clase3

guada

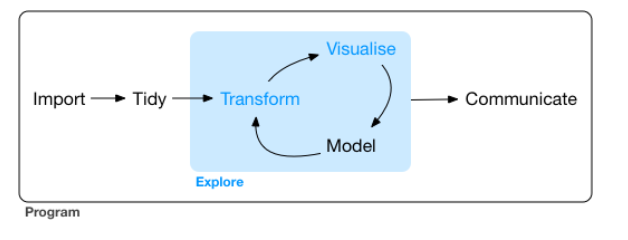
26/2/2020

# CLASE 2: VISUALIZACION DE DATOS EN R

## Introduccion

Recordando un poco de la clase anterior habiamos dicho que todo lo que no se comunica, es laburo que se pierde. Y para esto la visualizacion representa un elemento escencial en el proceso del trabajo con datos.

knitr::include\_graphics('https://github.com/Guadag12/R4RRII/raw/master/Clase%203/images/r4ds\_hadley.png', 1)



Hay muchas recomendaciones a la hora de trabajar pero por sobre todo hay que tener en cuenta que *“Una buena gramática nos permitirá conocer la composición de gráficos complicados y revelar conexiones inesperadas entre gráficos aparentemente diferentes”* (Cox 1978).

Siendo que en general trabajamos con datasets que tienen muchas variables, la idea es descubrir patrones, encontrar relaciones con el objetivo de comunicarlos y generar cambios.

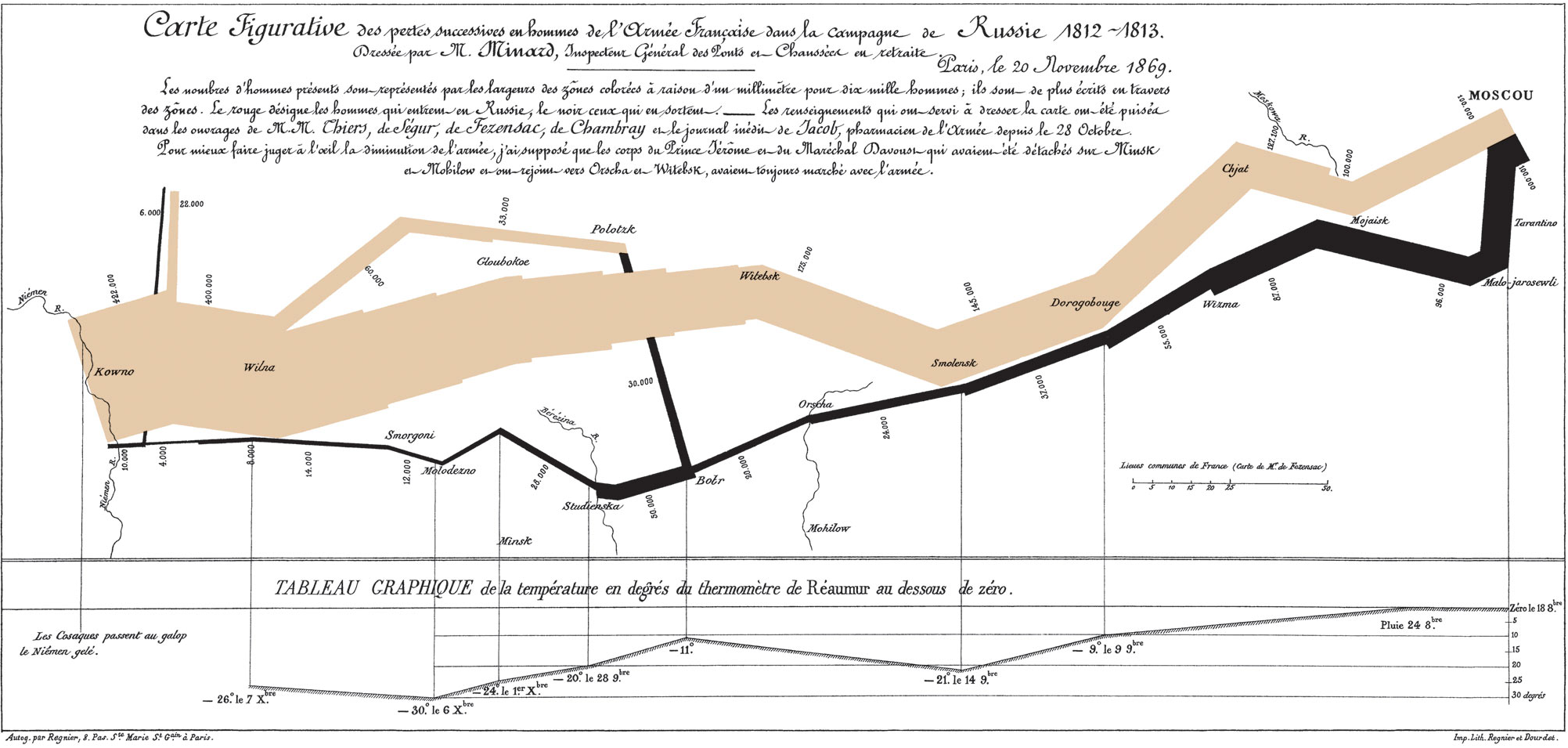
knitr::include\_graphics('https://github.com/Guadag12/R4RRII/raw/master/Clase%203/images/img\_def.jpg', 1)



El primer gráfico que se conoce en la historia exclusivamente a partir de datos es de [Charles Minard](https://en.wikipedia.org/wiki/Charles_Joseph_Minard) que grafico en 1812 la campaña Napoleonica -fallida- para invadir Rusia y cómo se da la evolución de la conquista conforme van llegando a Moscú. Y lo maravilloso del grafico es que muestra en dos dimensiones 6 tipos de variables diferentes: la cantidad de tropas, la distancia, la temperatura, la longitud, la latitud, la direccion del viaje y las fechas donde se sufren bajas.

Dicho grafico [es reproducible en R](https://github.com/andrewheiss/fancy-minard).

knitr::include\_graphics('https://github.com/Guadag12/R4RRII/raw/master/Clase%203/images/Minard.png', 1)



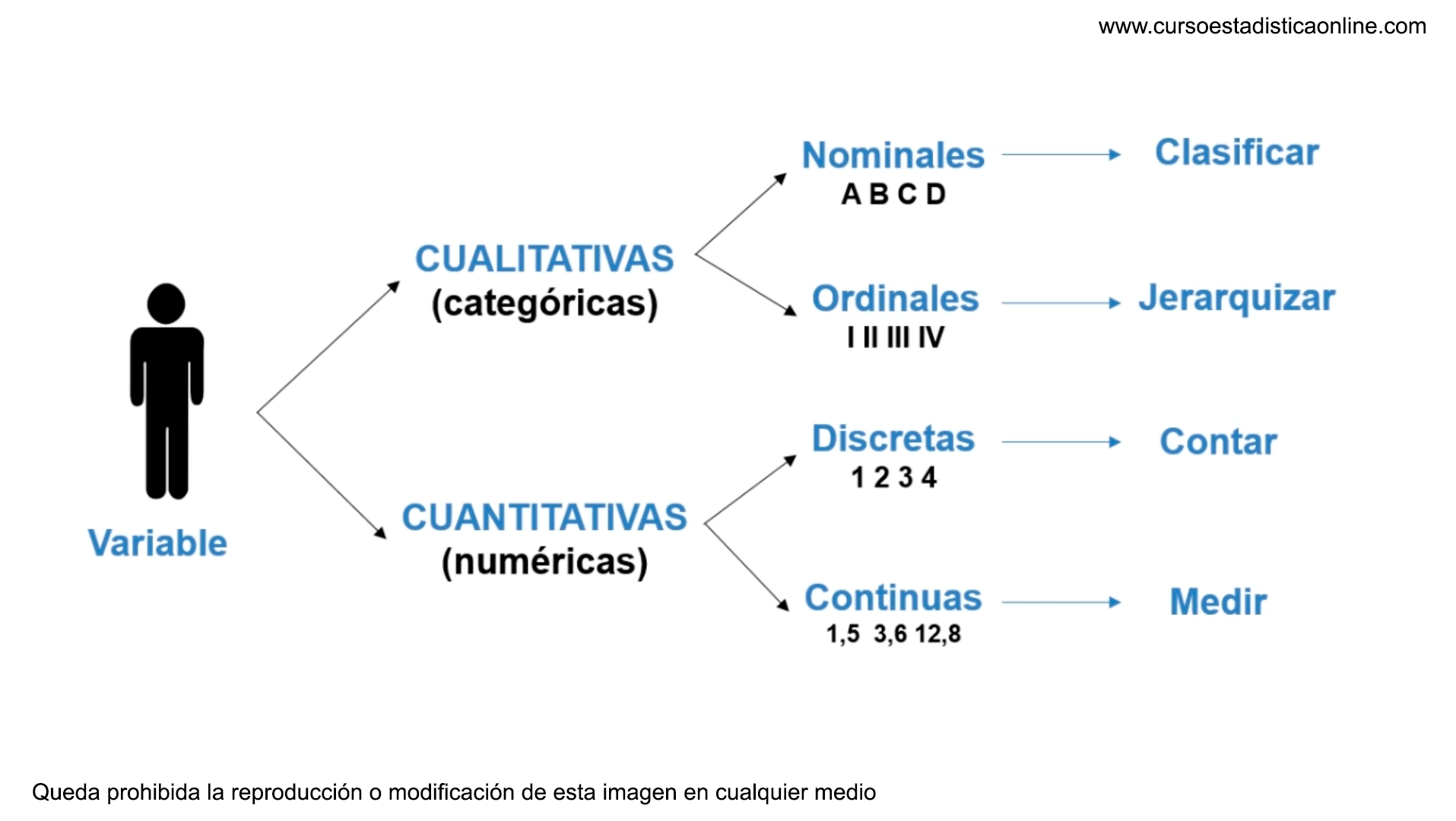
## Metodologia

Al trabajar con grandes cantidades de datos resulta necesario que seamos precisos con la cuestión metodológica. Hagamos un repaso!

knitr::include\_graphics('https://github.com/Guadag12/R4RRII/raw/master/Clase%203/images/variable-dependiente.png', 1)

![](data:text/html; charset=utf-8;base64,)

knitr::include\_graphics('https://github.com/Guadag12/R4RRII/raw/master/Clase%203/images/tiposVariables.png', 1)



Una vez entendido esto, es posible ver que en R hay múltiples gráficos que se pueden realizar:

knitr::include\_graphics('https://github.com/Guadag12/R4RRII/raw/master/Clase%203/images/flowchartmod.png', 1)

![](data:text/html; charset=utf-8;base64,)

## R BASE

#cargamos todas las librerias que necesitamos antes de arrancar  
library(viridis)

## Warning: package 'viridis' was built under R version 3.6.2

## Loading required package: viridisLite

## Warning: package 'viridisLite' was built under R version 3.6.2

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.2

## -- Attaching packages ------------------------------------------------------------------ tidyverse 1.3.0 --

## <U+2713> ggplot2 3.2.1 <U+2713> purrr 0.3.3  
## <U+2713> tibble 2.1.3 <U+2713> dplyr 0.8.3  
## <U+2713> tidyr 1.0.0 <U+2713> stringr 1.4.0  
## <U+2713> readr 1.3.1 <U+2713> forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.6.2

## Warning: package 'tibble' was built under R version 3.6.2

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'readr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2

## Warning: package 'dplyr' was built under R version 3.6.2

## Warning: package 'stringr' was built under R version 3.6.2

## Warning: package 'forcats' was built under R version 3.6.2

## -- Conflicts --------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

En esta ocasión vamos a trabajar con Indicadores de la [FUNDACION GAPMINDER](gapminder) que desarrolla estadisticas globales para Naciones Unidas desde 2005 y fue conocida a partir de una [Ted Talk](https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen?language=es) realizada por uno de sus co-fundadores Hans Rosling.

data <- read.csv("https://github.com/Guadag12/R4RRII/raw/master/Clase%203/data\_clase3.csv")

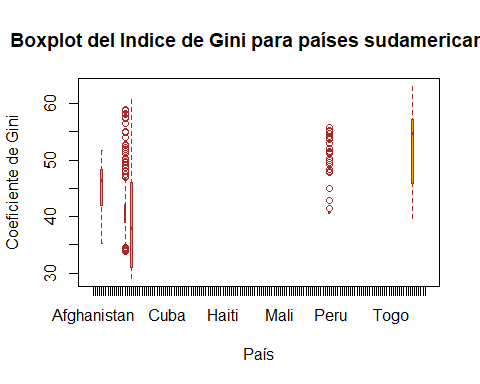
En primer lugar observemos nuestras variables:

head(data)

## country year gini\_coeff Code Continent Sub\_Continent child\_mortality  
## 1 Afghanistan 1800 30.5 AFG Asia Southern Asia 469  
## 2 Albania 1800 38.9 ALB Europe Southern Europe 375  
## 3 Algeria 1800 56.2 DZA Africa Northern Africa 460  
## 4 Angola 1800 57.2 AGO Africa Middle Africa 486  
## 5 Argentina 1800 47.7 ARG Americas South America 402  
## 6 Armenia 1800 31.5 ARM Asia Western Asia 371  
## PBI\_per\_capita  
## 1 603  
## 2 667  
## 3 715  
## 4 618  
## 5 1510  
## 6 514

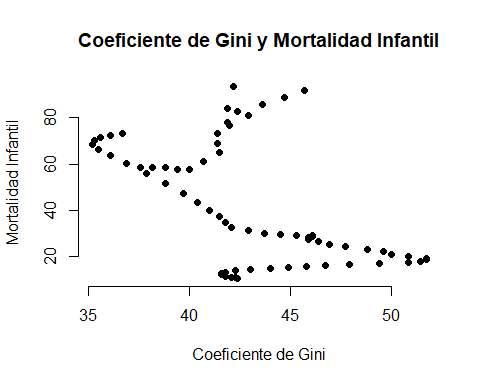
data\_sudamerica <- data %>% filter( Code == "ARG" | Code == "BOL" | Code == "BRA" |  
 Code == "PRY" | Code == "URY")

boxplot(data\_sudamerica$gini\_coeff~data\_sudamerica$country,  
 main="Boxplot del Indice de Gini para países sudamericanos",  
 xlab="País",  
 ylab="Coeficiente de Gini",  
 col="orange",  
 border="brown")

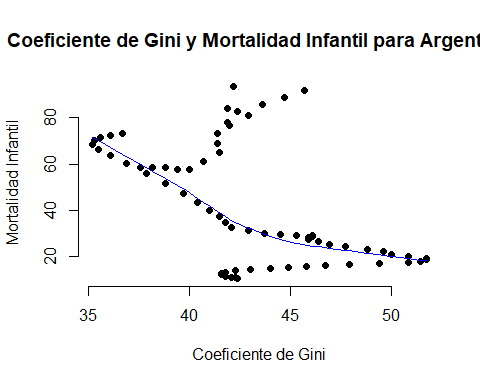


¿Qué pasa si nos preguntamos por la mortalidad en argentina según el grado de redistribucion de la riqueza?

data\_argentina <- data %>% filter(country == "Argentina" & year > 1950)  
  
plot(data\_argentina$gini\_coeff, data\_argentina$child\_mortality, main = "Coeficiente de Gini y Mortalidad Infantil",  
 xlab = "Coeficiente de Gini", ylab = "Mortalidad Infantil",  
 pch = 19, frame = FALSE)



plot(data\_argentina$gini\_coeff, data\_argentina$child\_mortality, main = "Coeficiente de Gini y Mortalidad Infantil para Argentina",  
 xlab = "Coeficiente de Gini", ylab = "Mortalidad Infantil",  
 pch = 19, frame = FALSE) +  
lines(lowess(data\_argentina$gini\_coeff, data\_argentina$child\_mortality), col = "blue")



## integer(0)

## GGPLOT

Sin embargo, cuando de Visualizacion se trata, el mejor paquete para trabajarlo en R se llama [ggplot2](https://ggplot2.tidyverse.org/) y lo tenemos cargado ya en nuestra libreria *TIDYVERSE*

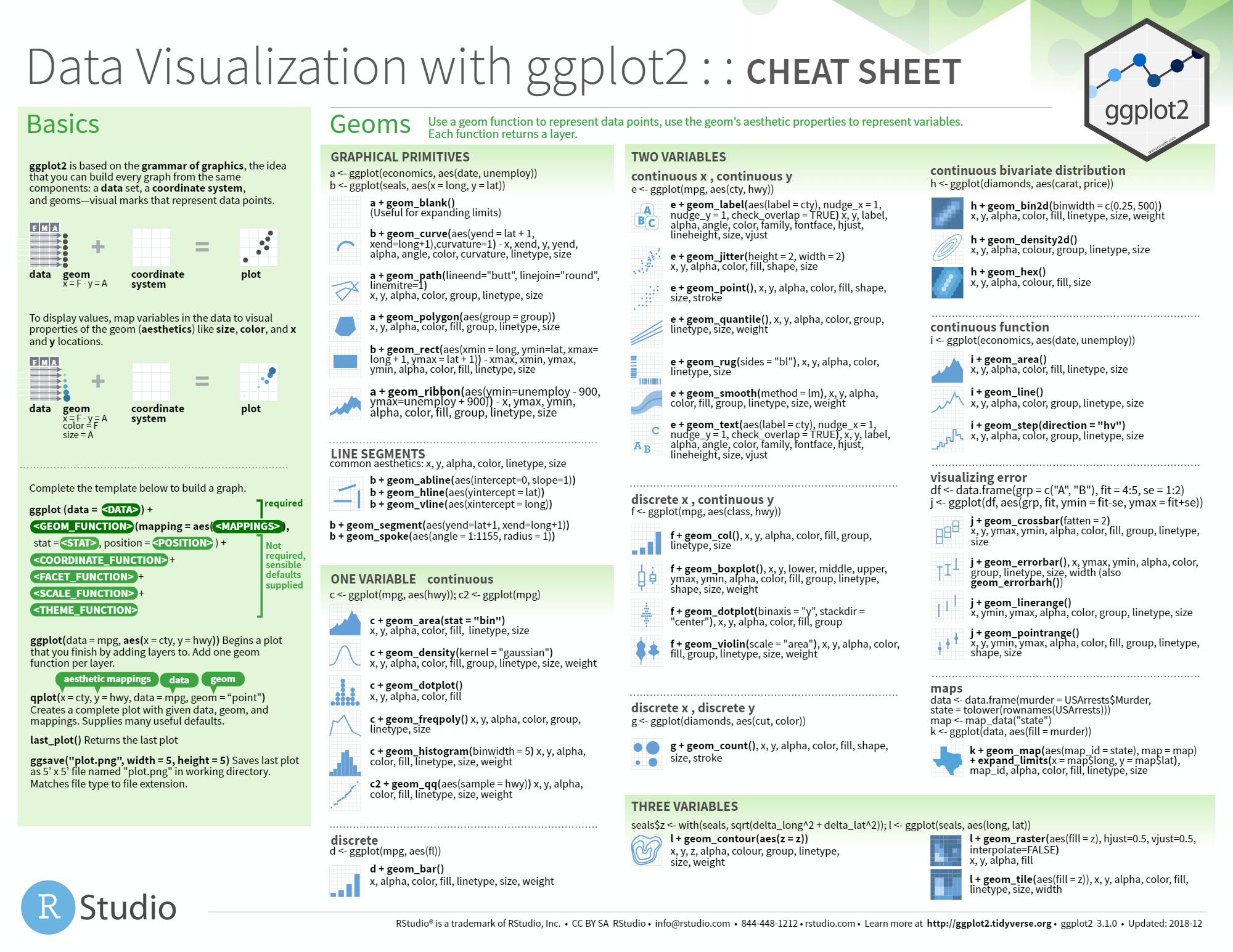
knitr::include\_graphics('https://github.com/Guadag12/R4RRII/raw/master/Clase%203/images/ggplotmeme.jpeg', 1)



ggplot2() es una herramienta superpoderosa para visualizar graficos en R y su creador es el mismo de Tidyverse: [Hadley Wickham](https://twitter.com/hadleywickham). Y la magia consiste en *“dividir los gráficos en componentes semánticos como escalas y capas”*.

Por lo cual tenemos múltiples posibilidades y e infinitas combinaciones a realizar en un mismo paquete.

knitr::include\_graphics('https://github.com/Guadag12/R4RRII/raw/master/Clase%203/images/cheatsheet.png', 1)

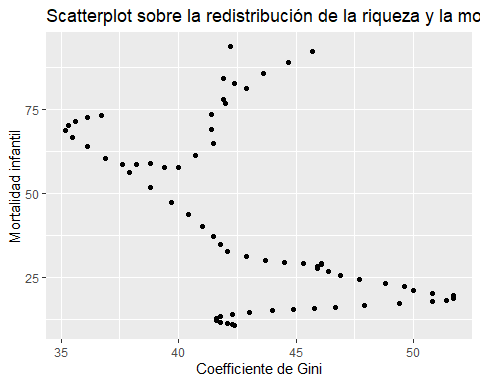


library(ggplot2)

Se basa en capas que se agregan a partir del signo *“+”* - asi como en tidyverse teniamos al *“%>%”*. Y consta de un fondo gris ya que múltiples investigaciones hablan de que ayuda y fomenta el contraste. Las formas geometricas y la estetica son los aspectos principales para distinguir capas y escalas.

Podemos hacer un scatterplot como el que teniamos arriba, por ejemplo asi:

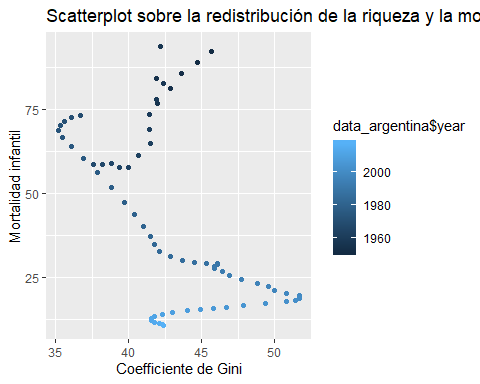
ggplot(data = data\_argentina) +   
 geom\_point(mapping = aes(x = data\_argentina$gini\_coeff, y = data\_argentina$child\_mortality)) +  
 ggtitle("Scatterplot sobre la redistribución de la riqueza y la mortalidad infantil") +  
 xlab("Coefficiente de Gini") + ylab("Mortalidad infantil")



### Aesthetic mappings

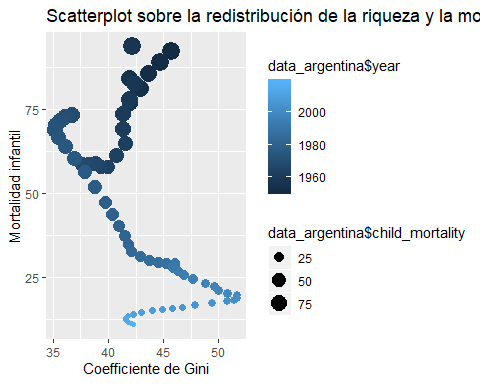
Y de lo que se distingue ggplot2() de otros paquetes es en la capacidad de hacer visualizaciones más lindas y simples. La funcion **aes()** es muy importante para esto. Podemos cambiarle el color:

ggplot(data = data\_argentina) +   
 geom\_point(mapping = aes(x = data\_argentina$gini\_coeff, y = data\_argentina$child\_mortality, color = data\_argentina$year)) +  
 ggtitle("Scatterplot sobre la redistribución de la riqueza y la mortalidad infantil") +  
 xlab("Coefficiente de Gini") + ylab("Mortalidad infantil")

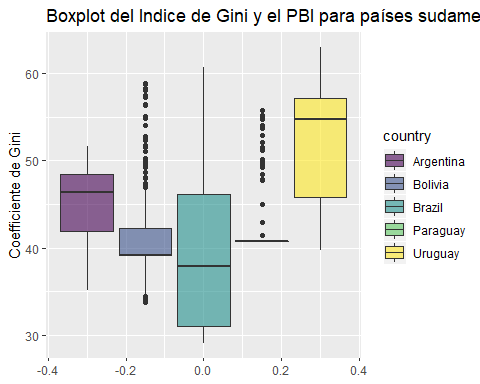


El tamaño:

ggplot(data = data\_argentina) +   
 geom\_point(mapping = aes(x = data\_argentina$gini\_coeff, y = data\_argentina$child\_mortality,color = data\_argentina$year, size = data\_argentina$child\_mortality)) +  
 ggtitle("Scatterplot sobre la redistribución de la riqueza y la mortalidad infantil") +  
 xlab("Coefficiente de Gini") + ylab("Mortalidad infantil")

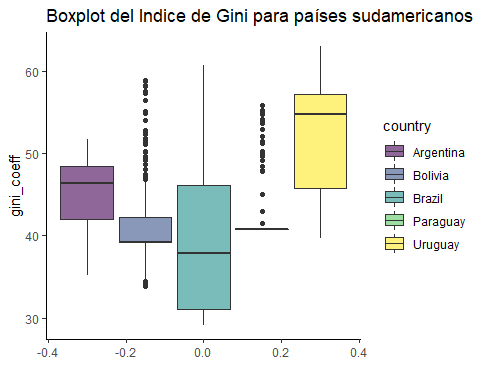
 Podemos hacer boxplots:

data\_sudamerica %>%  
 ggplot( aes( y=gini\_coeff, fill=country)) +  
 geom\_boxplot() +  
 scale\_fill\_viridis(discrete = TRUE, alpha=0.6) +  
 ggtitle("Boxplot del Indice de Gini y el PBI para países sudamericanos") +  
 ylab("Coefficiente de Gini")



Podemos incluso cambiarle el fondo gris con la funcion *theme()*:

data\_sudamerica %>%  
 ggplot( aes( y=gini\_coeff, fill=country)) +  
 geom\_boxplot() +  
 scale\_fill\_viridis(discrete = TRUE, alpha=0.6) +  
 ggtitle("Boxplot del Indice de Gini para países sudamericanos") +  
 xlab("") +  
 theme\_classic()

 Podemos hacer diferentes graficos para cada pais con la funcion *facet\_grid()*:

data\_sudamerica %>%  
 ggplot( aes(y=gini\_coeff, fill=country)) +  
 geom\_boxplot() +  
 scale\_fill\_viridis(discrete = TRUE, alpha=0.6) +  
 ggtitle("Boxplot del Indice de Gini para países sudamericanos") +  
 xlab("") +  
 facet\_wrap(~ country, nrow =2)

