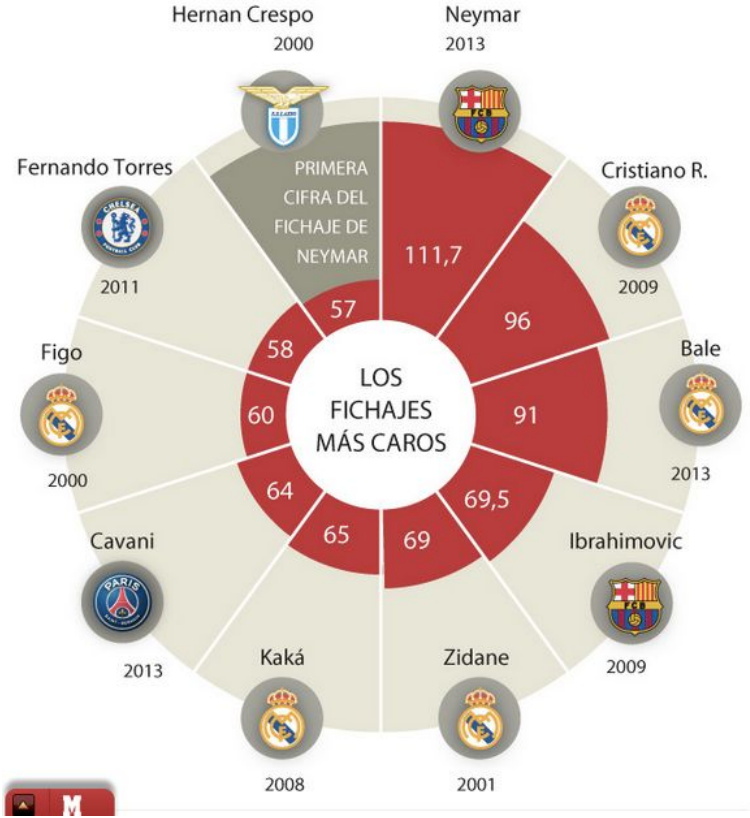
Visualización de múltiples variables

Diplomatura CDAAyA 2018

¿Qué vamos a ver hoy?

- Alteraciones comunes y gráficos incorrectos
- Cómo hacer gráficos de correlaciones en seaborn
- Otras librerías de visualización y cómo elegir
- Presentación del trabajo práctico

The lie factor



¿Qué problemas tiene la imagen?

- No hay explicación de qué significan los números
- ¿Qué significan las barras? ¿Estamos comparando radio del arco rojo o el área coloreada? ¿Son proporcionales?

Neymar: ¿el más caro o uno de los más baratos?

Consistencia visual

The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented.

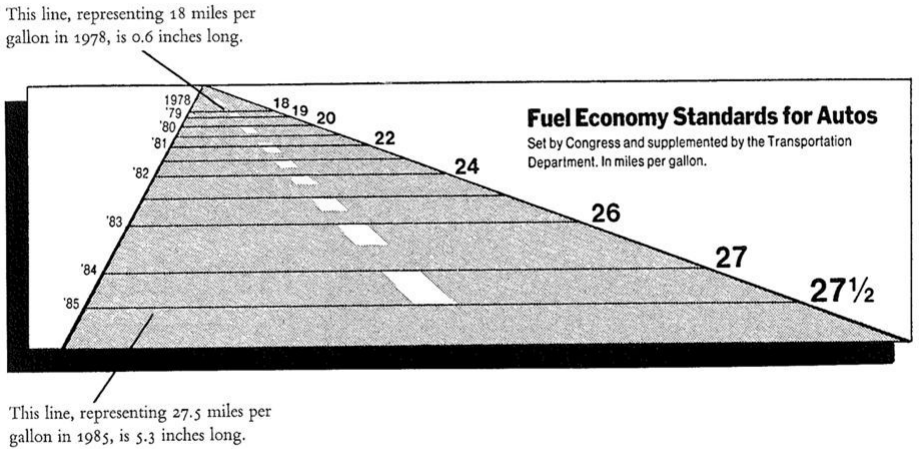
- Edward Tufte

The lie factor

$$Lie\ Factor = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

Lie factors de más de 1.5 o menos de 0.95 son considerados distorciones

Lie factor: 14.8



Line increase: 783%
Actual increase: 53%

***New York Times,* 9th August 1978, p D-2**

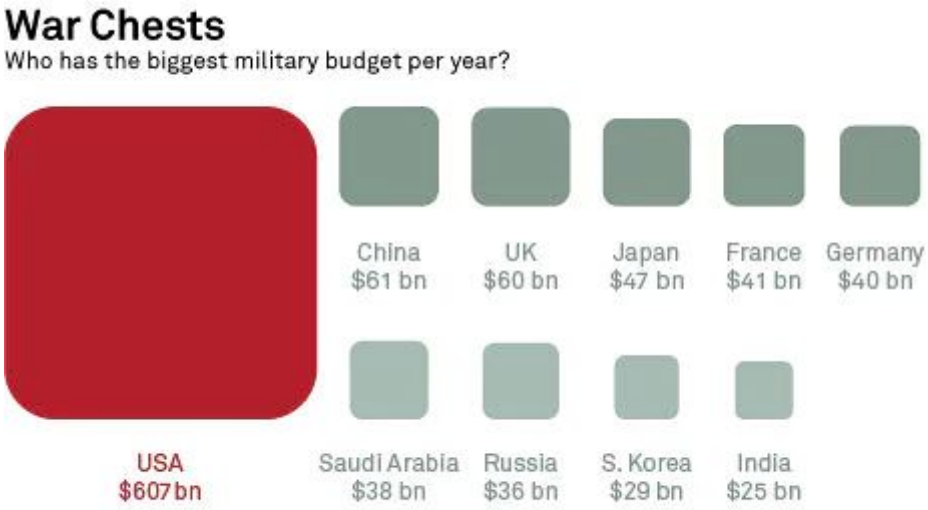
http://slideplayer.com/slide/8269500/

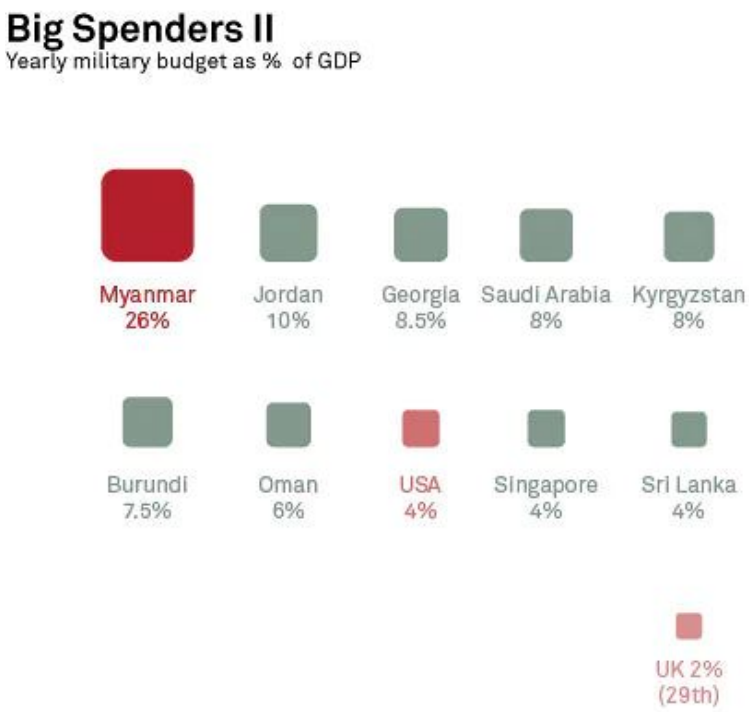
Coherencia de los datos

Biased vs unbiased information

- ¿Cuál es el impacto real de estos números?
- ¿Por qué India tiene un color diferente?

Biased vs unbiased information

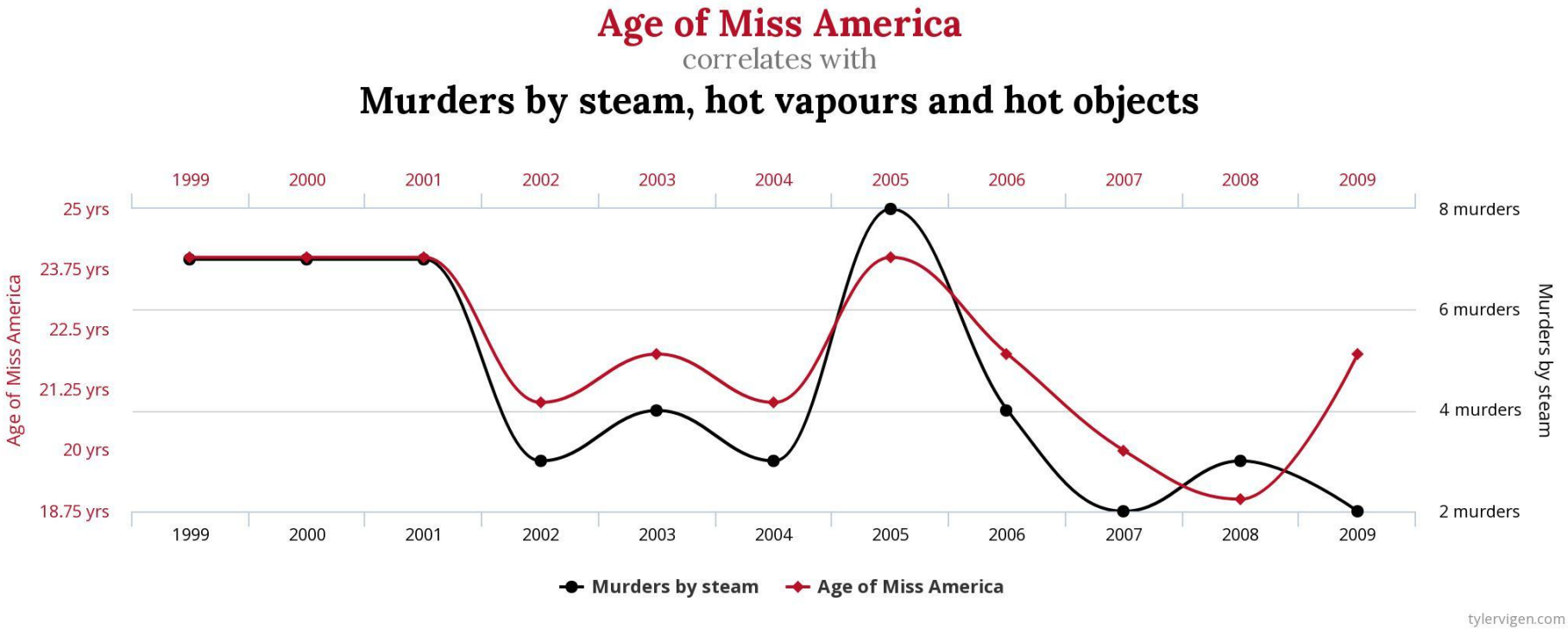




Information is beautiful: war games

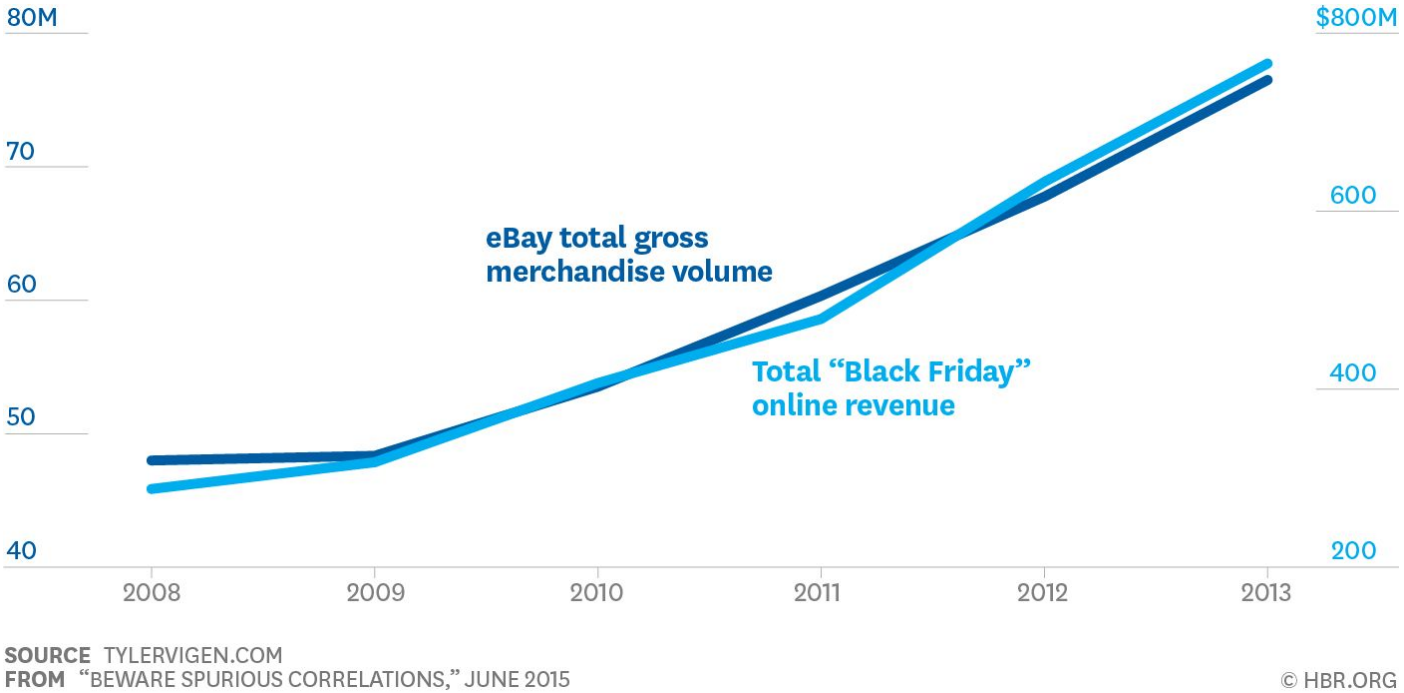
Correlación no implica causalidad!

Variables no relacionadas



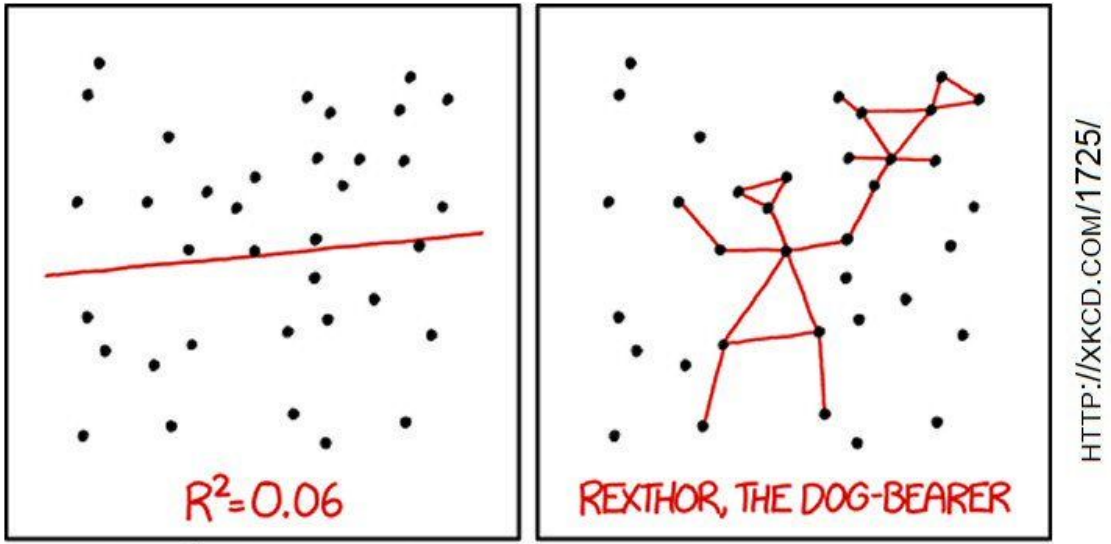
Spurious correlations

Variables relacionadas, datasets distintos



Beware of spurious correlations

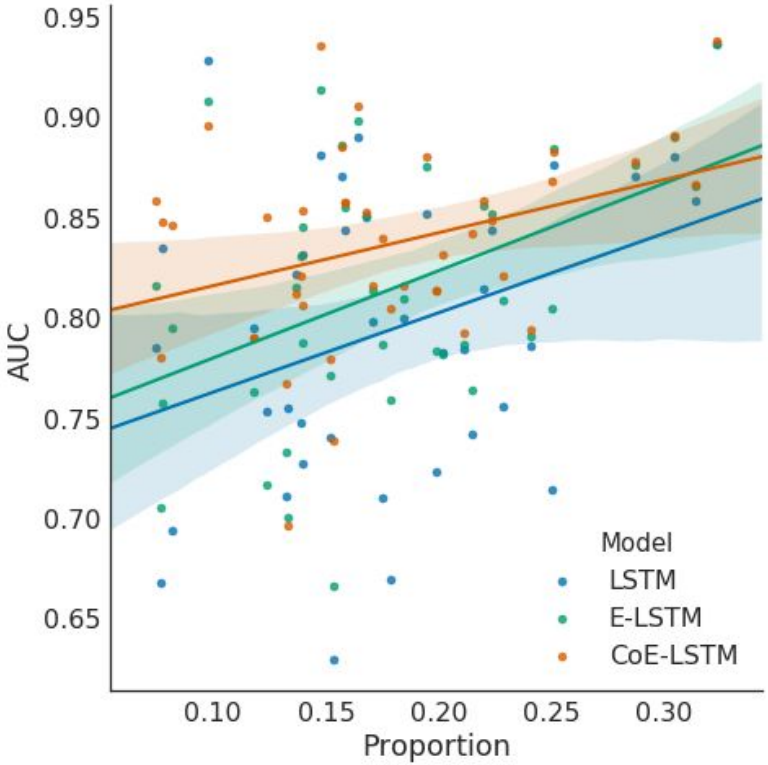
Correlaciones forzadas



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

XKCD

Correlaciones forzadas



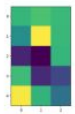
http://callingbullshit.org/

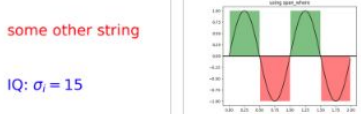
Gráficos de correlaciones - Notebook

Matplotlib

- Fácil integración con notebooks
- Muy versátil
- Es relativamente simple para gráficos comunes

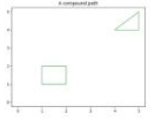
Watermark image

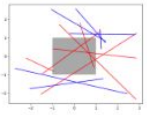

Modifying the coordinate formatter

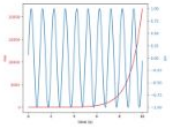


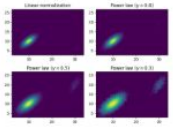
A mathtext image as numpy array

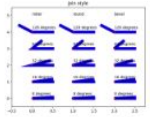
Using span_where

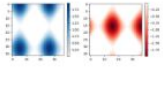

Compound path

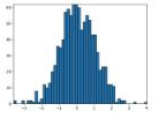

Changing colors of lines intersecting a box

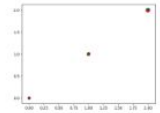

Plots with different scales

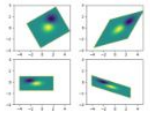

Exploring normalizations

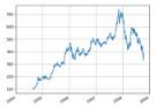

Join styles


Colorbar

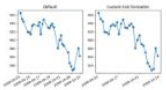

Building histograms using Rectangles and PolyCollections


Scatter plot with pie chart markers


Affine transform of an image

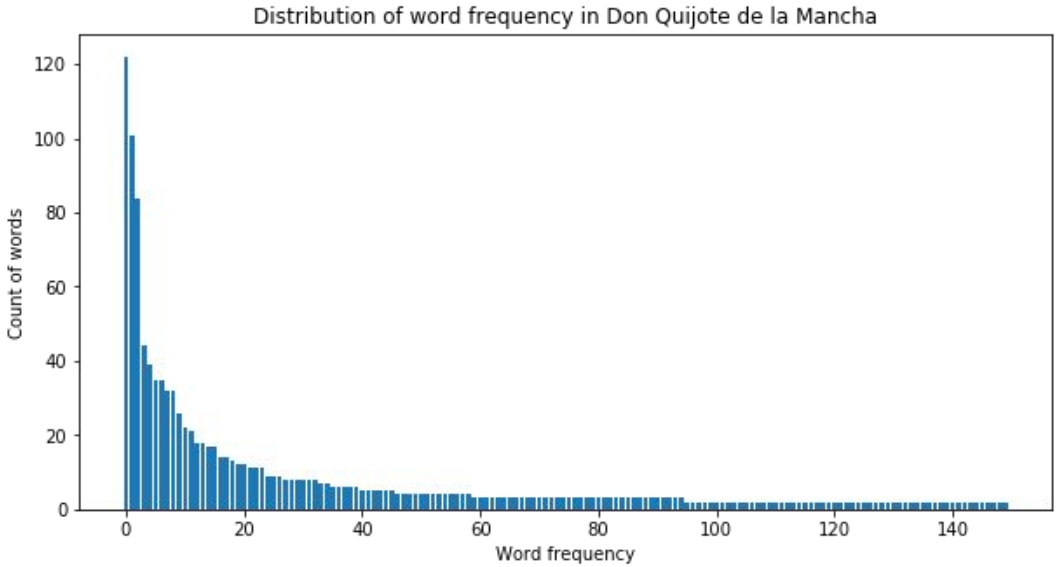

Date tick labels


Labeling ticks using engineering notation


Custom tick formatter for time series

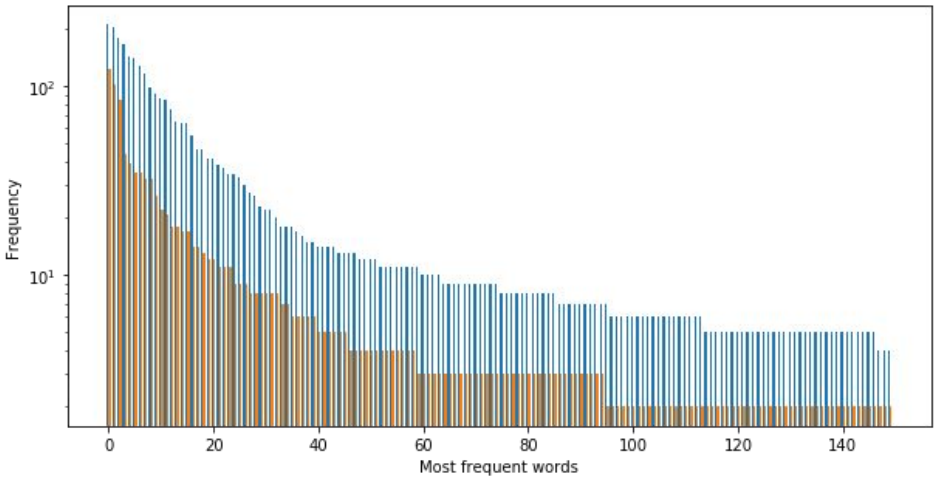
Matplotlib - Ejemplo

Tenemos en dos arreglos la frecuencia con la que aparecen palabras del castellano en dos obras: el Martín Fierro y Don Quijote de la Mancha. ¿Cómo comparamos las distribuciones?



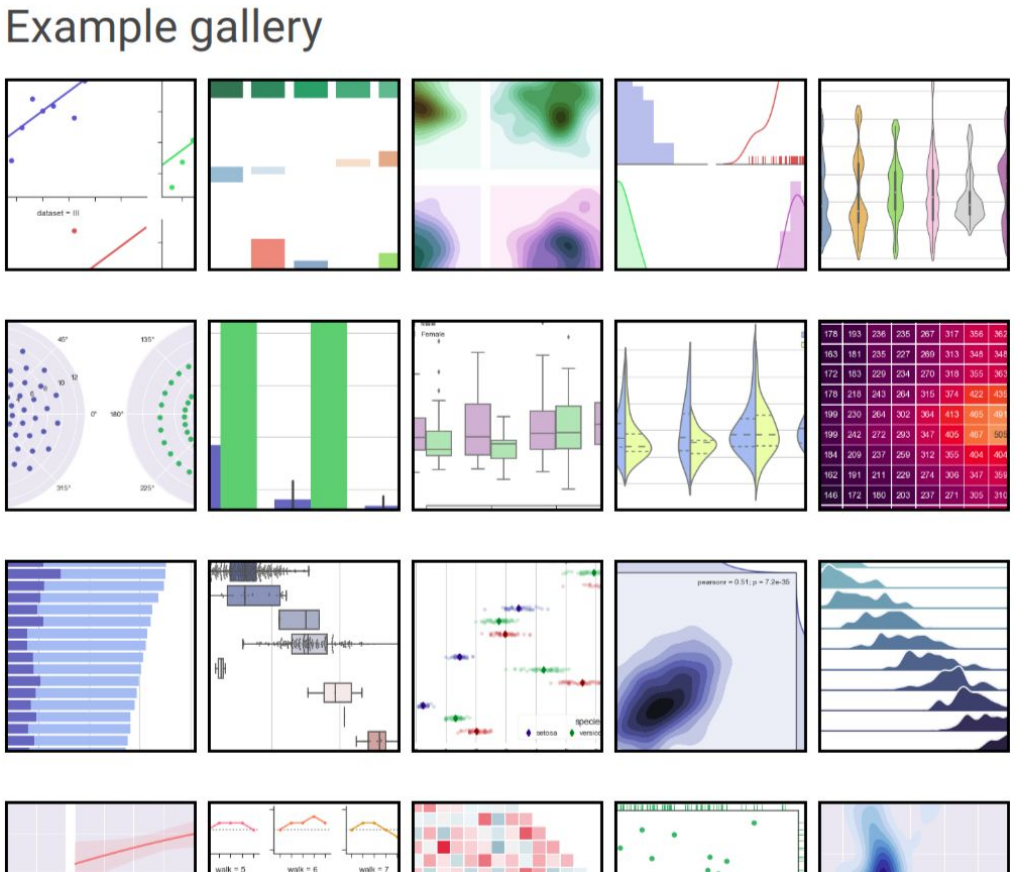
Matplotlib - Ejemplo

```
import matplotlib.pyplot as plt
import numpy as np

data = [_ for _ in zip(counts_m, counts_q)][:20]
dimw = 0.75 / len(data[0])  # Width of the bars

fig, ax = plt.subplots()
x = np.arange(len(data))
for i in range(len(data[0])):
    y = [d[i] for d in data]
    b = ax.bar(x + (i * dimw) - dimw / 2, y, dimw)

ax.set_yscale('log')
ax.set_xlabel('Most frequent words')
ax.set_ylabel('Frequency')
```



Seaborn

- Fácil integración con notebooks
- Fácil de crear gráficos con múltiples variables
- Integración con **pandas**



Seaborn - Ejemplo

Tenemos el dataset del titanic, con múltiples variables de distintos tipos. ¿Cómo hacemos para graficar **más de dos columnas** al mismo tiempo?

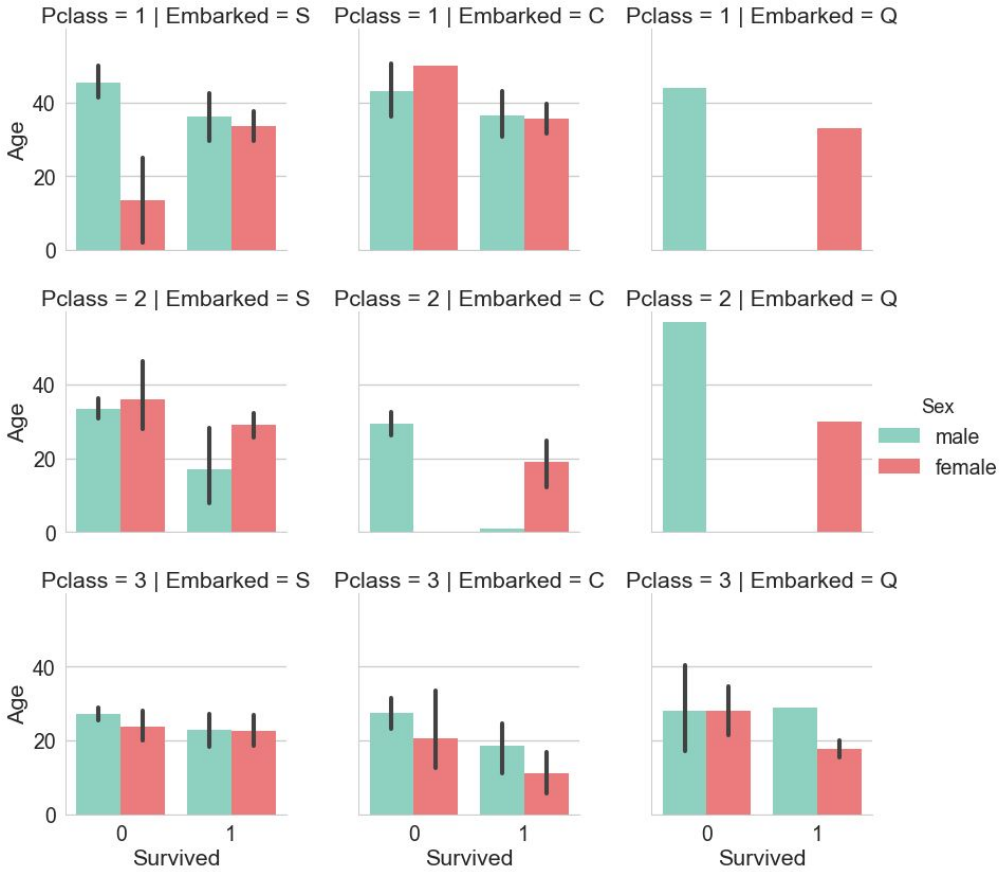
In [38]:
```
titanic = pandas.read_csv(
    'https://cs.famaf.unc.edu.ar/~mteruel/datasets/diplodatos/titanic_train.csv')
titanic[:5][['Survived', 'Age', 'Sex', 'Pclass', 'Embarked']]
```

Out[38]:

	Survived	Age	Sex	Pclass	Embarked
0	0	22.0	male	3	S
1	1	38.0	female	1	C
2	1	26.0	female	3	S
3	1	35.0	female	1	S
4	0	35.0	male	3	S

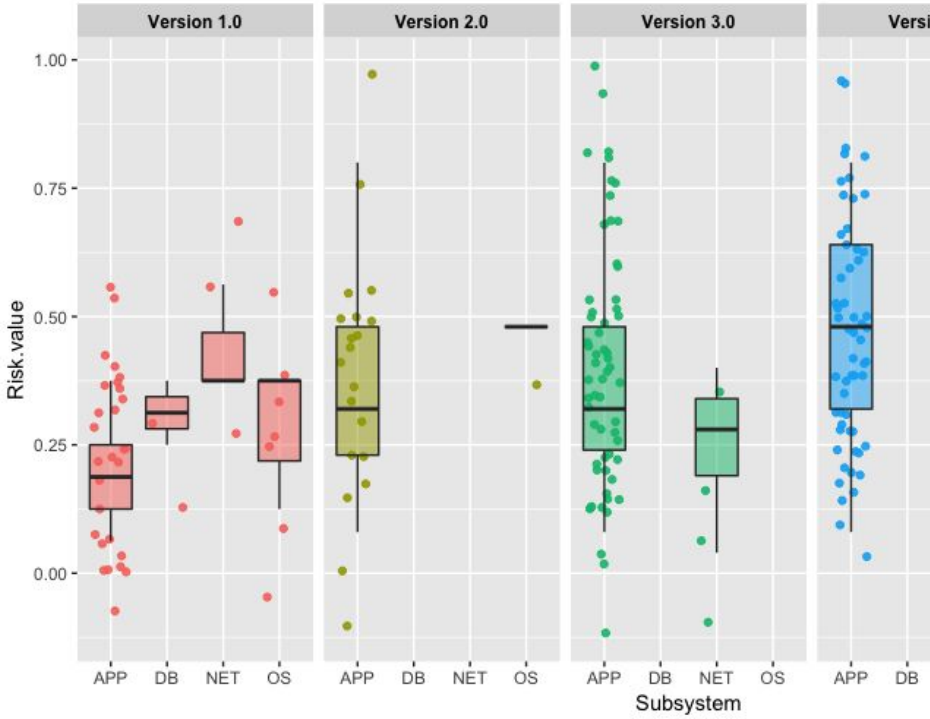
Seaborn - Ejemplo

```
palette ={'female':'#FF686B', 'male':'#84DCC6'}
seaborn.factorplot(
    'Survived', 'Age', data=titanic, hue='Sex',
    row='Pclass', col='Embarked', kind='bar',
    palette=palette)
seaborn.despine()
```



GGPlot

- No tan fácil integración con notebooks
- Equivalente a seaborn, **con más control** sobre configuración
- Separa Aesthetics y Geometrics



GGplot - Ejemplo

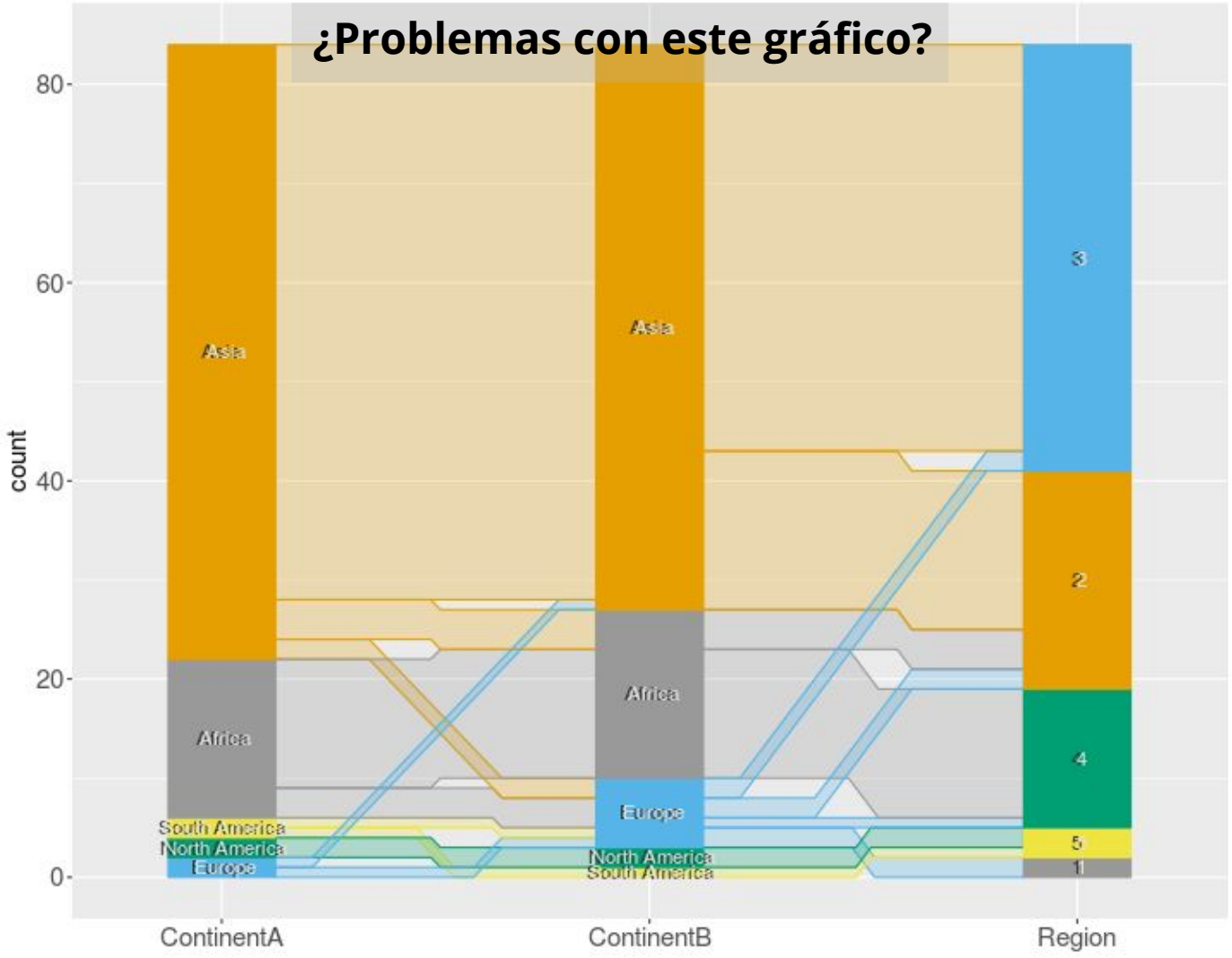
Con el dataset de los **conflictos mundiales**, junto con datos de los países actuales y sus respectivos continentes, queremos ver si la mayoría de los conflictos son dentro del mismo continente o se traspasan fronteras. ¿Hipótesis?

```
import rpy2
%load_ext rpy2.ipython
```

```
%%R

library(ggplot2)
library(reshape2)
```

```
%%R -i country_conflicts -w 10 -h 8 -u in
```

¿Problemas con este gráfico?



Descripción del dataset

GGplot - Ejemplo

```
%%R -i country_conflicts -w 10 -h 8 -u in

colors <- c("#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442")
ggparallel(list("ContinentA", "ContinentB", "Region"),
           data=country_conflicts, text.angle=0, alpha=0.25) +
    theme(legend.position="none") +
    scale_fill_manual(values = rep(colors, 14)) +
    scale_colour_manual(values = rep(colors, 14)) +
    theme(text=element_text(size=15), axis.text=element_text(size=15))
```

Aestetics + Graphics

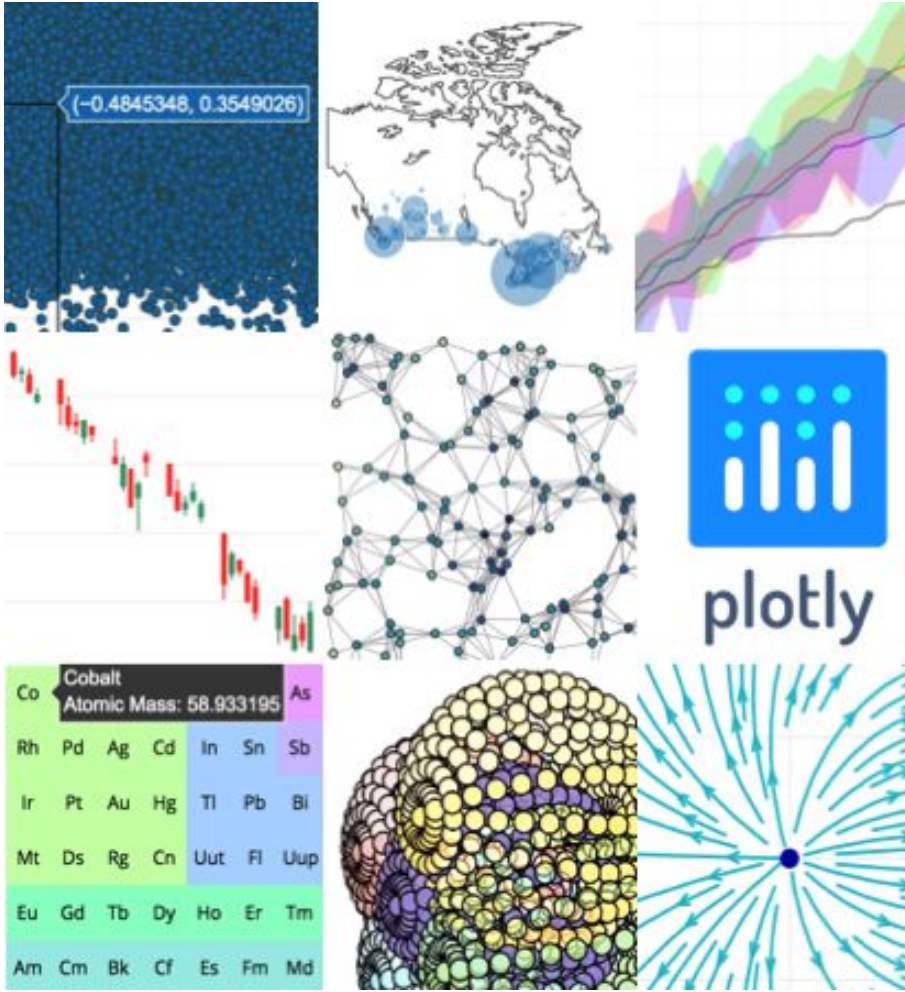
GGplot - Ejemplo

```
%%R -i country_conflicts -w 10 -h 8 -u in

colors <- c("#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442")
ggparallel(list("ContinentA", "ContinentB", "Region"),
           data=country_conflicts, text.angle=0, alpha=0.25) +
    theme(legend.position="none") +
    scale_fill_manual(values = rep(colors, 14)) +
    scale_colour_manual(values = rep(colors, 14)) +
    theme(text=element_text(size=15), axis.text=element_text(size=15))
```

Aestetics + Graphics

Plotly

- Fácil integración con notebooks
- Requiere tener una cuenta
- Los gráficos son mucho más interactivos

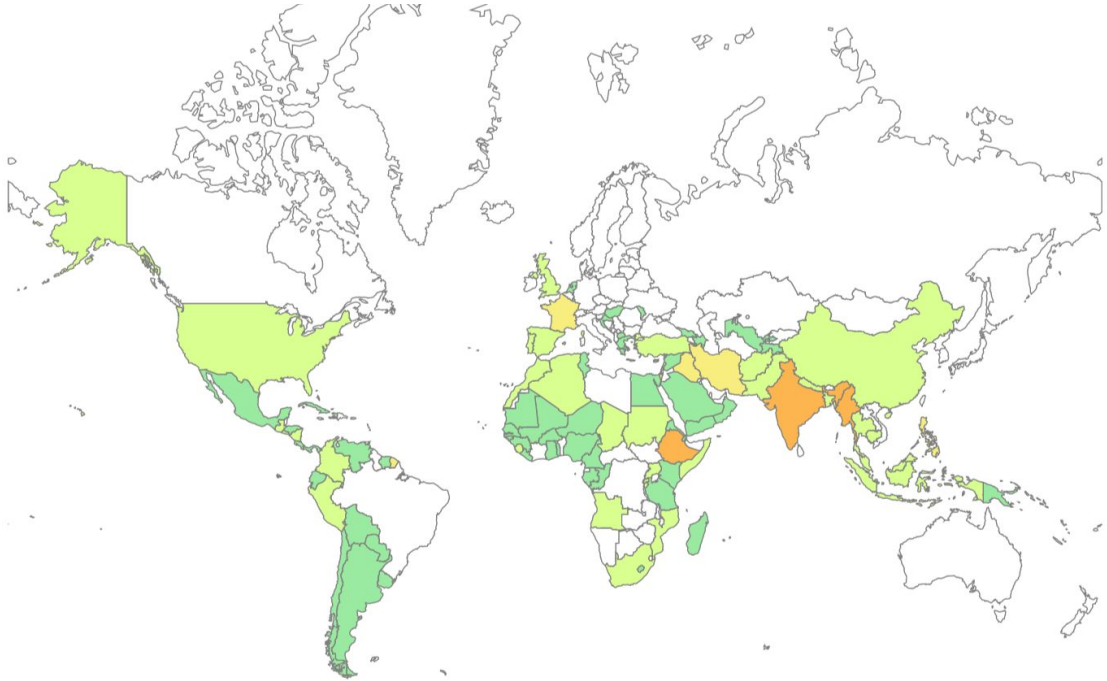


D3 y librerías javascript

- Mucha variedad de **interacciones**
- Código **más complejo**
- Dependiente de los datos
- Completamente flexible
- No para Big Data

D3 - Ejemplo

Con el dataset de los **conflictos mundiales** (sin filtrar), mostrar la cantidad de conflictos por país.



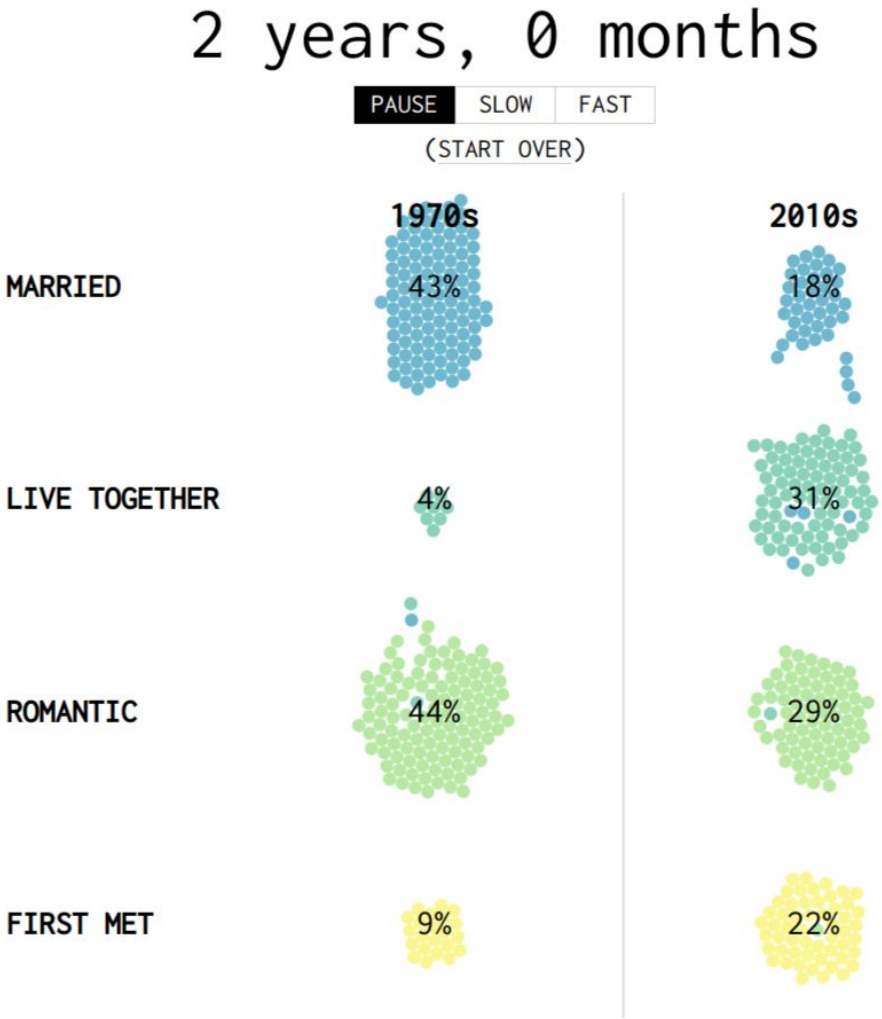
D3 - Animaciones

[The stages of relationships]

Las animaciones son útiles para representar patrones temporales

No siempre son gráficos

https://www.gapminder.org/dollar-street/matrix

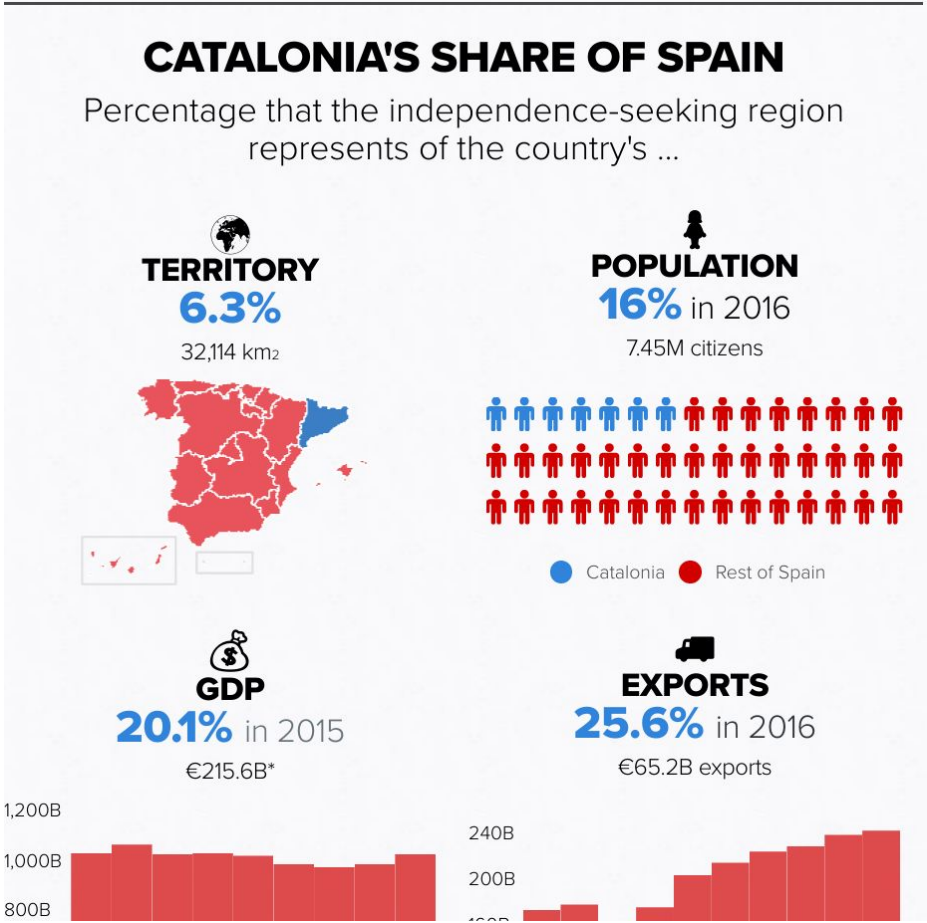


Otras librerías javascript

- chart.js
- Google visualizations
- plotly.js
- Sugerencias?

Software integrales

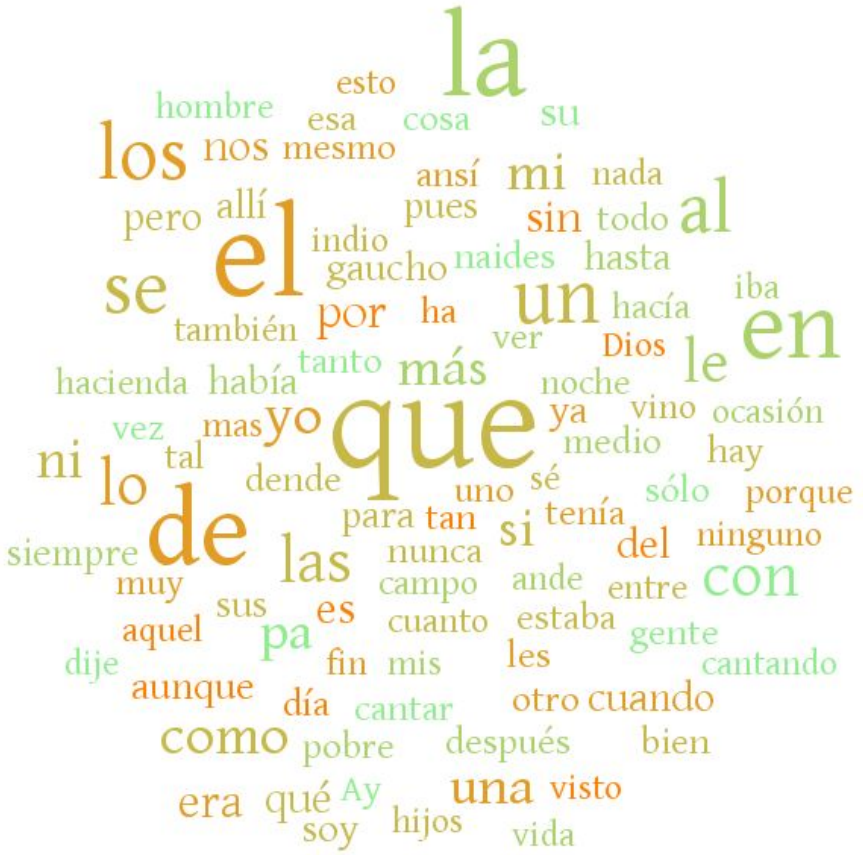
- Más estéticos y simples de manejar
- Menos flexibilidad
- Más preproceso de datos necesario



What Spain has to lose

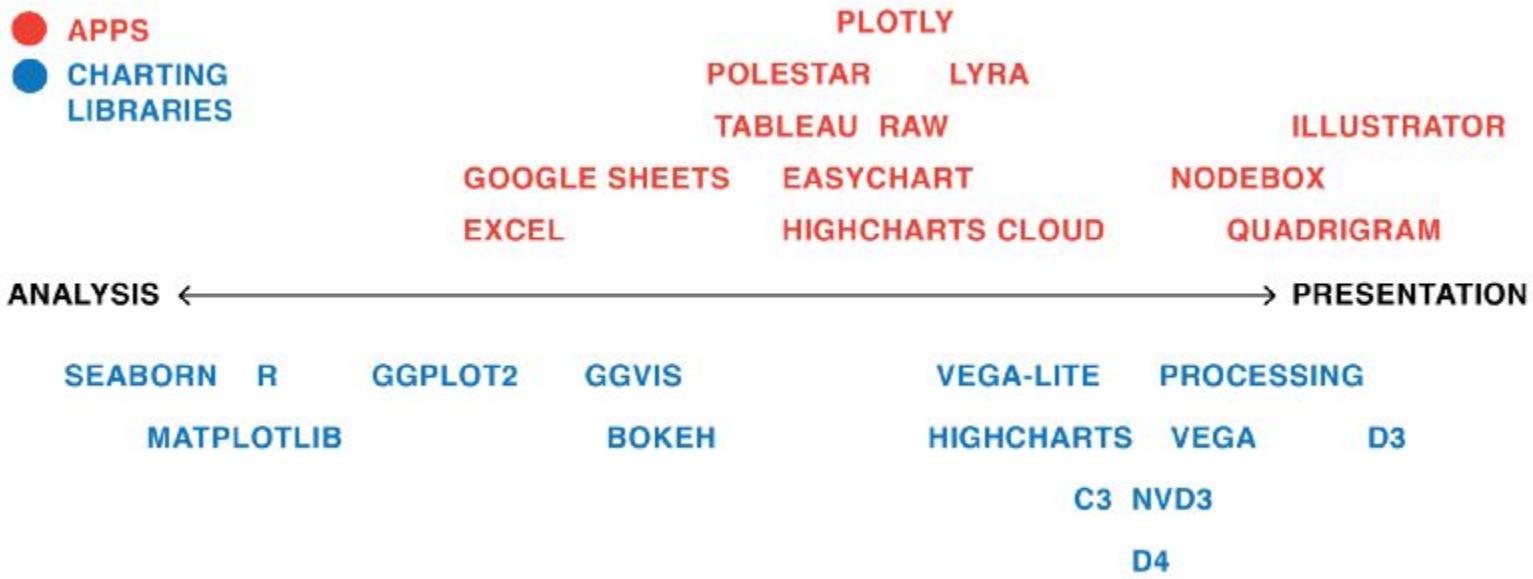
Software integrales

- Infogram
- Tableau
- Worditout



What Spain has to lose

Análisis vs presentación



What I Learned Recreating One Chart Using 24 Tools

¿Preguntas?

“We are demanding a visual aspect to our information”

- David McCandless