of the reasons that R is so successful: the chances are that someone has already solved a problem that you’re working on, and you can benefit from their work by downloading their package.

Hadley (2015). R packages

Y otras más potentes... El límite es tu imaginación

Analytic Health Demo Sep 18



Fuente: Analytic Health

Conceptos: Paquetes (35-40)

Hay paquetes para todo!

- ➡ **Temas académicos:** tesis, investigación reproducible, libros, papers, manejar bibliografías, etc.
- ➡ **Producción:** Desarrollo de aplicaciones, servidores, etc.
- ➡ **Ciencias:** Economía, Psicología, Biología, Medicina, Finanzas, etc.
- ➡ **Acceso a APIs:** Banco Mundial, FMI, Bloomberg, etc.

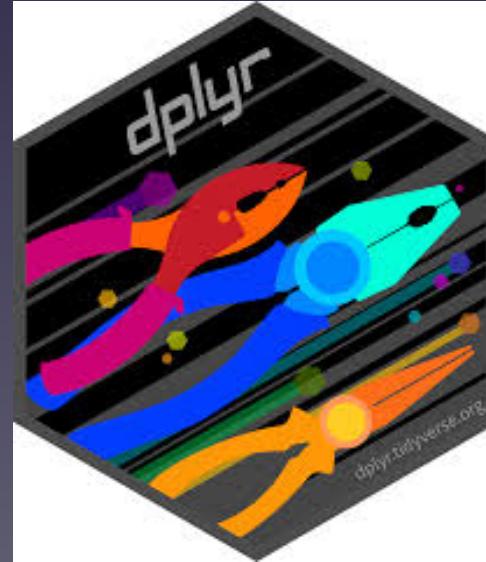
Conceptos: Paquetes

‘

But packages are useful even if you never share your code. As Hilary Parker says in her introduction to packages: “Seriously, it doesn’t have to be about sharing your code (although that is an added benefit!). It is about saving yourself time.” Organising code in a package makes your life easier because packages come with conventions.

Hadley (2015). R packages

dplyr



Programación funcional (40-55)

```
arrange( #ORDENA
  summarize(#CREA UNA VARIABLE
    group_by( # AGRUPA VARIABLES
      filter(mtcars, carb > 1), #FILTRA
      cyl
    ),
    Avg_mpg = mean(mpg)
  ),
  desc(Avg_mpg)
)
```

Programación basada en objetos

```
a <- filter(mtcars, carb > 1)#FILTRA
b <- group_by(a, cyl)# AGRUPA VARIABLES
c <- summarise(b, Avg_mpg = mean(mpg))#CREA UNA VARIABLE
d <- arrange(c, desc(Avg_mpg))#ORDENA
print(d)
```

Programación con pipas

```
library(magrittr)
library(dplyr)

mtcars %>%
  filter(carb > 1) %>%
  group_by(cyl) %>%
  summarise(Avg_mpg = mean(mpg)) %>%
  arrange(desc(Avg_mpg))
```

Una introducción a Dplyr

- ➡ **filtrar:** `filter()` Utiliza == , <, >, >=, <=, %in% c("elemento1","elemento2").
- ➡ **agrupar por:** `group_by()` Crea grupo para tener resúmenes de datos luego con `summarise`.
- ➡ **crea nuevas variables resumidas:** `summarise()` Crea resúmenes de un grupo. Ejm promedio de notas por salón de clase.
- ➡ **creación de variables sin resumir:** `mutate()` Crea una variable de cualquier tipo.
- ➡ **Ordenar datos:** `arrange()` y `arrange(desc())` Ordena una base de datos de forma ascendente o descendente.
- ➡ Hay muchas más funciones usa `?dplyr`

Práctica 2: Trabajar por primera vez en R (55-70)

```
library(tidyverse) # carga el paquete
municipalidades <- readr::read_csv('https://raw.githubusercontent.com/BESTDAT/')

municipalidades %>%
  group_by(Municipalidad) %>%
  summarise(avance=median(`Avance %`)) %>% # cuando una variable tiene nombre
  arrange(avance) %>% # ordenar por avance por default de menor a mayor
  View()

municipalidades %>%
  group_by(Municipalidad) %>%
  summarise(avance=median(`Avance %`)) %>% # cuando una variable tiene nombre
  arrange(desc(avance)) %>% # ordenar por avance por default de mayor a menor
  View()
```

Mardown vs Rmarkdown (70-75)

- ➡ Pueden encontrar muchos detalles en [la sección de referencias de la web del curso](#) por lo que se les sugiere leerlos
- ➡ Markdown es un tipo especial de lenguaje de markup que permite formatear de forma sencilla el texto. Uno puede convertir el markdown usando por ejemplo un programa convertidor como pandoc en cualquier formato que uno quiere: HTML, PDF, Word, PowerPoint, etc.
- ➡ R Markdown es regular Markdown con código de R y el resultado puede ser HTML, Word, PPT, etc. Tú puedes hacer todo lo que sabes de regular Markdown, así como añadir gráficos, tablas y otros resultados de R directamente en tu documento.

Práctica 3: Trabajar por primera vez en Rmarkdown (Parte I) (75-90)

☒ Primero hay que crear un RMD.

☒ modifiquemos el texto!

Práctica 3: Trabajar por primera vez en Rmarkdown (Parte II) (75-90)

➡ Pongamos el siguiente script!

```
municipalidades %>%
  group_by(Municipalidad) %>%
  summarise(avance=median(`Avance %`)) %>% # cuando una variable tiene nombre
  arrange(avance) %>% # ordenar por avance por default de menor a mayor
  DT::datatable()
```

Práctica 3: Trabajar por primera vez en Rmarkdown (Parte III) (75-90)

➡ Pongamos la siguiente configuración en el chunk!

```
# ``{r, warning=FALSE,message=FALSE,echo=FALSE}
```

Práctica 3: Trabajar por primera vez en Rmarkdown (Parte IV) (75-90)

↳ Crea otro subtítulo usando ## o ### seguido del nombre (Ejm: Gráfico de tendencia).

```
municipalidades<-read.csv("https://raw.githubusercontent.com/BESTDATASCIENCE/1  
municipalidades2 <- municipalidades %>%  
  filter(PROVINCIA=="MUNICIPALIDAD METROPOLITANA DE LIMA")  
  
p5 <- ggplot(municipalidades2, aes(x = periodo, y = avance))  
(p5 <- p5 + geom_line() +  
  facet_wrap(~Municipalidad, ncol = 2)+  
  
  theme( axis.text = element_text( size = 14 ),  
        axis.text.x = element_text( size = 12 ),  
        axis.title = element_text( size = 14, face = "bold" ),  
        legend.position="none",  
        strip.text = element_text(size = 6)) +  
  labs(title = "Avance presupuestal", subtitle = "Histórico 2007-2019", ca|  
    x="Periodo", y="Avance presupuestal (%))+ stat_smooth(method=lm))45 / 53
```

Práctica 3: Trabajar por primera vez en Rmarkdown (Parte IV) (75-90)

```
p5 <- ggplot(municipalidades2, aes(x = periodo, y = avance))  
(p5 <- p5 + geom_line() +  
  facet_wrap(~Municipalidad, ncol = 2)+  
  
  theme( axis.text = element_text( size = 14 ),  
         axis.text.x = element_text( size = 12 ),  
         axis.title = element_text( size = 14, face = "bold" ),  
         legend.position="none",  
         strip.text = element_text(size = 6)) +  
  labs(title = "Avance presupuestal", subtitle = "Histórico 2007-2019", ca|  
    x="Periodo", y="Avance presupuestal (%))+ stat_smooth(method=lm))
```

Git y Github (90-95)

Es crítico para proyectos de mayor envergadura y para aprovechar manos a la data!



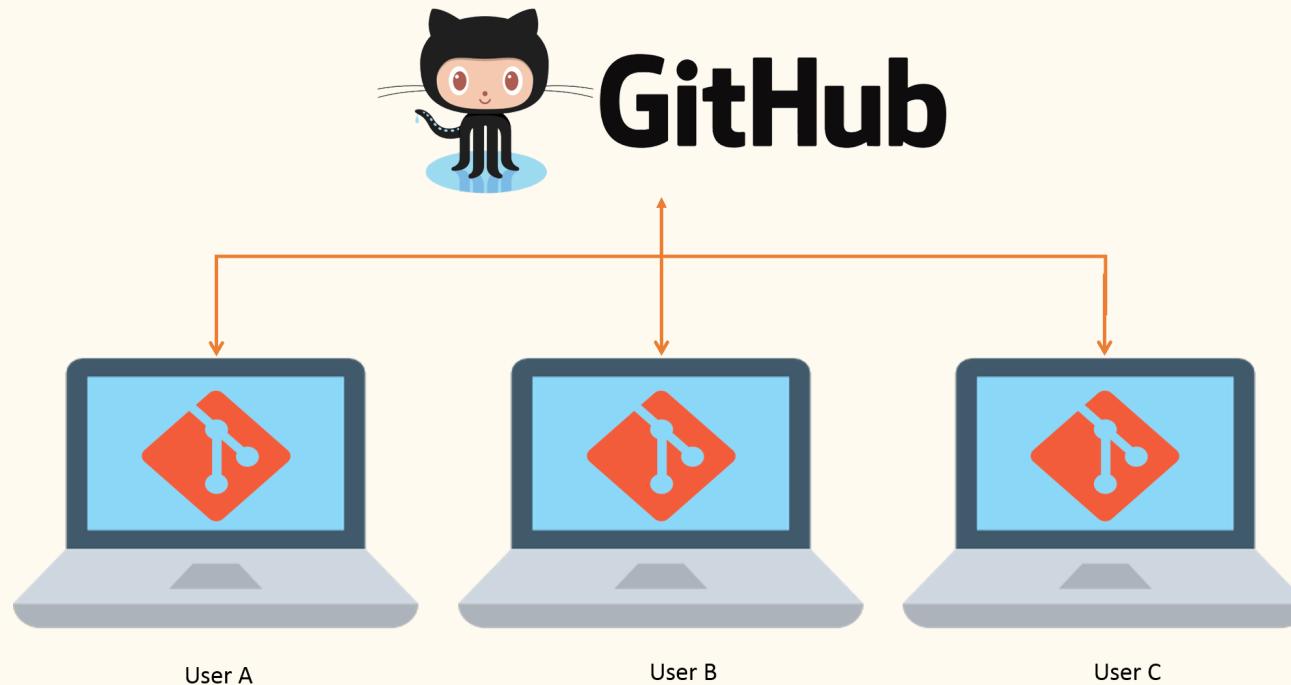
Git y Github (90-95)

Así evitamos los clásicos, final, final final, final final esta es, final final esta es 2, etc.

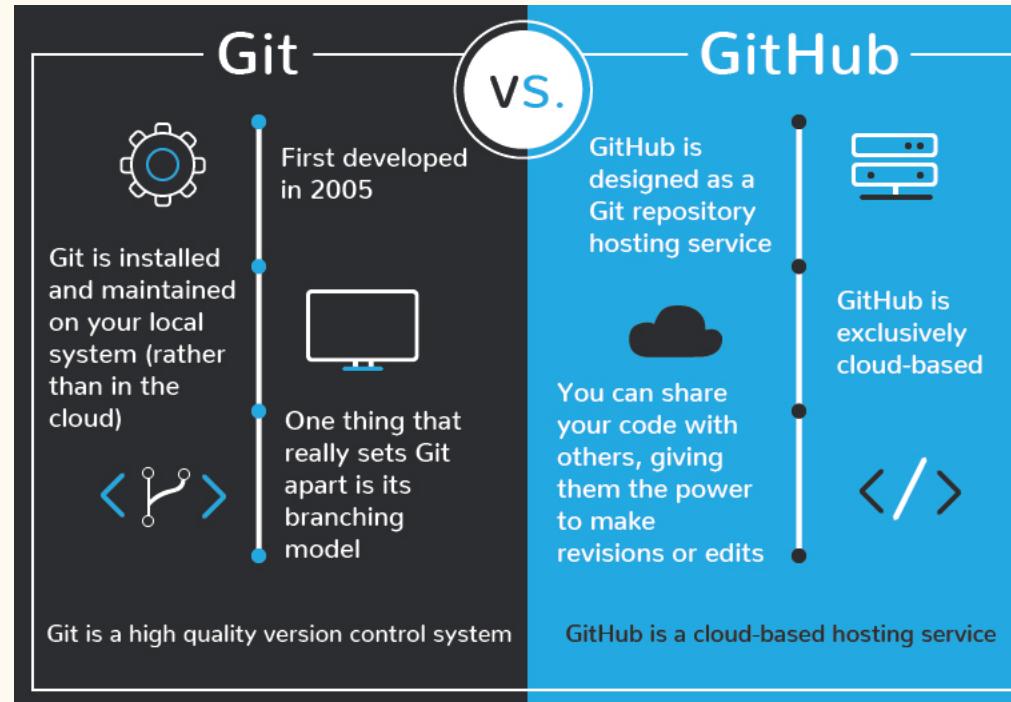


Git y Github (90-95)

- Git es un sistema de control de versiones.
- Github es un servicio que usa Git. Hay otros como Gitlab por ejemplo.



Git y Github (90-95)



Práctica 4: poner gráficos específicos (95-120)

- ➡ Descarga el siguiente [ejemplo 1](#) y córrelo en RStudio!
- ➡ Descarga el siguiente [ejemplo 2](#) y córrelo en RStudio!

Tarea (95-120)

- ➡ Crearse su cuenta de Github (25%).
- ➡ Comentar qué bases de datos les gustaría tener en el futuro en Manos a la Data (25%) [Link](#)
- ➡ ¡Graficar 3 distritos, poner el logo del distrito y calcular la media, mediana, máximo, mínimo, desviación estándar y ponerlo como texto! (50%)

¿Consultas adicionales?

➡ Web del curso

➡ El whatsapp del grupo.

➡ ¿Más consultas? Separa una **cita!**