


Dra. Diana Diaz

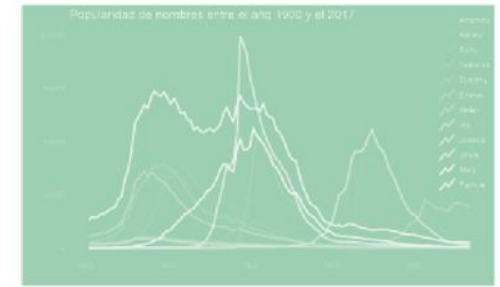
Profesora en el Departamento de
Ciencias de la Computacion
Universidad de Illinois en
Chicago (UIC)

 DIANA DIAZ

 @dianadiazh



WOMEN IN DATA SCIENCE



Hoy hablamos sobre...

PARTE 1:

- **¿De qué hablamos cuando hablamos de visualización de datos?**
¿Qué es la visualización de datos? ¿Para qué visualizamos?
Importancia de la visualización dentro del proceso de la ciencia de datos
Cualidades de una buena visualización
- **Representando mis datos**
Componentes visuales: ¿Cuáles son los ingredientes de una visualización?
Visualizando con claridad
- **Un recorrido por los gráficos más típicamente utilizados**
Gráficos para representar: **cantidad**, **distribución**, **relación** y **dispersión**.
Gráficos elementales: Bar chart- Line Plot - Scatter Plot- Density Plot.



Hoy hablamos sobre...

PARTE 2:

- **El paquete ggplot2**

El paquete ggplot2 y la gramática de gráficos en capas

Capas de un gráfico

Sintaxis - ¿Cómo hacer un gráfico paso a paso?



- **¿Por dónde empiezo?**

Comunidades de aprendizaje: #Rladies #R4DSEs #DatosdemierRcoles

¿Cómo Participo?

- **Hands-on con ggplot2**

Practicamos con los datasets de Gapminder y Propina





Parte 1

¿De qué hablamos
cuando hablamos de
visualización de datos?

Podemos entender a la visualización como un medio que puede ser usado como una **herramienta** y a la vez como una **forma de expresar datos**.

Source: *Data Points, Visualization that Means Something* de Nathan Yau.



Visualización



El mundo
real

Datos

Formas y
colores



Interpretación

Source: *Data Points, Visualization that Means Something* de Nathan Yau.



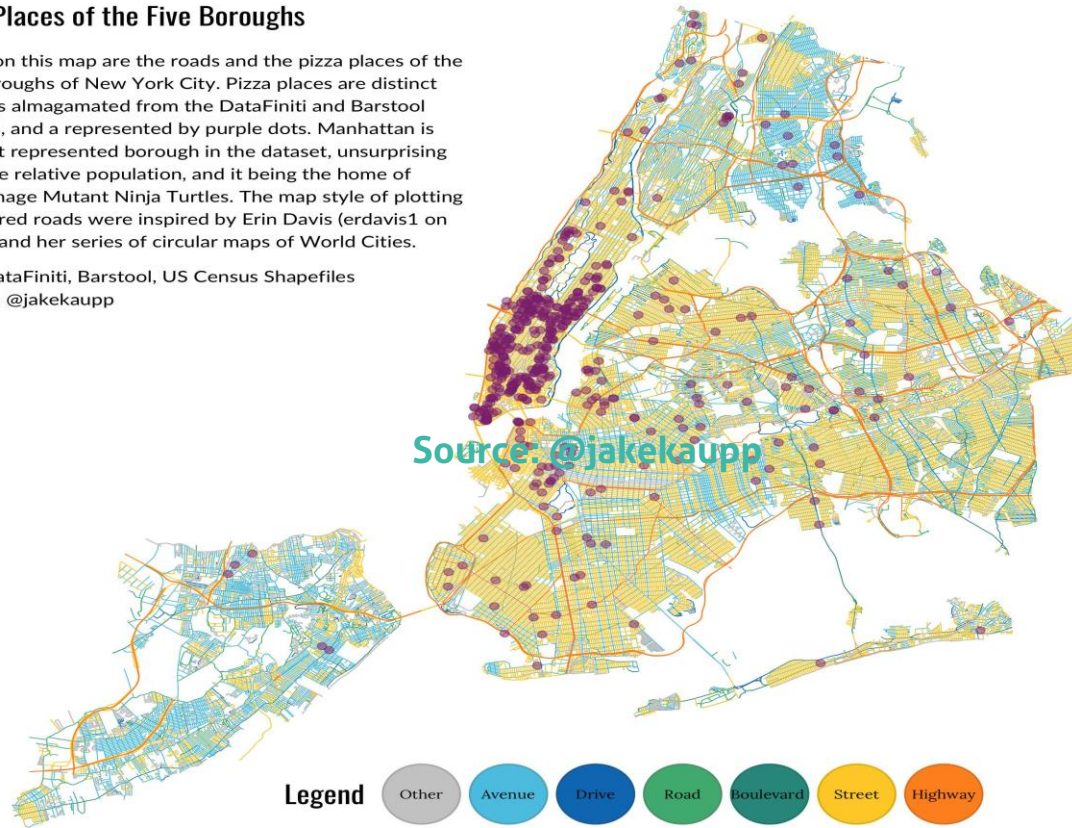
Ejemplos

Pizza Places of the Five Boroughs

Shown on this map are the roads and the pizza places of the Five Boroughs of New York City. Pizza places are distinct locations amalgamated from the DataFiniti and Barstool datasets, and are represented by purple dots. Manhattan is the most represented borough in the dataset, unsurprising given the relative population, and it being the home of the Teenage Mutant Ninja Turtles. The map style of plotting the colored roads were inspired by Erin Davis (erdavis1 on github), and her series of circular maps of World Cities.

Data: DataFiniti, Barstool, US Census Shapefiles

Graphic: @jakekaupp



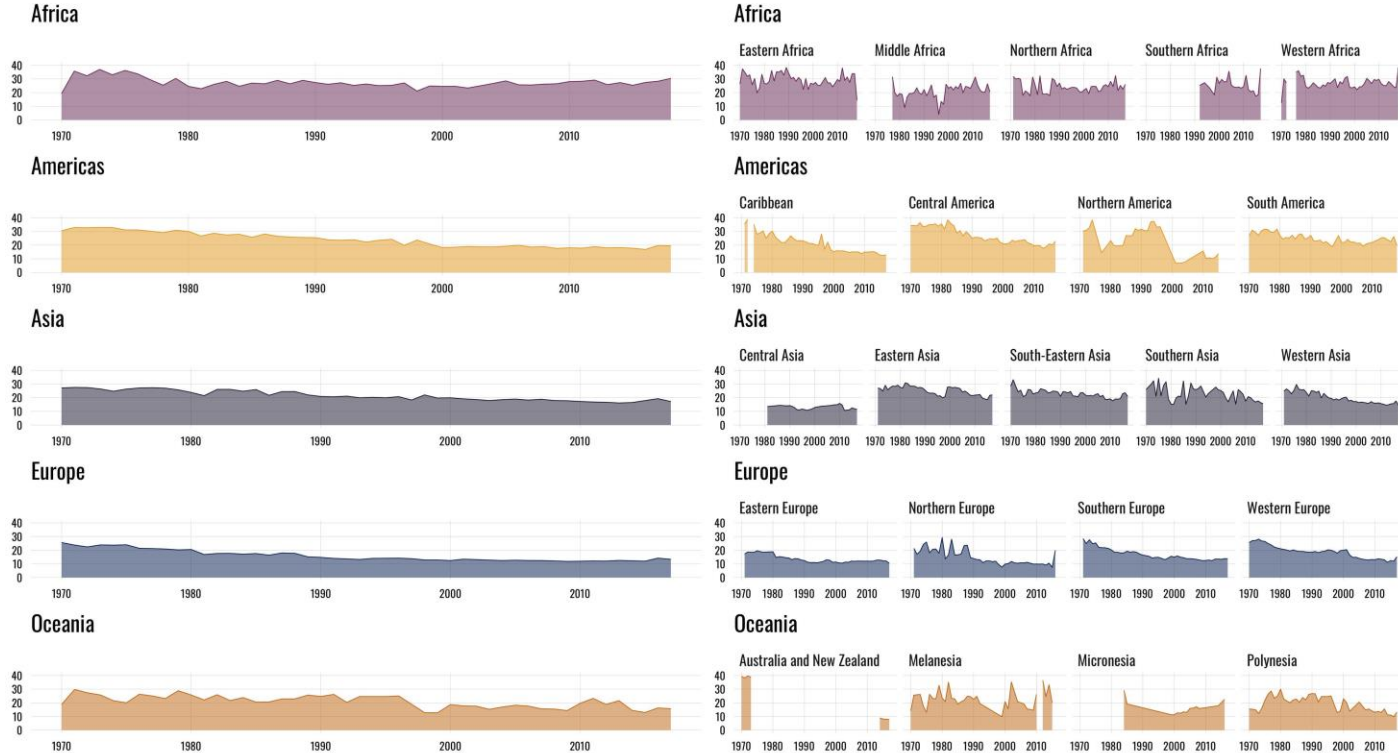
Legend



Source: @jakekaupp

Working to Two Sigma: Student Teacher Ratios Improving Since the 1970s

Illustrated below is the average student to teacher ratio across each continent (left column) and region (right column). Continent and region assigned from iso3c coding of country name and are consistent with the World Bank Development Indicators.



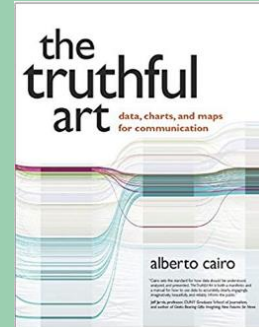
Data: UNESCO Institute of Statistics | Graphic: @jakekaupp

Source: @jakekaupp



¿Y Para qué visualizamos?

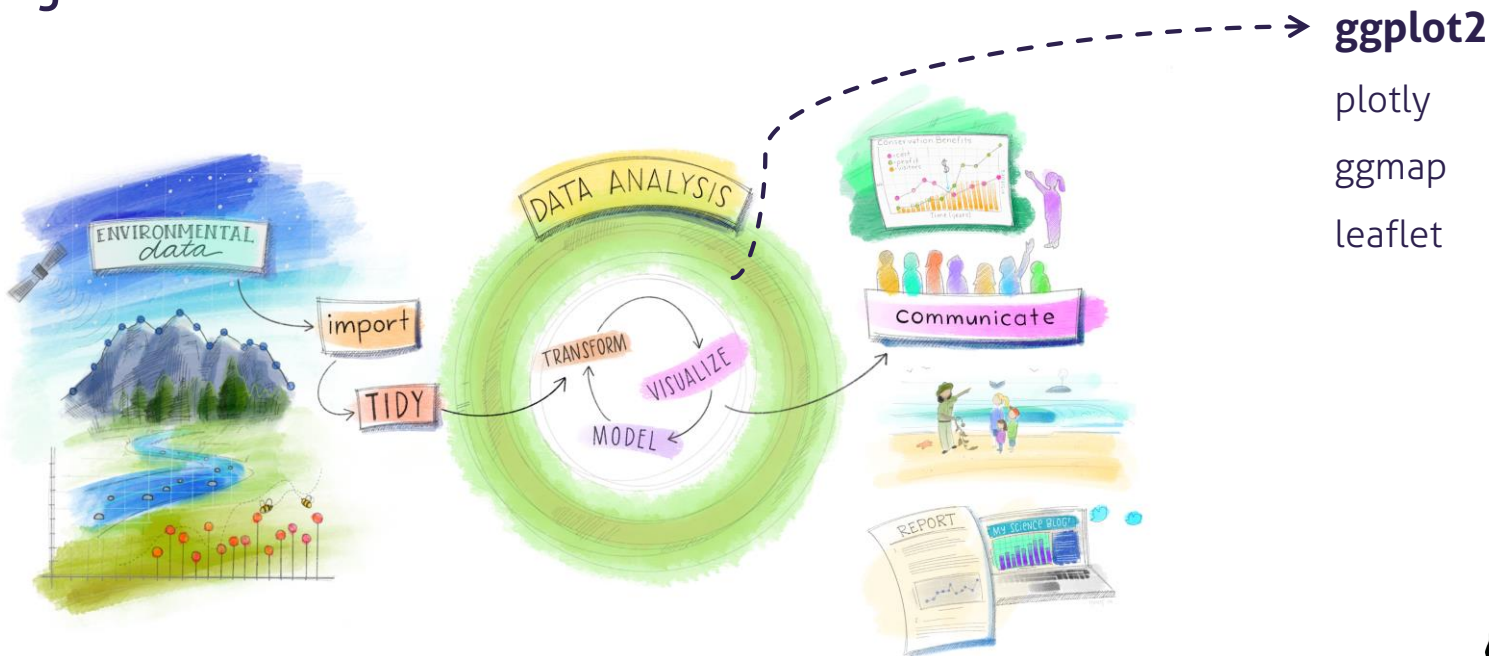
Para expresar y transmitir el significado de los datos de manera precisa, clara, atractiva, imaginativa, bella y confiable con el objetivo de informar al público, a nuestro público.



Source: *The Truthful Art* de Alberto Cairo.



La visualización dentro del esquema de trabajo en ciencia de datos



Art by Allison Horst

A close-up photograph of a hand holding a piece of white chalk, writing on a green chalkboard. The hand is positioned on the right side of the frame, with the thumb and index finger visible. The chalkboard has some faint, handwritten text in white chalk, including the word 'Konni' and some numbers. A dark purple rectangular box is overlaid on the left side of the image, containing the text 'Una imagen vale más que mil palabras pero...'.

Una imagen vale más que mil
palabras pero...

Cualidades de una buena visualización

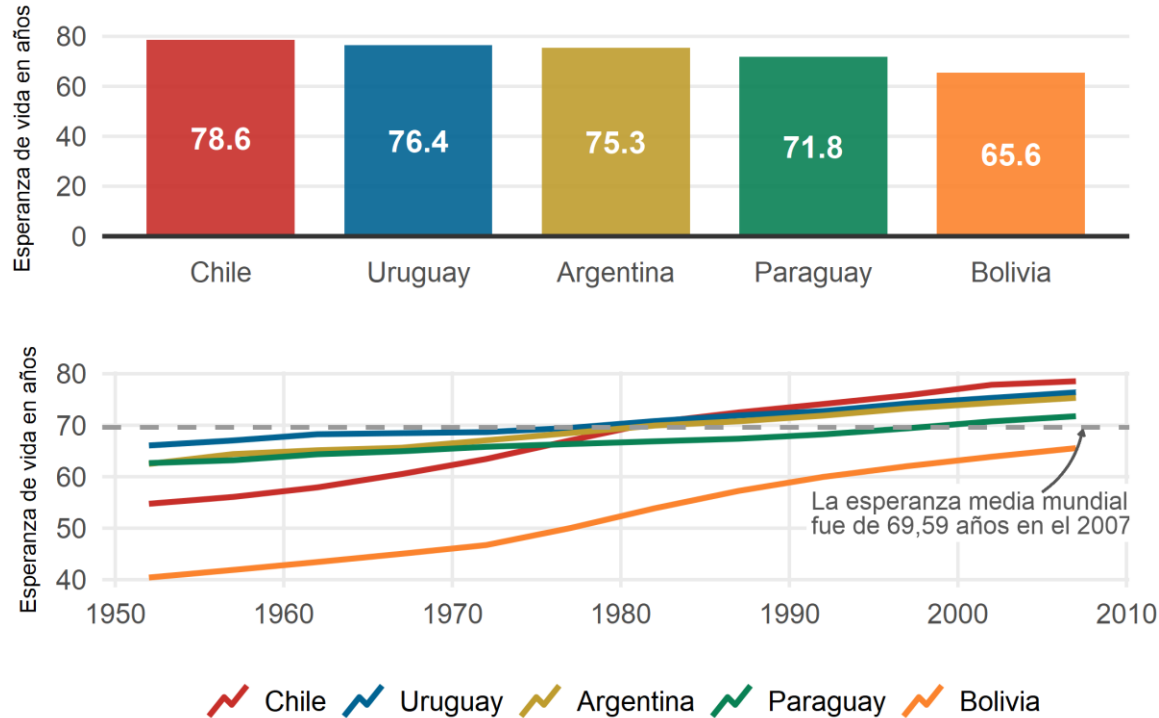
Las cinco cualidades de una visualización memorable:

- Que sea agradable a la vista
- Que sea funcional
- Que muestre hallazgos
- Que esté basada en datos confiables, es decir, que transmita la verdad



Expectativa de vida de Argentina y países limítrofes

Período 1950-2017



Data: países del paquete datos | Por Patricia Loto



WOMEN IN DATA SCIENCE

Representando mis datos



¿Cuáles son los ingredientes de una **visualización**?

- Cada visualización se construye sobre **datos** y **cuatro componentes**:



¿Cuáles son los ingredientes de una visualización?

1. Señales visuales +

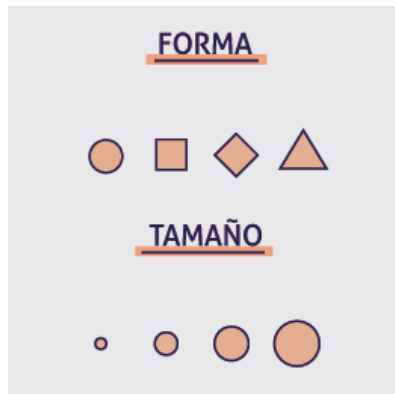
Involucran la codificación de datos mediante:

- **Formas**
- **Tamaños**
- **Colores, etc.**



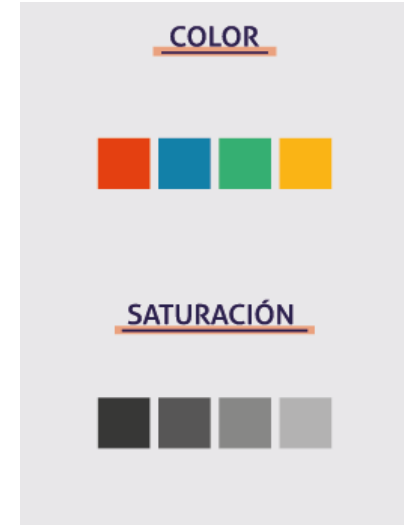
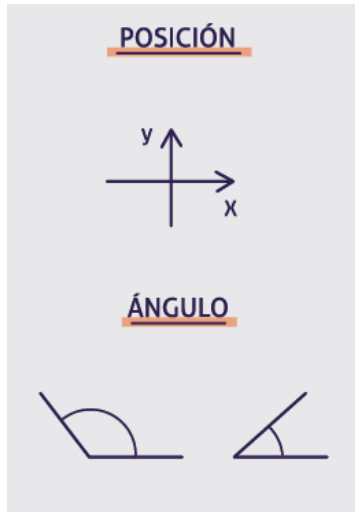
Source: *Data Points, Visualization that Means Something* de Nathan Yau.

1. Señales visuales



Source: *Data Points, Visualization that Means Something* de Nathan Yau.

1. Señales visuales



Source: *Data Points, Visualization that Means Something* de Nathan Yau.



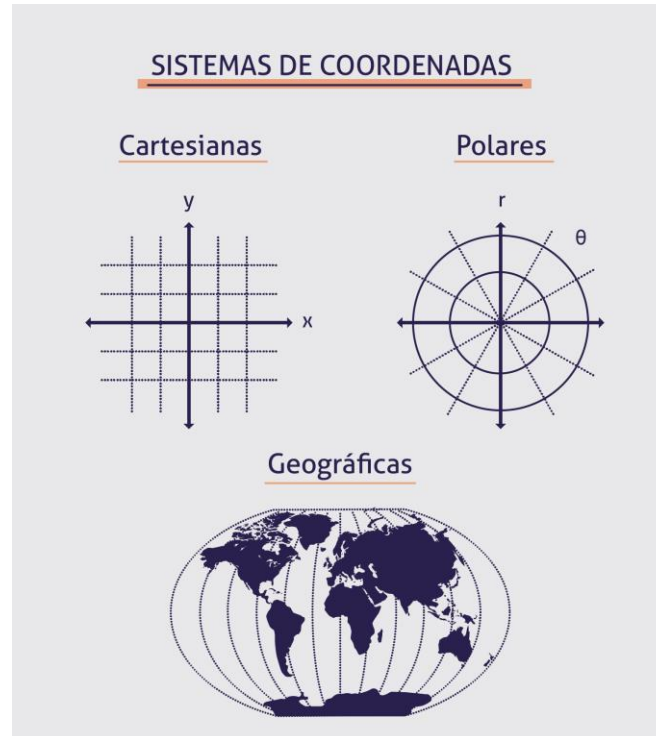
¿Cuáles son los ingredientes de una visualización?

1. Señales visuales +
2. Sistemas de coordenadas +

Source: *Data Points, Visualization that Means Something* de Nathan Yau.



2. Sistemas de coordenadas





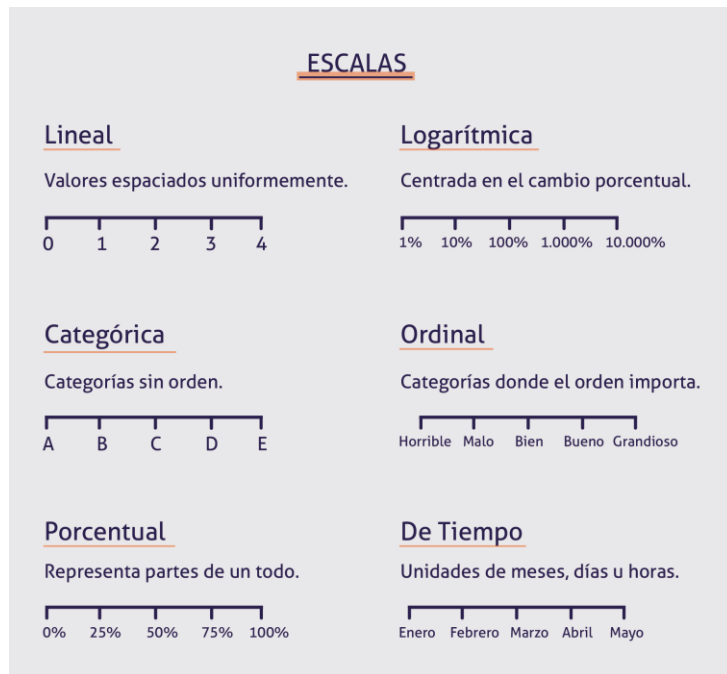
¿Cuáles son los ingredientes de una visualización?

1. Señales visuales +
2. Sistemas de coordenadas +
3. Escalas +

Source: *Data Points, Visualization that Means Something* de Nathan Yau.



3. Escalas



Source: *Data Points, Visualization that Means Something* de Nathan Yau.

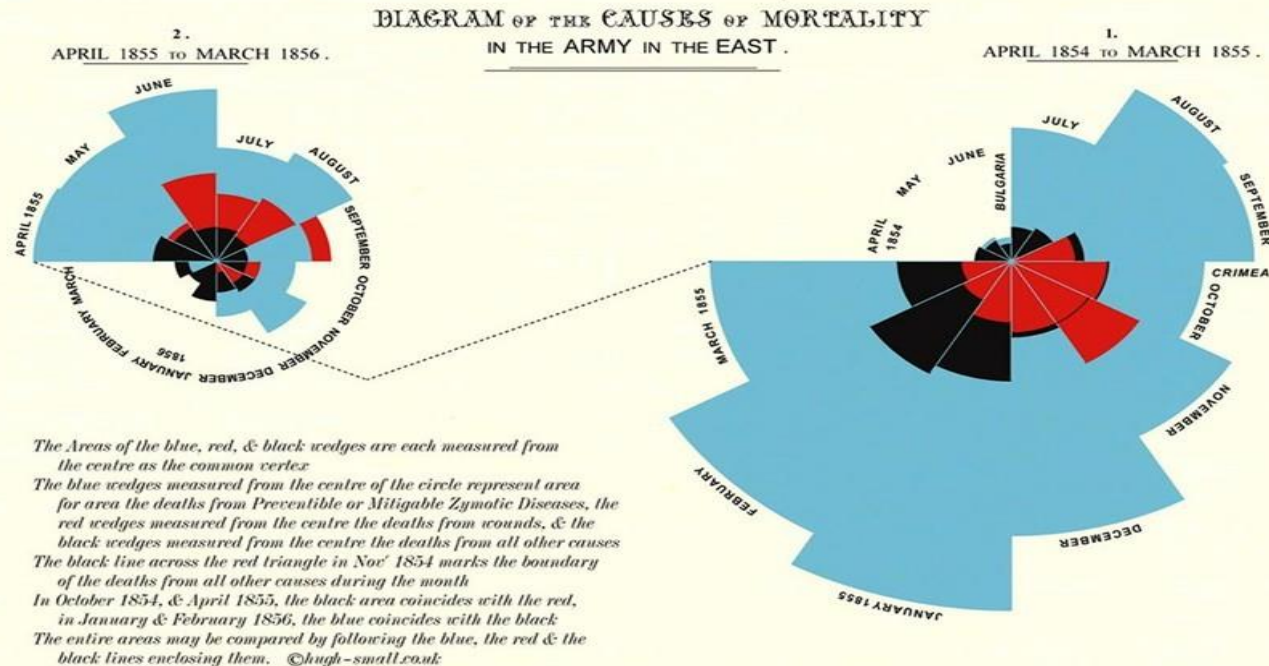


¿Cuáles son los ingredientes de una visualización?

1. Señales visuales +
2. Sistemas de coordenadas +
3. Escalas +
4. Contexto



¿Qué sucede si unimos todos los ingredientes?





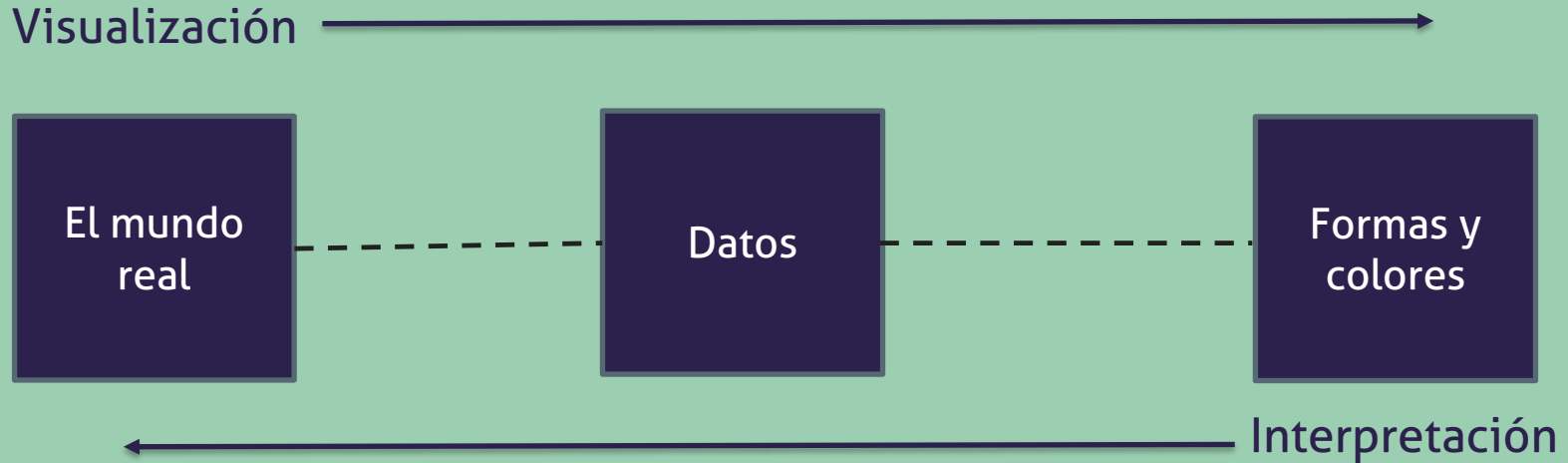
Visualizando con claridad

“When you use graphics to present results to other people, you must make your graphics readable to those who don’t know your data as well as you do.”

Source: Data Points, Visualization that Means Something de Nathan Yau.



Visualizando con claridad

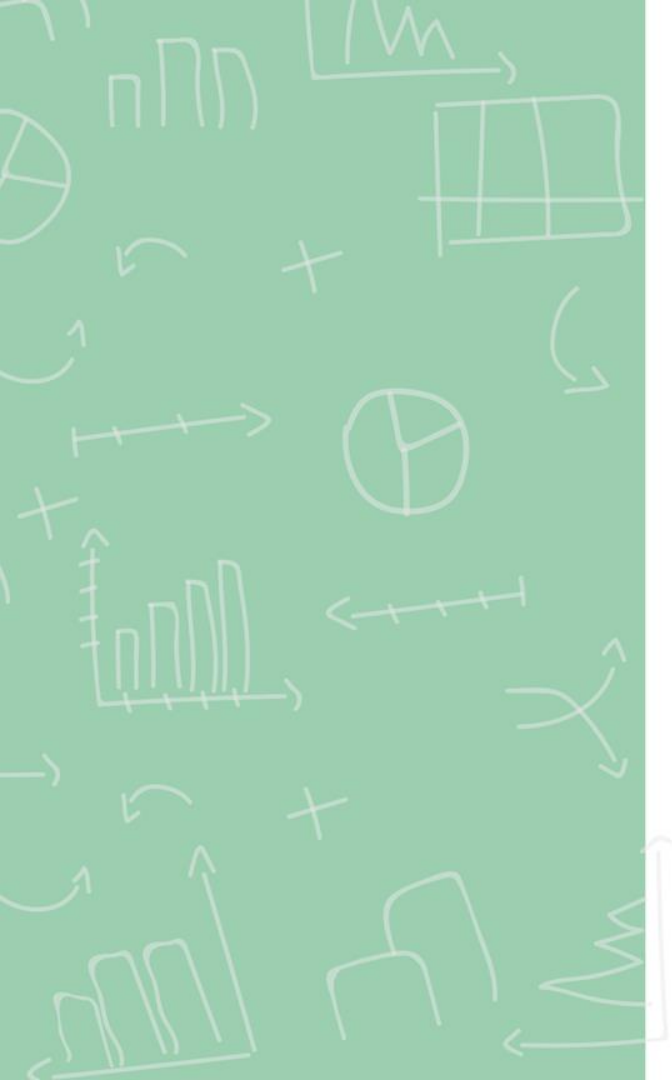


Source: *Data Points, Visualization that Means Something* de Nathan Yau.



Recursos que favorecen la interpretación de los gráficos

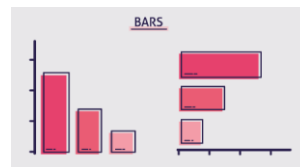
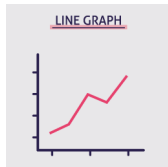
- Jerarquía en la visualización
- Resaltado de la información más importante
- Anotaciones: brindando contexto a los datos
- Medidas estadísticas: media, quartiles, etc.
- Color



Recorrido por los gráficos más típicamente utilizados

Tipos de gráficos

- Cantidad
- Distribución
- Proporción
- Relación entre 'x' e 'y'
- Dispersión
- Datos geoespaciales



Source: *Fundamentals of Data Visualization* de Claus Wilke



Cantidad (una variable)

Gráfico de barras

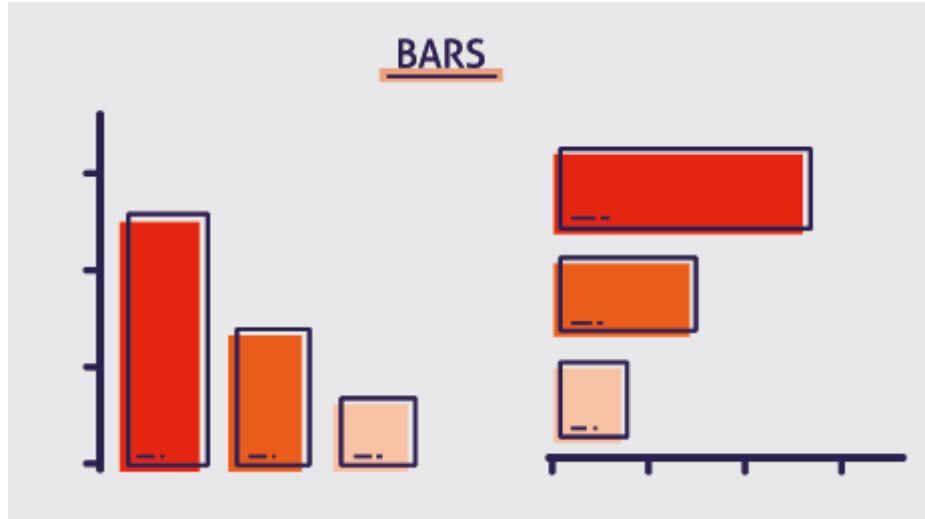
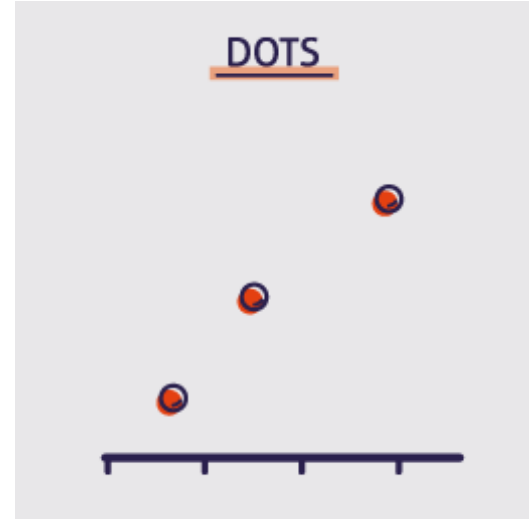


Gráfico de puntos



Cantidad (múltiples variables)

Gráfico de barras agrupadas

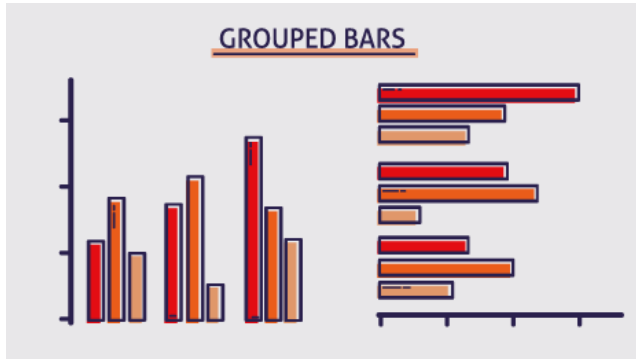
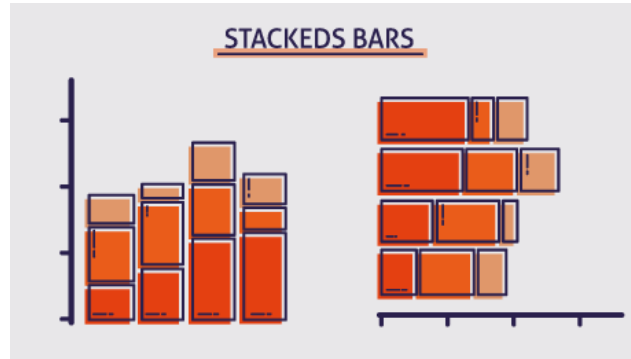
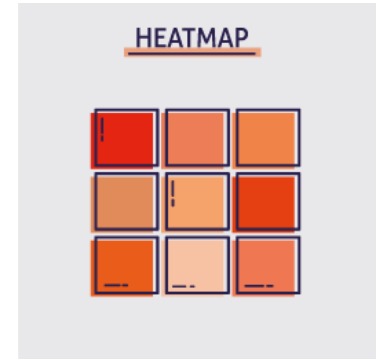


Gráfico de barras apiladas



Mapa de calor



Distribución Simple

Histograma

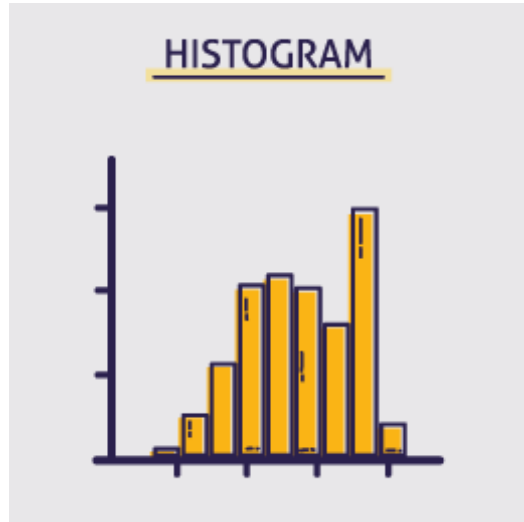
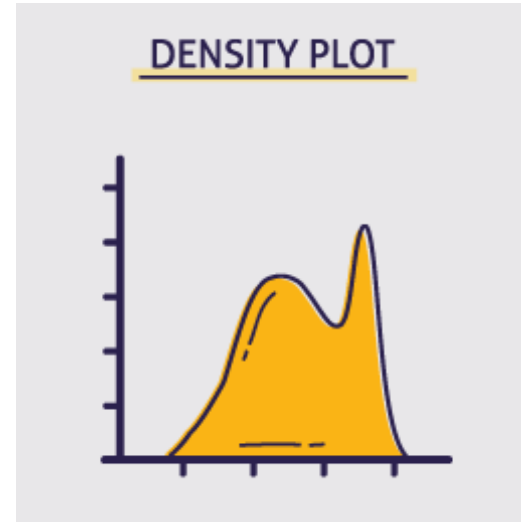
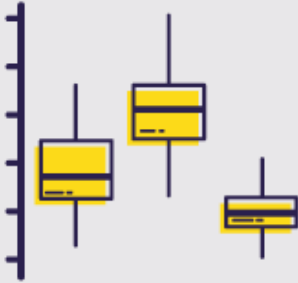


Gráfico de Densidad

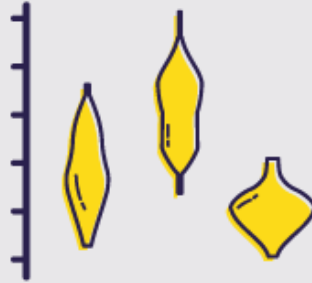


Distribución Múltiple

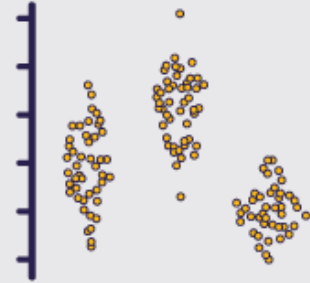
BOXPLOTS



VIOLINS

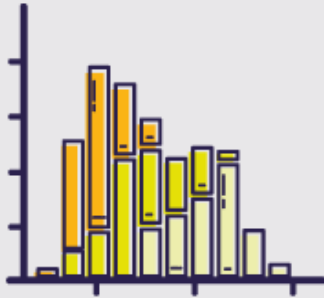


SINA PLOTS

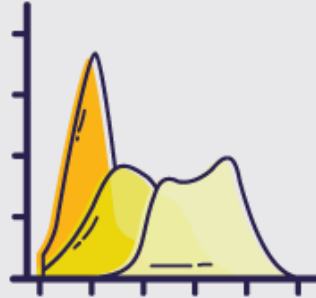


Distribución Múltiple

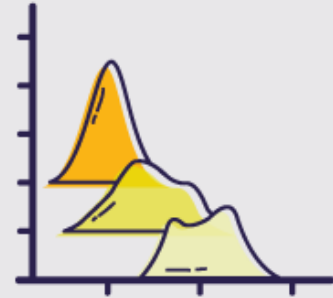
STACKED HISTOGRAMS



OVERLAPPING DENSITIES



RIDGELINE PLOT



Proporción (una variable)

Gráfico de torta

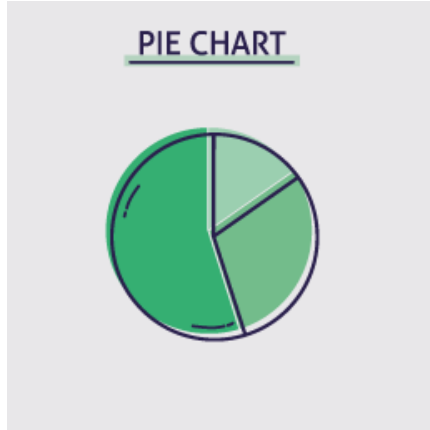
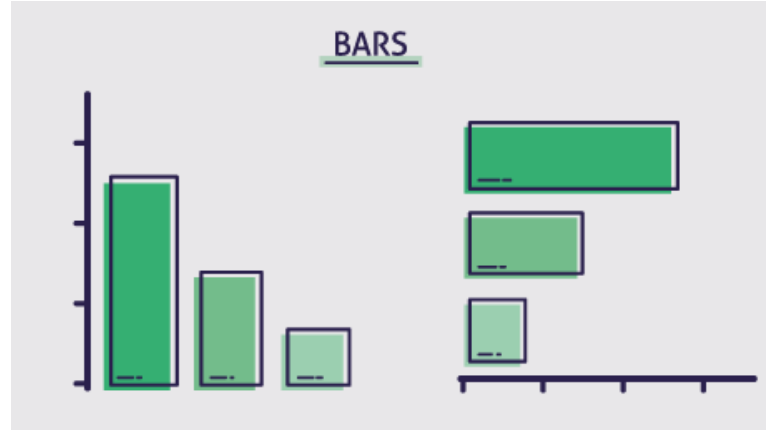


Gráfico de barras



Proporción (múltiples variables)

Múltiples gráficos de torta

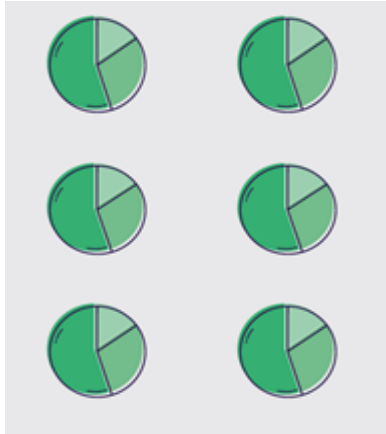
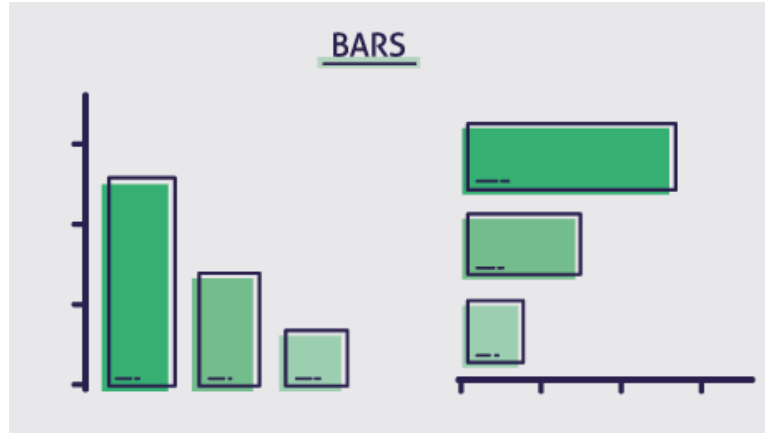
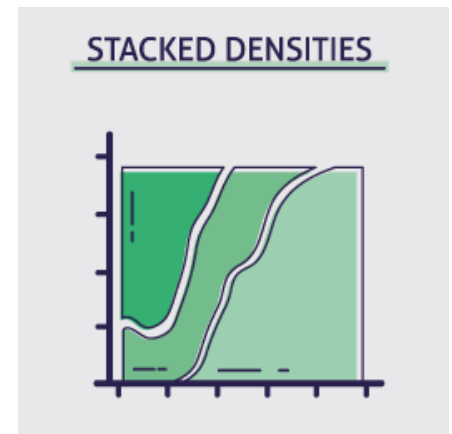


Gráfico de barras

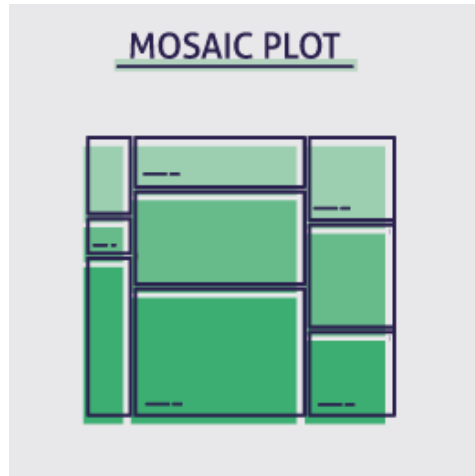


Densidades apiladas



Proporción (múltiples variables)

Gráfico de Mosaico



Mapa de árboles

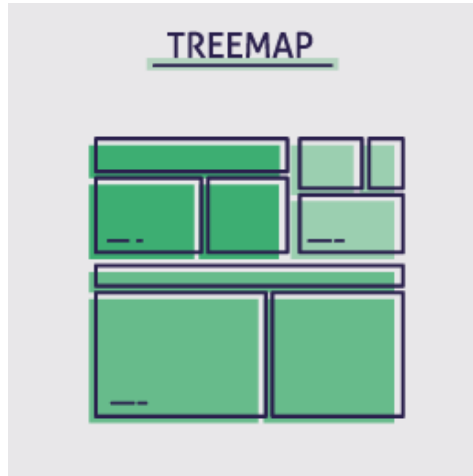
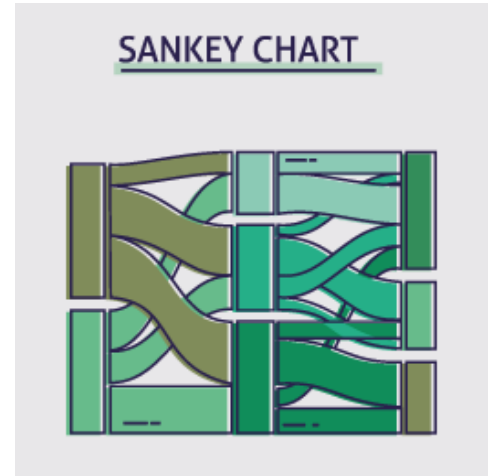
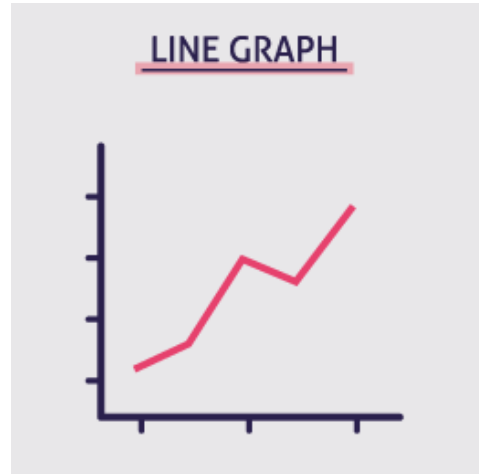


Diagrama de sankey



Relación entre 'x' e 'y' (una variable)

Gráfico de líneas



Scatterplot conectado

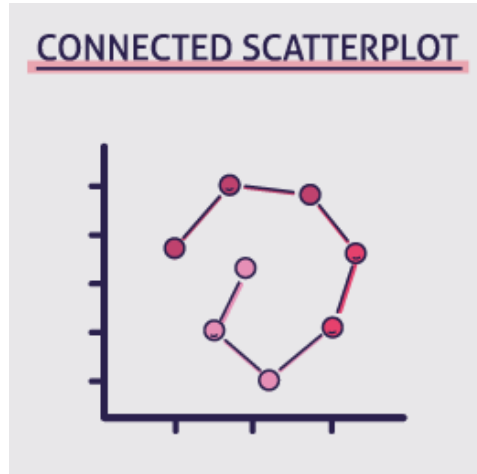
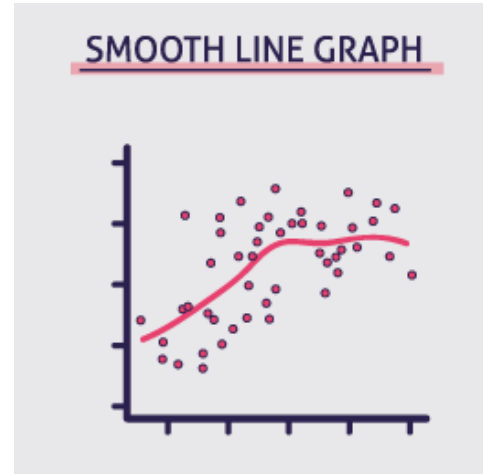


Gráfico dispersión



Relación entre más de una variable

Hex Bins

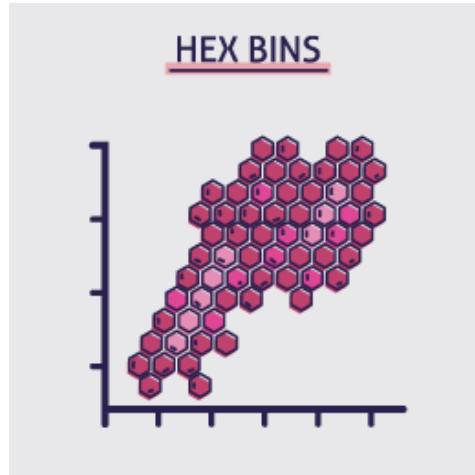


Gráfico de líneas

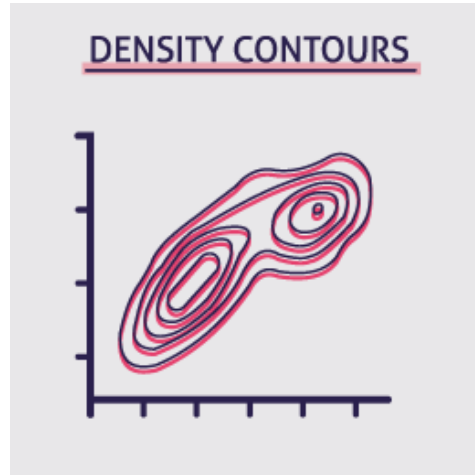
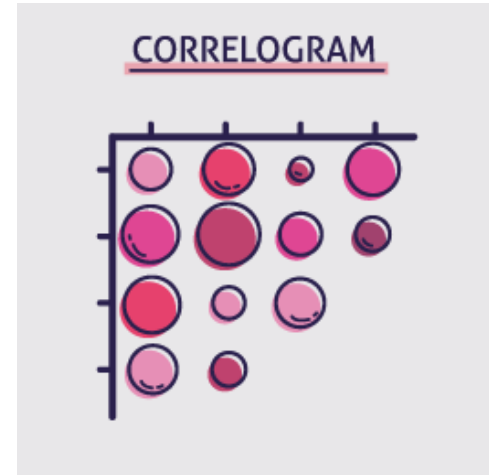


Gráfico de correlación



Dispersión

Gráfico de burbujas

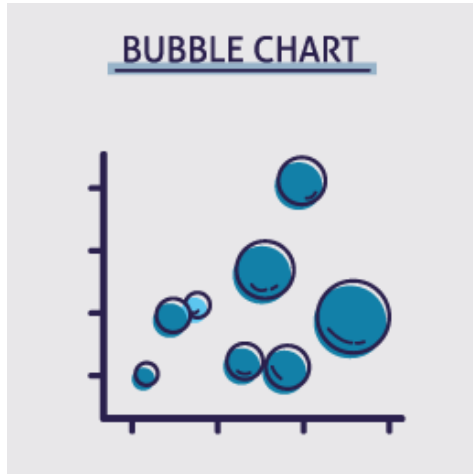


Gráfico de dispersión

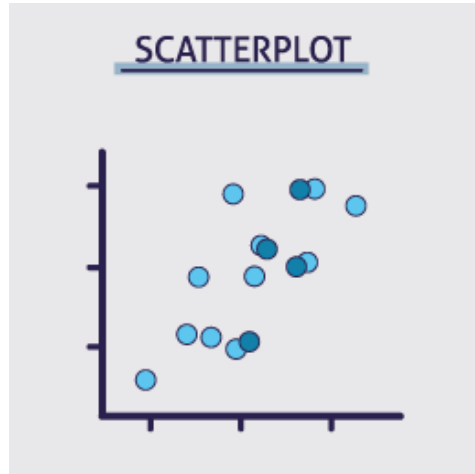
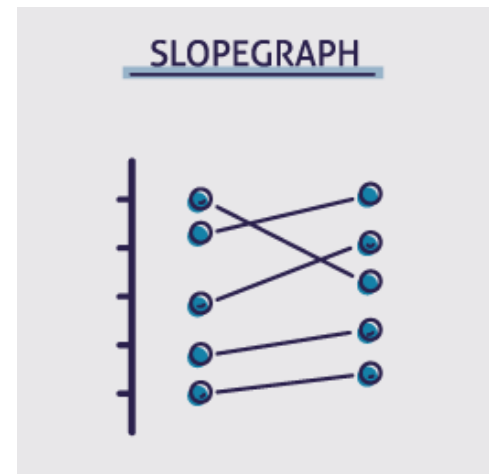
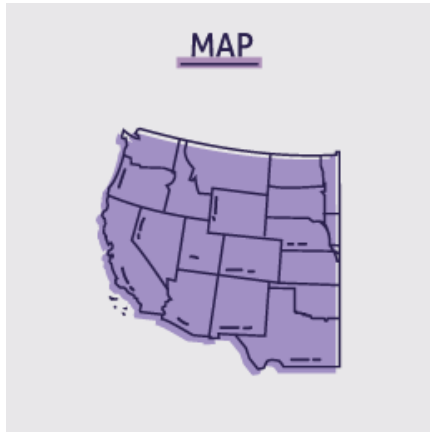


Gráfico de líneas suavizada

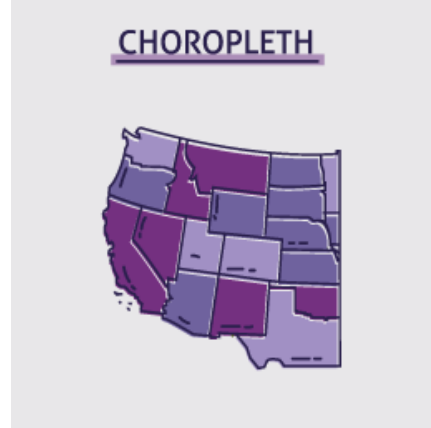


Datos geoespaciales

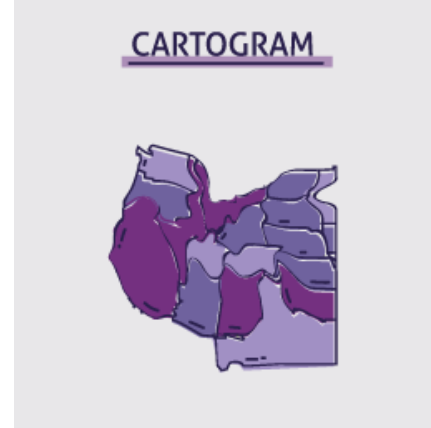
Mapa



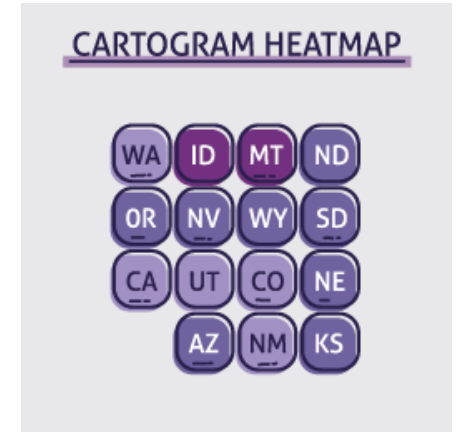
Mapa coroplético



Cartograma

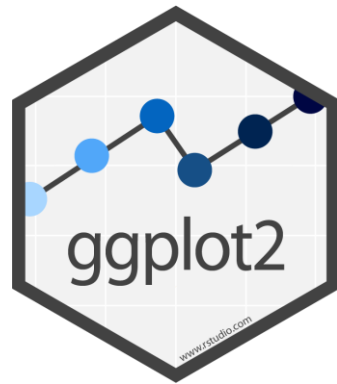


Cartograma Heatmap




Source: *Fundamentals of Data Visualization* de Claus Wilke

Parte 2



El paquete ggplot2



ggplot2 es uno de los  más populares para visualización de datos dentro de la comunidad R.

Fue desarrollado por **Hadley Wickham** (2008) y está basado en la **gramática de gráficos en capas**.

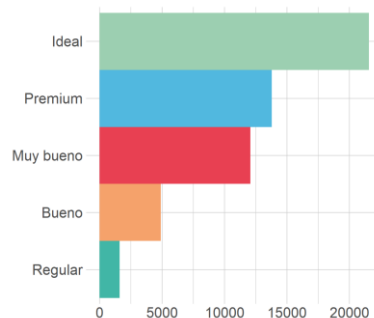
Es parte de un conjunto de paquetes que tiene foco en la ciencia de datos llamado **Tidyverse**.



Gramática de gráficos en capas o layered grammar of graphics

La Gramática de gráficos en capas (basada en “The Grammar of Graphics” by Wilkinson, Anand, and Grossman) nos permite conocer:

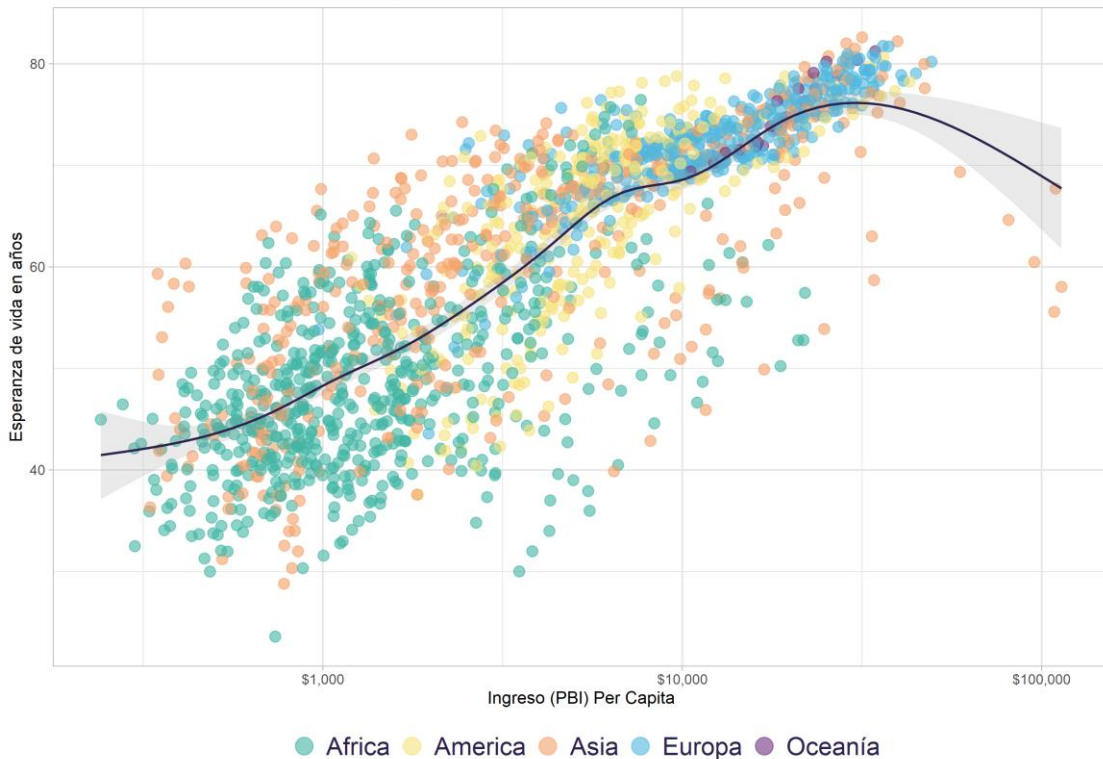
- ¿Qué es un gráfico?
- ¿Cuáles son los componentes de un gráfico?
- ¿Cómo describir y crear un gráfico?



¿Qué hay detrás de un gráfico?

Crecimiento económico y esperanza de vida

Los puntos se representan por año-país



DataSource: Gapminder- Link: <https://www.gapminder.org>.



WOMEN IN DATA SCIENCE

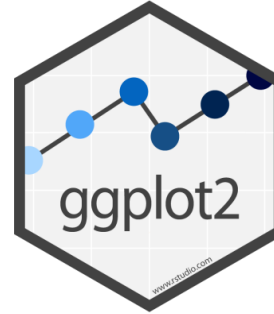
¿Qué hay atrás de cada gráfico?

- ✓ Cada **observación** está representada con un **punto**, cuya posición está dada por **dos variables** (posición horizontal y vertical).
- ✓ Cada punto tiene tamaño y color, estos atributos son denominados elementos estéticos o aes.
- ✓ Los aes son propiedades que pueden ser percibidas en el gráfico.
 - Cada **aes** puede ser mapeado a una variable o fijado en un valor constante.





Ggplot2: primeros pasos



Sintaxis de ggplot2

```
P <- ggplot(data= <DATOS>,  
           mapping = aes( <MAPEOS>)) +
```

```
< FUNCTION_GEOM> (  
  mapping = aes( <MAPEOS>),  
  stat= < STAT>,  
  position= <POSITION>) +
```

```
<ESCALAS> +  
<COORDENADAS> +  
<ETIQUETAS>  
<FACETAS> +  
<TEMAS>
```

Todo objeto de ggplot2 tiene al menos 3 componentes principales:

1. Datos (data)
nuestro set de datos.
2. Atributos o elementos estéticos (aes)
un conjunto de mapeos estéticos entre las variables de nuestro set de datos y las propiedades visuales (color, tamaño, forma, etc.)
3. Capas (layers):
al menos una capa que describe cómo representar cada observación, usualmente creada con una **función geom.** Además cada capa puede tener una **transformación estadística (stat)**, una **posición** y opcionalmente un conjunto de **mapeos estéticos**.

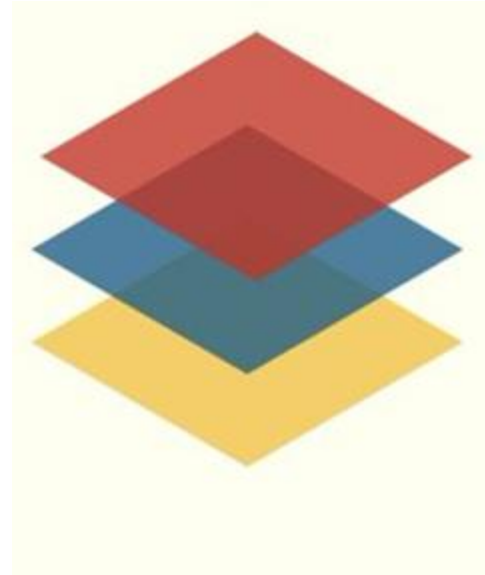
Capas básicas de un gráfico

Gramática de gráficos

Función geom

Atributos estéticos

Datos



Función geom u objetos geométricos

- **geom_bar()**



d + geom_bar()

x, alpha, color, fill, linetype, size, weight

- **geom_boxplot()**



f + geom_boxplot(), x, y, lower, middle, upper, ymax, ymin, alpha, color, fill, group, linetype, shape, size, weight

- **geom_histogram()**



c + geom_histogram(binwidth = 5) x, y, alpha, color, fill, linetype, size, weight

- **geom_line():**



i + geom_line()

x, y, alpha, color, group, linetype, size

- **geom_point():**



e + geom_point(), x, y, alpha, color, fill, shape, size, stroke

- **geom_smooth():**



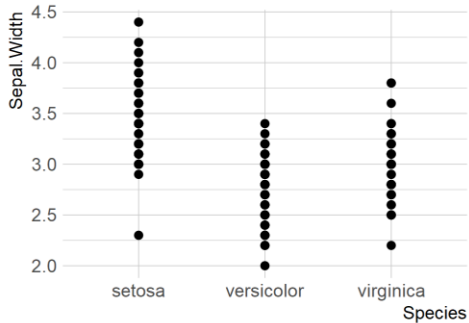
e + geom_smooth(method = lm), x, y, alpha, color, fill, group, linetype, size, weight

Source: *CheatSheet de ggplot2 de Rstudio.*

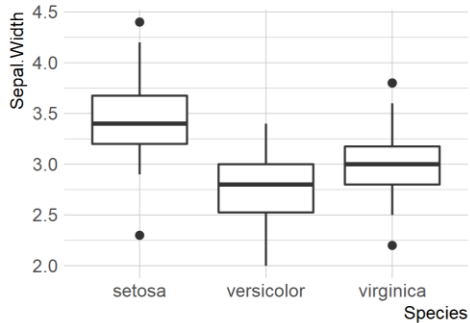
Función geom u objetos geométricos

```
p1 <- ggplot(iris, mapping=aes(Species, Sepal.Width))
```

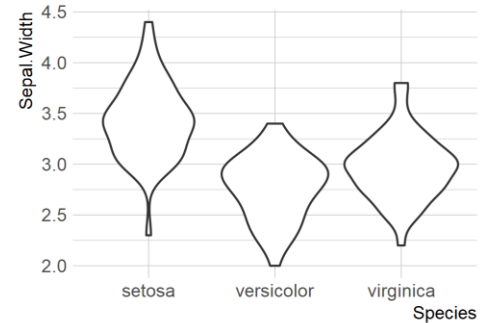
p1 + geom_point()



p1 + geom_boxplot()



p1 + geom_violin()



Otros Componentes

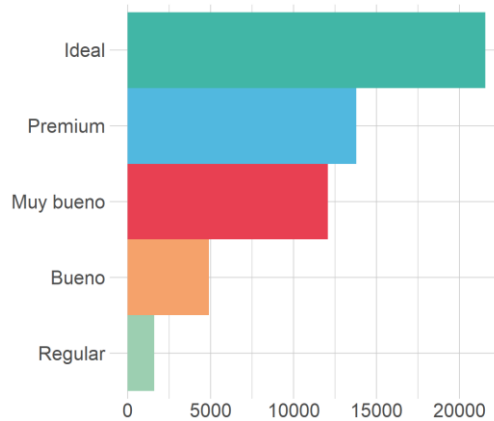
- **Escalas (Scales):** mapea valores en el espacio de los datos a valores en el espacio estético (ej. color, tamaño, forma o posición).
- **Sistema de coordenadas (coord):** por defecto, `ggplot2` utiliza coordenadas cartesianas (`coord._cartesian`).
- **Facetas (facets):** definen cómo se arregla el display cuando son muchos gráficos
- **Temas (themes),** items para mejorar el gráfico como fuente, tamaño, color, background, entre otros.

Source: *R para Ciencia de datos* de Hadley Wickham.

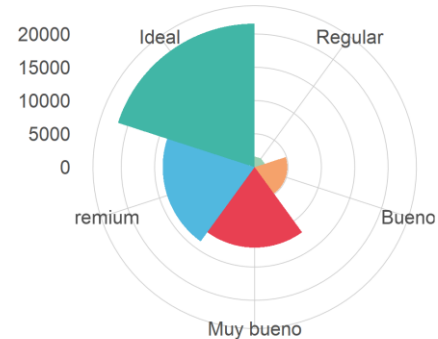


Sistema de coordenadas

■ Cartesianas



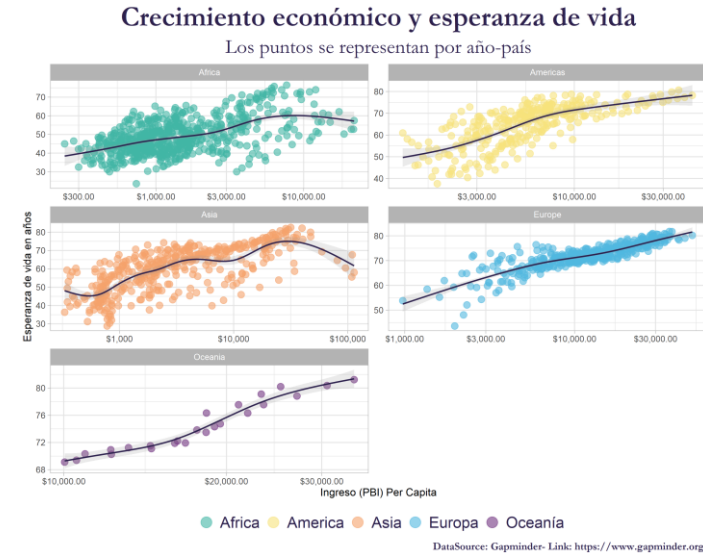
■ Polares



Source: **R4DS** de Hadley Wickham.

Facetas

`facet_wrap`

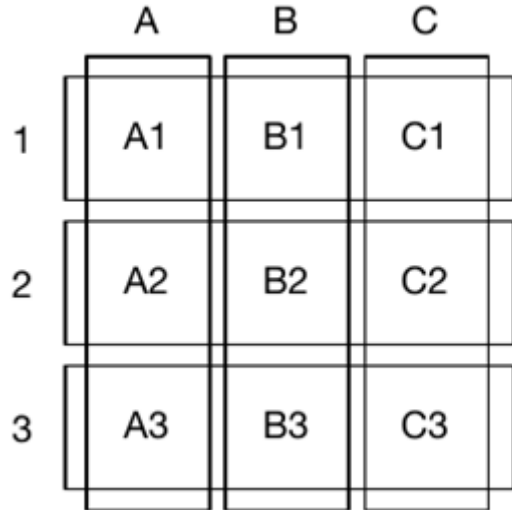


Source: *R4DS* de Hadley Wickham.



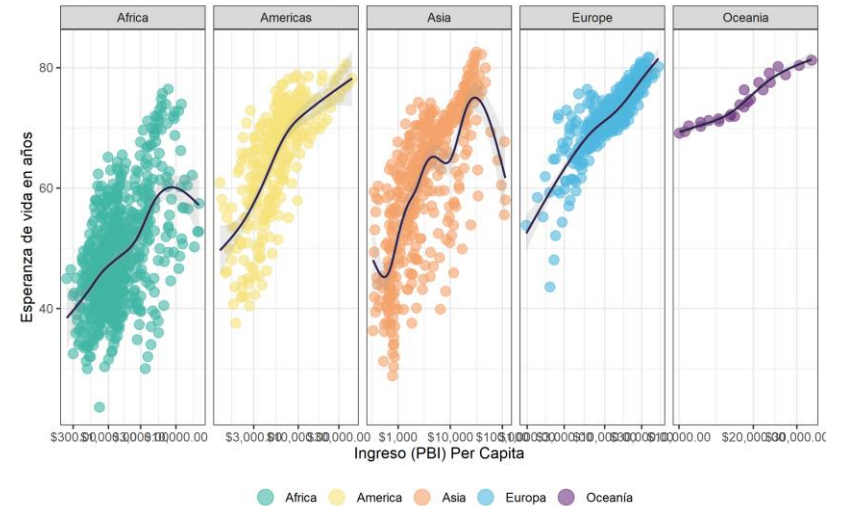
Facetas

facet_grid



Crecimiento económico y esperanza de vida

Los puntos se representan por año-país



DataSource: Gapminder- Link: <https://www.gapminder.org>

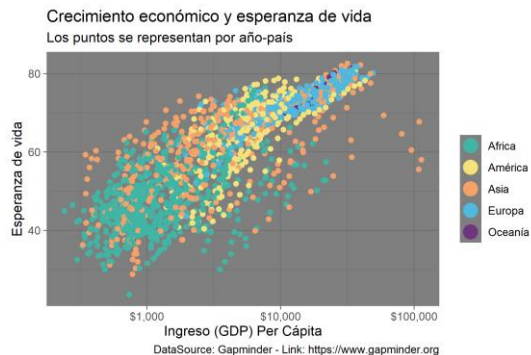


Temas o themes

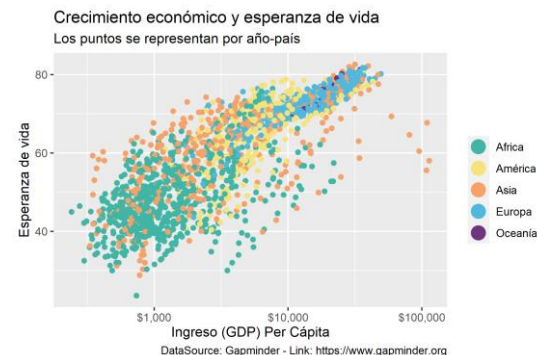
`theme_bw()`



`theme_dark()`



`theme_gray()`



Veamos un ejemplo

Pero antes recordemos la sintaxis



```
P <- ggplot(data= <DATOS>,  
            mapping = aes( <MAPEOS>)) +
```

```
< FUNCTION_GEOM> (  
  mapping = aes( <MAPEOS>),  
  stat= < STAT>,  
  position= <POSITION>) +
```

```
<ESCALAS> +  
<COORDENADAS> +  
<ETIQUETAS>  
<FACETAS> +  
<TEMAS>
```



1 er Paso: datos

1 DATOS ORDENADOS

p ← **ggplot** (data = paises , ...

pib_per_capita	esperanza_de_vida	poblacion	continente
340	65	31	Europa
227	51	200	América
909	81	80	Europa
126	40	20	Asia

2do paso: Mapeos estéticos

3er paso: Función geom

2 MAPEOS

```
p ← ggplot (data = paises ,  
            mapping = aes( x = gdp ,  
                           y = lifexp , size = poblacion ,  
                           color = continente ) )
```

3 FUNCIÓN GEOM



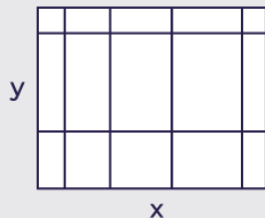
```
p + geom_point( )
```

Imagen original: Data Visualization de Kieran Healy.

4to paso: Coordenadas y escalas

4 SISTEMAS DE COORDENADAS Y ESCALAS

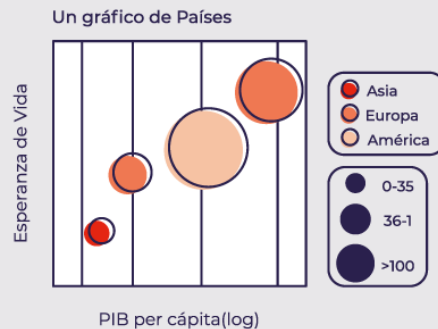
p + `coord_cartesian()` +
`scale_x_log10()`



5to paso: Etiquetas y guías

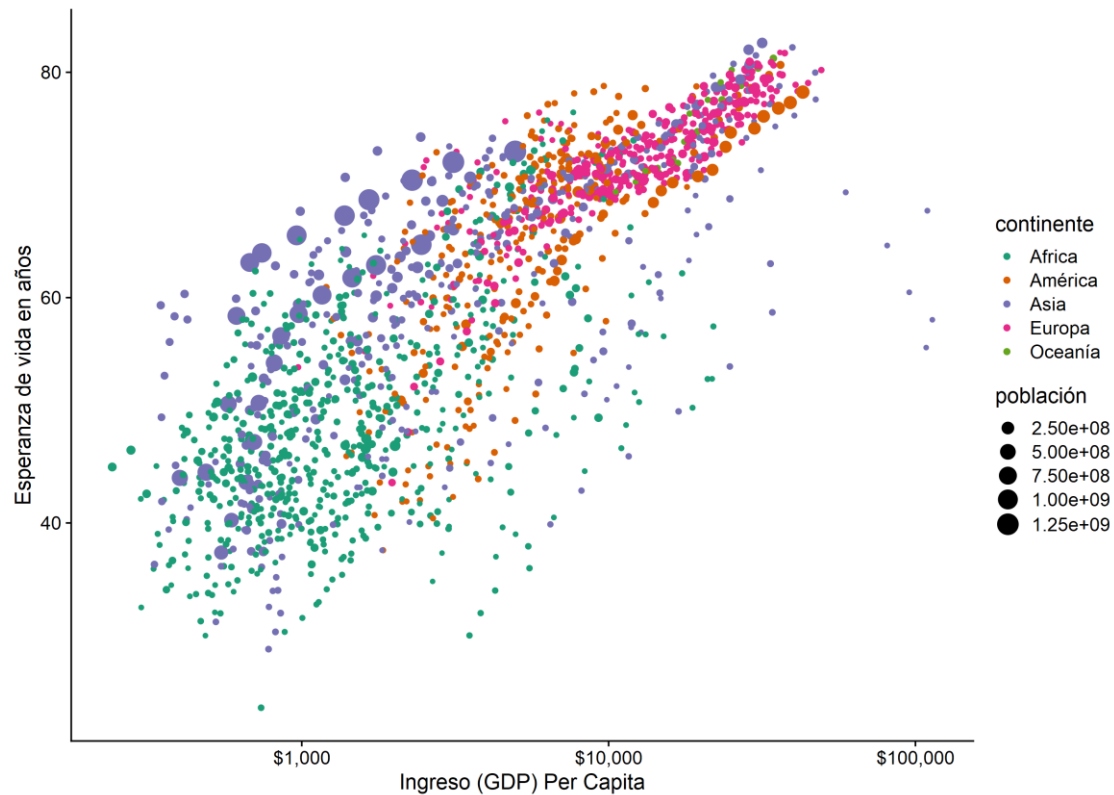
5 ETIQUETAS Y GUÍAS

p + `labs` (x = "PIB per cápita(log)",
Y = "Esperanza de vida",
title = "Un gráfico de Países")



Source: Data Visualization de Kieran Healy.

Crecimiento económico y esperanza de vida



DataSource: Gapminder.

¿Por dónde empiezo?

¿Cómo visualizar y no frustrarse en el intento?

- **Sé perseverante:** la única manera de aprender es practicando y experimentando.
- **Sé paciente contigo y con R.**
- Trabaja de manera **incremental**, comienza por un pequeño gráfico y luego en cada iteración mejóralo.
- No estás sólo, busca una **comunidad** abierta e inclusiva de la que puedas aprender, ejemplo **Rladies** .



¿Cómo visualizar y no frustrarse en el intento?

- Seguí en twitter a gente de la comunidad de R que se dedica a lo que vos querés aprender.

@CedScherer

@jbkunst

@r0mymendez

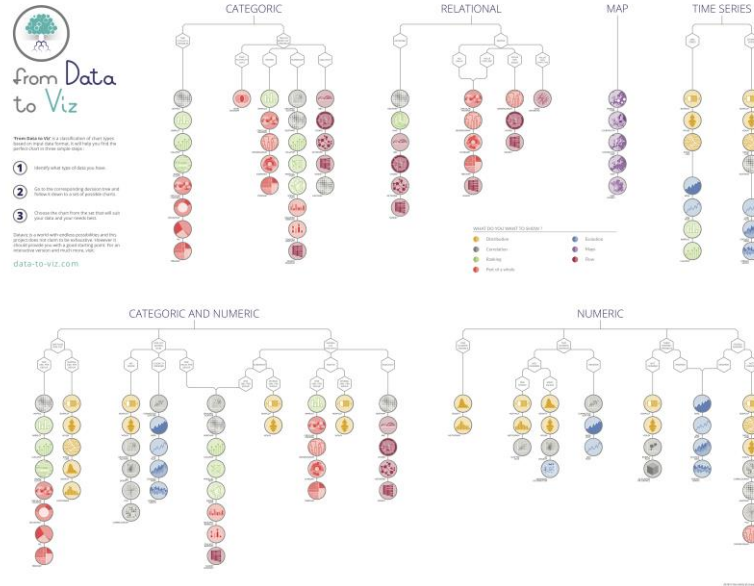
@watzoevery

@committedtotape



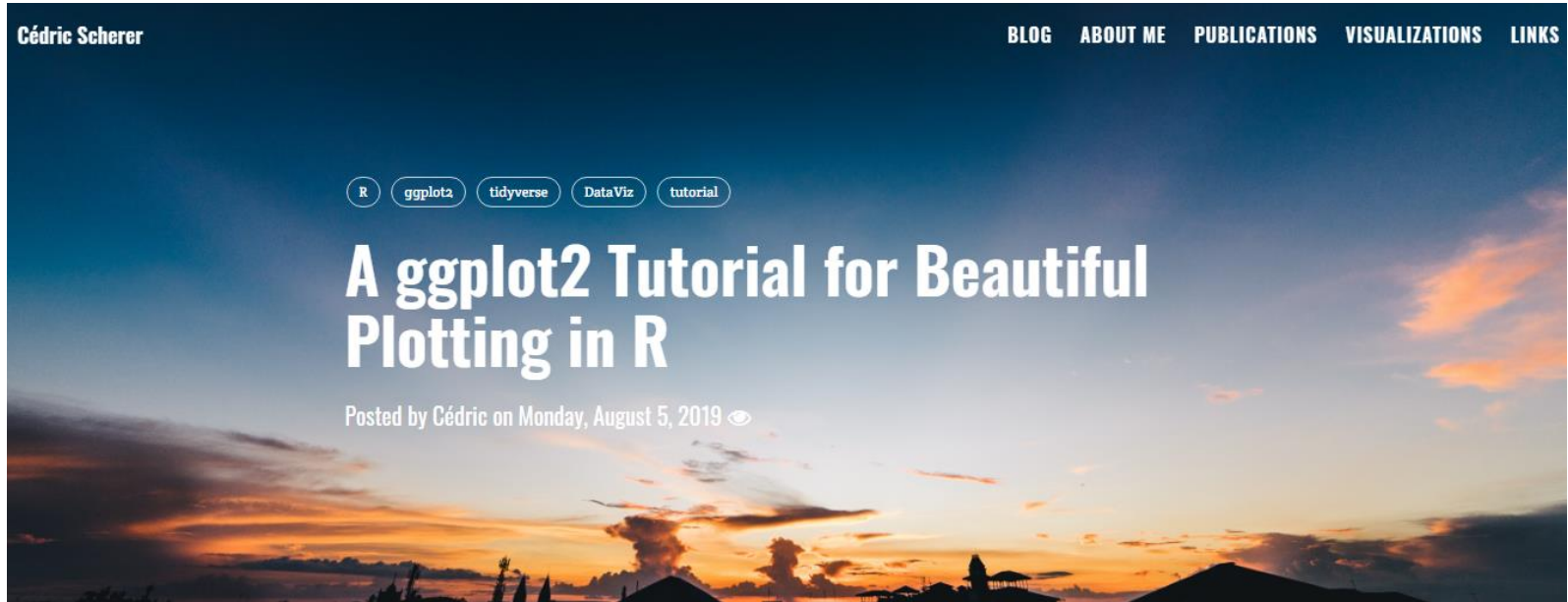
Art by Allison Horst

Blogs y sitios web con mucho para aprender



<https://www.data-to-viz.com>

Blogs y sitios web con mucho para aprender




<https://cedricscherer.netlify.com/>



Blogs y sitios web con mucho para aprender

HIGHCHARTER

Sections ▾  Github

WELCOME

SHOWCASE

API

hchart FUNCTION

SHORTCUTS

THEMES

SHINY

HIGHCHARTS

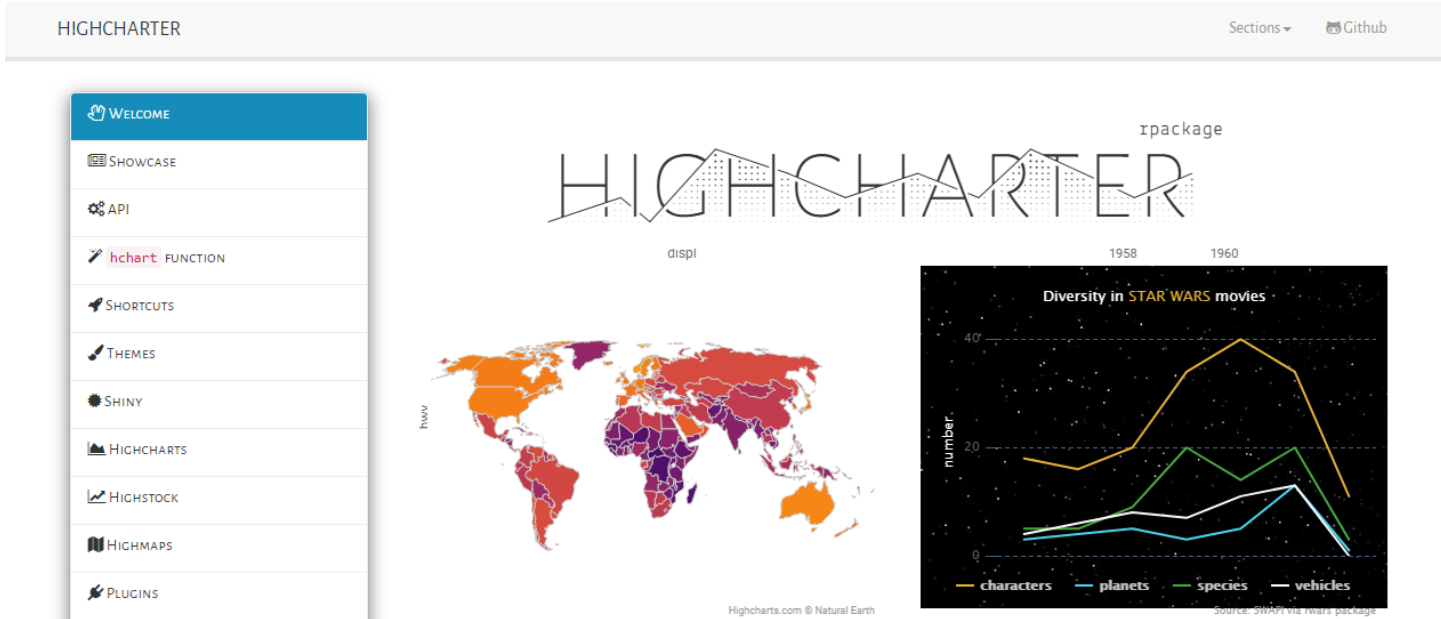
HIGHSTOCK

HIGHMAPS

PLUGINS

displ

1958 1960

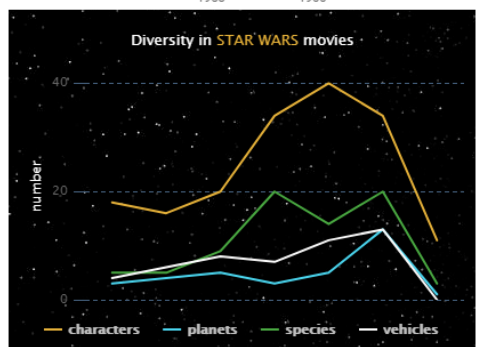


Highcharts.com © Natural Earth

ipackage

HIGHCHARTER

Diversity in STAR WARS movies



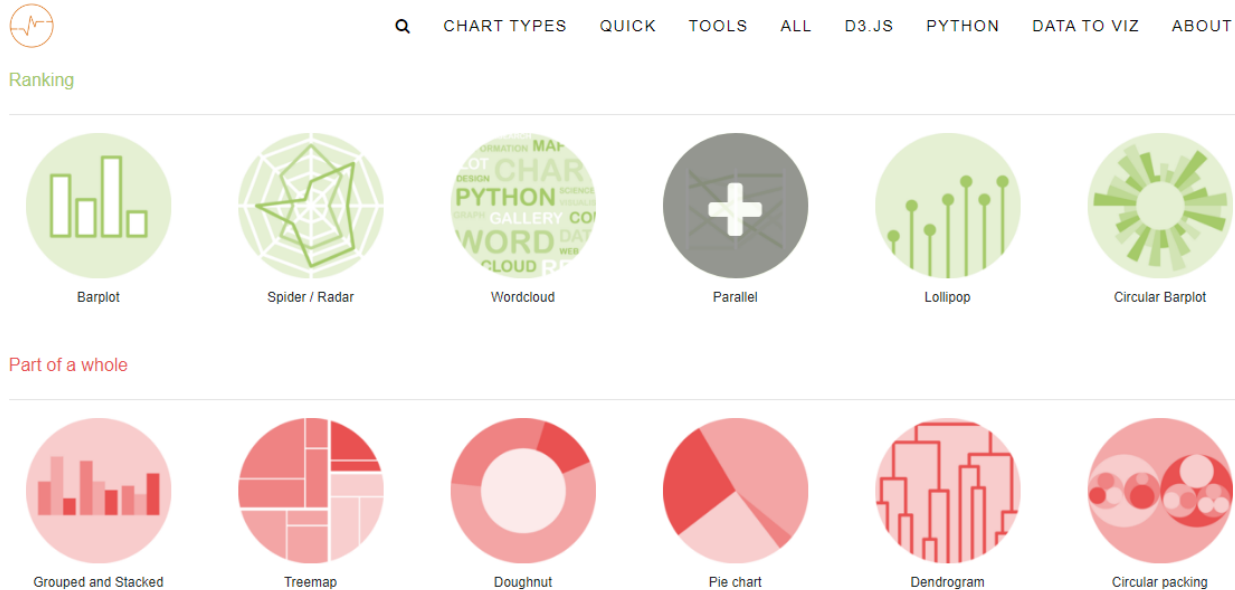
number.

characters planets species vehicles

Source: SHINY via rCharts package

<http://jkunst.com/highcharter/>

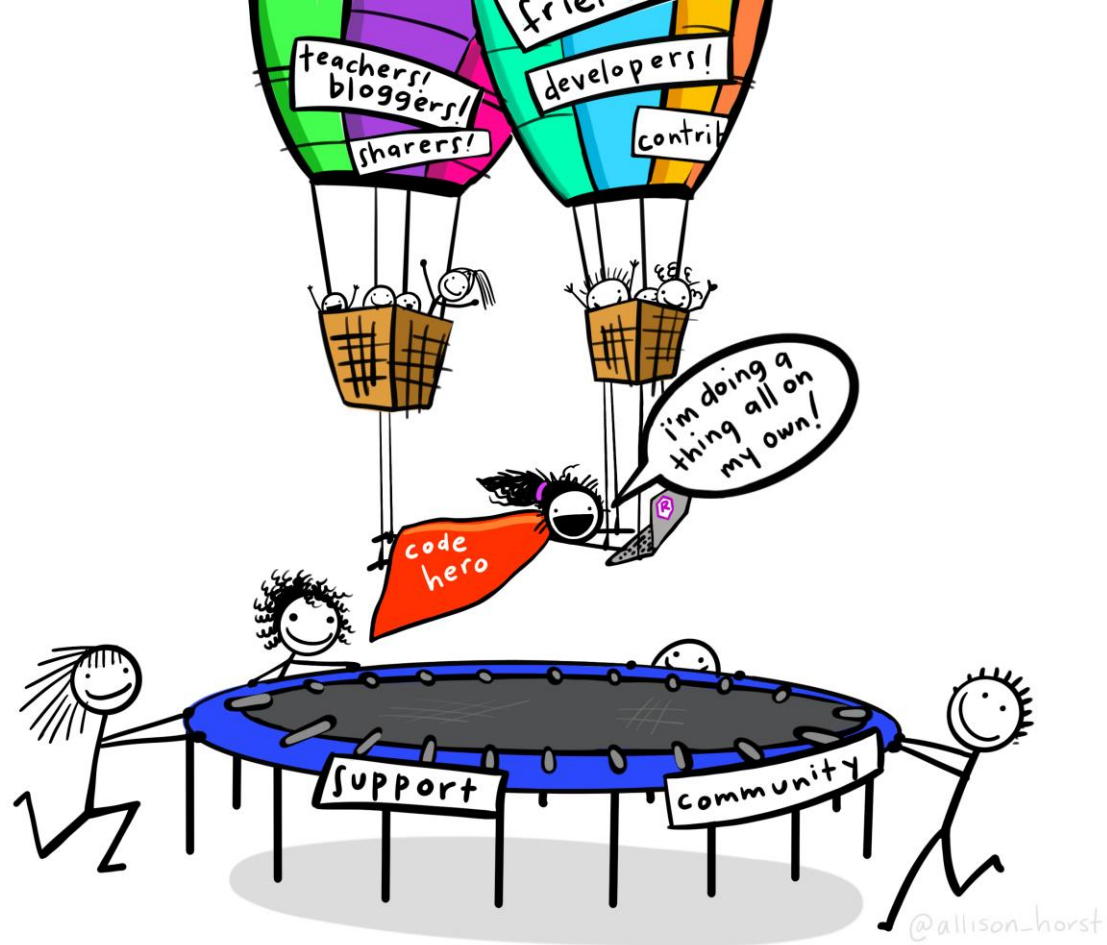
Blogs y sitios web con mucho para aprender



<https://www.r-graph-gallery.com/>



Comunidades



Art by Allison Horst



Comunidad de aprendizaje por proyecto



@R4DS_es

“R for Data Science”
en español



@LatinR_Conf



Comunidad RLadies



Organización global que promueve la diversidad de género en la comunidad de R mediante meetups y mentorías en un espacio amigable y seguro.

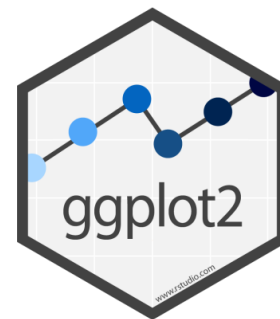
@RLadiesGlobal

@RLadies_rciacte

@RLadiesCba



Tu turno



1. Sigue el tutorial en:
<https://www.rpubs.com/dlee8/ggplot2>
2. En la carpeta Material existen dos archivos rmarkdown de práctica:
 - Graficando barcharts
 - Graficando scatterplots con Gapminder

Referencias:

Libros utilizados para armar el material del curso:

- ***R4DS*** de Hadley Wickham
- ***Data Visualization: A practical introduction*** de Kieran Healy
- ***Fundamentals of Data Visualization*** de Claus Wilke
- ***Data Points, Visualization that Means Something*** de Nathan Yau.
- **Cookbook for R: Practical Recipes for Visualizing Data** de Winston Chang.



¡Gracias!

¿Alguna pregunta?

@dianadiazh

dmdh@uic.edu

