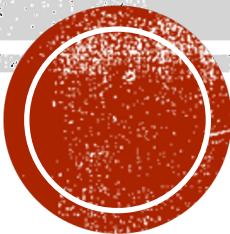




UNIVERSIDAD
COMPLUTENSE
DE MADRID



ORIENTACIONES PARA LA VISUALIZACIÓN EFFECTIVA DE DATOS



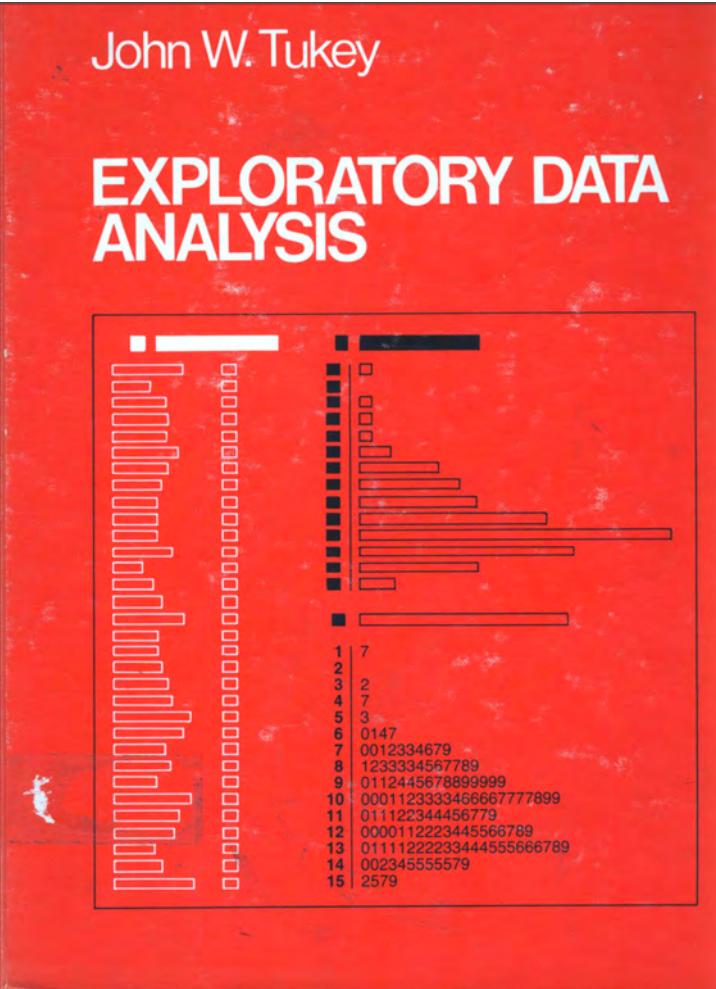
¿EDA?

¿De qué hablamos?



ESTO SERÁ NUEVO, ¿NO?

Preface



This book is based on an important principle:

It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.

Learning first what you can do will help you to work more easily and effectively.

This book is about exploratory data analysis, about looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights. Its concern is with appearance, not with confirmation.

Examples, NOT case histories

The book does not exist to make the case that exploratory data analysis is useful. Rather it exists to expose its readers and users to a considerable variety of techniques for looking more effectively at one's data. The examples are not intended to be complete case histories. Rather they show isolated techniques in action on real data. The emphasis is on general techniques, rather than specific problems.

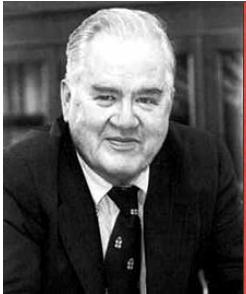
**It is important to understand what you CAN DO
before you learn to measure how WELL you seem to
have DONE it.**

John W. Tukey 1976

In particular:

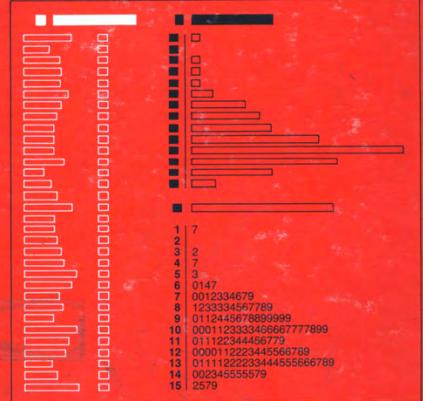
- ❖ to be able to say that we looked one layer deeper, and found nothing, is a definite step forward--though not as far as to be able to say that we looked deeper and found thus-and-such.
- ❖ to be able to say that "if we change our point of view in the following way . . . things are simpler" is always a gain--though not quite as much as to be able to say "if we don't bother to change our point of view (some other) things are equally simple".

ESTO SERÁ NUEVO, ¿NO?



John W. Tukey

EXPLORATORY DATA ANALYSIS

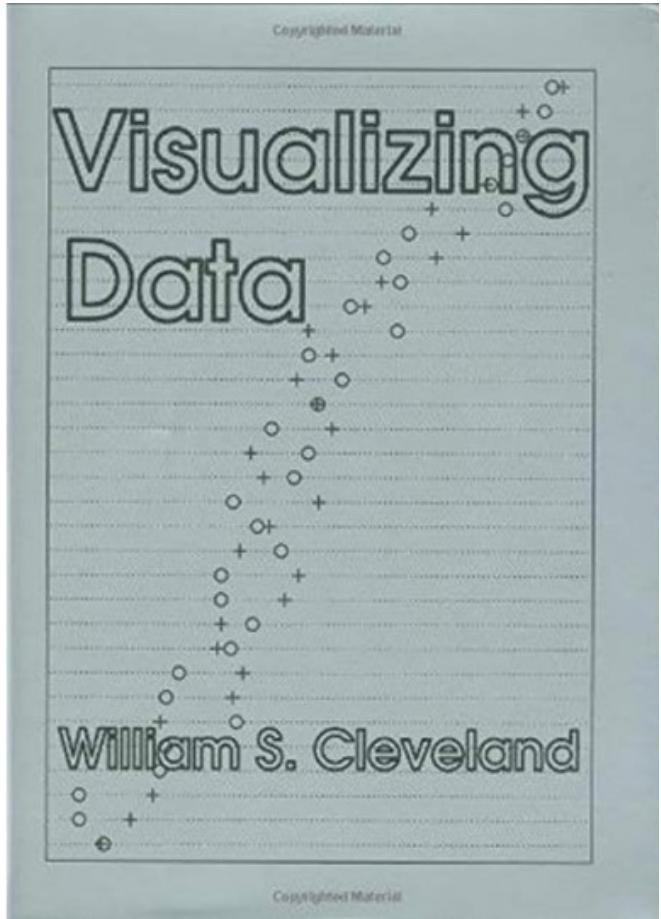


“‘Exploratory data analysis’ is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there.”

— John Tukey, Author of Exploratory Data Analysis

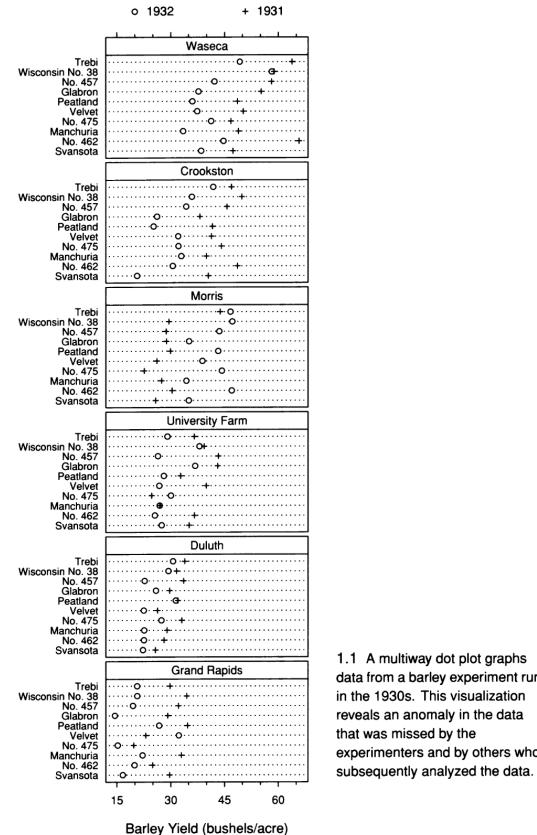


A HOMBROS DE GIGANTES



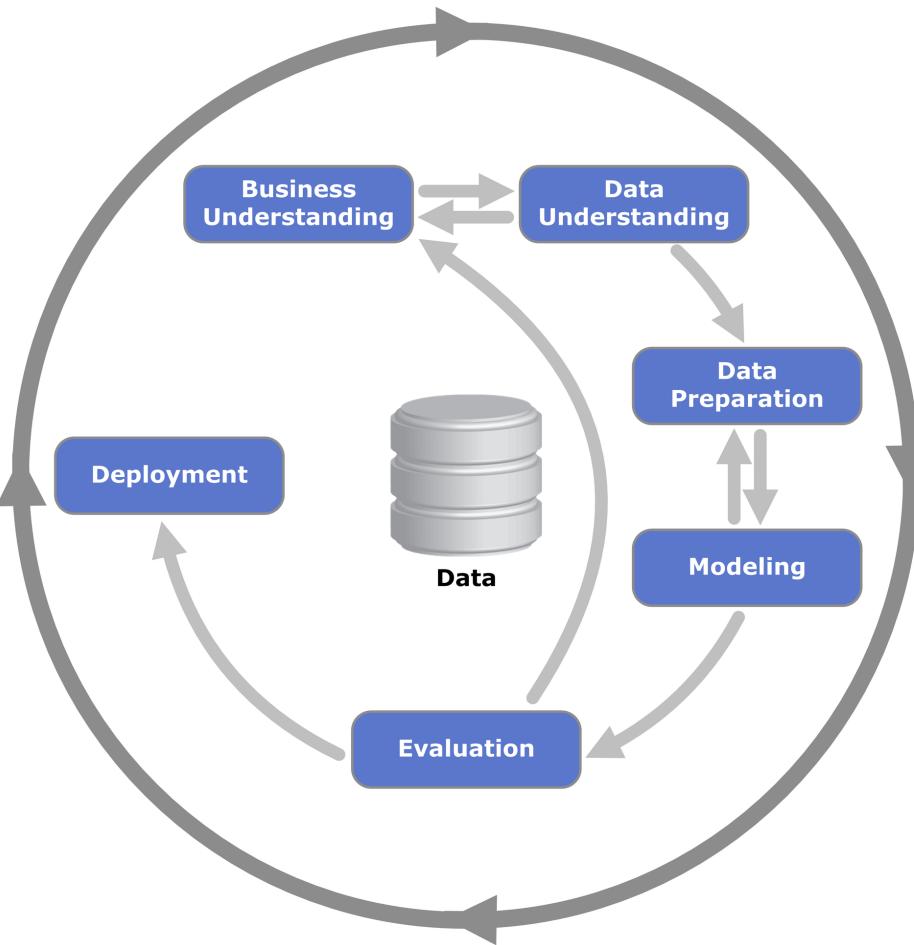
1993

Visualization is critical to data analysis. It provides a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way. We discover unimagined effects, and we challenge imagined ones.



1.1 A multiway dot plot graphs data from a barley experiment run in the 1930s. This visualization reveals an anomaly in the data that was missed by the experimenters and by others who subsequently analyzed the data.

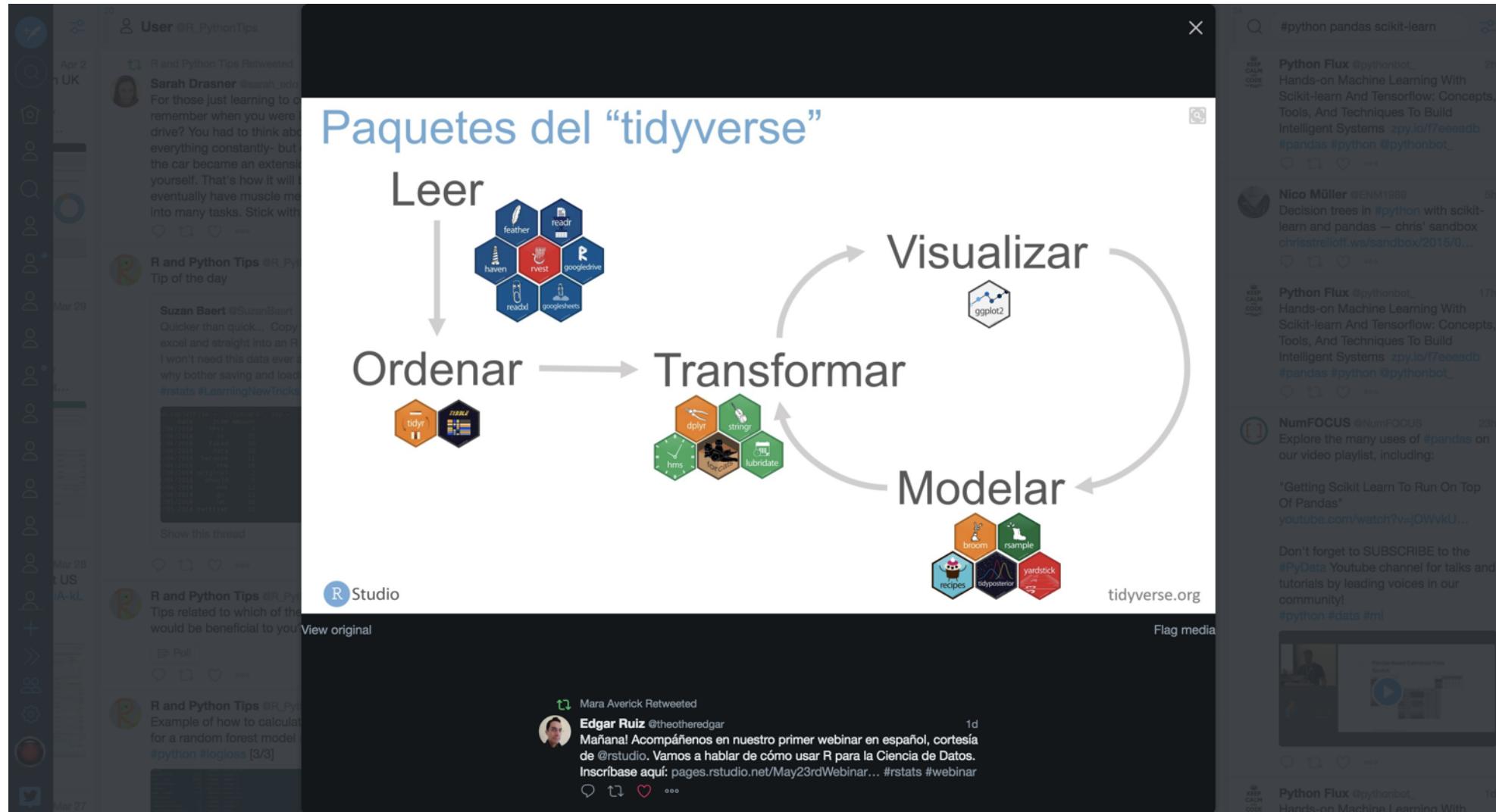
¿EDA? ¿ESTO QUÉ ES? – CICLO METODOLÓGICO (1 DE 3)



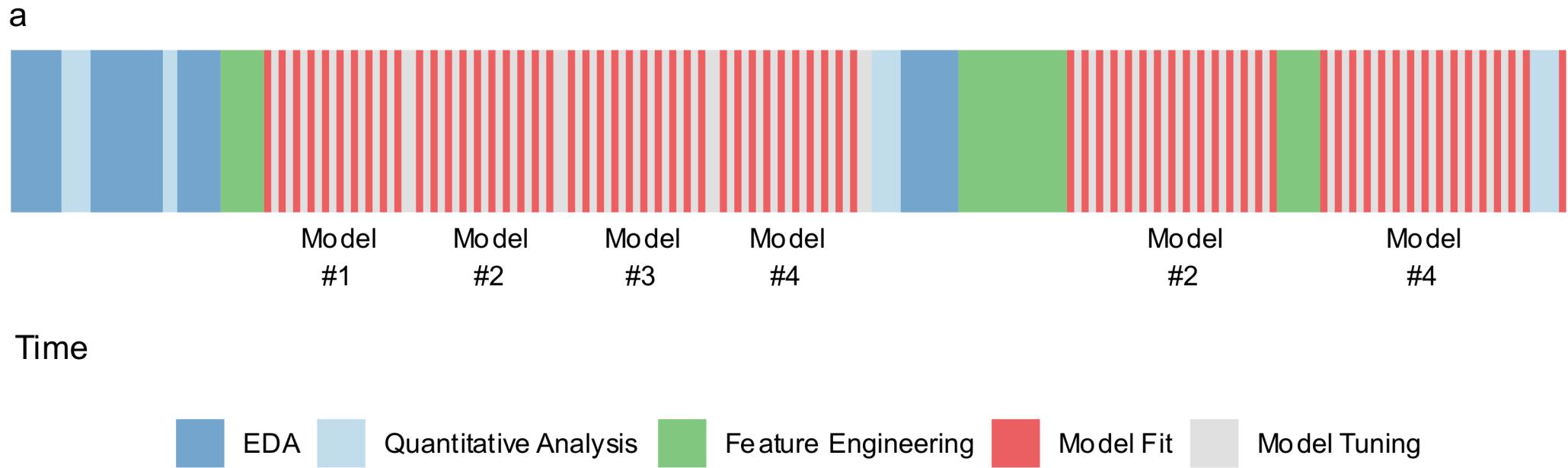
https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining



¿EDA? ¿ESTO QUÉ ES? – UN POCO MÁS CLARO... (2 DE 3)



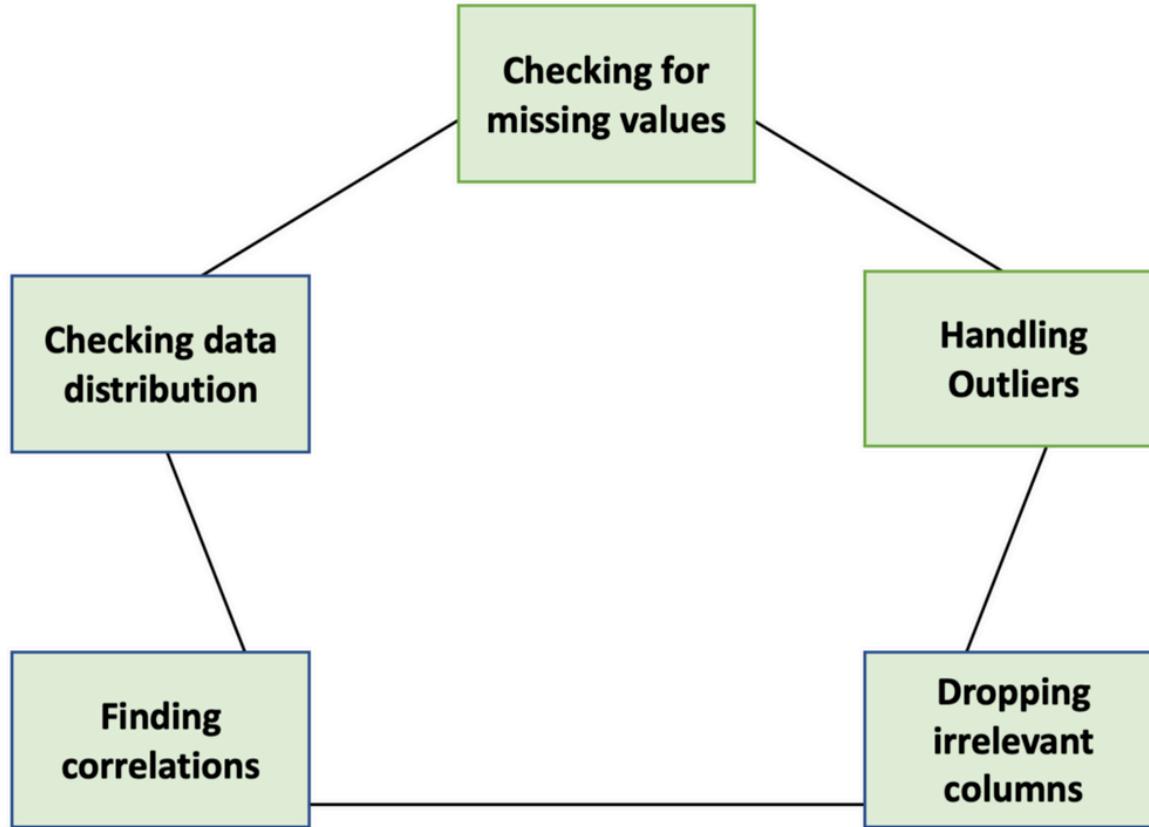
¿EDA? ¿ESTO QUÉ ES?... A MÍ ME GUSTA ESTE ... (3 DE 3)



Feature Engineering (Max Kuhn and Kjell Johnson - 2019-06-21)
<http://www.feat.engineering/index.html>



EN LA FASE DE EDA, ¿QUÉ VAMOS A COMPROBAR?



Ejemplo - R DataExplorer



DATAEXPLORER UNA LIBRERÍA DE R PARA EDA

Introduction to DataExplorer

Boxuan Cui

2019-03-17

- Data
- Exploratory Data Analysis
 - Missing values
 - Distributions
 - Bar Charts
 - Histograms
 - QQ Plot
 - Correlation Analysis
 - Principal Component Analysis
 - Slicing & dicing
 - Boxplots
 - Scatterplots
- Feature Engineering
 - Replace missing values
 - Group sparse categories
 - Dummify data (one hot encoding)
 - Drop features
 - Update features
- Data Reporting

This document introduces the package [DataExplorer](#), and shows how it can help you with different tasks throughout your data exploration process.

There are 3 main goals for **DataExplorer**:

1. [Exploratory Data Analysis \(EDA\)](#)
2. [Feature Engineering](#)
3. Data Reporting

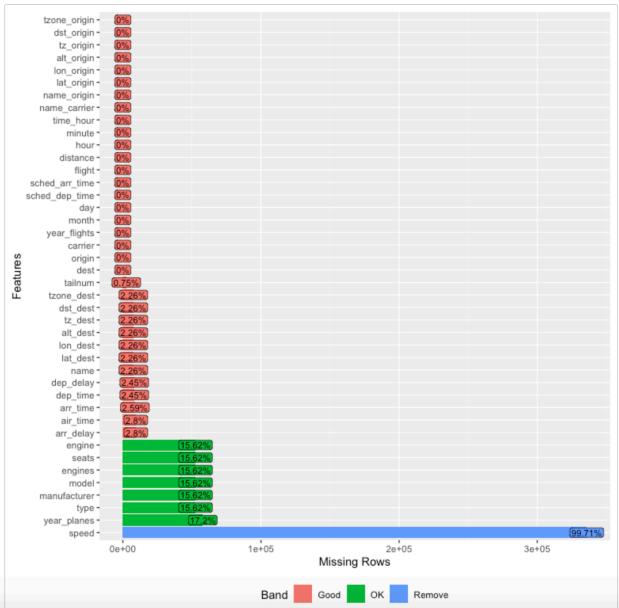
The remaining of this guide will be organized in accordance with the goals. As the package evolves, more content will be added.

<https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html>



DATAEXPLORER UNA LIBRERÍA DE R PARA EXP

plot_missing(final_data)

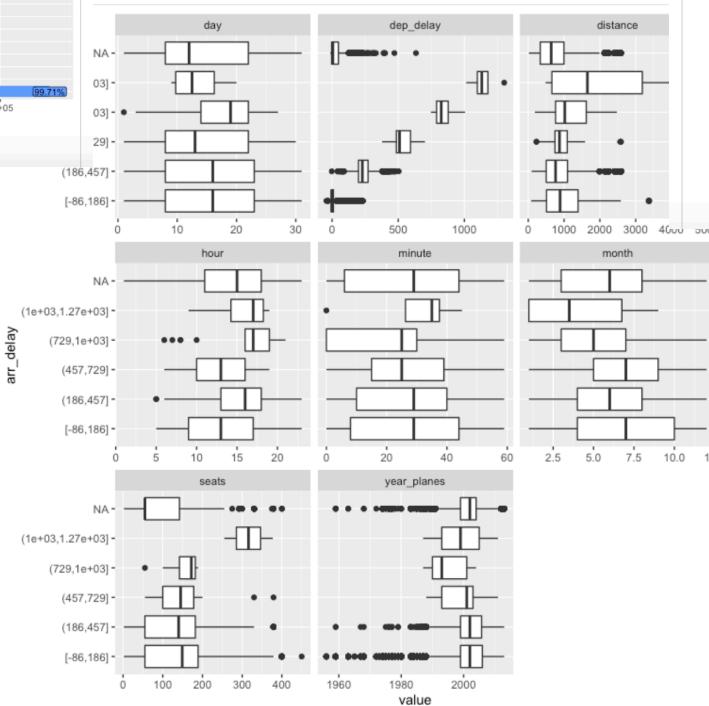


a size for demo purpose

<- final_data[, c("arr_delay", "month", "day", "hour", "minute", "dep_delay", "distanc

ot function

arr_delay_df, by = "arr_delay")



Among all the subtle changes in correlation with arrival delays, you could immediately spot that planes with 300+ seats tend to have much longer delays (16 ~ 21 hours). You may now drill down further to verify or generate more hypotheses.

```
## 5 columns ignored with more than 50 categories.
## dest: 105 categories
## tailnum: 4044 categories
## time_hour: 6936 categories
## model: 128 categories
## name: 102 categories
```

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

name: 102 categories

5 columns ignored with more than 50 categories.

dest: 105 categories

tailnum: 4044 categories

time_hour: 6936 categories

model: 128 categories

HAY OTRAS...

DITCH THE OLD WAYS. THE FUTURE IS AUTOMATED.

Modern Exploratory Data Analysis

Review of 4 libraries for automatic EDA



Karim Lahrichi [Follow](#)

May 2 · 6 min read ★

[...](#)



<https://towardsdatascience.com/modern-exploratory-data-analysis-29fdbcecc957>

Ejemplo - Python Pandas-profiling



PANDAS-PROFILING UNA LIBRERÍA DE PYTHON PARA EDA



Overview

Overview Reproduction Warnings 6

Dataset statistics

Number of variables	14
Number of observations	32561
Missing cells	4262
Missing cells (%)	0.9%
Duplicate rows	25
Duplicate rows (%)	0.1%
Total size in memory	18.1 MiB
Average record size in memory	583.0 B

Variable types

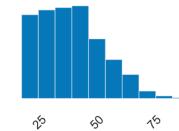
CAT	8
NUM	6

Variables

age

Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	73
Unique (%)	0.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Mean	38.58164675532078
Minimum	17
Maximum	90
Zeros	0
Zeros (%)	0.0%



https://pandas-profiling.github.io/pandas-profiling/examples/census/census_report.html



PANDAS-PROFILING UNA LIBRERÍA DE PYTHON PARA EDA

Acceleration											
Numeric											
Distinct count	87	Mean	64.614								
Unique (%)	0.5%	Minimum	12								
Missing (%)	0.3%	Maximum	97								
Missing (n)	48	Zeros (%)	0.0%								
Infinite (%)	0.0%										
Infinite (n)	0										
Toggle details											
Age											
Numeric											
Distinct count	29	Mean	25.122								
Unique (%)	0.2%	Minimum	16								
Missing (%)	0.0%	Maximum	45								
Missing (n)	0	Zeros (%)	0.0%								
Infinite (%)	0.0%										
Infinite (n)	0										
Toggle details											
Aggression											
Numeric											
Distinct count	86	Mean	55.869								
Unique (%)	0.5%	Minimum	11								
Missing (%)	0.3%	Maximum	95								
Missing (n)	48	Zeros (%)	0.0%								
Infinite (%)	0.0%										
Infinite (n)	0										
Toggle details											
Agility											
Numeric											
Distinct count	82	Mean	63.504								
Unique (%)	0.5%	Minimum	14								
Missing (%)	0.3%	Maximum	96								
Missing (n)	48	Zeros (%)	0.0%								
Infinite (%)	0.0%										
Infinite (n)	0										
Toggle details											

PANDAS-PROFILING PERO HAY NUEVAS..."DATAPREP.EDA"

Dataprep.eda: Accelerate your EDA

Everything you need to know about dataprep.eda.

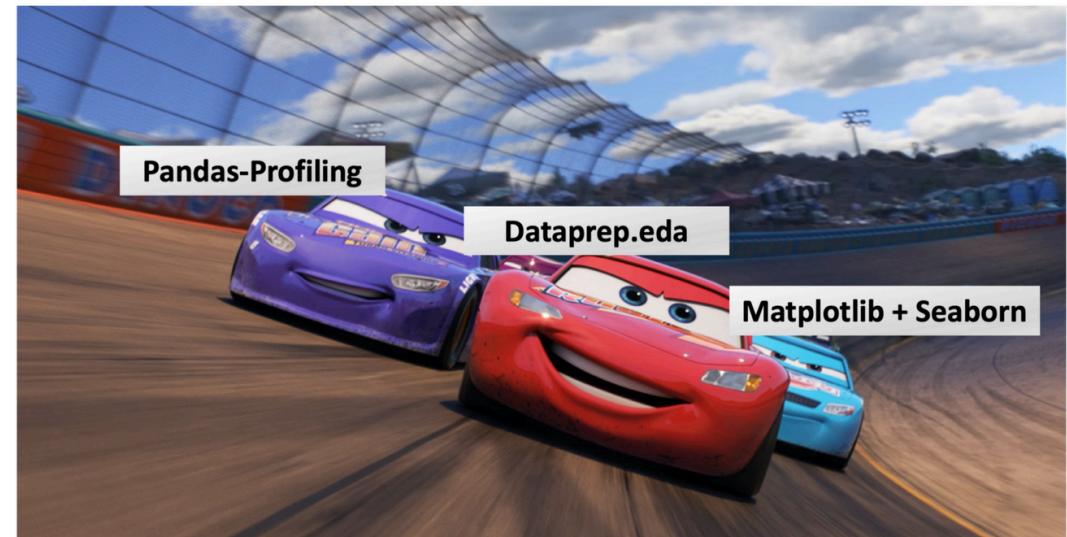


Slavy Coelho Following
Apr 14 · 7 min read ★

...



source: MicroStockHub, via: [Getty Images/iStockphoto](#)

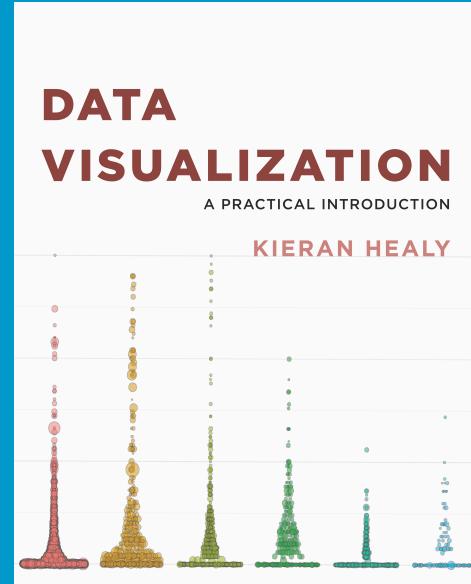


<https://towardsdatascience.com/dataprep-eda-accelerate-your-eda-eb845a4088bc>

¿Qué estilo de Gráficos ?

Una guía de estilo

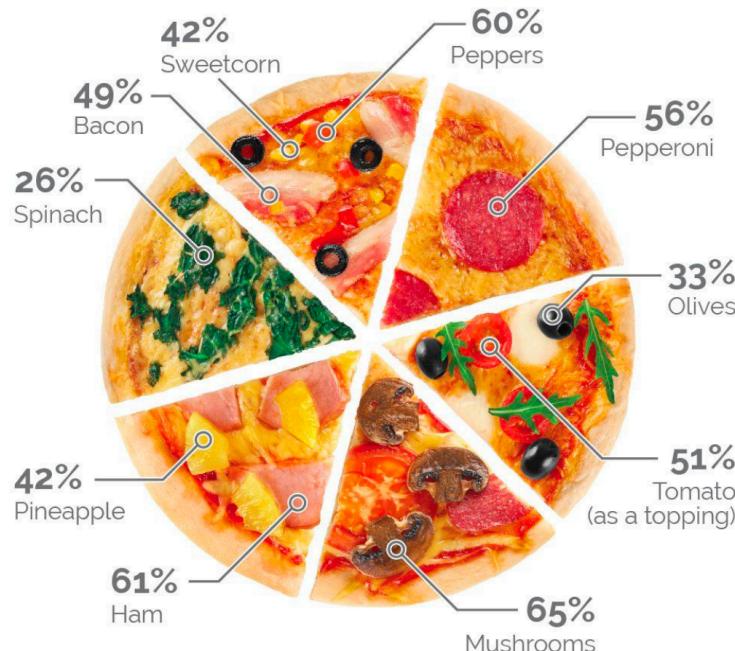
<https://socviz.co/index.html>



“¡ UN POCO DE POR FAVOR !”

Mushroom is the UK's most liked pizza topping

Generally speaking, which of the following toppings do you like on a pizza? Select as many as you like



Other items not depicted include: onions (62%), chicken (56%), beef (36%), chillies (31%), jalapeños (30%), pork (25%), tuna (22%), anchovies (18%). 2% of people say they only like Margherita pizzas

YouGov | yougov.com

February 26-28, 2017

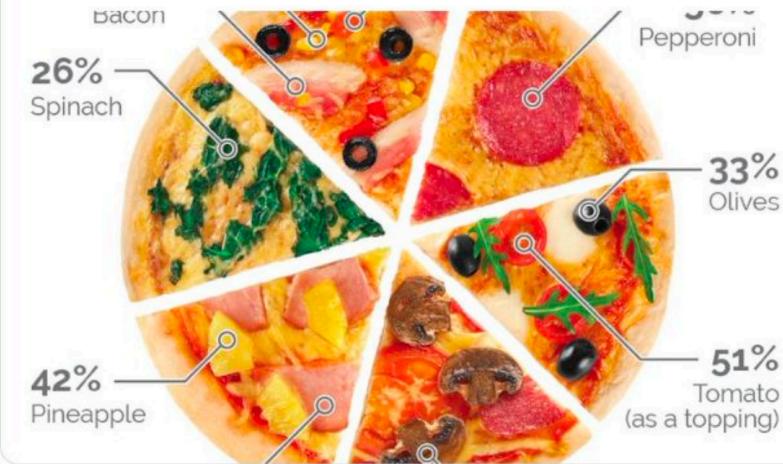


YouGov
@YouGov

We're very sorry for the confusion, but this is NOT a pie chart - it is just a top-down photo of a pizza with some topping stats pointed out

YouGov @YouGov · Mar 6, 2017

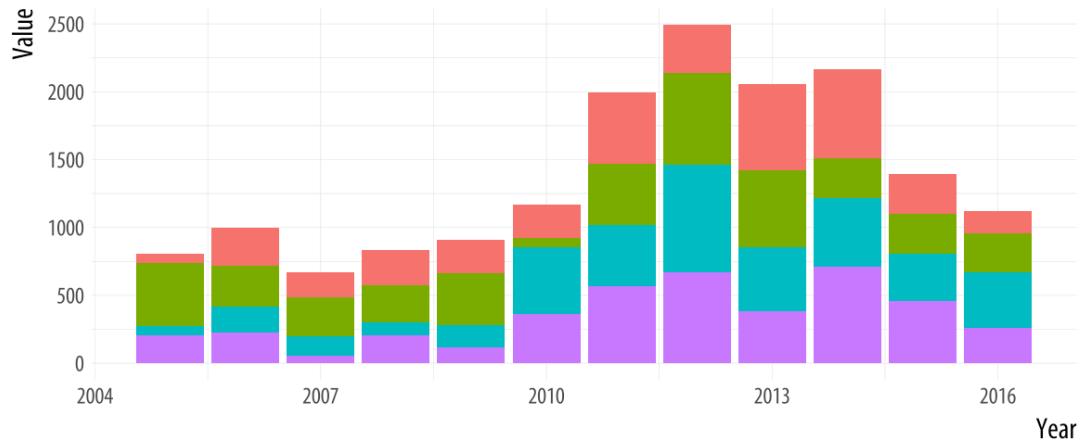
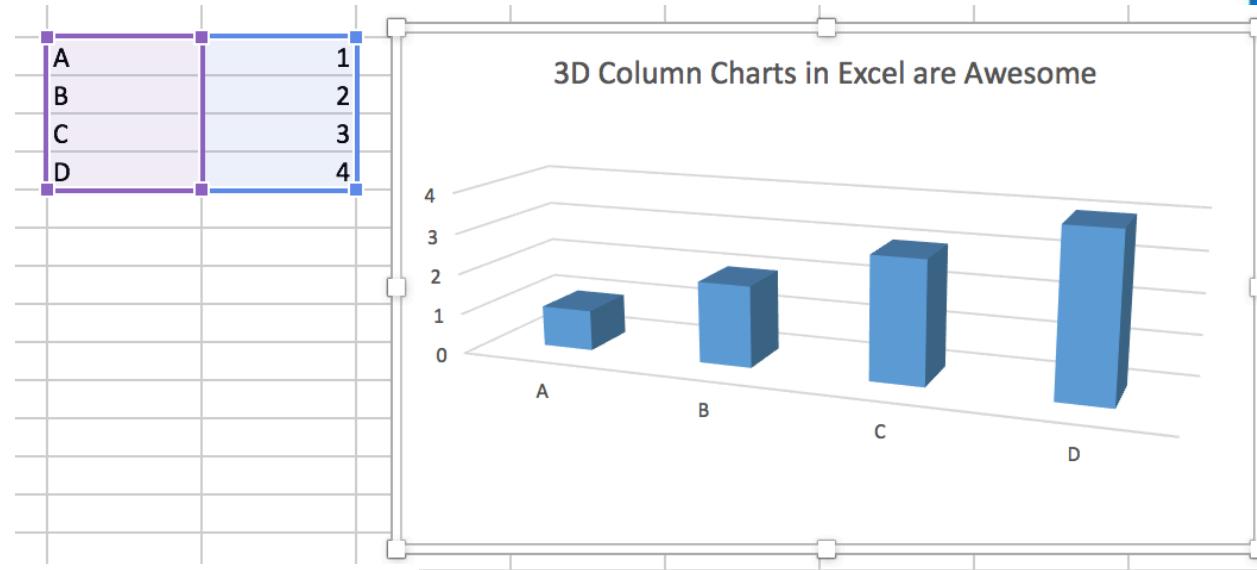
Forget pepperoni - mushroom is Britain's most liked pizza topping (65%), followed by onion (62%) and then ham (61%) yougov.co.uk/news/2017/03/0...



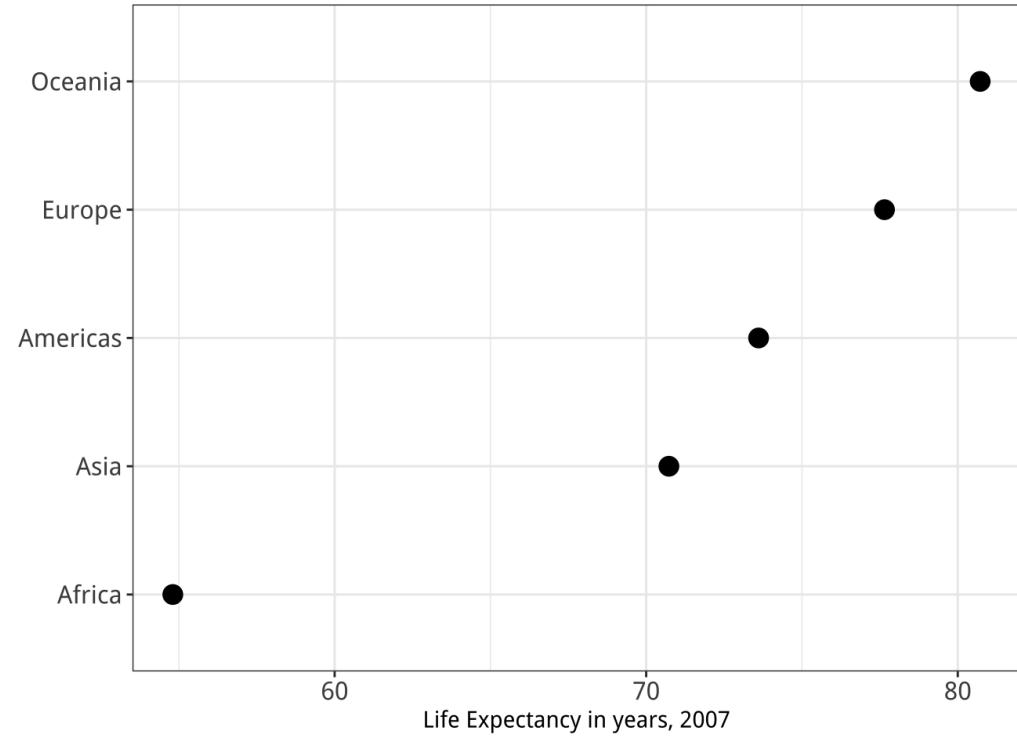
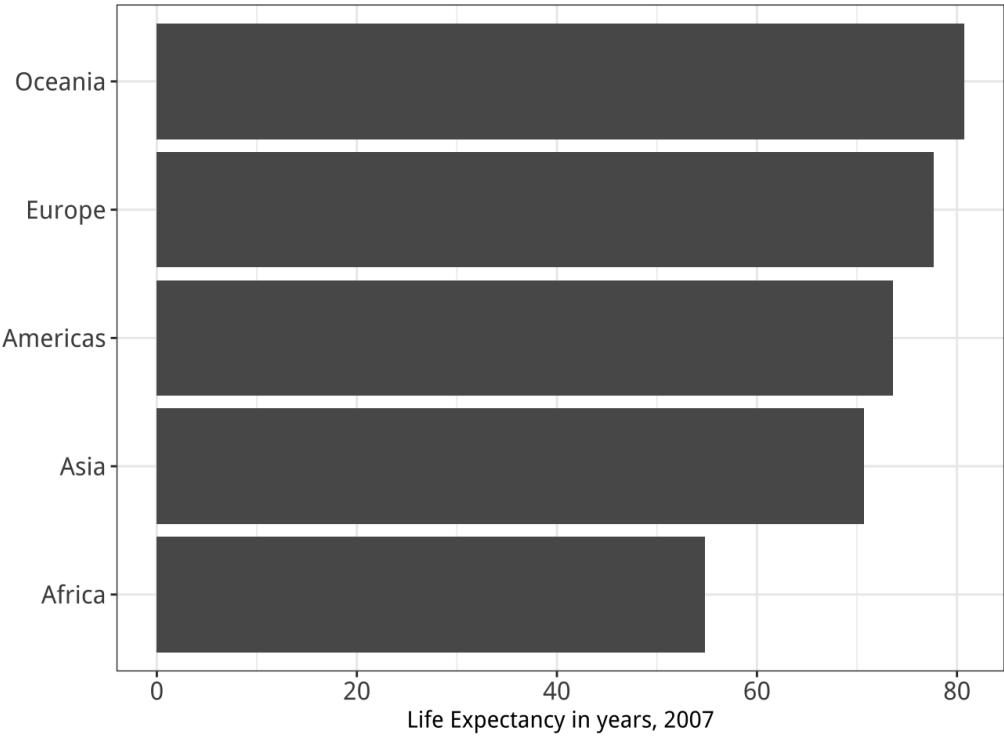
2:56 PM · Mar 6, 2017 · Twitter Web Client



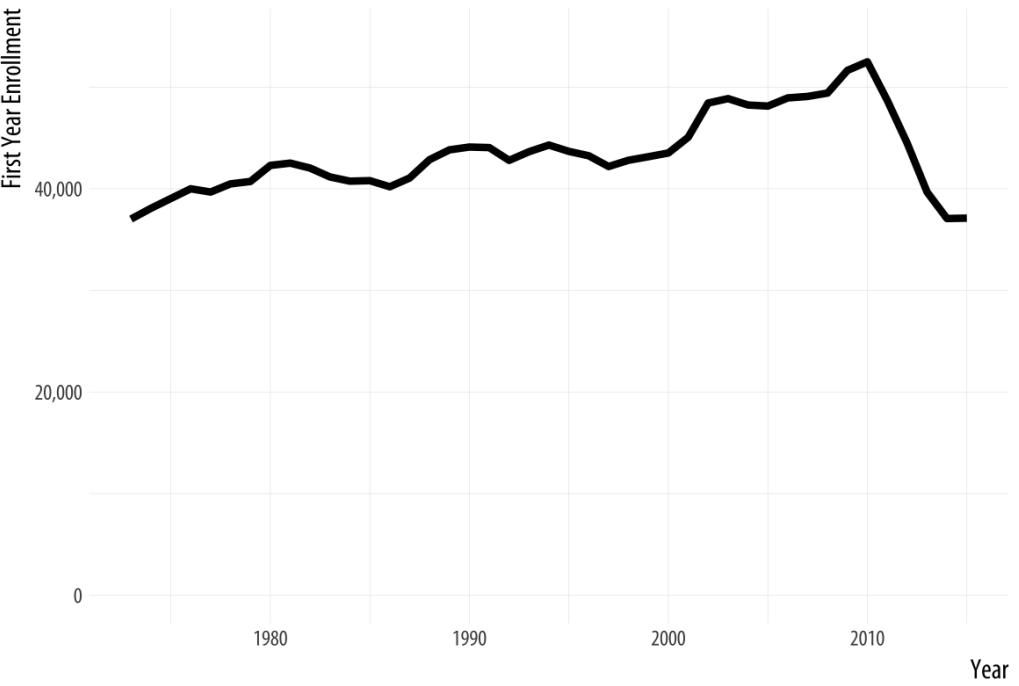
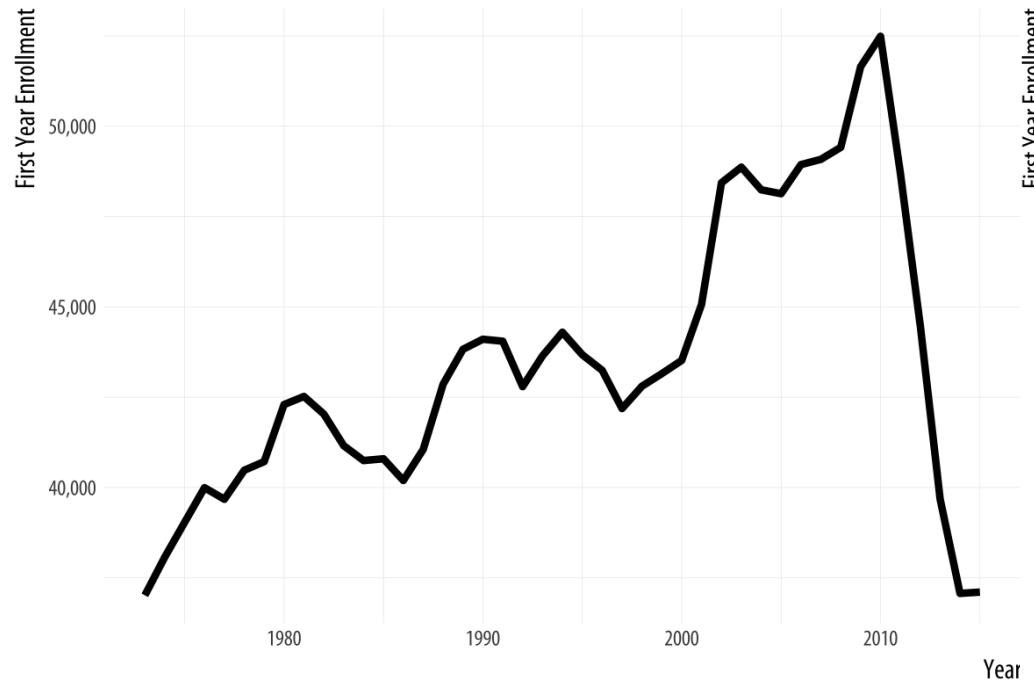
“¡ UN POCO DE POR FAVOR !”



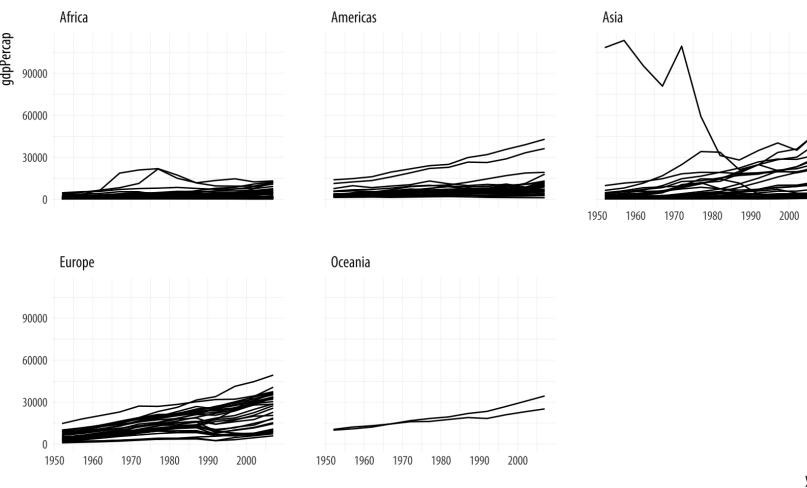
ATENCIÓN AL CERO (1 DE 2)



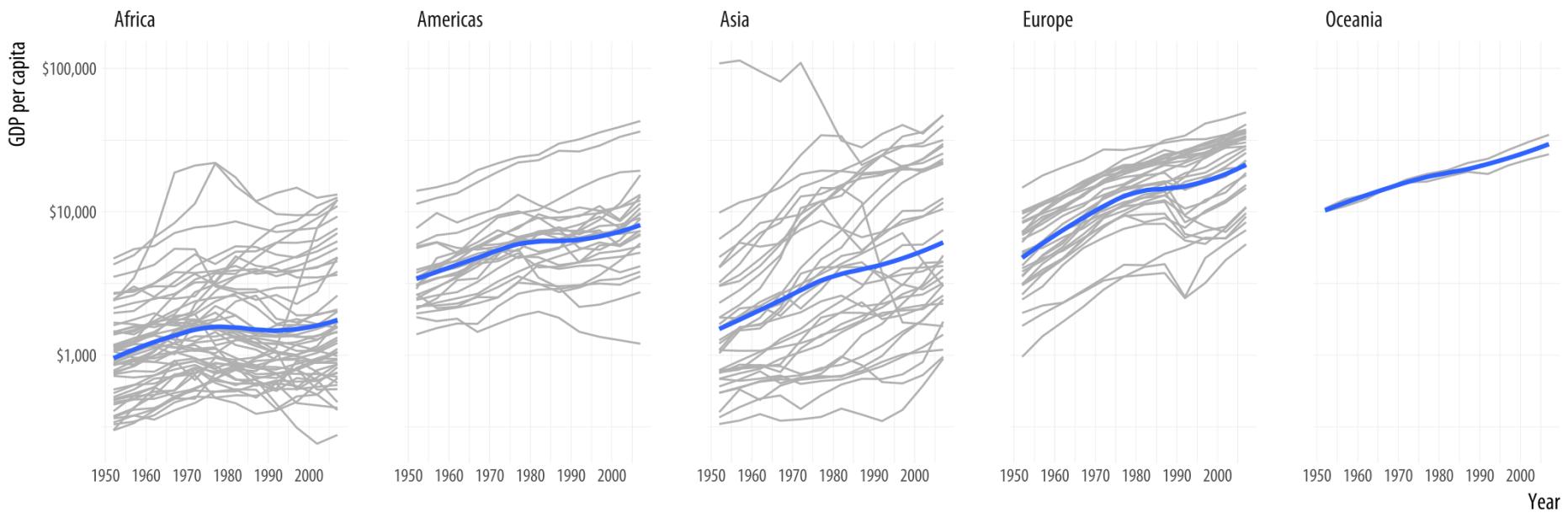
ATENCIÓN AL CERO (2 DE 2)



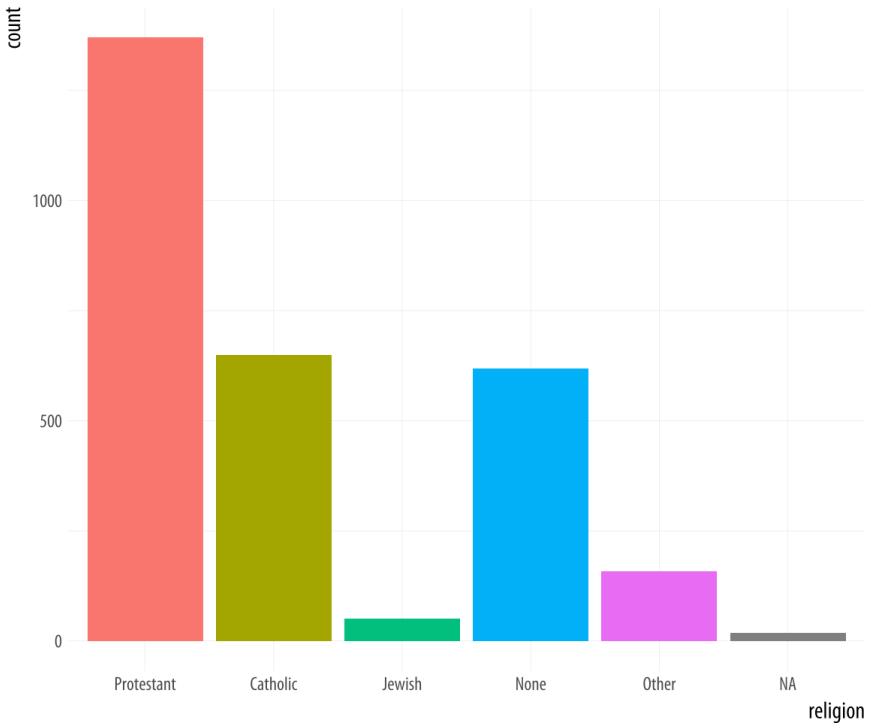
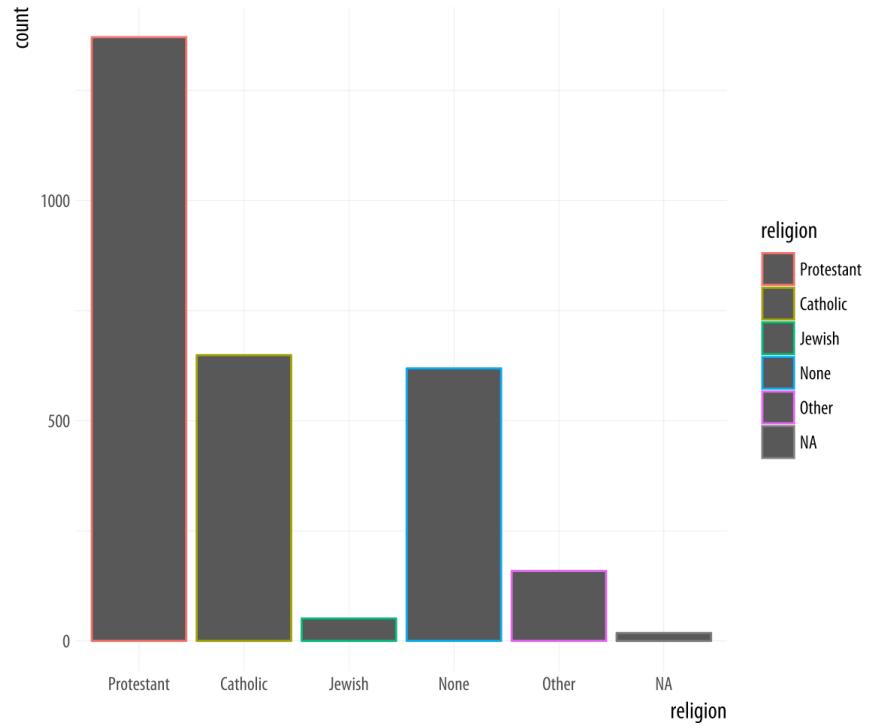
GRÁFICOS MÚLTIPLES



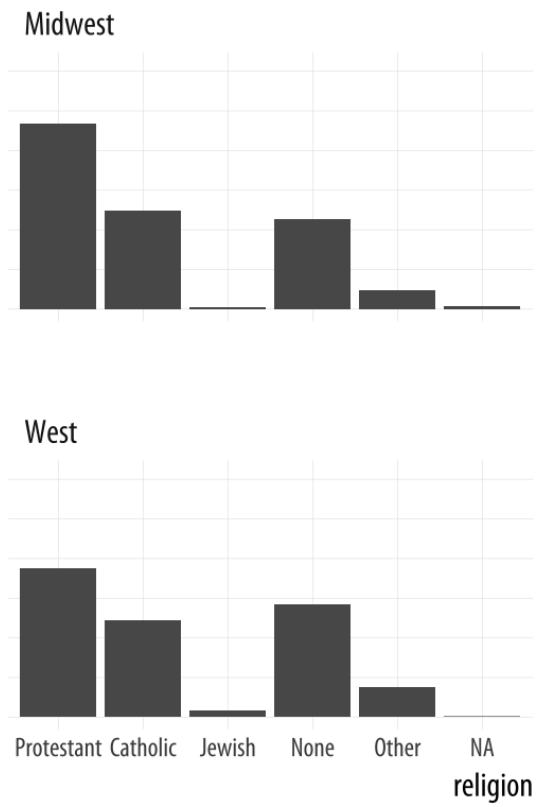
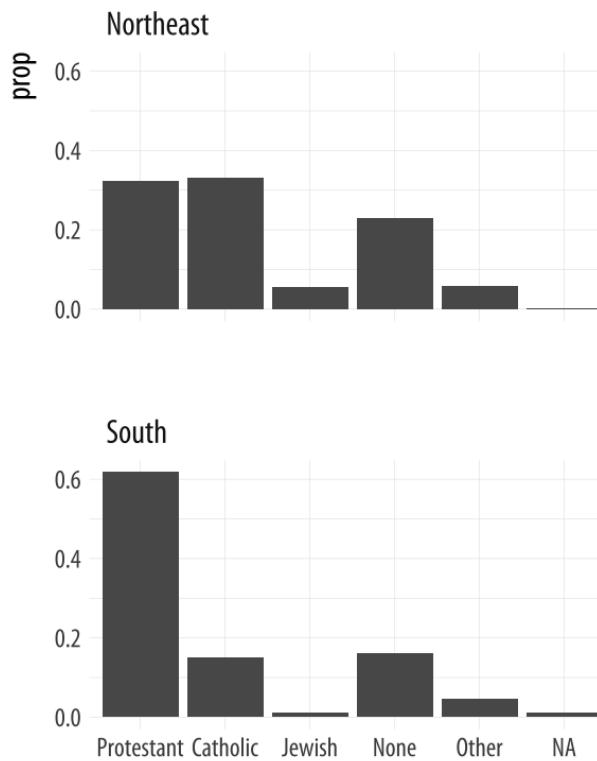
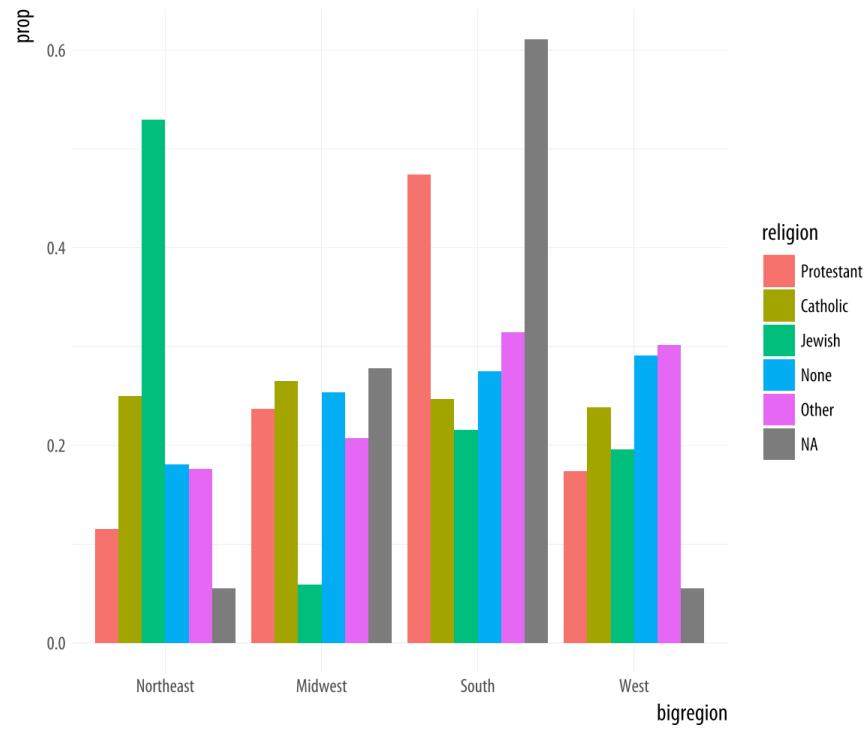
GDP per capita on Five Continents



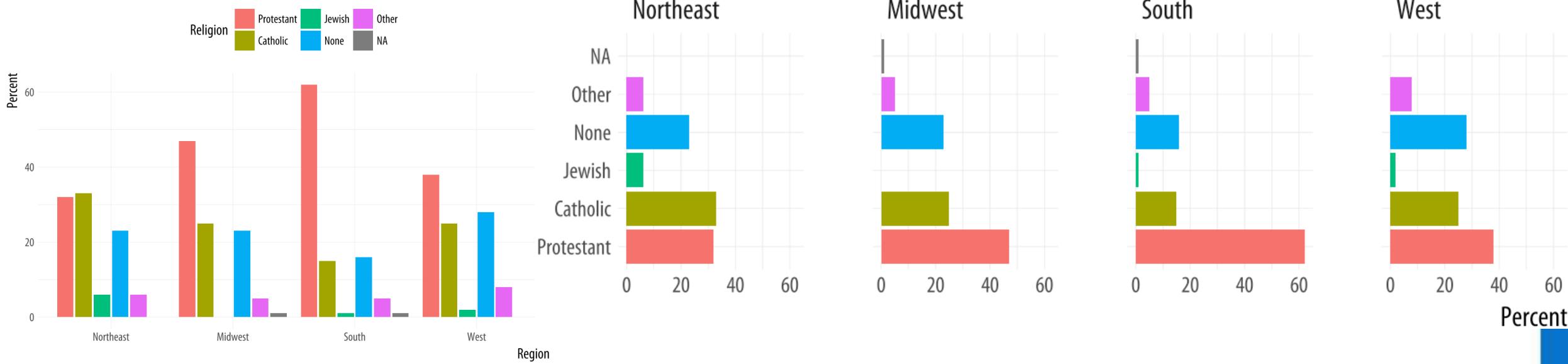
EL “COLOR” AYUDA, ÚSALO!



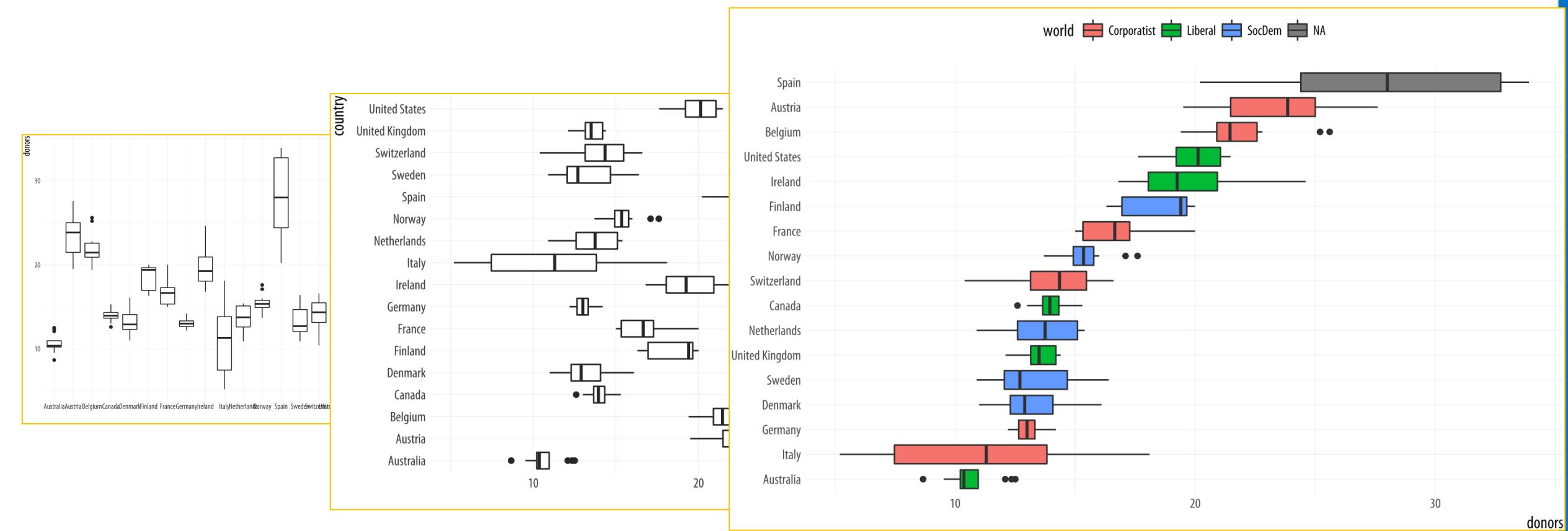
POR FAVOR, NO MUCHAS BARRAS JUNTAS... "SMALL MULTIPLES".



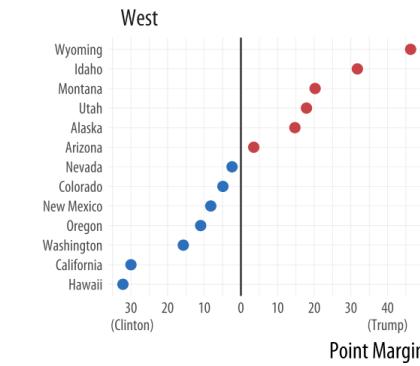
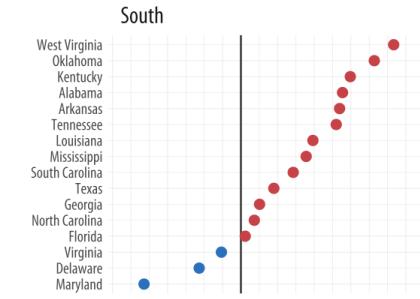
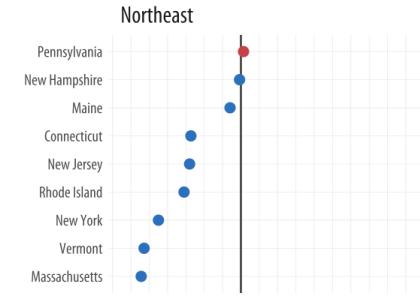
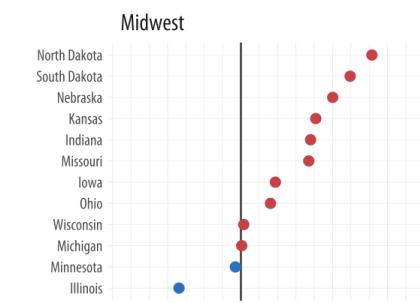
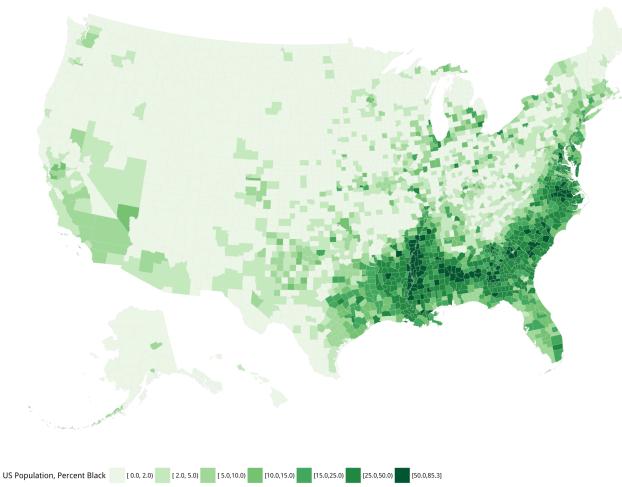
SE PUEDE HACER UN POCO MEJOR... (1 DE 2)



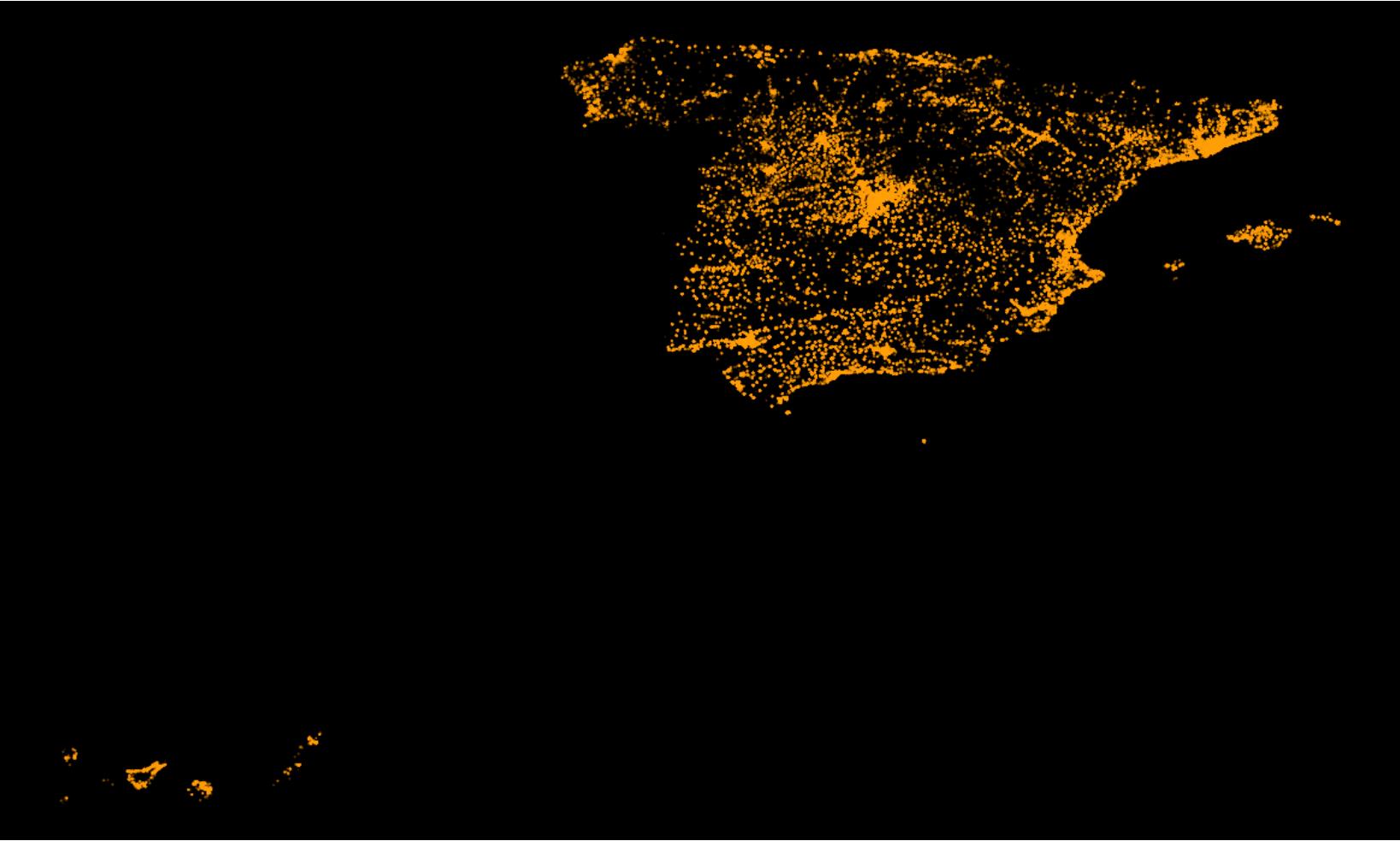
SE PUEDE HACER UN POCO MEJOR... (2 DE 2)



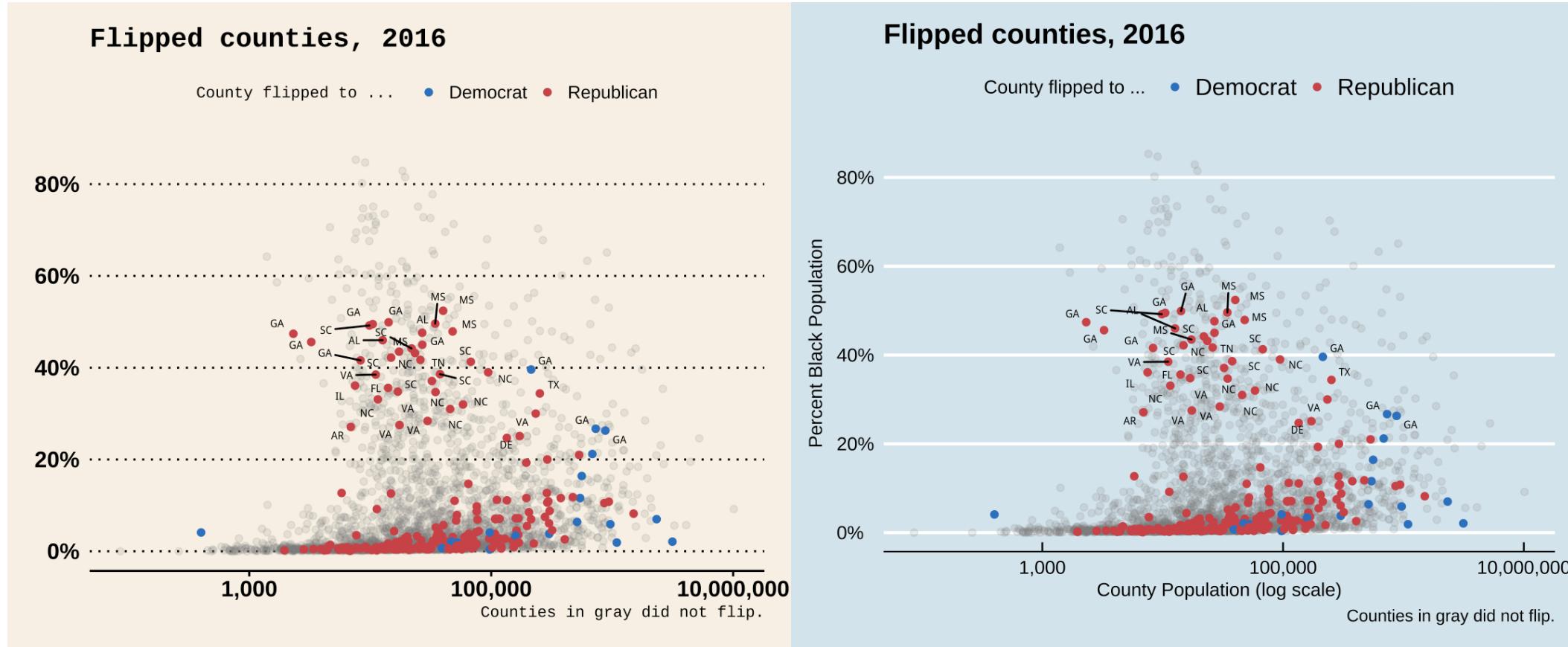
MAPAS...



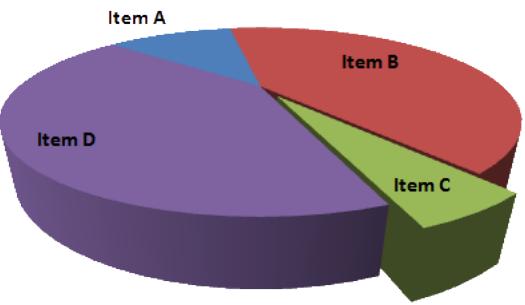
MAPAS...



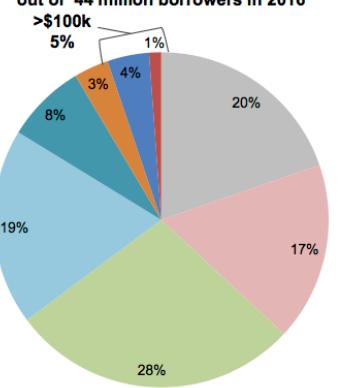
COLORES, FONDOS... GRACIAS POR LA BELLEZA...



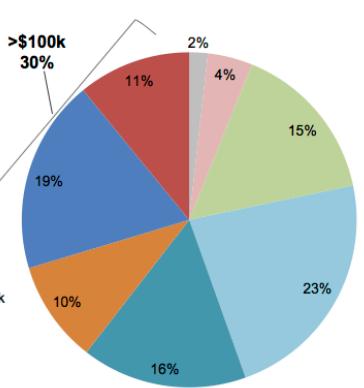
Noooooooooooo!!!!!!



Borrower Distribution by Outstanding Balance out of 44 million borrowers in 2016



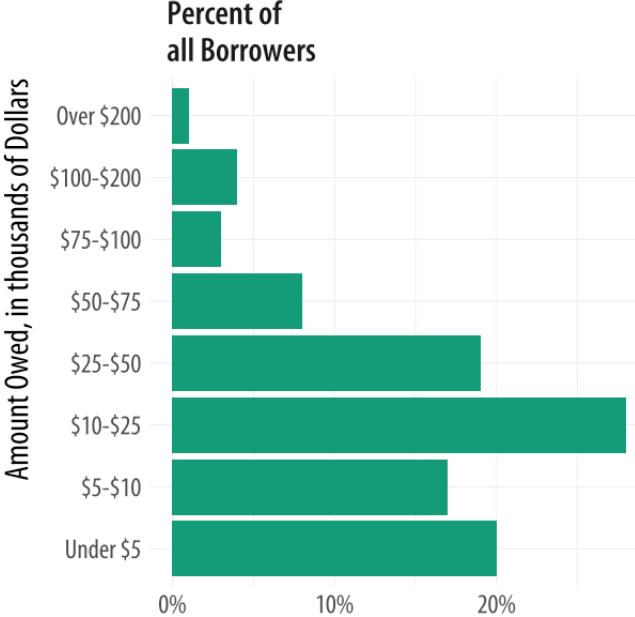
Debt Distribution by Outstanding Balance out of \$1.3 trillion in 2016



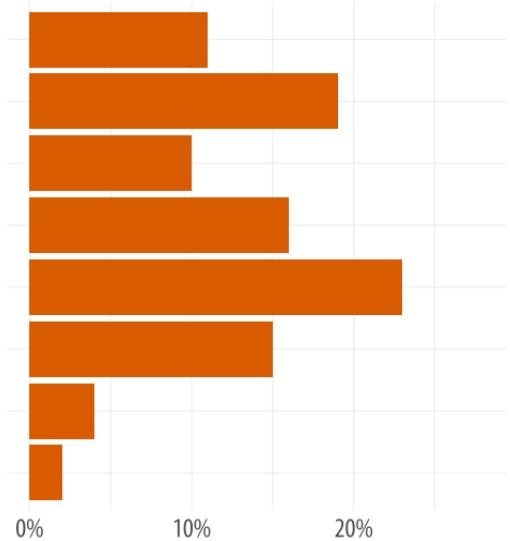
#porunaempresasingraficosdetarta

Outstanding Student Loans

44 million borrowers owe a total of \$1.3 trillion



Percent of all Balances



Source: FRB NY



ENTONCES... PROPÓSITO, CLARIDAD Y MENSAJE..

Three laws for improving **visual** communication

Have a clear purpose

- Know the purpose of creating the graph
- Identify the quantitative evidence to support the purpose
- Identify the audience and focus the design to support their needs

Show the data clearly

- Avoid misrepresentation (use appropriate scales)
- Choose the appropriate graph type to display your data
- Maximize data to ink ratio (reduce distraction, less is more)

Make the message obvious

- Use proximity and alignment to aid in comparisons
- Minimize mental arithmetic (e.g. plot the difference)
- Use colors and annotations to highlight important details



<https://rstudio.com/assets/img/visR-Rviews-slides.pdf>

<https://rstudio.com/resources/webinars/effective-visualizations-for-data-driven-decisions/>

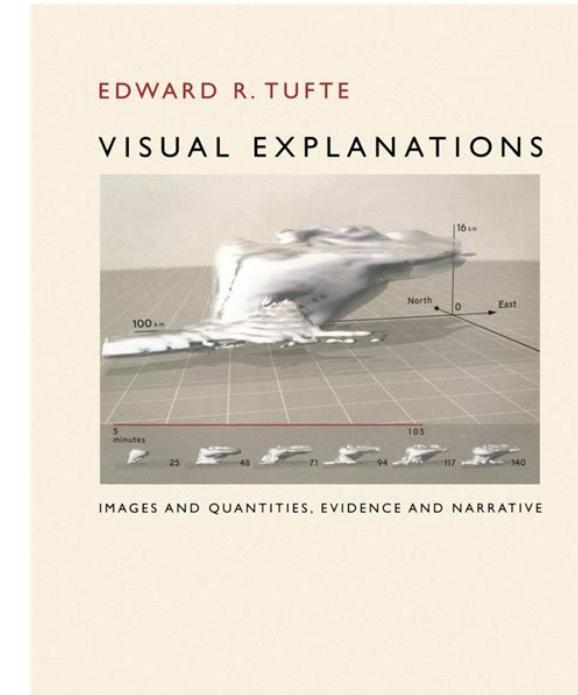
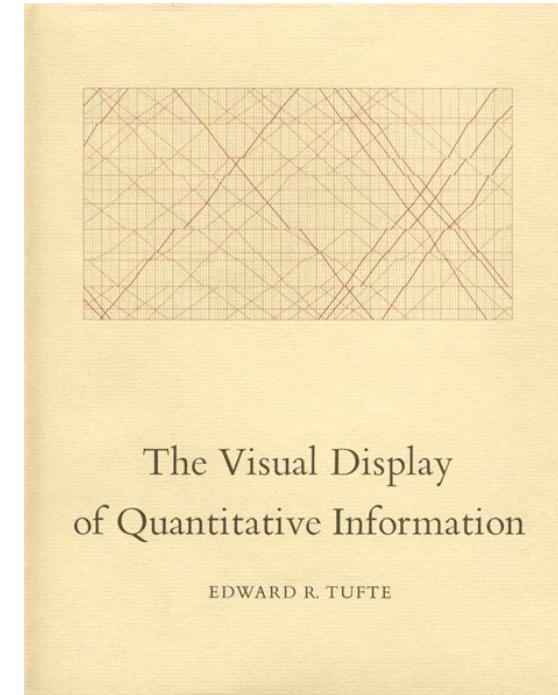
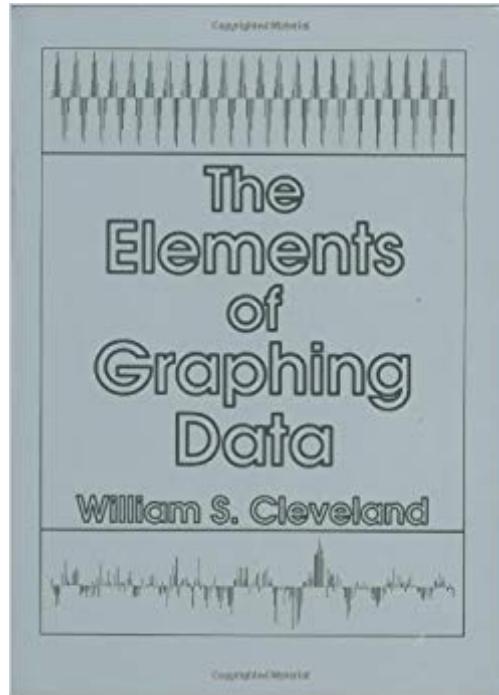
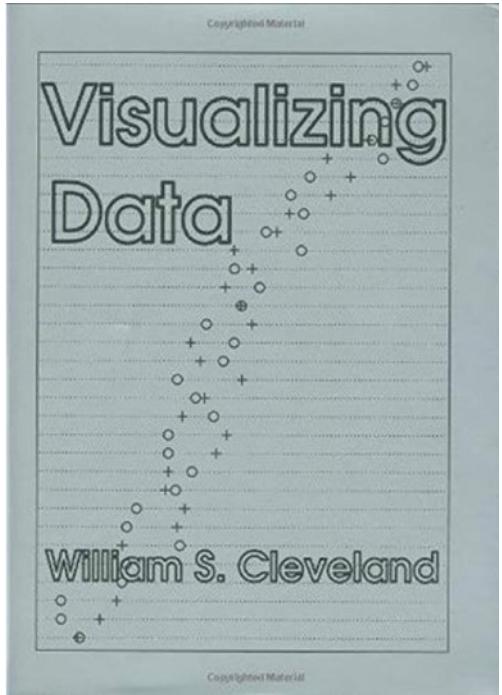
<https://github.com/GraphicsPrinciples/CheatSheet/blob/master/NVSCheatSheet.pdf>



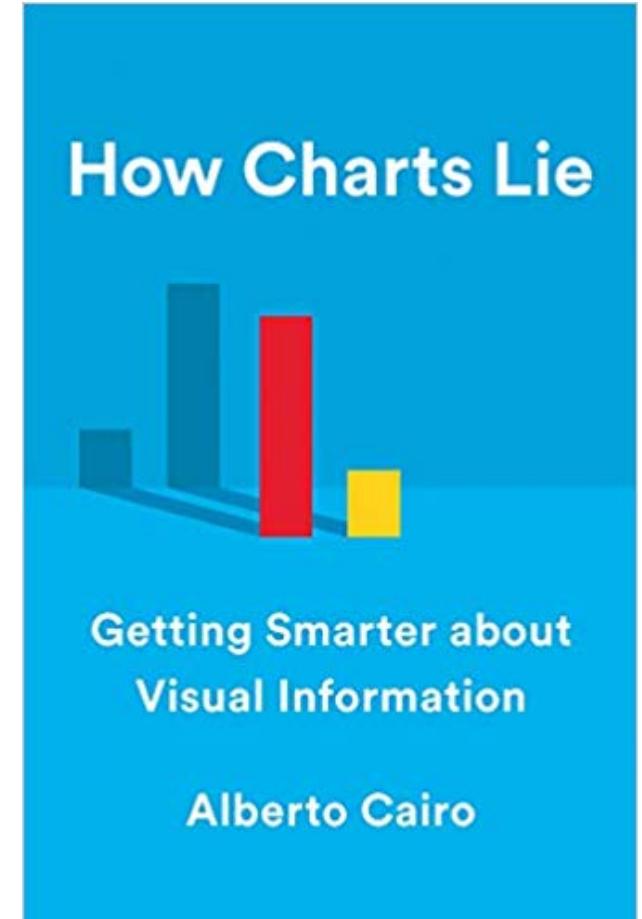
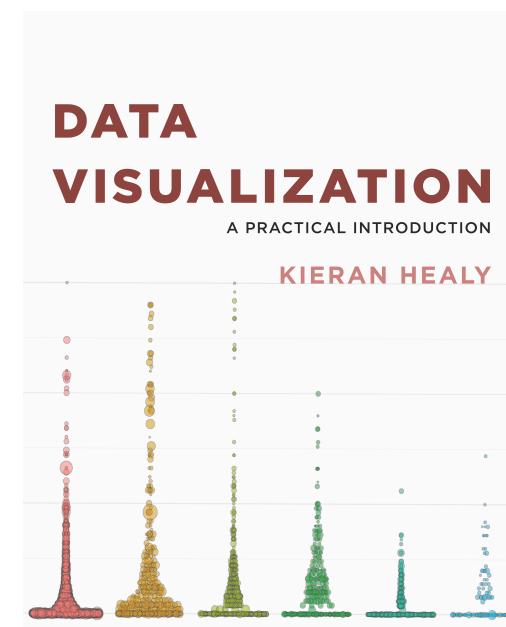
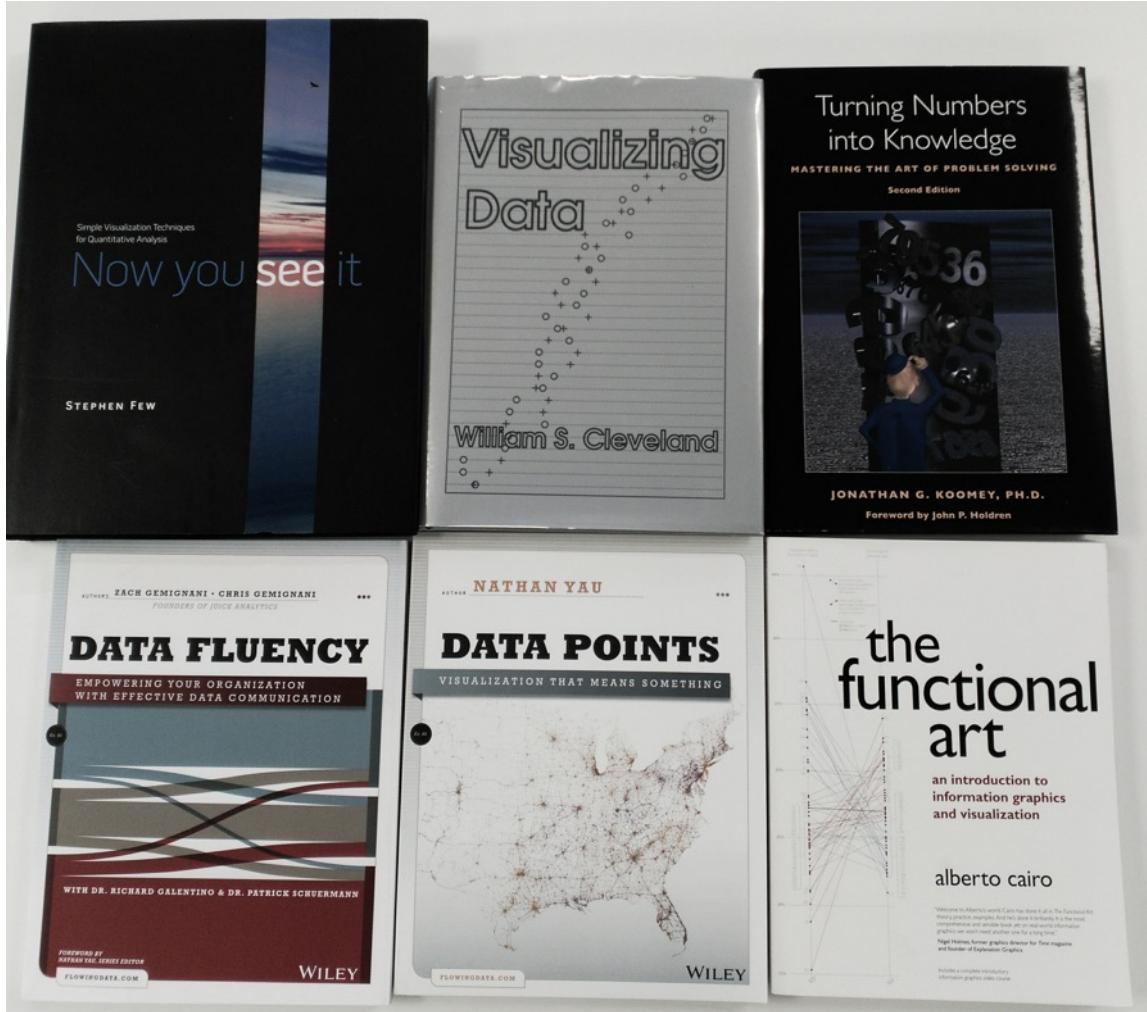
Referencias

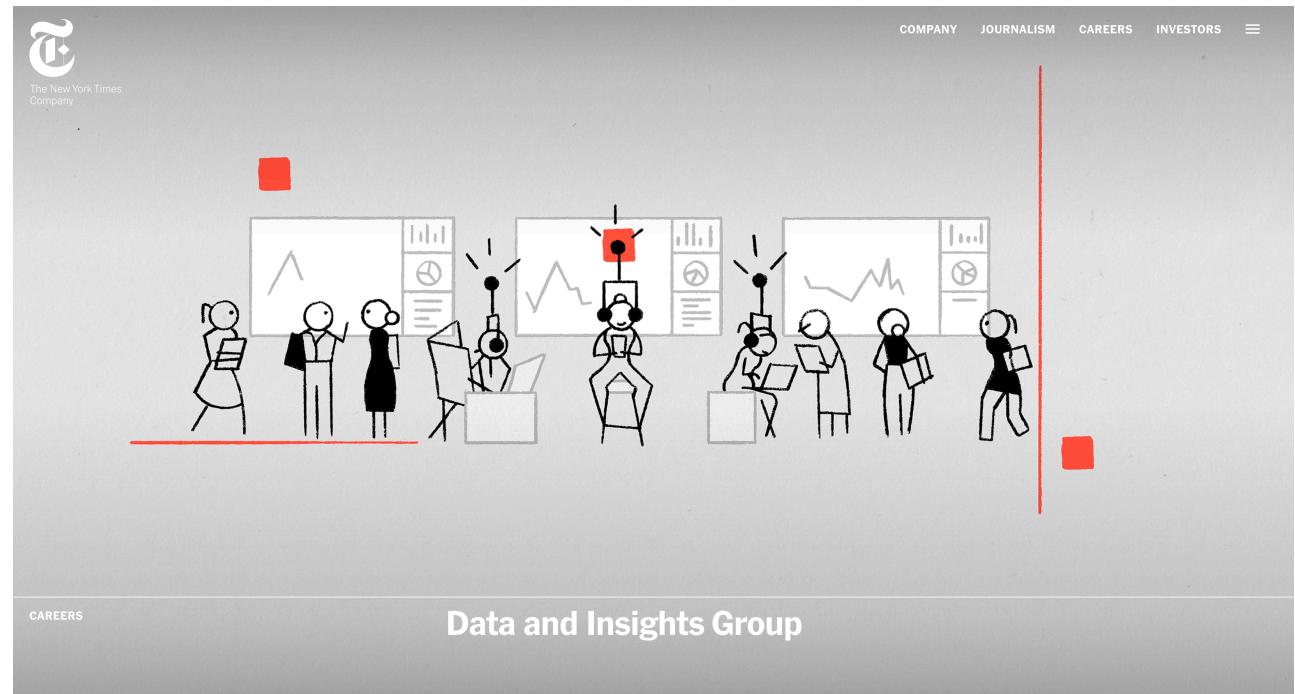
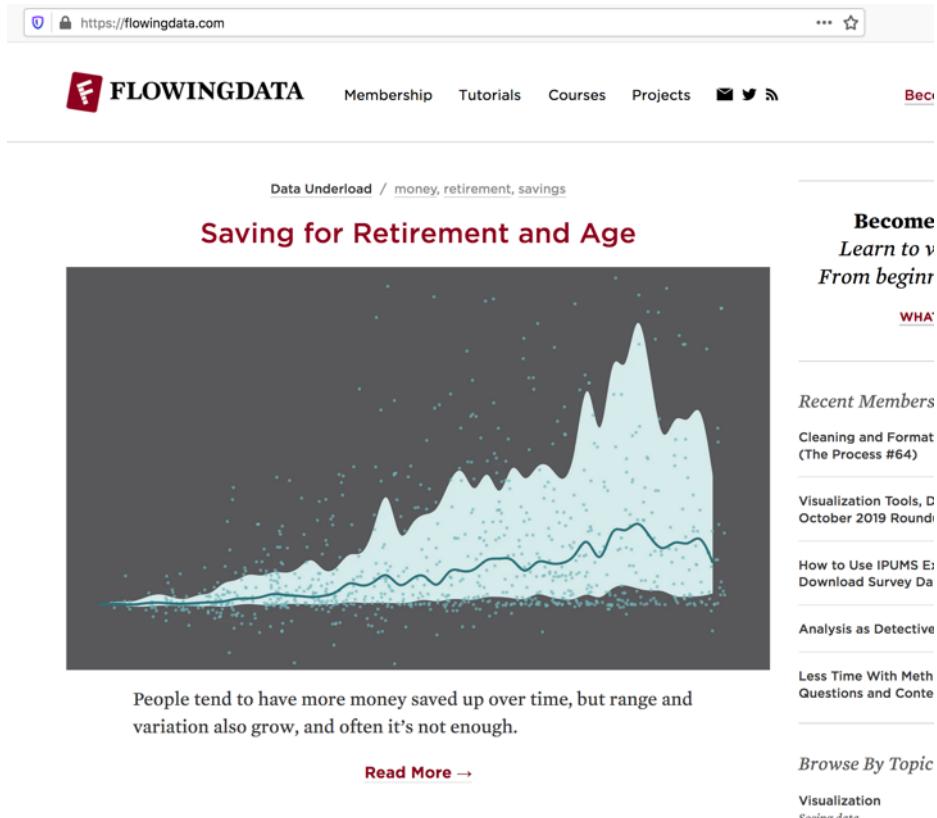
Visualización - EDA

LIBROS – VISUALIZACIÓN (CLÁSICOS)



LIBROS – VISUALIZACIÓN (ACTUALES)





PRESENTACIÓN..



DOWNLOAD SUPPORT COMMUNITY



Products Resources Pricing About Blogs

- Webinars
- RStudio Essentials
- Data Science Essentials
- Advanced Data Science
- Working with Spark
- RStudio Pro Administration
- Materiales en Español
- Additional Talks
- rstudio::conf by year



Effective Visualizations for Credible, Data-Driven Decision Making

Presented by:  NOVARTIS 

WEBINARS

Effective Visualizations for Data Driven Decisions

Charlotta Früchtenicht | Diego Saldana | Mark Baille | Marc Vandemeulebroecke | April 1, 2020



ELIGE EL QUE MÁS TE GUSTE.....

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

The R Graph Gallery

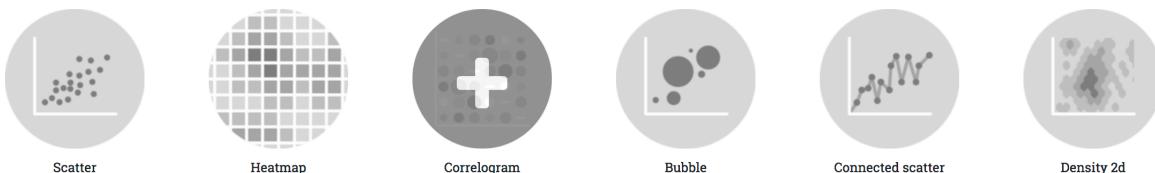


Welcome the R graph gallery, a collection of charts made with the [R programming language](#). Hundreds of charts are displayed in several sections, always with their reproducible code available. The gallery makes a focus on the tidyverse and [ggplot2](#). Feel free to suggest a chart or report a bug, any feedback is highly welcome. Stay in touch with the gallery by following it on [Twitter](#) or [Github](#). If you're new to R, consider following this [course](#).

Distribution



Correlation



Ranking



.html

© 2018 Teradata

<https://python-graph-gallery.com/>

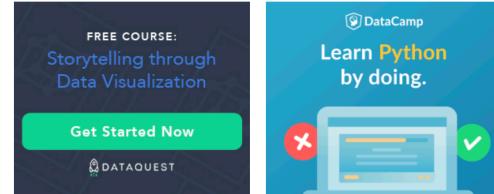


THE PYTHON
GRAPH GALLERY

HOME CHART TYPES TOOLS ALL CHARTS R GALLERY D3JS DATA TO VIZ ABOUT

Welcome to the Python Graph Gallery. This website displays hundreds of charts, always providing the reproducible python code! It aims to showcase the awesome dataviz possibilities of python and to help you benefit it. Feel free to [propose](#) a chart or [report](#) a bug. Any feedback is highly welcome. Get in touch with the gallery by following it on [Twitter](#), [Facebook](#), or by [subscribing](#) to the blog. Note that [this online course](#) is another good resource to learn dataviz with python.

Sponsors



DISTRIBUTION



¡GRACIAS!
(cof@qualityexcellence.es)