



DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA



Visualización y Análisis Exploratorio

Laboratorio de Datos

2^{do} Cuatri, 2025

Hasta ahora:

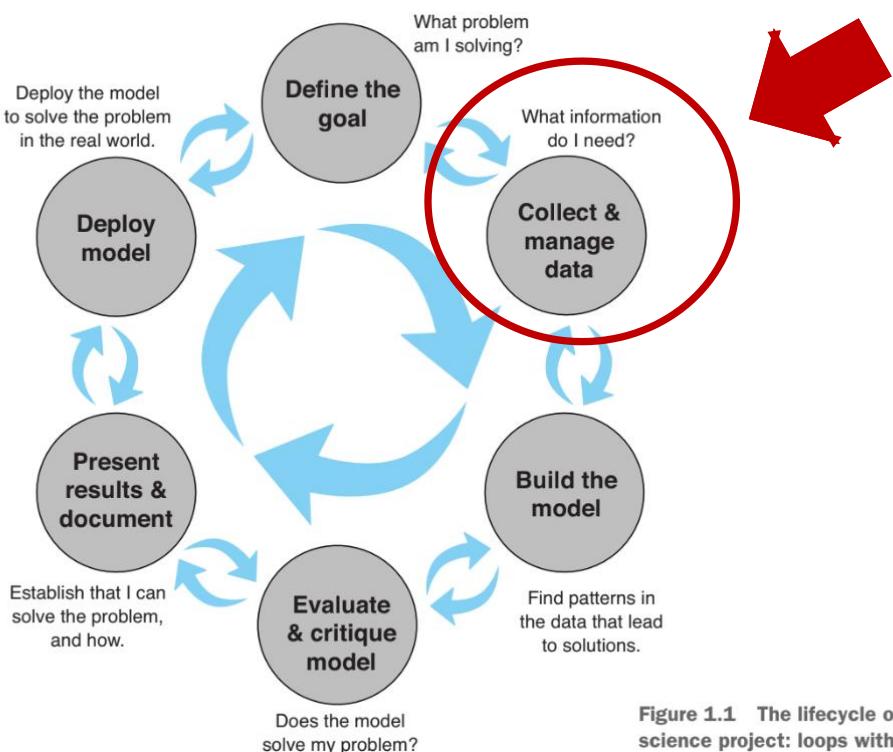


Figure 1.1 The lifecycle of a data science project: loops within loops

1ra parte de la materia

- Lenguaje de programación (Python)
- Modelado conceptual de los datos (DER)
- Representación de los datos (modelo relacional)
- Formas de consulta a DB (SQL)
- Recomendaciones para el diseño de DB (Normalización)
- Calidad de datos

Hoy:

- Visualización y exploración de los datos del paso anterior

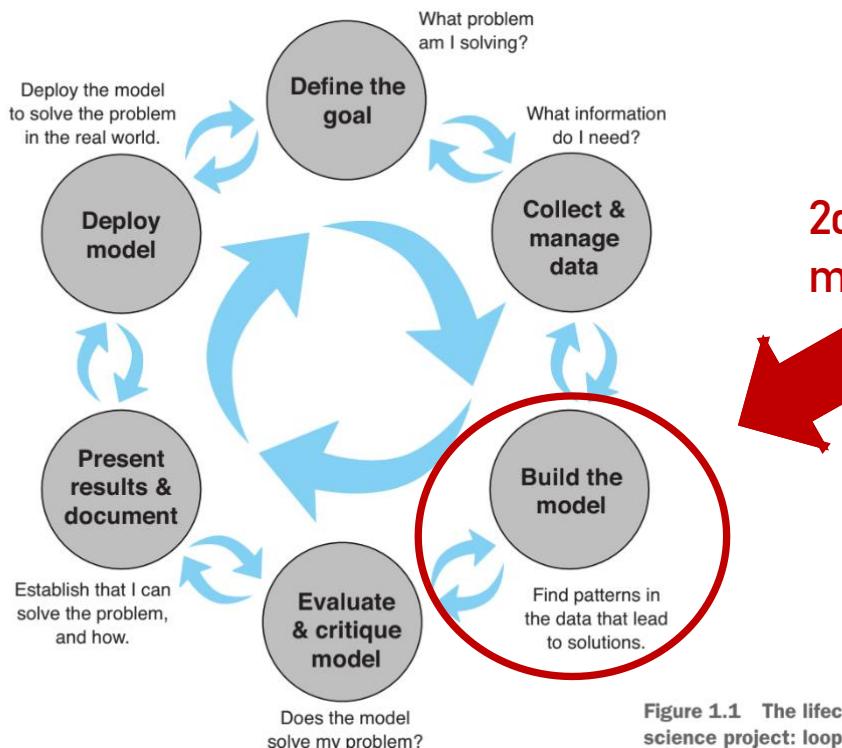


Figure 1.1 The lifecycle of a data science project: loops within loops

2da parte de la materia

Visualización de datos

¿Para qué sirve graficar?

Dataset: [CO₂ Emissions Across Countries.](#)

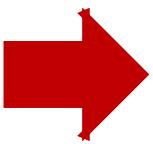
Name	# year	iso_code	# population	# gdp	# co2	# coal_co2	# oil_co2	# gas_co2	# methane
Algeria	2022	DZA	45477391.0	595820000000.0	184.558	0.775	57.407	100.626	96.622
Argentina	2022	ARG	45407904.0	854914000000.0	204.081	4.78	99.135	92.908	129.569
Australia	2022	AUS	26200987.0	1344250000000.0	384.362	146.069	133.807	80.412	126.538
Austria	2022	AUT	9064679.0	398815000000.0	61.489	11.666	30.734	15.967	9.886
Azerbaijan	2022	AZE	10295307.0	175371000000.0	40.391	0.0	12.148	26.791	23.905
Bangladesh	2022	BGD	169384890.0	858144000000.0	113.863	16.915	38.093	57.303	111.407
Belarus	2022	BLR	9173241.0	175774000000.0	59.384	4.246	16.348	35.594	17.243
Belgium	2022	BEL	11641813.0	489678000000.0	89.002	11.941	41.814	31.087	12.721
Brazil	2022	BRA	210306411.0	3187410000000.0	483.841	56.078	319.935	60.085	601.386
Bulgaria	2022	BGR	6825863.0	138390000000.0	46.966	24.577	13.793	5.14	9.941

¿Qué países emiten más?

¿Qué fuente de energía emite más a nivel mundial?

Las políticas contra el cambio climático, ¿hicieron disminuir las emisiones?

¿Se relaciona la población y el PBI con las emisiones?



Todo esto (y más) se puede ver graficando

¿Para qué sirve graficar?

Los datos se hacen más comprensibles y fáciles de usar:

- Resumir información
- Encontrar patrones o tendencias
- Detectar valores anómalos (outliers)
- Encontrar relaciones entre variables
- Guiar la investigación (o ver por donde arrancar)
- Probar una hipótesis
- Presentar resultados

Tipos de datos

Tipos de datos

Según el tipo de cada variable

Pozo	Yacimiento	Formacion	Tipo	Subtipo	Termino	Nfracturas	ArenaBombeadaNaciona	ArenaBombeadalmport	Aguinalnyectada
154744	EL OREJANO	vaca muerta	NO CONVENCIONAL	SHALE	Tapón disp	15	985.23	2442.96	15526.6
155372	EL SALITRAL	los molles	NO CONVENCIONAL	TIGHT	Punzado	3	0	109	40
136369	EL TRAPIAL	agrio	CONVENCIONAL		Punzado	2	0	31.6	131
153770	RINCON DEL MANGRULLO	mulichinco	NO CONVENCIONAL	TIGHT	Punzado	2	0	255.87	1174.52
153949	ESTANCIA LA MARIPOSA	comodoro rivadavia	CONVENCIONAL		Punzado	3	0	51.3	230.4
133184	AGUADA PICHANA	mulichinco	CONVENCIONAL		Jetteo	5	183.015	0	0
137832	EL SALITRAL	lajas	CONVENCIONAL		Punzado	4	0	45	220
161495	CAMPO INDIO	magallanes	NO CONVENCIONAL	TIGHT	Punzado	1	0	139	288.72

Nominales

Categóricas

Discretas

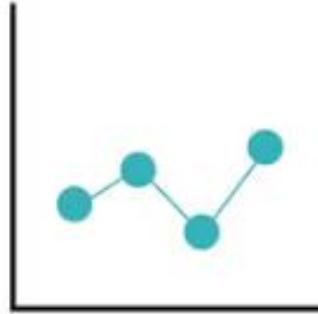
Continuas

Variables cualitativas

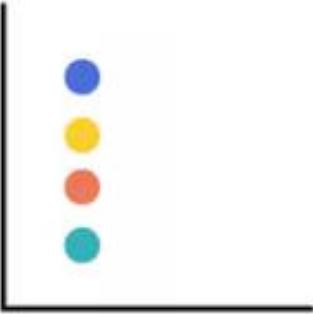
Variables cuantitativas

Tipos de datos

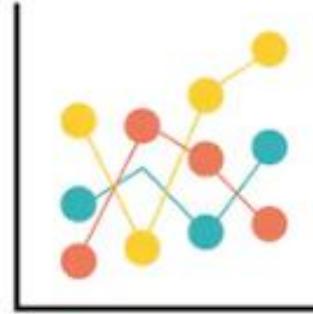
Según el momento y tipo de estudio



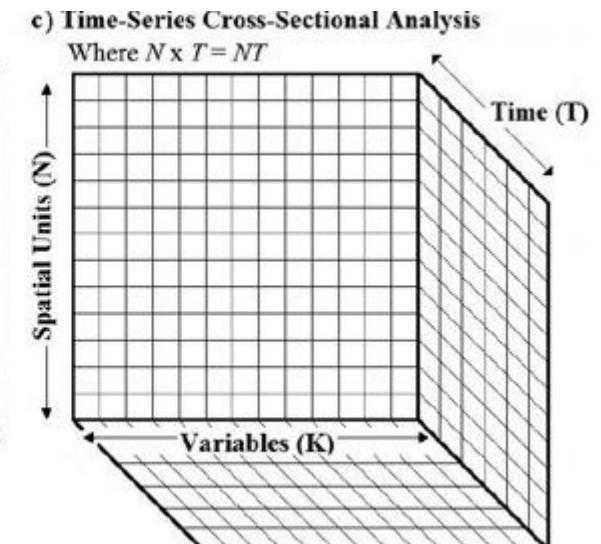
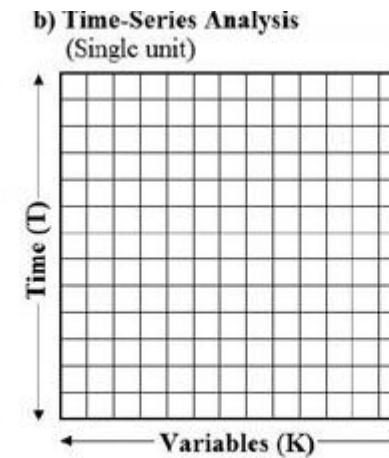
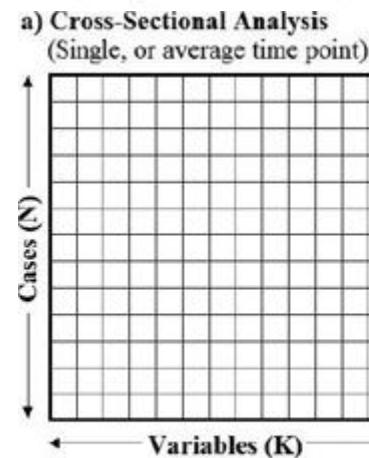
Time series data



Cross-Sectional data



Panel Data
(Longitudinal Data)



Gráficos comunes

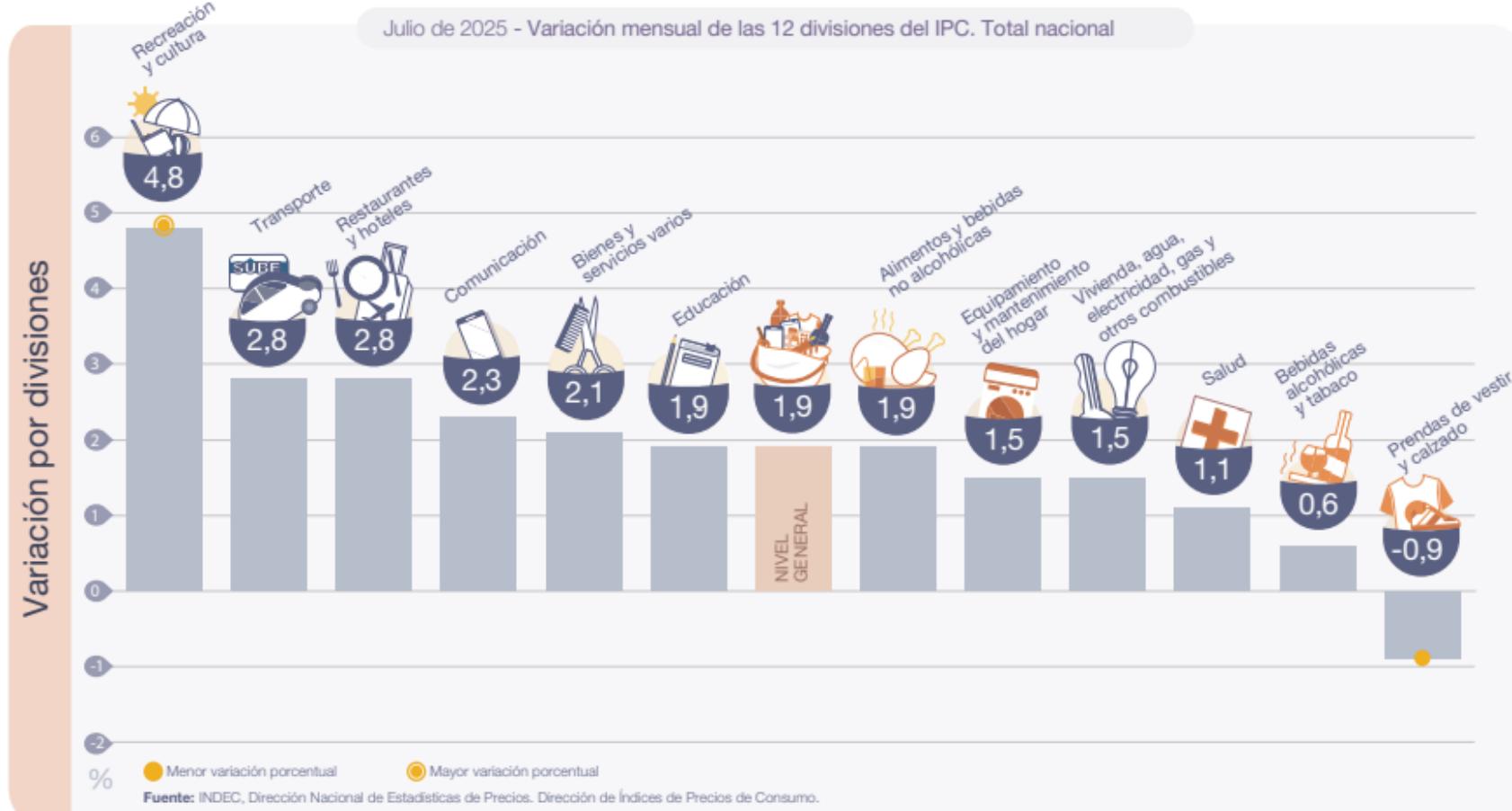
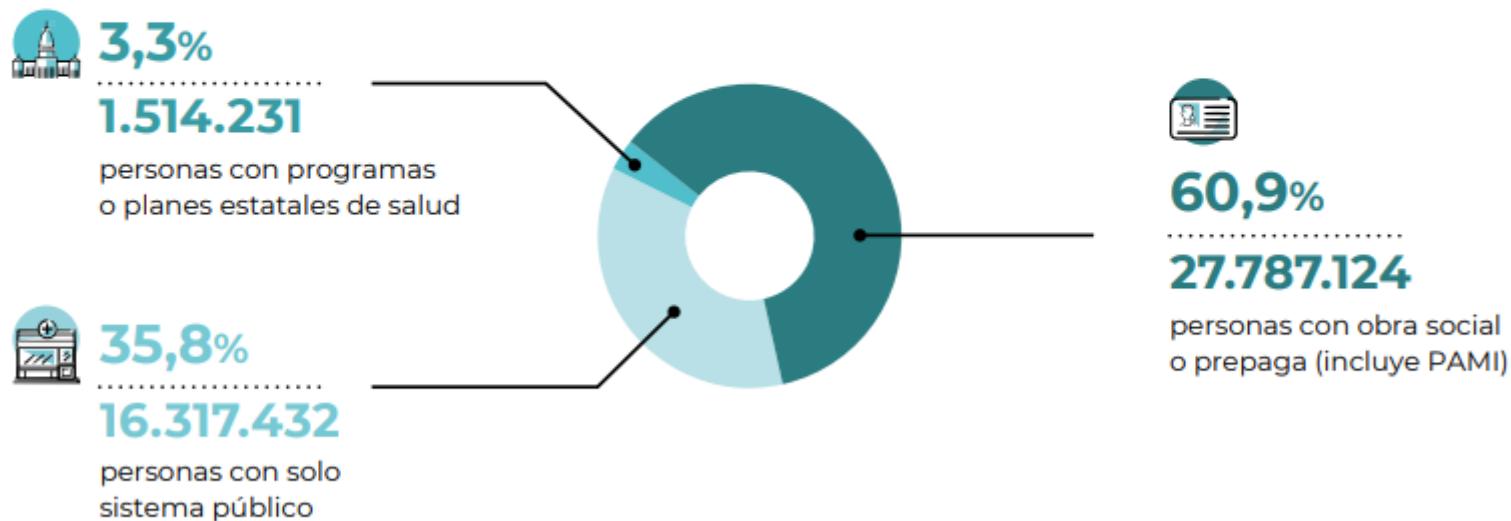


Gráfico de barras

Gráficos comunes

- 7.1 Distribución porcentual de la población, según tipo de cobertura de salud.
Total del país. Año 2022



Fuente: INDEC, Censo Nacional de Población, Hogares y Viviendas 2022.

Gráfico de torta (piechart)

Gráficos comunes

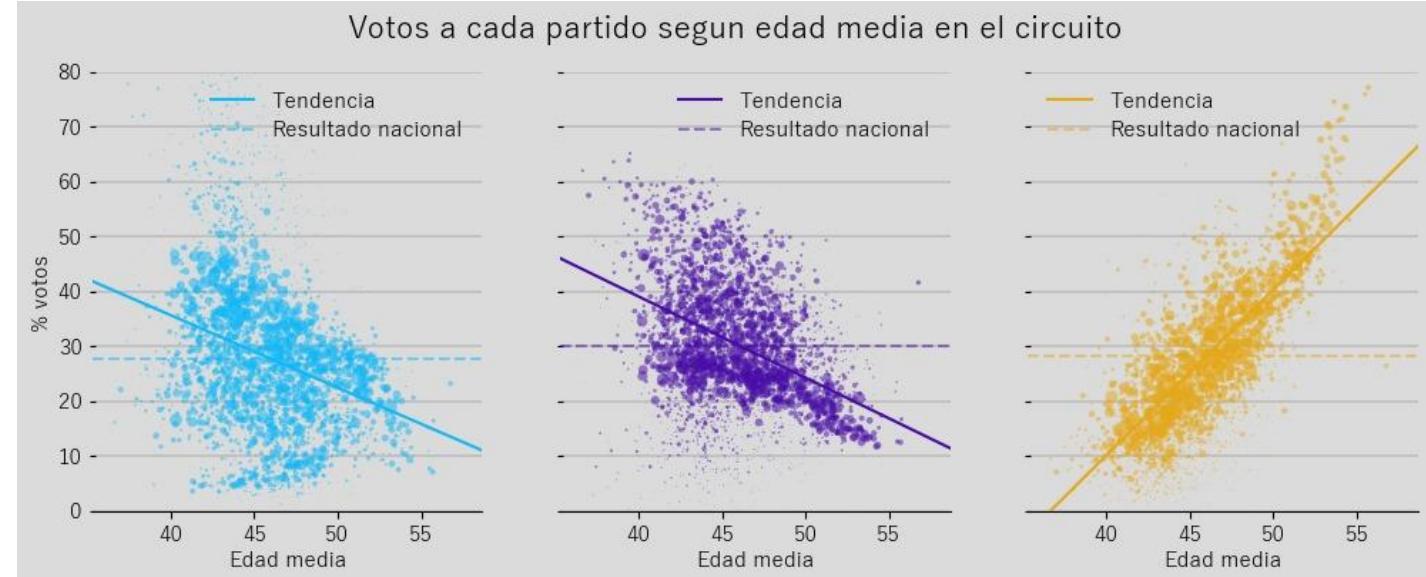
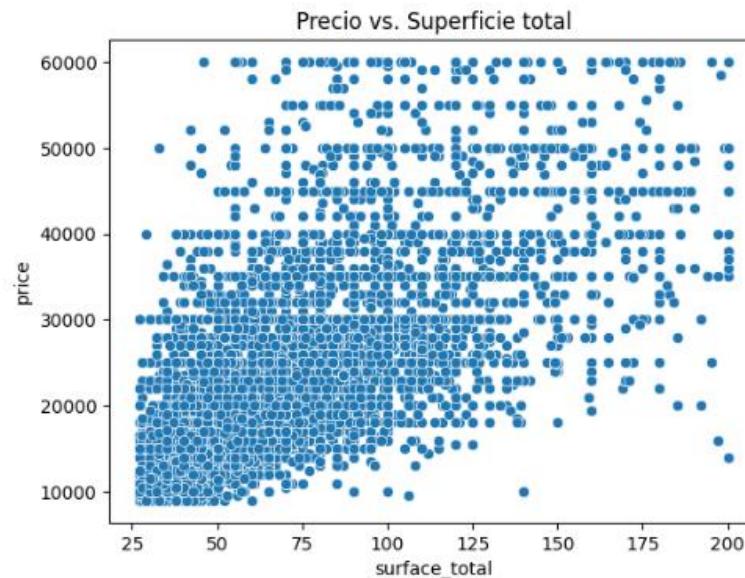


Gráfico de dispersión (scatterplot)

Gráficos comunes

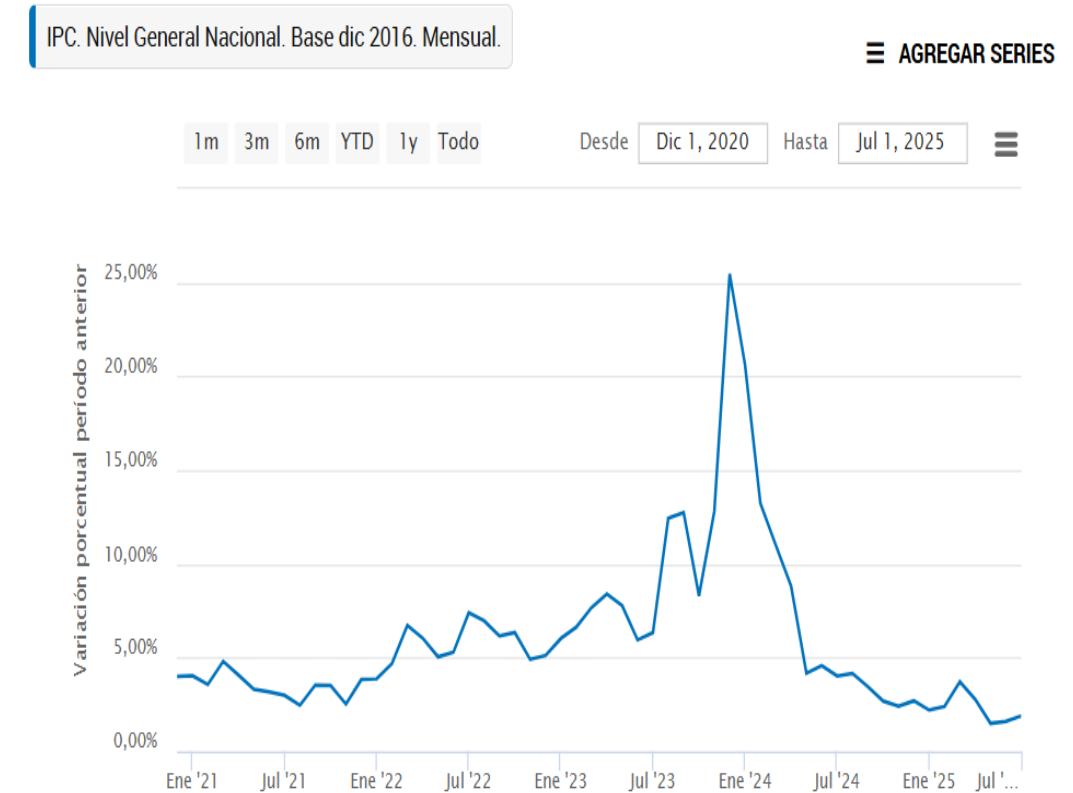
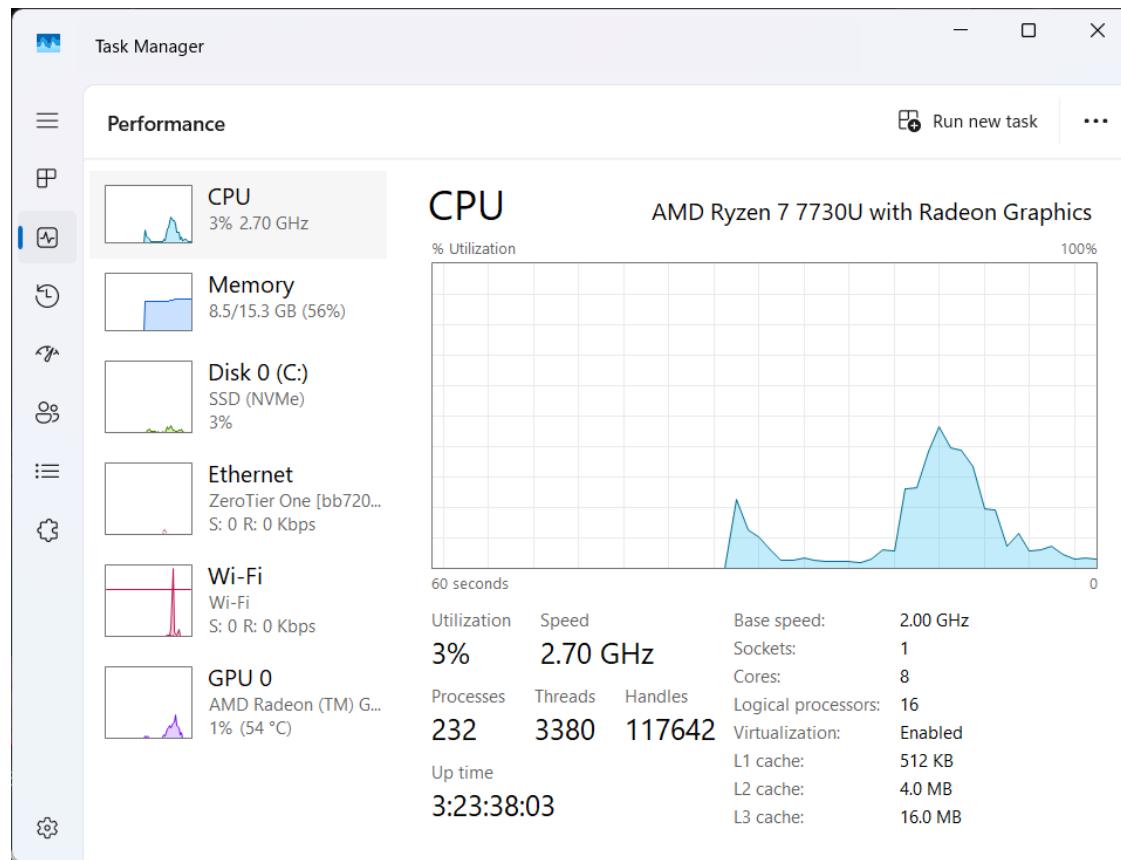
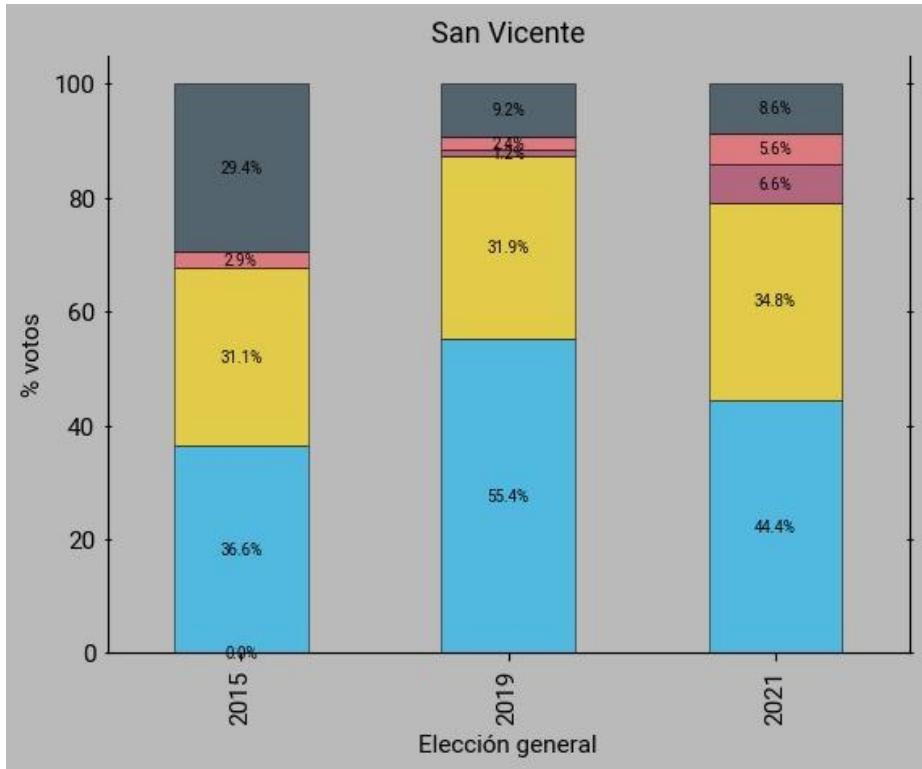
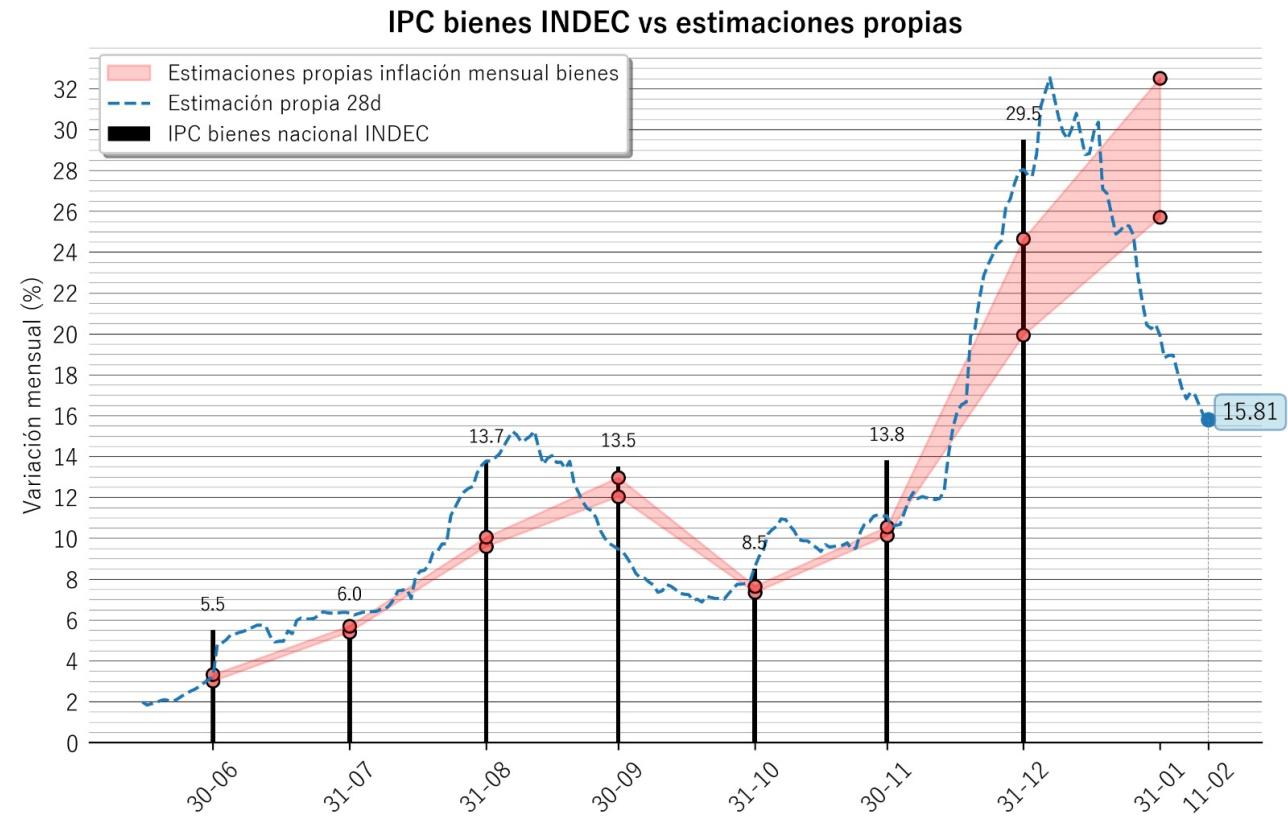


Gráfico de línea (lineplot)

Gráficos más complejos

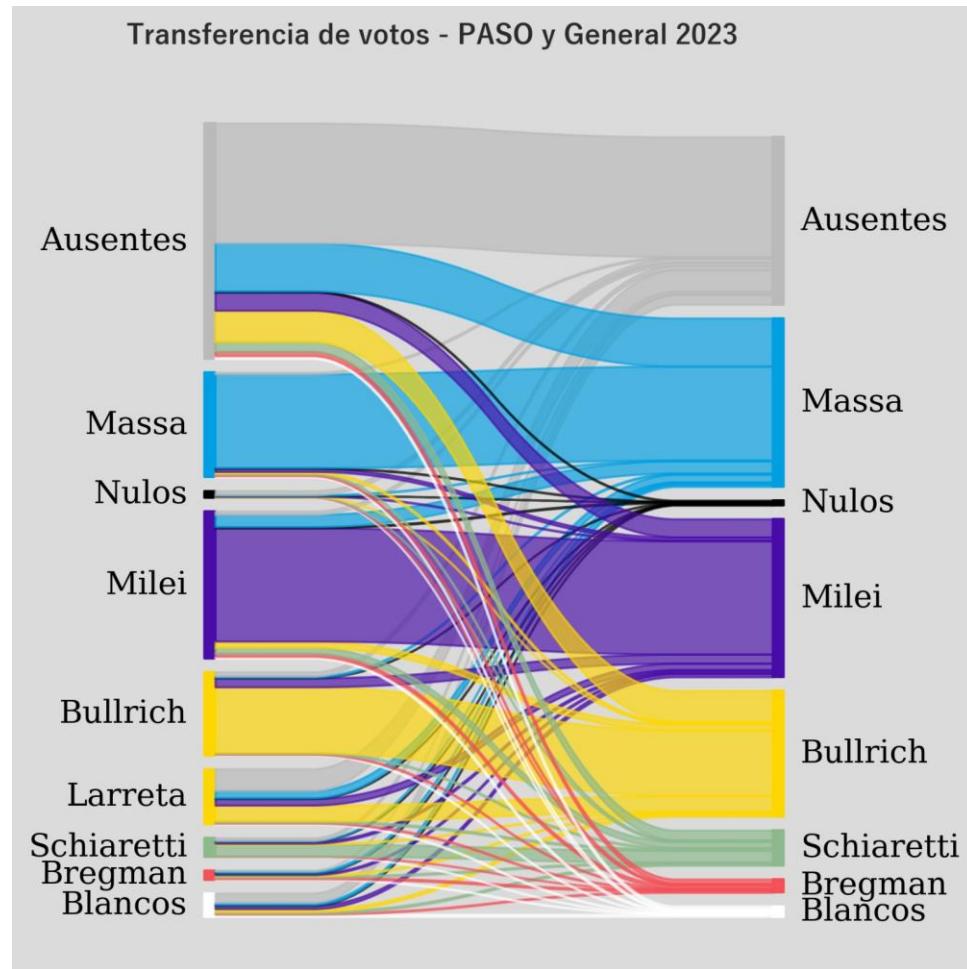


Barras apiladas

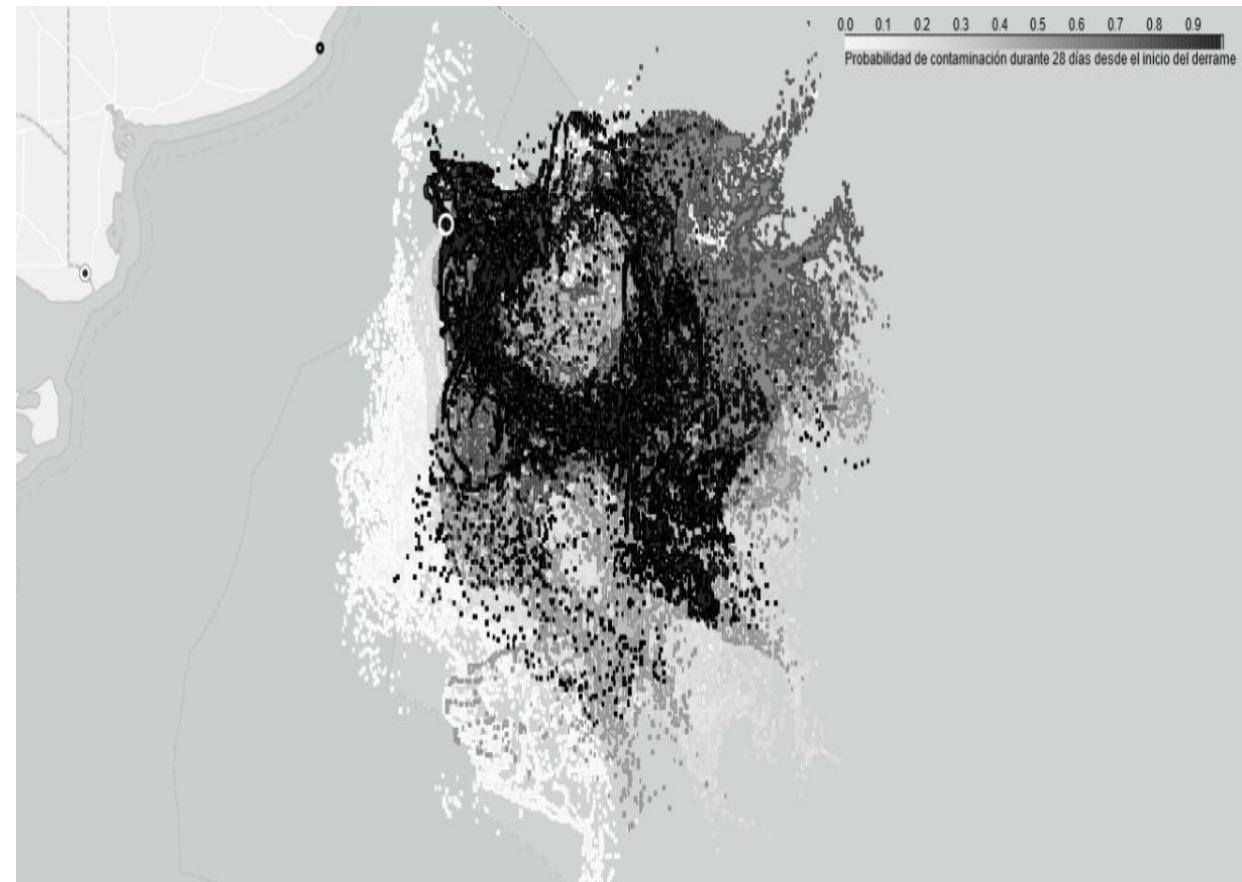


Barras + línea + área

Gráficos más complejos

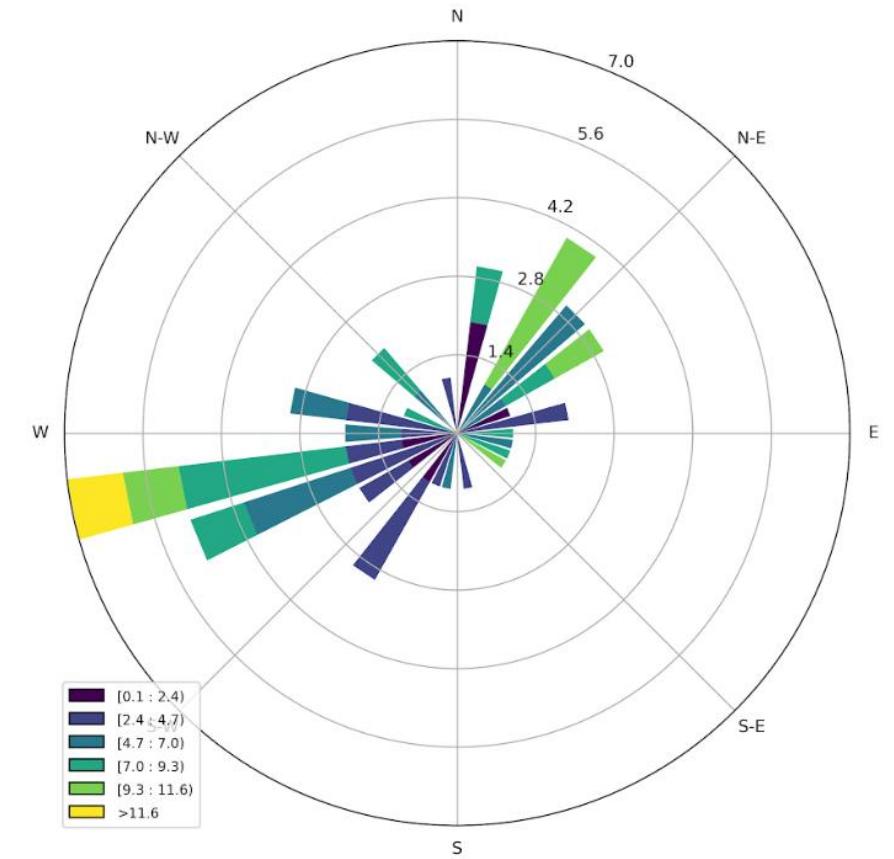
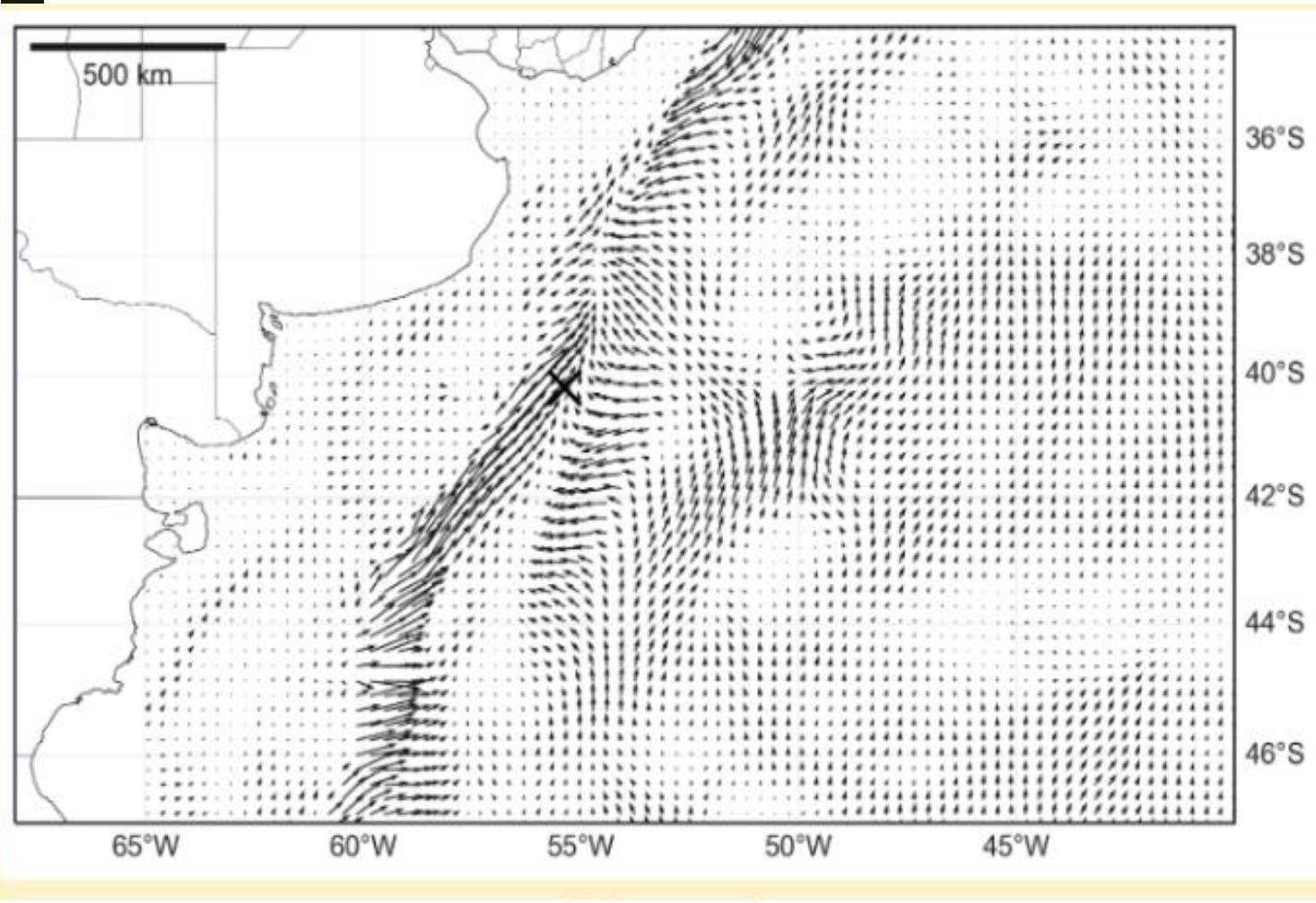


Sankey (transferencia)



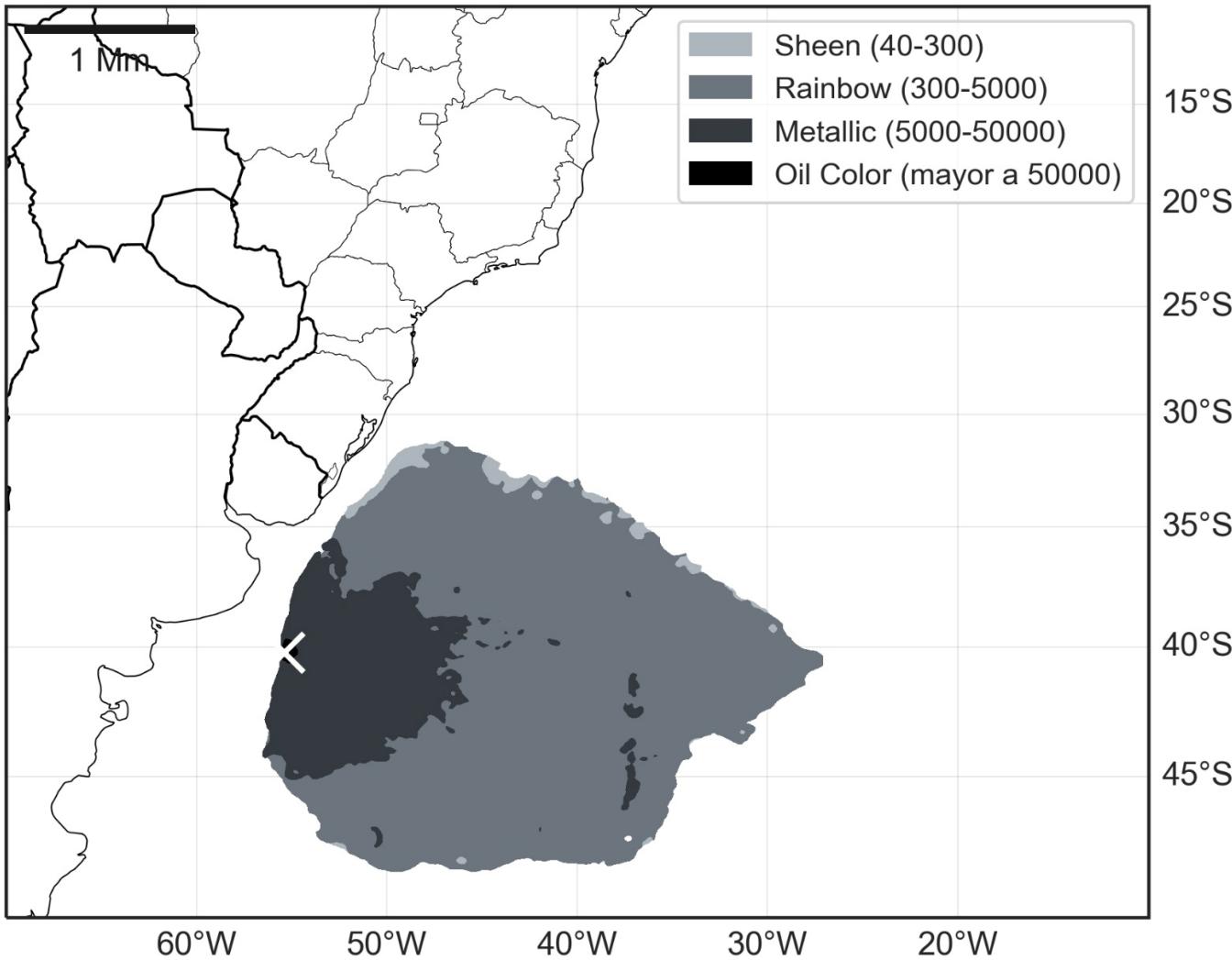
Scatterplot (sobre un mapa)

Gráficos más complejos



Fuente: Elaboración propia.

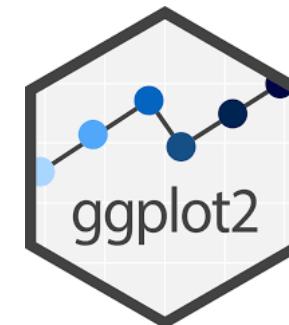
Gráficos más complejos



Las posibilidades para hacer gráficos son prácticamente infinitas, va a depender del tipo de datos que tengamos, y qué es lo que queremos mostrar

¿Y cómo los hacemos?

Hay infinidad de herramientas para hacer gráficos



Además, cada lenguaje de programación tiene su paquete para graficar

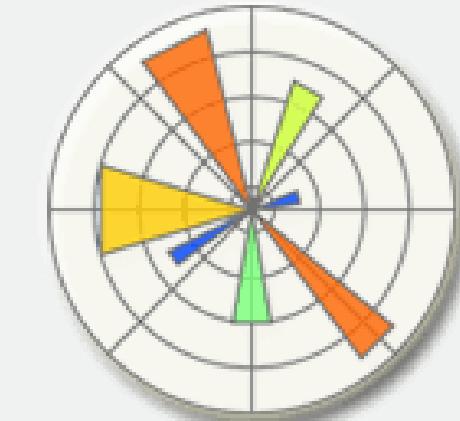
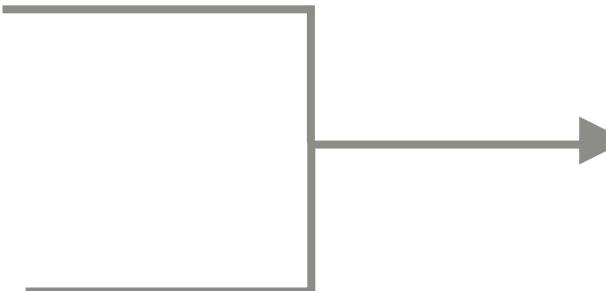
¿Y cómo los hacemos en Python?

Matplotlib, Seaborn y Pandas



seaborn

pandas



Matplotlib

Se puede usar matplotlib directamente, o generar objetos de matplotlib desde otros paquetes (en nuestro caso, seaborn y pandas)

Ejemplo: wine dataset

type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	pH	sulphates	alcohol	quality
white	8.4	0.17	0.31	6.7	0.038	29	132	3.1	0.32	10.6	7
white	6	0.18	0.31	1.4	0.036	14	75	3.34	0.58	11.1	8
white	8.6	0.36	0.26	11.1	0.03	43.5	171	3.03	0.49	12	5
white	6.9	0.4	0.17	12.9	0.033	59	186	3.08	0.49	9.4	5
red	6.8	0.785	0	2.4	0.104	14	30	3.52	0.55	10.7	6
red	10.8	0.29	0.42	1.6	0.084	19	27	3.28	0.73	11.9	6
white	7.1	0.21	0.32	2.2	0.037	28	141	3.2	0.57	10	7
white	6.1	0.17	0.21	1.9	0.09	44	130	3.07	0.41	9.7	5
white	9.2	0.28	0.46	3.2	0.058	39	133	3.14	0.58	9.5	5
red	11.5	0.59	0.59	2.6	0.087	13	49	3.18	0.65	11	6



```
1 import pandas as pd
2 import seaborn as sns
3 import matplotlib.pyplot as plt
4
5 df = pd.read_csv(r"wine.csv", sep=";")
```

Importamos los tres paquetes y cargamos el dataset (no siempre usamos todos, depende que querramos hacer)

Matplotlib

- Principal paquete para visualizar en Python
- En general se puede usar de dos formas:

Interactuando con el “gráfico actual”

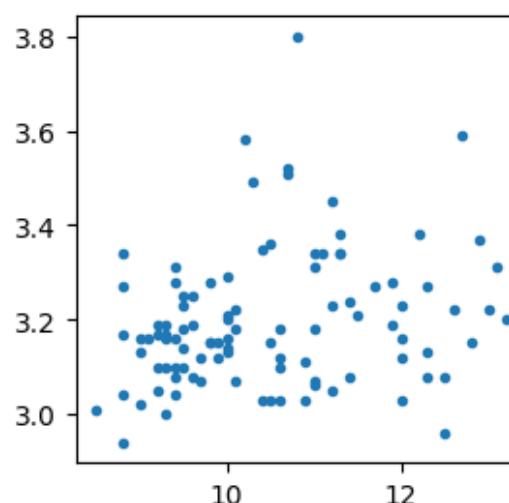


```
1 plt.plot(df['alcohol'], df['pH'], '.')
```

Generando un objeto con el que interactuar



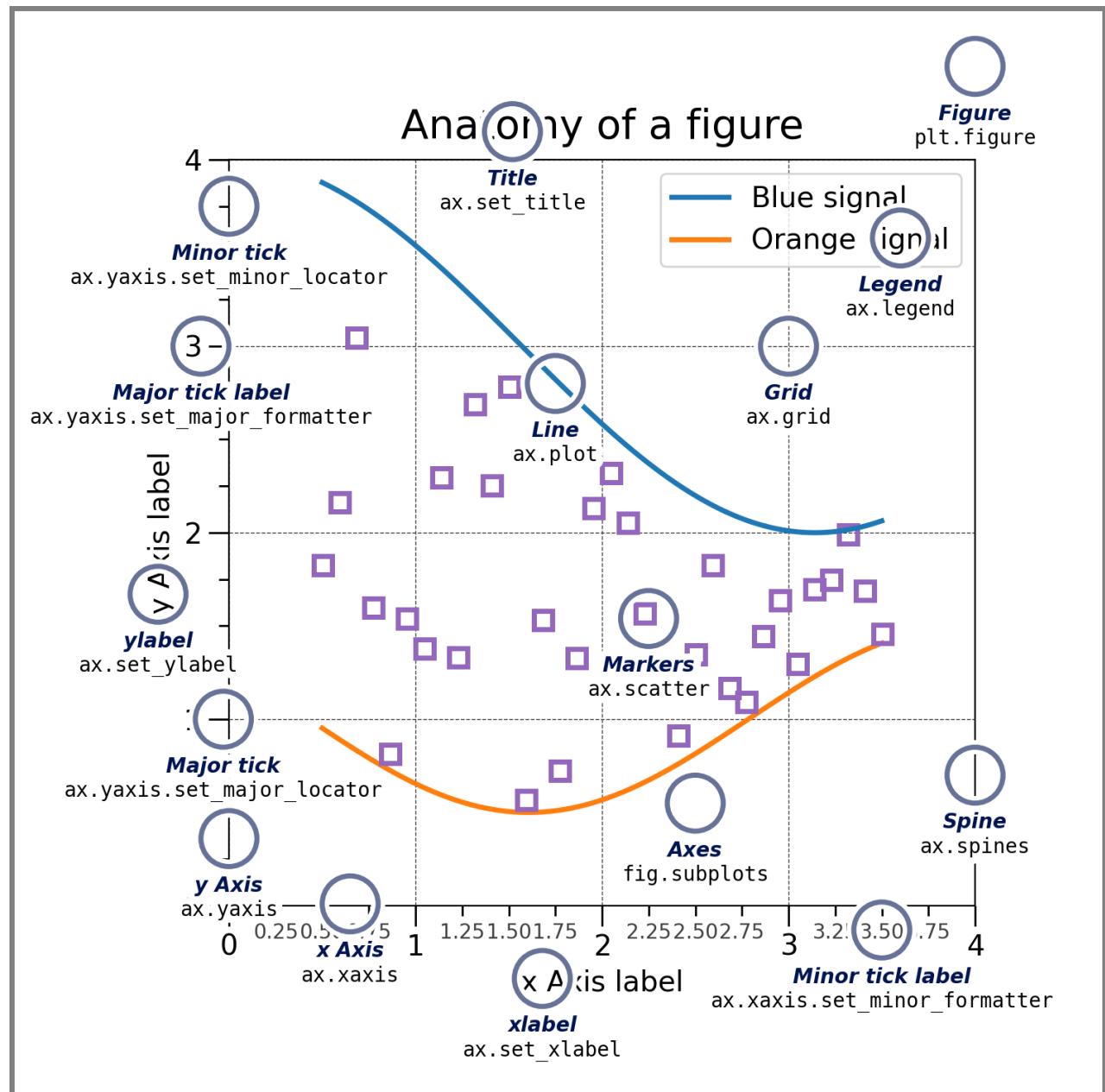
```
1 fig, ax = plt.subplots()  
2 ax.plot(df['alcohol'], df['pH'], '.')
```



Matplotlib

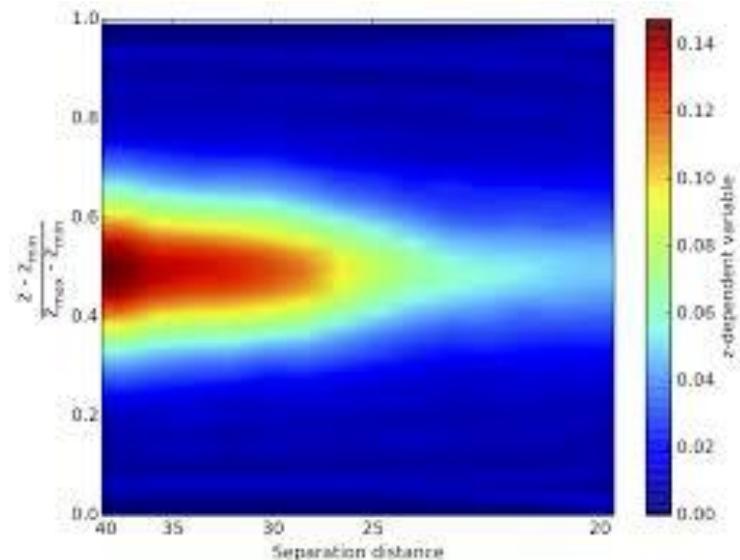
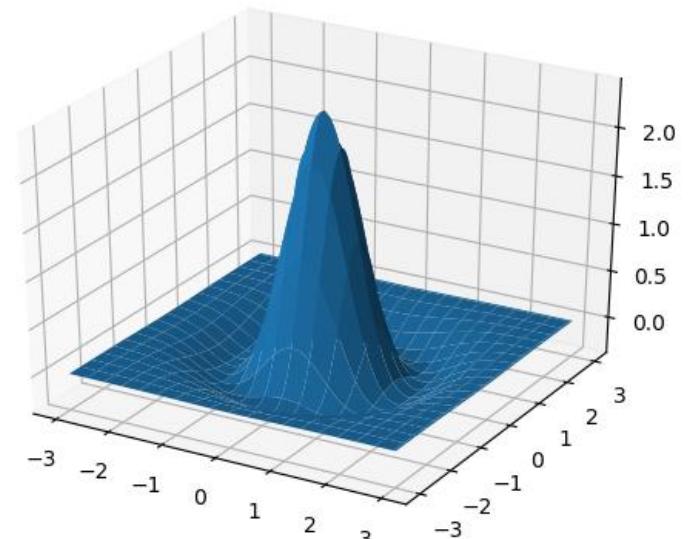
Un gráfico es una figura (plt.figure) y uno o muchos recuadros (plt.axes), a su vez compuesto por muchos objetos.

Al igual que creando gráficos, se puede usar desde el objeto del gráfico (ax.set_title) o desde el "gráfico actual" (plt.title)



Matplotlib

- `plt.plot` es la función general para gráficos de x, y. Hay otras para cosas más específicas (`scatter`, `pie`, `bar`, `boxplot`, etc.)
 - Cada una tiene sus parámetros.
Googleen y revisen la documentación!
- Se puede hacer casi cualquier gráfico con matplotlib, pero no siempre es fácil
 - De hecho, en general es más fácil hacerlos con cosas que “usan” matplotlib, sin usarlo nosotros directamente

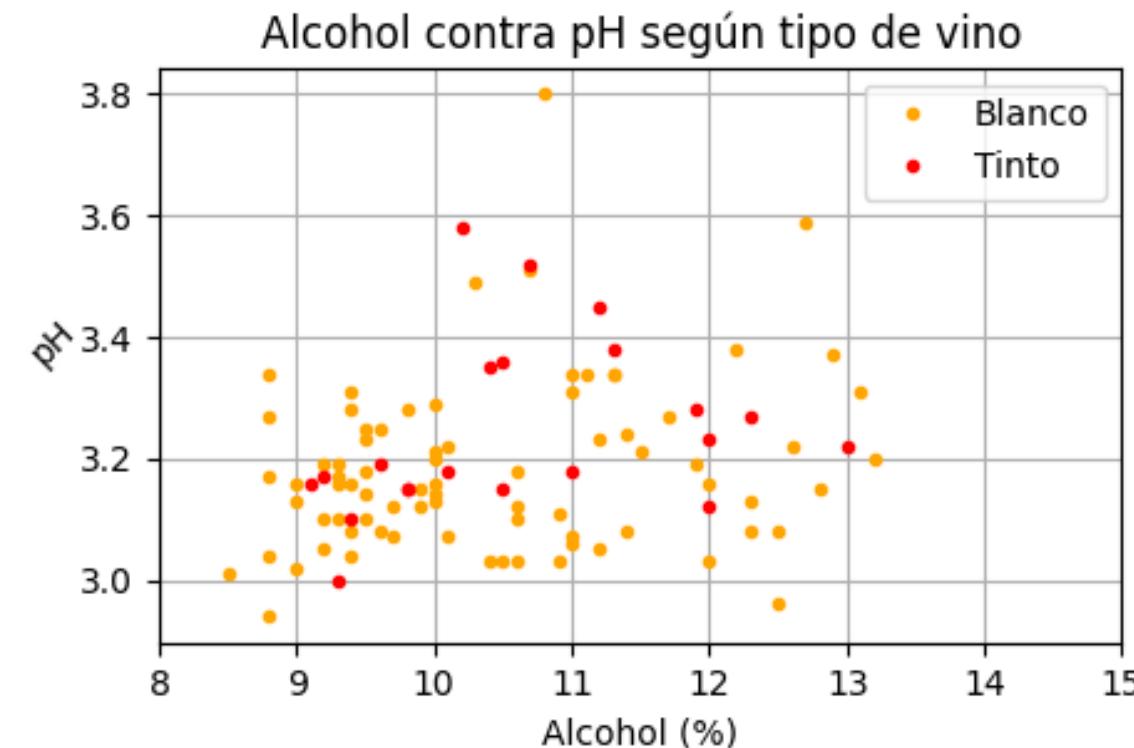


Matplotlib

“Mejoremos” el gráfico que hicimos antes



```
1 blancos = df[df['type'] == 'white']
2 rosados = df[df['type'] == 'red']
3
4 plt.plot(blancos['alcohol'], blancos['pH'], '.', color="orange", label="Blanco")
5 plt.plot(rosados['alcohol'], rosados['pH'], '.', color="red", label="Tinto")
6 plt.xlabel("Alcohol (%)")
7 plt.ylabel("pH", rotation=45)
8 plt.title("Alcohol contra pH según tipo de vino")
9 plt.grid()
10 plt.legend()
11 plt.xlim(8, 15)
12
```

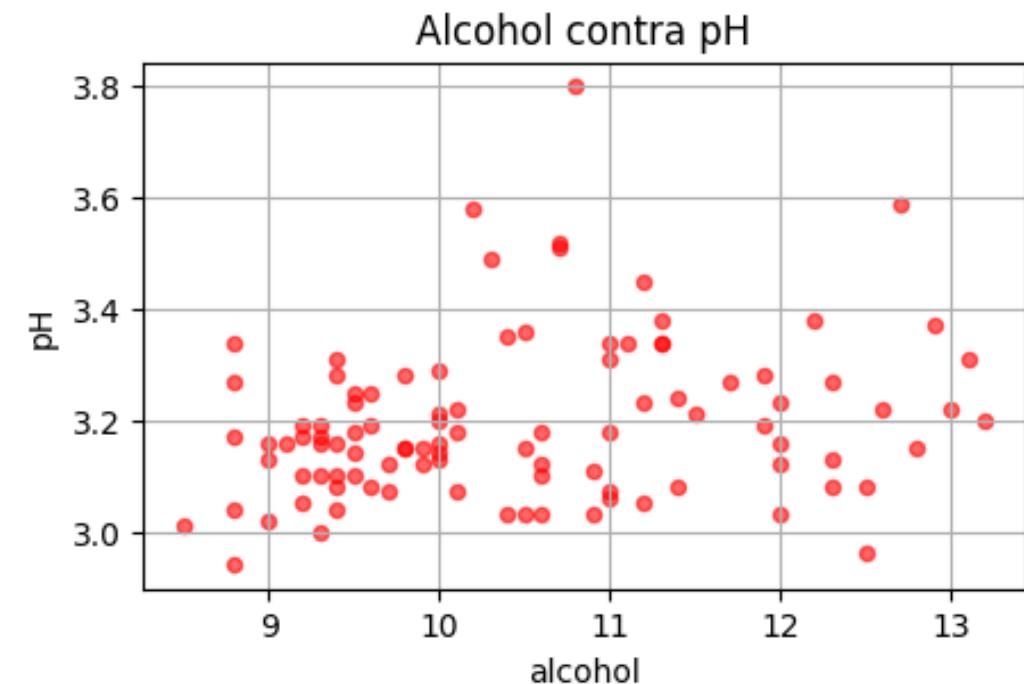


Tuvimos que generar dos dataframes, dos .plot , y acomodar las cosas para cada uno. ¿Se podrá más fácil?: Seaborn o Pandas

Pandas

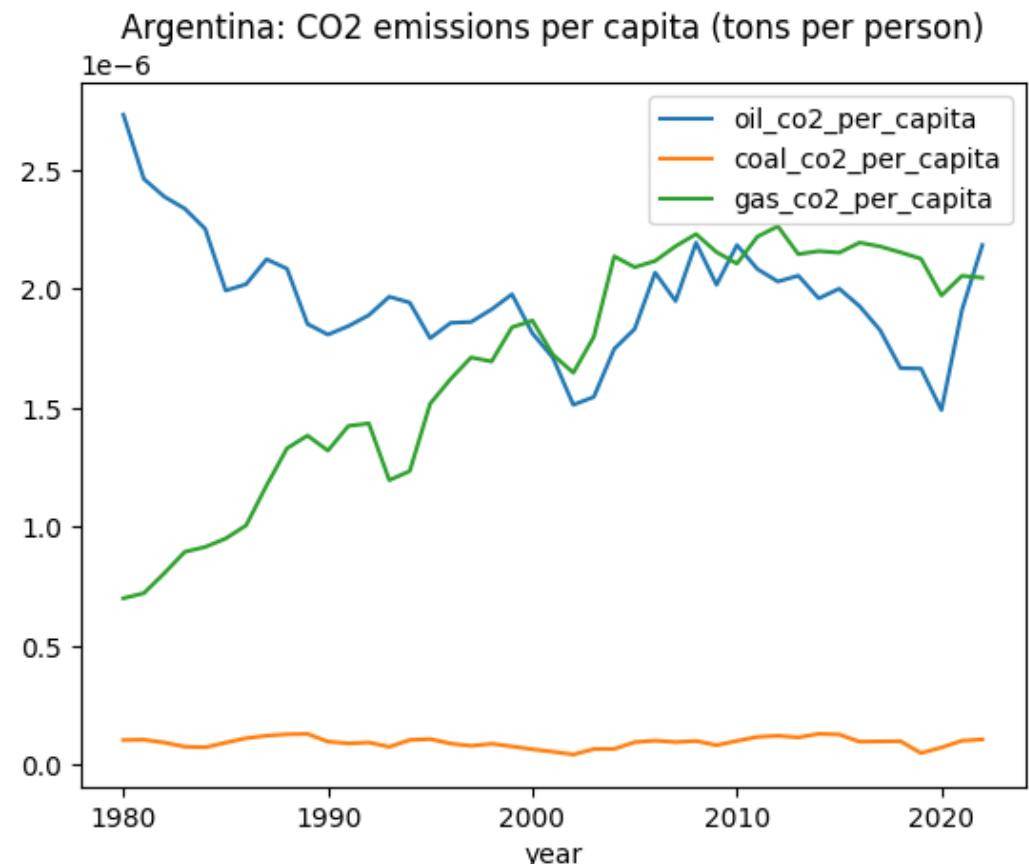
- Ya lo conocen de todo lo que hicieron al principio de la materia
- No es un paquete hecho “para graficar”
 - Por ejemplo, colorear según otra columna no es fácil
- Se usa siempre .plot, cambiando el tipo de gráfico según lo que se ponga en kind
- Maneja bien los nan (NULL)

```
df.plot(  
    kind="scatter",  
    x="alcohol",  
    y="pH",  
    title="Alcohol contra pH",  
    legend=True,  
    grid=True,  
    color="red",  
    alpha=0.6,  
)
```



Pandas

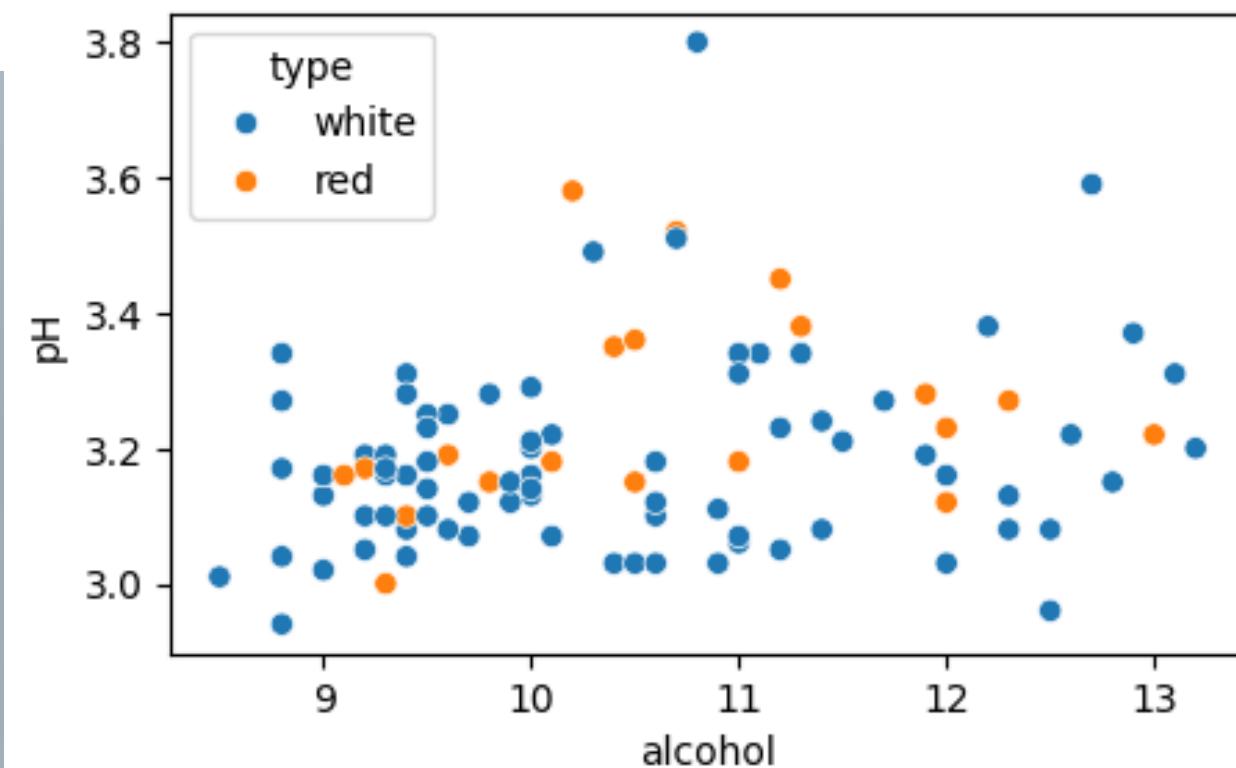
- ¿Lo notan parecido?
 - Pandas “usa” matplotlib para generar el gráfico (de hecho, devuelve un axes con el que después podemos interactuar)



Seaborn

- Está diseñado para gráficos estadísticos y análisis de datos
- También usa matplotlib para graficar
- Una de las principales ventajas es que permite el parámetro “hue”, que separa según la variable que le asignas:

```
1 sns.scatterplot(  
2     data=df,  
3     x="alcohol",  
4     y="pH",  
5     hue="type"  
6 )
```

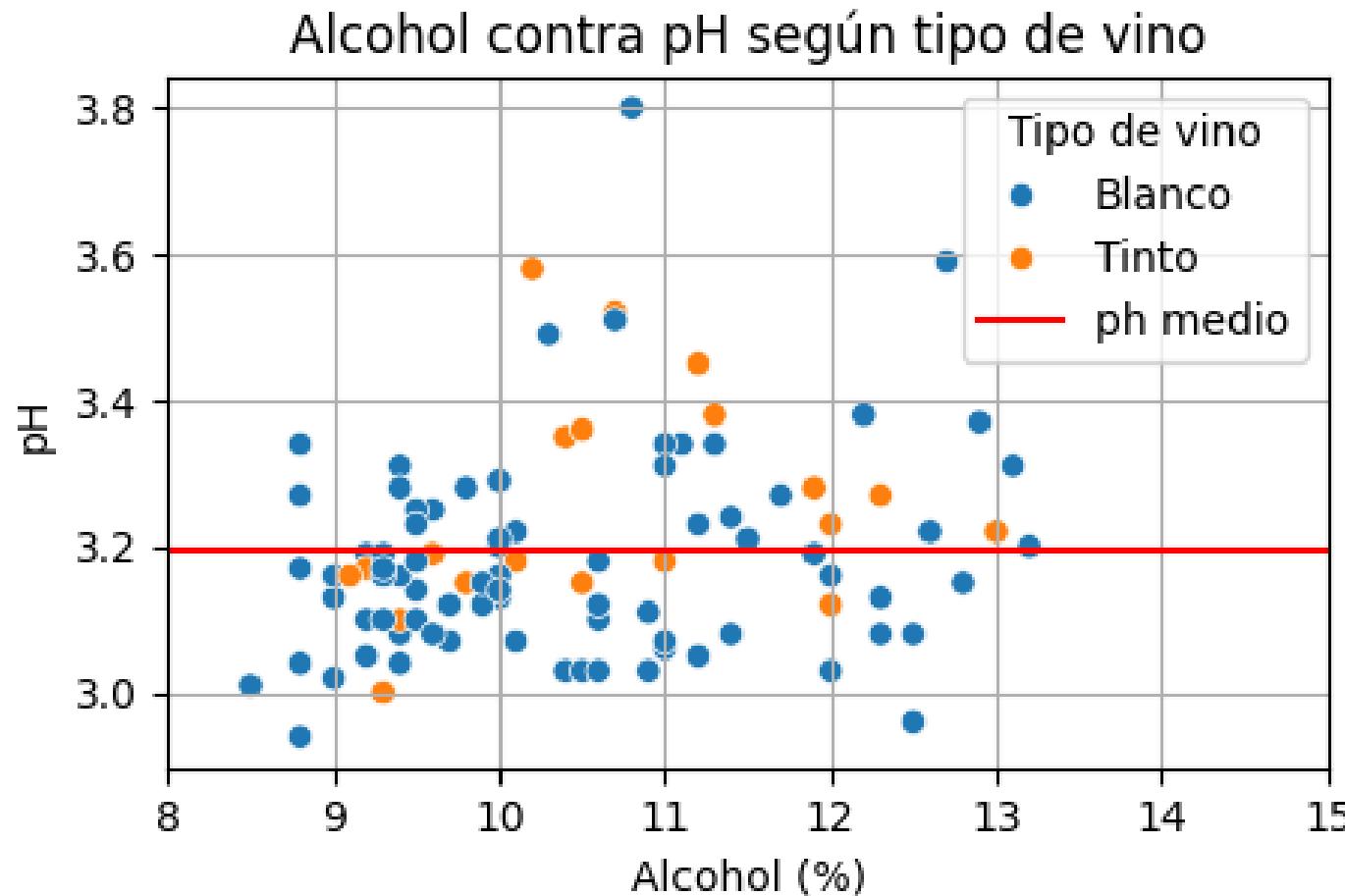


Juntando todo

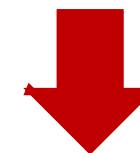
- La ventaja de que todo, en el fondo, use matplotlib, es que se puede hacer un gráfico con el paquete que les guste, y después retocarlo con los métodos de matplotlib

```
● ● ●  
1 ax = sns.scatterplot(data=df, x="alcohol", y="pH", hue="type")  
2 ax.axhline(y=df["pH"].mean(), color="red", label="ph medio")  
3 ax.set_title("Alcohol contra pH según tipo de vino")  
4 ax.set_xlim(8, 15)  
5 ax.grid()  
6 handles, labels = ax.get_legend_handles_labels()  
7 ax.legend(handles, ["Blanco", "Tinto", "ph medio"], title="Tipo de vino")
```

Juntando todo



Así como seaborn, hay otros paquetes que usan matplotlib para hacer gráficos específicos (mapas, matrices, sankeys, heatmaps, etc.)



Aprendan a usarlo!



```
1
2  ax = sns.boxplot(
3      r_plt,
4      x="n_features",
5      y="roc_auc",
6      showmeans=True,
7      meanprops={"marker": "o", "markerfacecolor": "white", "markeredgecolor": "black"},
8      label="Selected features",
9  )
10 ax.set_title(f"Internal validation - FASTAI")
11 ax.spines["top"].set_visible(False)
12 ax.spines["right"].set_visible(False)
13 ax.set_xlabel("Number of features")
14 ax.set_ylabel("AUC")
15 plt.ylim(0.6, 1)
16 plt.grid(axis="y", linestyle="--", alpha=0.7)
17
18
19 sns.boxplot(
20     FASTAI_benchmark,
21     x="n_features",
22     y="roc_auc",
23     ax=ax,
24     color="red",
25     fliersize=0,
26     label="Decipher features",
27     boxprops=dict(alpha=.7),
28     showmeans=True,
29     meanprops={"marker": "o", "markerfacecolor": "grey", "markeredgecolor": "black"},
30
31 )
32
33 ax.set_xticks(range(2, 26, 3))
34
35
36 # benchmark
37 ax.axhline(0.76, color="red", linestyle="--", label="Erho et al. 2013")
38 plt.legend()
```

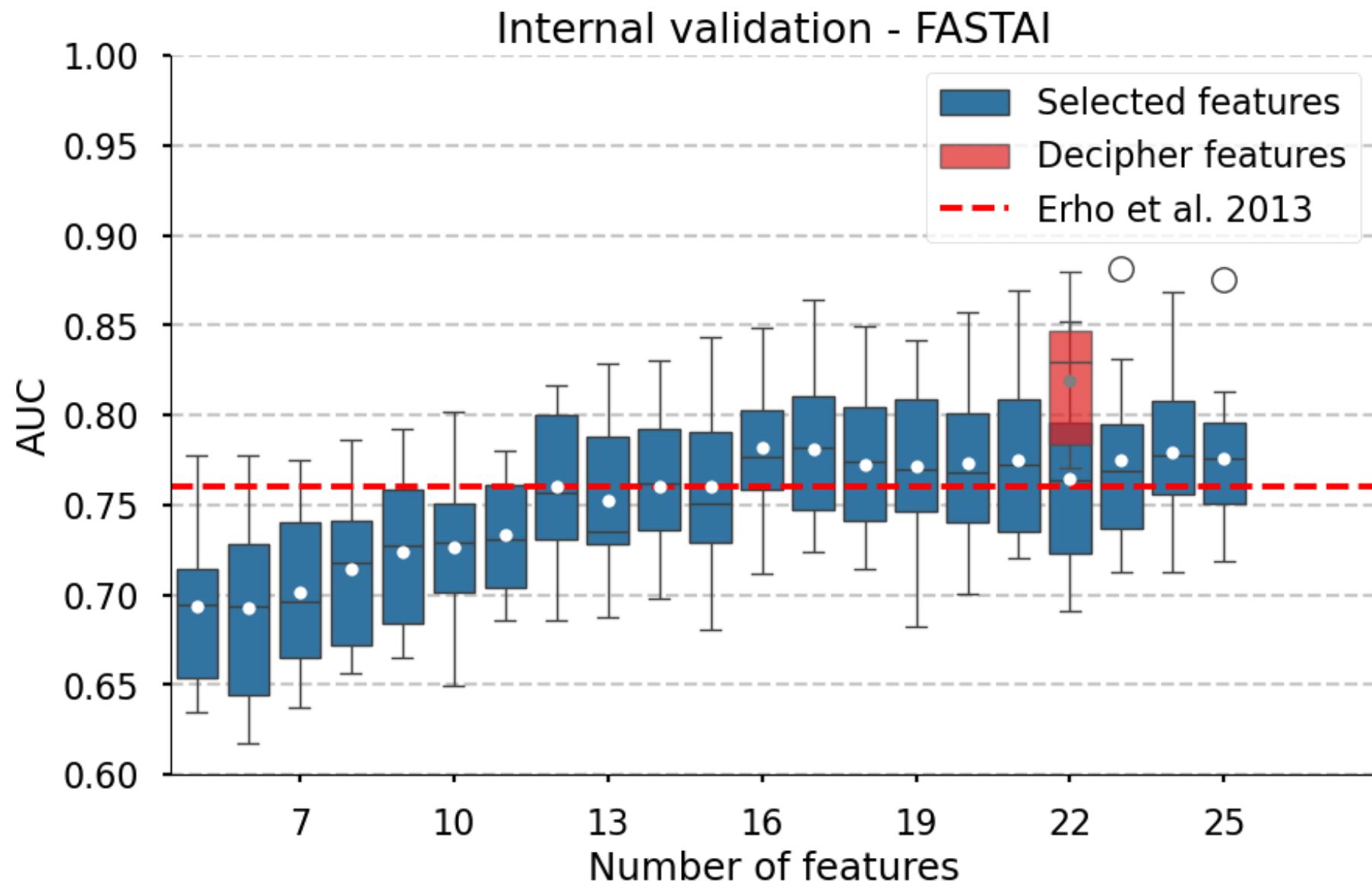
Seaborn

Matplotlib

(y antes se usó pandas
para generar los df)

Seaborn

Matplotlib



¿Cuándo uso cada cosa?

(no hay una respuesta única, les cuento lo que a mi me sirve)

Matplotlib

- Si venís con datos por fuera de pandas (listas, outputs de funciones)
- Después de hacer un gráfico desde otro paquete, para ponerlo lindo

Pandas

- Si venís desde un dataframe, y es un gráfico “común” (que usa los datos que ya tenés en el df)
- Suelen ser más simples en código, así que conviene probar y ver si sale fácil, y sino a otro paquete

Seaborn

- Si querés separar por una variable externa (agregar una tercer variable como color de los datos)
- Si querés hacer un gráfico que use cálculos derivados (i.e. boxplot)
- Si no está implementado en los otros paquetes

A dibujar!



A dibujar!

Dataset: CO₂ Emissions Across Countries.

Datos de panel, 1980-2022 (está en el campus)

Name	# year	iso_code	# population	# gdp	# co2	# coal_co2	# oil_co2	# gas_co2	# methane
Algeria	2022	DZA	45477391.0	595820000000.0	184.558	0.775	57.407	100.626	96.622
Argentina	2022	ARG	45407904.0	854914000000.0	204.081	4.78	99.135	92.908	129.569
Australia	2022	AUS	26200987.0	1344250000000.0	384.362	146.069	133.807	80.412	126.538
Austria	2022	AUT	9064679.0	398815000000.0	61.489	11.666	30.734	15.967	9.886
Azerbaijan	2022	AZE	10295307.0	175371000000.0	40.391	0.0	12.148	26.791	23.905
Bangladesh	2022	BGD	169384890.0	858144000000.0	113.863	16.915	38.093	57.303	111.407
Belarus	2022	BLR	9173241.0	175774000000.0	59.384	4.246	16.348	35.594	17.243
Belgium	2022	BEL	11641813.0	489678000000.0	89.002	11.941	41.814	31.087	12.721
Brazil	2022	BRA	210306411.0	3187410000000.0	483.841	56.078	319.935	60.085	601.386
Bulgaria	2022	BGR	6825863.0	138390000000.0	46.966	24.577	13.793	5.14	9.941

A dibujar!

Siguiendo visualizaciones_1.py: (y la [documentación de df.plot](#))

1. Abrir el dataset con Pandas
2. Empecemos analizando datos para el último año disponible (2022)
 - a) Crear **una copia** del dataframe, conservando solo 2022 (`df_2022 = df[filtro].copy()`)
3. Hacer un gráfico para mostrar la relación entre GDP (PBI) y emisiones de CO2
 1. Ciertos países están varios órdenes de magnitud por encima de la mayoría. Ver si pasando los ejes a escala logarítmica no queda más legible
 2. Ponerle título al gráfico y a los ejes
4. Ídem para emisiones de CO2 y población

¿Qué explica mejor el aumento en emisiones, más población o más PBI?

A dibujar!

Siguiendo visualizaciones_1.py:
(y la [documentación de df.plot](#))

5. Generar df_oil, usando la función agrupa_otros del script, definiendo el umbral en 3%
 - Esto junta los países que consumen menos del 3% en un mismo grupo, así evitamos tener 200 países en el gráfico
6. Hacer dos gráficos distintos que muestre los países que más CO2 emiten por consumo de petróleo
7. Repetir lo anterior (con el tipo de gráfico que les parezca mejor) para gas y carbón (coal)

¿Cuál de las fuentes de energía está más repartida globalmente? ¿Por qué creen que es?

A dibujar!

Siguiendo visualizaciones_1.py:
(y la [documentación de df.plot](#))

8. Pasamos a un corte longitudinal para Argentina
 - Generar df_arg, con los datos de Argentina para los años que esté disponible
9. Usando las variables de antes (y GDP), generarlas per capita, haciendo la división por population
10. Graficar la evolución per capita de las tres fuentes de energía (en un mismo gráfico)
11. Ahora graficarlas como porcentaje del total de emisiones
 - Hacerlo en barras y en líneas. ¿Cuál es más legible?

Ejercicio: ver si alguna de las fuentes de emisión se relaciona con años de alto PBI per capita

¿Cuándo usamos un gráfico de...?

Dispersión (scatter)

- Si quiero ver la relación entre dos variables (numéricas)
 - Pueden agregarse más variables, mediante color/tamaño del punto/3er eje

Línea

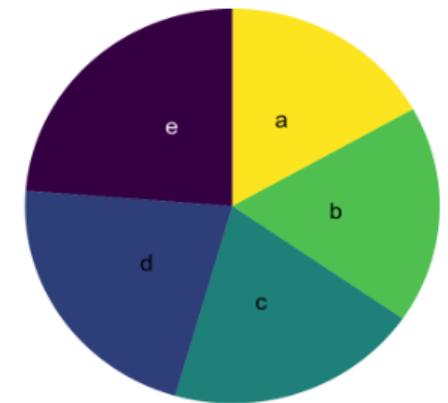
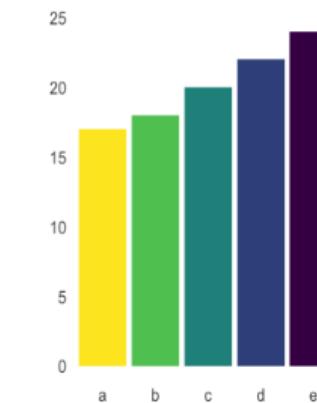
- Si quiero ver la evolución de una variable con respecto a otra (numéricas)
 - A diferencia del scatter, voy a tener un único valor en Y para cada X

Barras

- Si quiero comparar una numérica con una categórica
 - Se las puede apilar o agrupar para las categorías del eje X

Torta

- Si quiero comparar cantidad de cada categoría de una variable discreta
 - En general, un gráfico de barras tiene la misma info y es más fácil comparar categorías (evitar, salvo restricciones de espacio)



Atributos pre-atentivos

Atributos pre-atentivos

¿Cuántos 7 hay?

7	3	4	1	3	4	5	6	4	0
3	0	6	9	0	4	5	8	6	3
2	7	2	2	9	9	4	5	2	1
2	2	4	5	2	0	9	2	0	4
2	4	0	7	6	9	3	0	0	4
7	7	8	9	2	6	7	2	4	7
6	1	3	3	2	1	4	4	9	0
3	6	6	2	7	5	5	2	5	4
1	1	4	0	6	3	4	0	5	1
3	7	5	2	7	5	7	7	3	9
3	3	8	6	9	5	5	3	6	4
7	6	0	3	0	9	9	0	2	9
4	6	9	4	8	2	6	5	8	3
9	3	9	2	2	8	4	3	9	8
5	8	8	2	9	1	2	4	8	5
1	7	4	0	1	1	9	9	5	8

- Si sabemos que nos interesan los 7, podemos cambiar la visualización para que haga foco en ellos?

Atributos pre-atentivos

¿Acá se ve
más fácil?

7	3	4	1	3	4	5	6	4	0
3	0	6	9	0	4	5	8	6	3
2	7	2	2	9	9	4	5	2	1
2	2	4	5	2	0	9	2	0	4
2	4	0	7	6	9	3	0	0	4
7	7	8	9	2	6	7	2	4	7
6	1	3	3	2	1	4	4	9	0
3	6	6	2	7	5	5	2	5	4
1	1	4	0	6	3	4	0	5	1
3	7	5	2	7	5	7	7	3	9
3	3	8	6	9	5	5	3	6	4
7	6	0	3	0	9	9	0	2	9
4	6	9	4	8	2	6	5	8	3
9	3	9	2	2	8	4	3	9	8
5	8	8	2	9	1	2	4	8	5
1	7	4	0	1	1	9	9	5	8

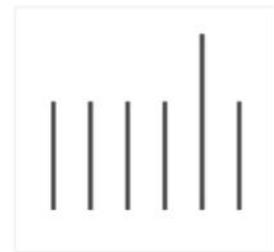
Atributos pre-atentivos

¿Y acá?

7	3	4	1	3	4	5	6	4	0
3	0	6	9	0	4	5	8	6	3
2	7	2	2	9	9	4	5	2	1
2	2	4	5	2	0	9	2	0	4
2	4	0	7	6	9	3	0	0	4
7	7	8	9	2	6	7	2	4	7
6	1	3	3	2	1	4	4	9	0
3	6	6	2	7	5	5	2	5	4
1	1	4	0	6	3	4	0	5	1
3	7	5	2	7	5	7	7	3	9
3	3	8	6	9	5	5	3	6	4
7	6	0	3	0	9	9	0	2	9
4	6	9	4	8	2	6	5	8	3
9	3	9	2	2	8	4	3	9	8
5	8	8	2	9	1	2	4	8	5
1	7	4	0	1	1	9	9	5	8

Atributos pre-atentivos

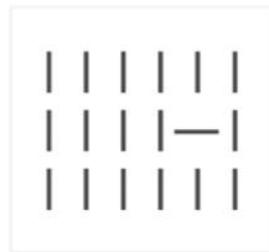
Se tarda menos en ver porque utiliza atributos que percibimos incluso antes que los datos propios:
Atributos pre-atentivos



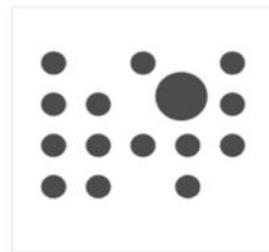
Length



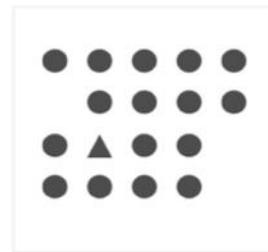
Width



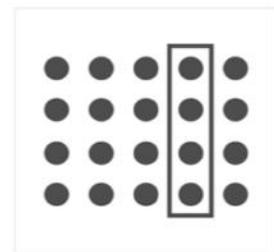
Orientation



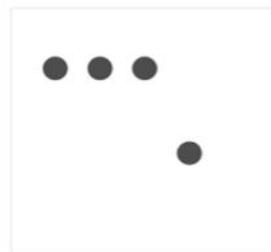
Size



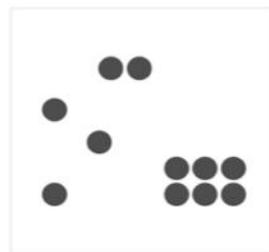
Shape



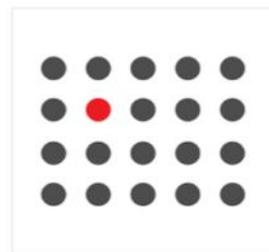
Enclosure



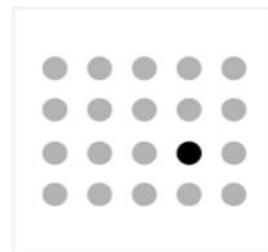
Position



Grouping



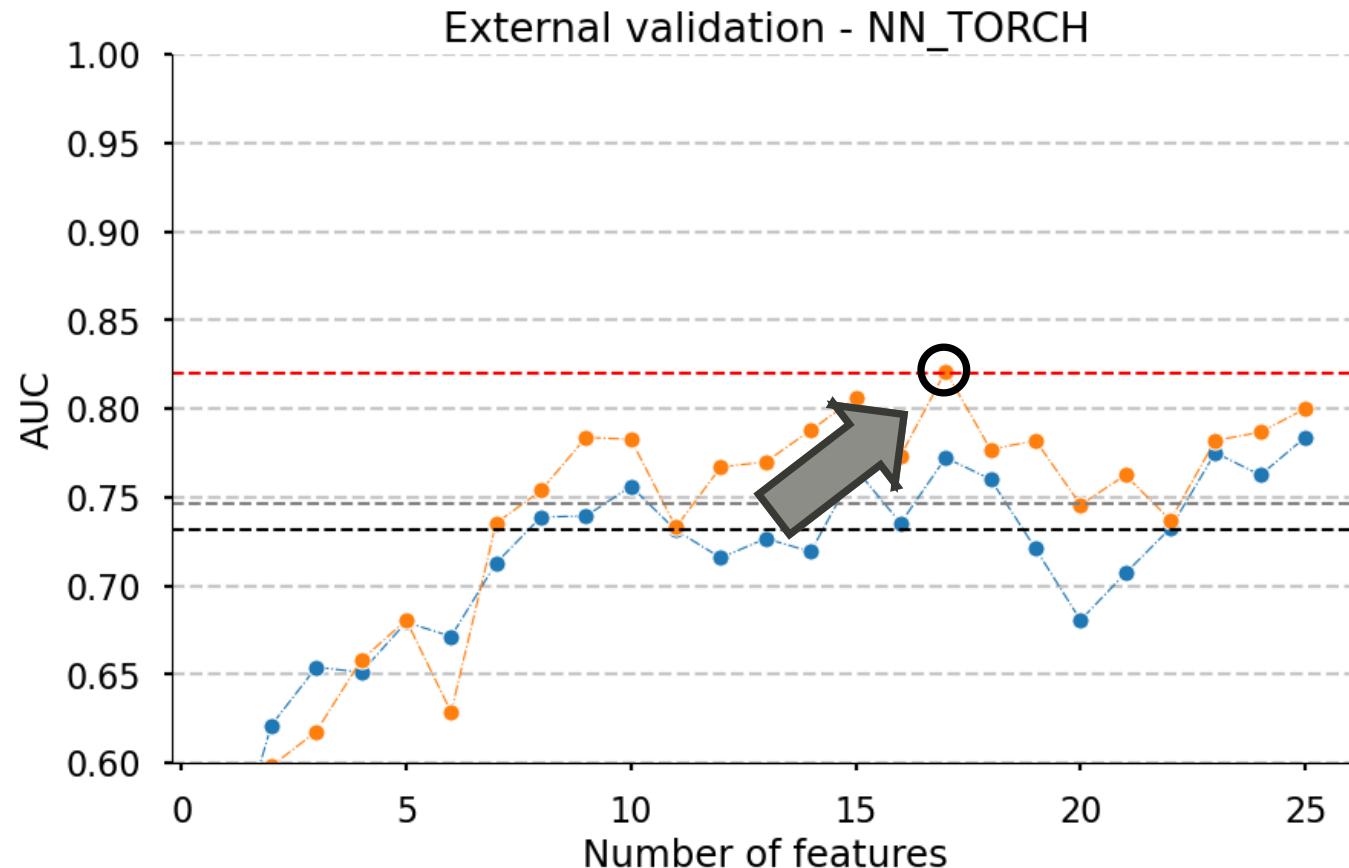
Color Hue



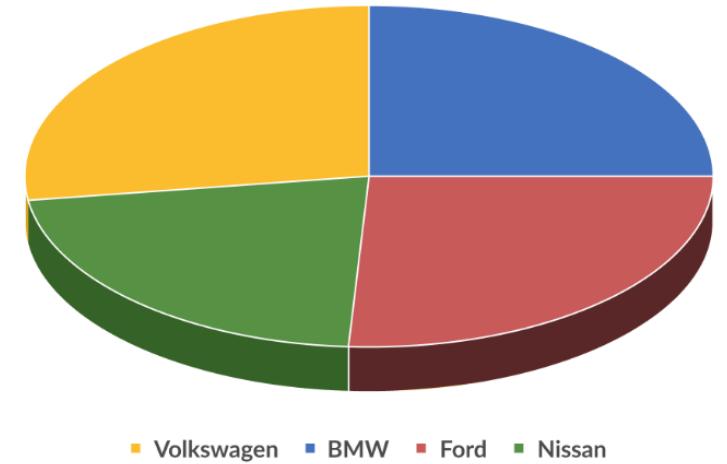
Color
Intensity

Atributos pre-atentivos

Es importante que los atributos que usemos para “priorizar” algo, no cambien la información del gráfico, ni lo vuelva “engañoso”

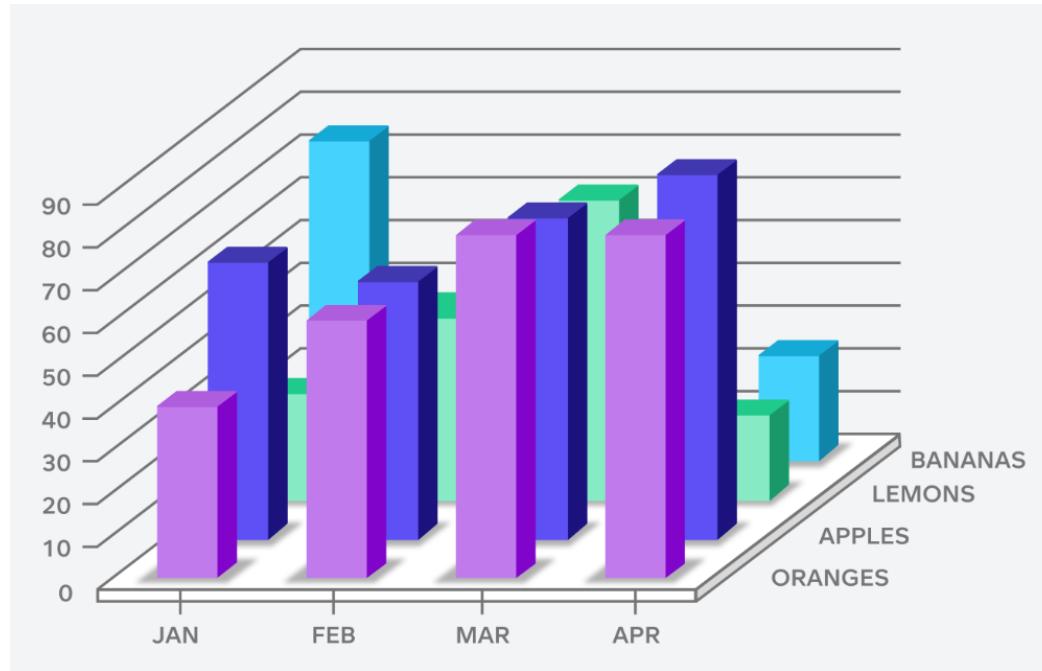


Sales

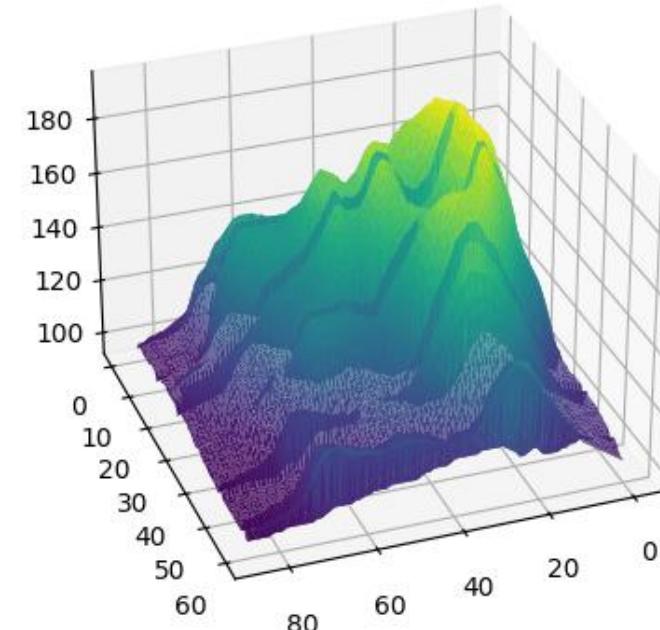


Lo que no hay que hacer

Gráficos 3D

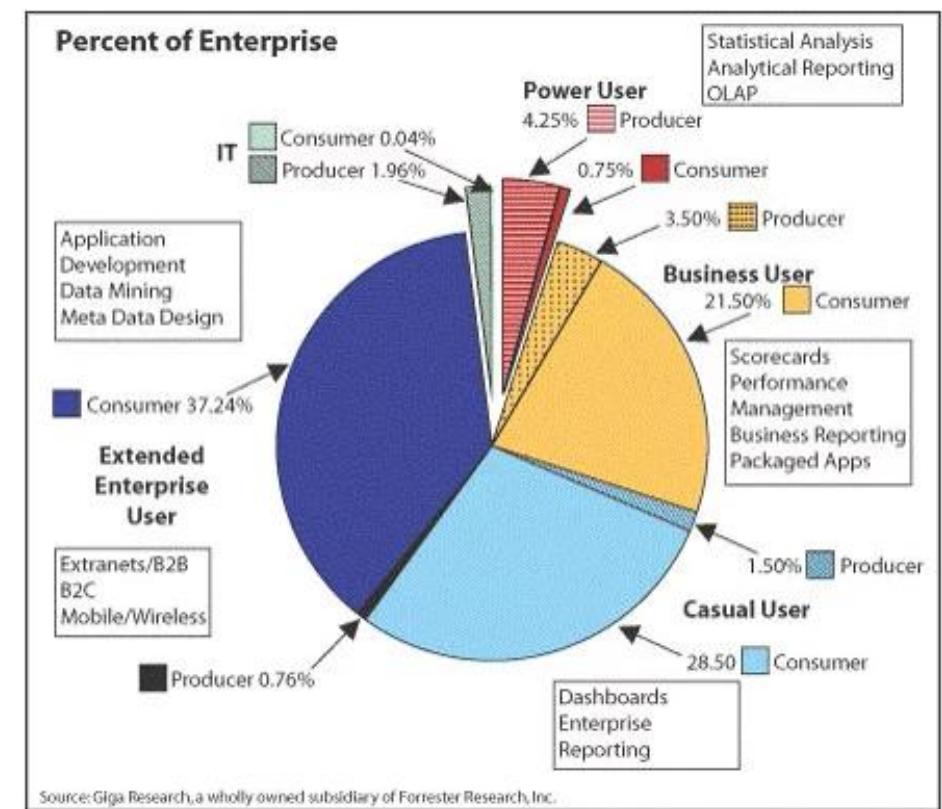
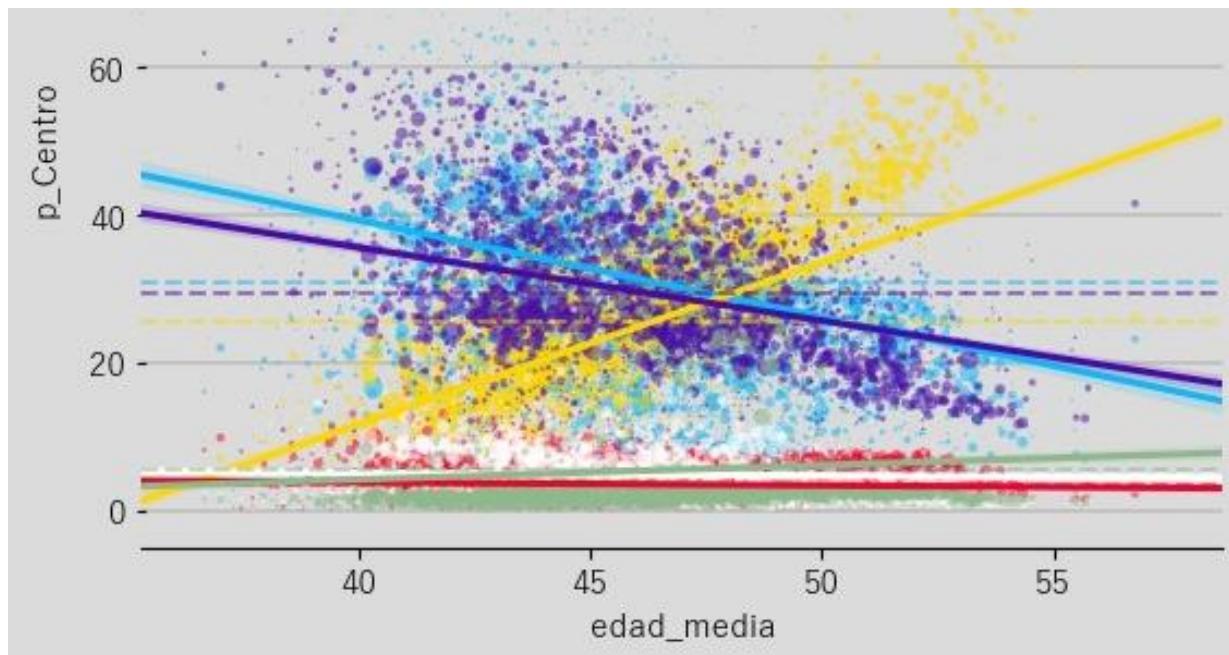


Salvo que sea interactivo:



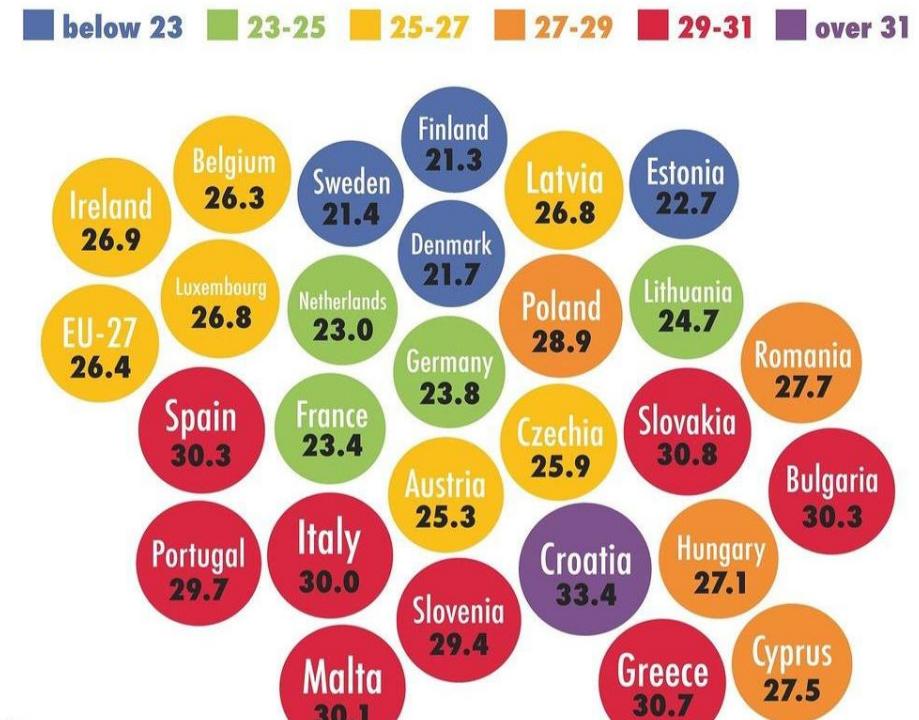
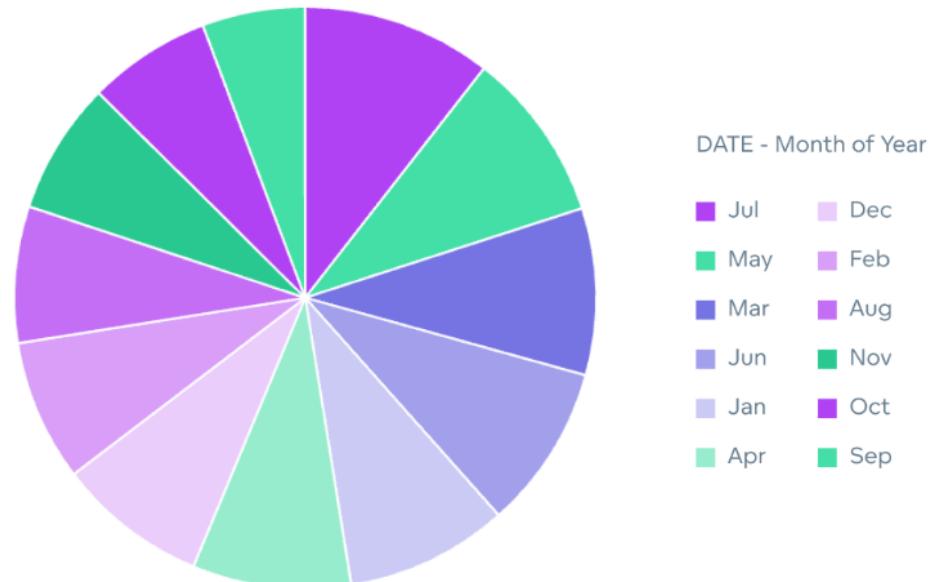
Lo que no hay que hacer

Llenarlos de cosas



Lo que no hay que hacer

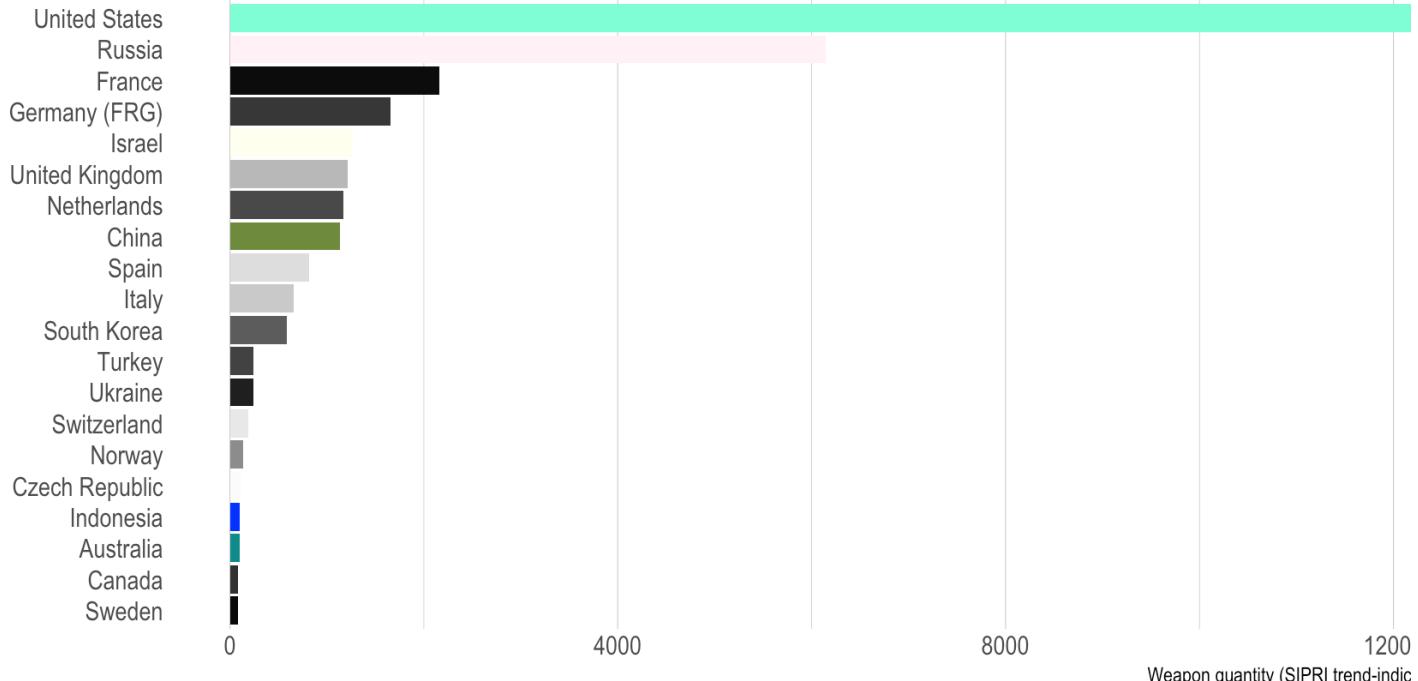
Usar escalas de color continuas en variables categóricas (y viceversa)



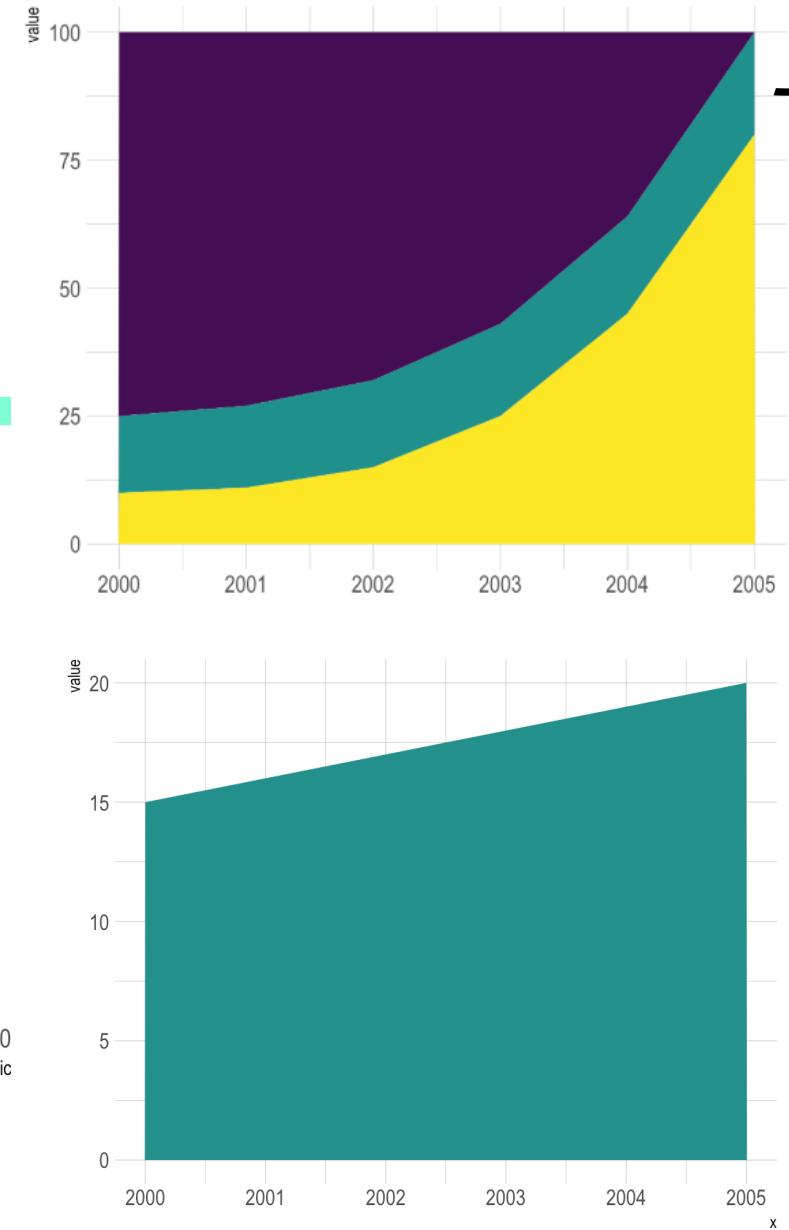
Source: Eurostat

Lo que no hay que hacer

Usar colores que no dan significado

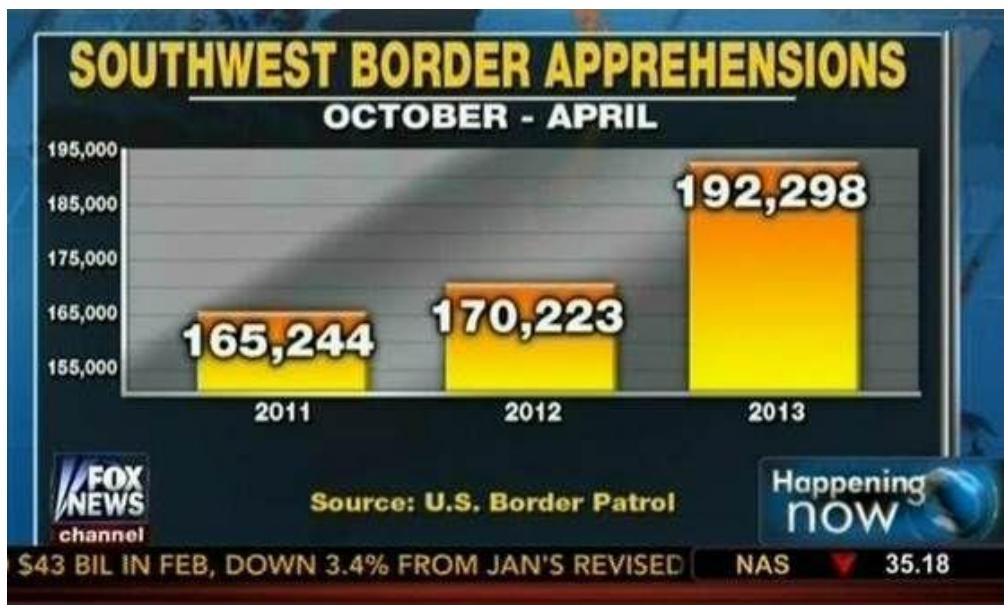


Apilar áreas que evolucionan



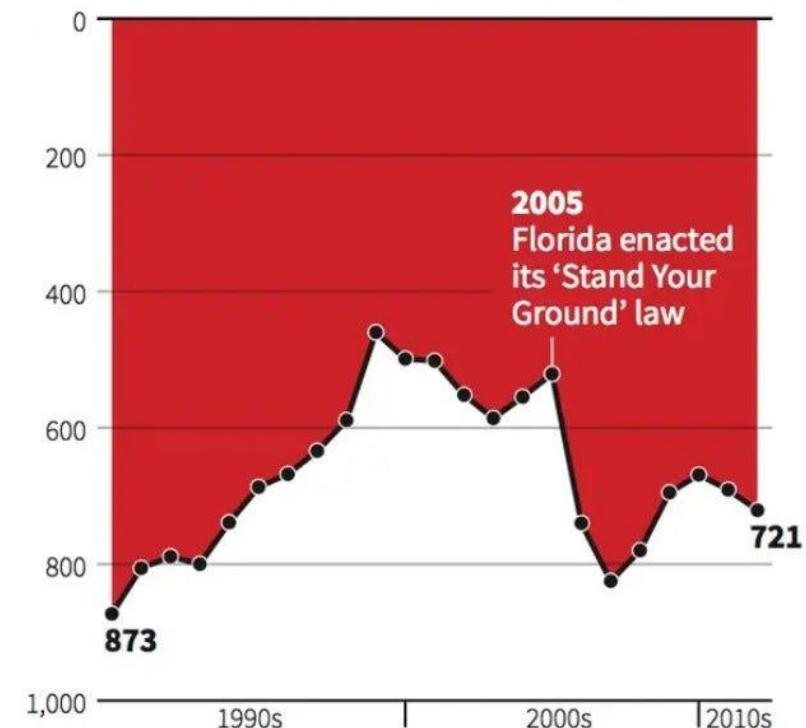
Lo que no hay que hacer

Modificar el eje Y (al menos no tanto)



Gun deaths in Florida

Number of murders committed using firearms



Source: Florida Department of Law Enforcement

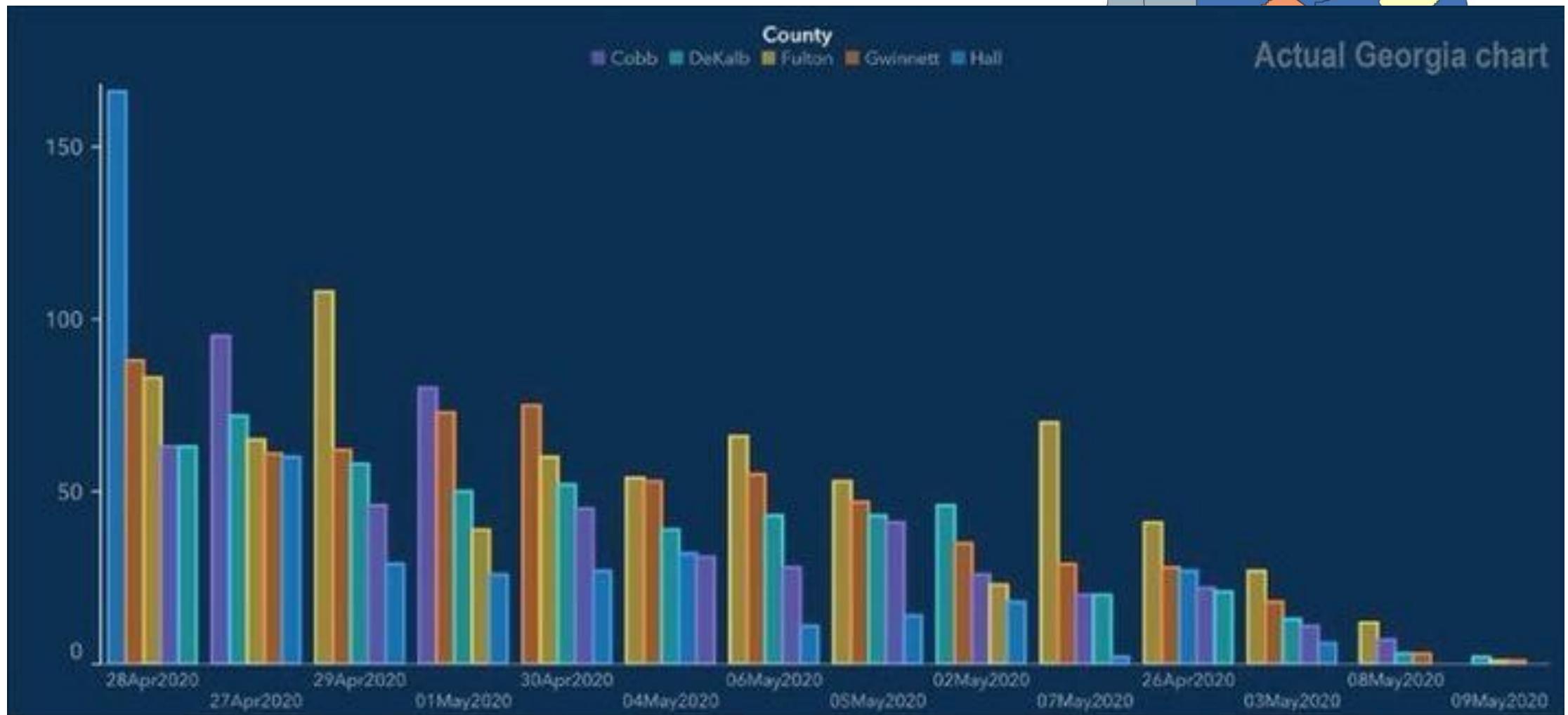
C. Chan 16/02/2014

Reuters

REUTERS

Lo que no hay que hacer

A ver si se dan cuenta



Distribuciones

Distribuciones

- Trabajan para Edelap (equivalente a EDESUR/EDENOR de La Plata).
- Un amigo les cuenta que una vez, un empleado que le fue a habilitar el medidor le ofreció “toqueteárselo” para que “frene” en los 400 kWh (sabiendo que si pasas los 400, subís de categoría y te cobran mucho más)
- Indignados (o no), vuelven a la oficina y quieren ver si hay muchos casos de estos.

calle	altura	piso	depto	telefono	subcategory	consumo
SAN CARLOS	0	0	0	NaN	.	328.00000
SAN CARLOS	0	0	0	NaN	.	41.00000
EL JACARANDA	408	0	0	NaN	.	17.00000
MONTEVIDEO	0	0	DRA	NaN	.	132.00000
MONTEVIDEO	0	0	0	NaN	.	383.00000
...
38	2306	0	0	NaN	.	194.00000
38	2306	0	0	NaN	.	85.00000
133	1871	0	5	NaN	.	4.00000
11	22	0	PA3	NaN	.	55.00000

Distribuciones

- Nos interesa la variable consumo
- Viendo 10 medidores, ninguno pasa los 400.
- En La Plata hay 300.000 viviendas, no alcanza con ver 10 para sacar conclusiones
- ¿Qué tipo de gráfico permite ver esto?

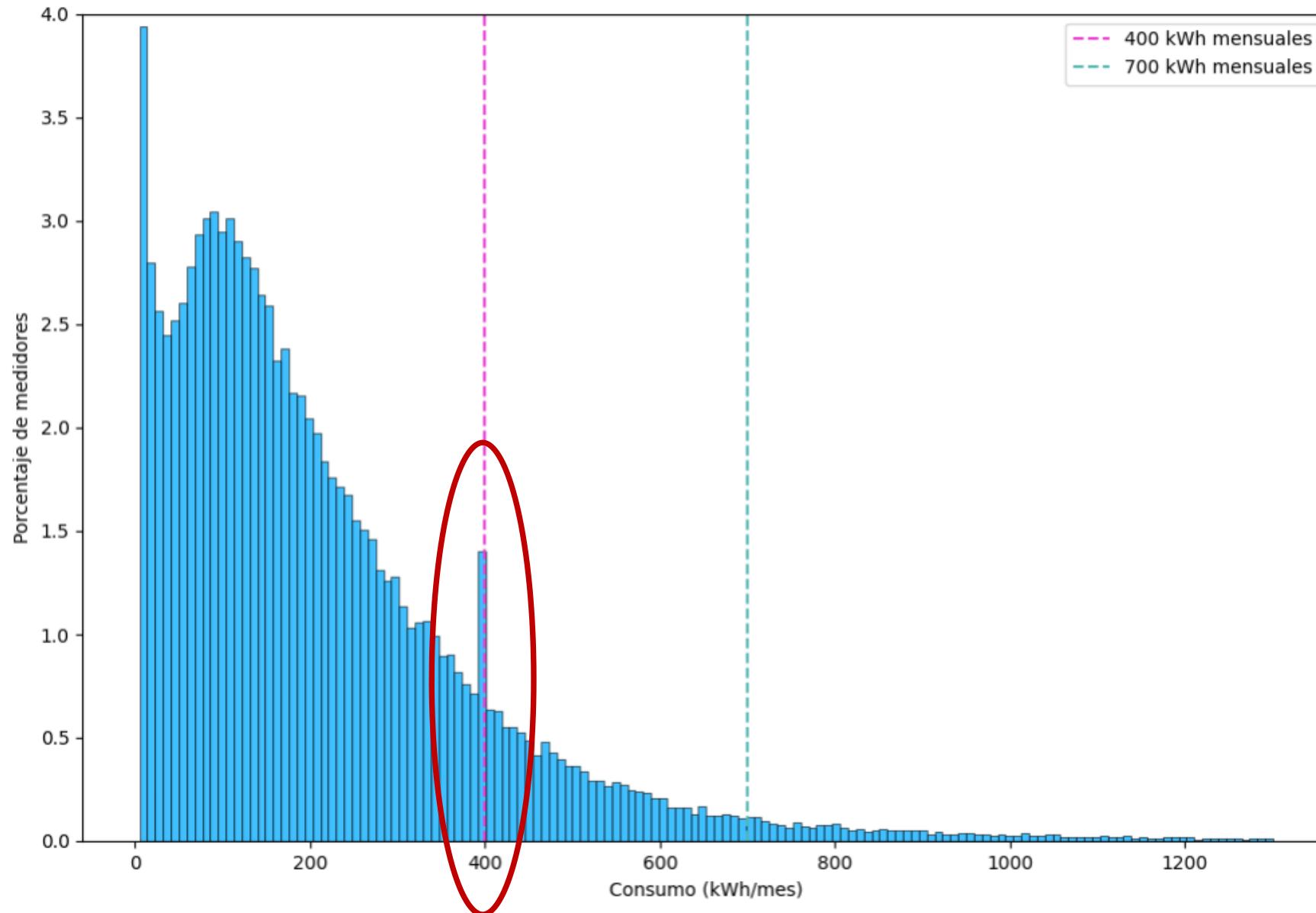


Histograma

calle	altura	piso	depto	telefono	subcategory	consumo
SAN CARLOS	0	0	0	NaN	.	328.00000
SAN CARLOS	0	0	0	NaN	.	41.00000
EL JACARANDA	408	0	0	NaN	.	17.00000
MONTEVIDEO	0	0	DRA	NaN	.	132.00000
MONTEVIDEO	0	0	0	NaN	.	383.00000
...
38	2306	0	0	NaN	.	194.00000
38	2306	0	0	NaN	.	85.00000
133	1871	0	5	NaN	.	4.00000
11	22	0	PA3	NaN	.	55.00000

Histograma

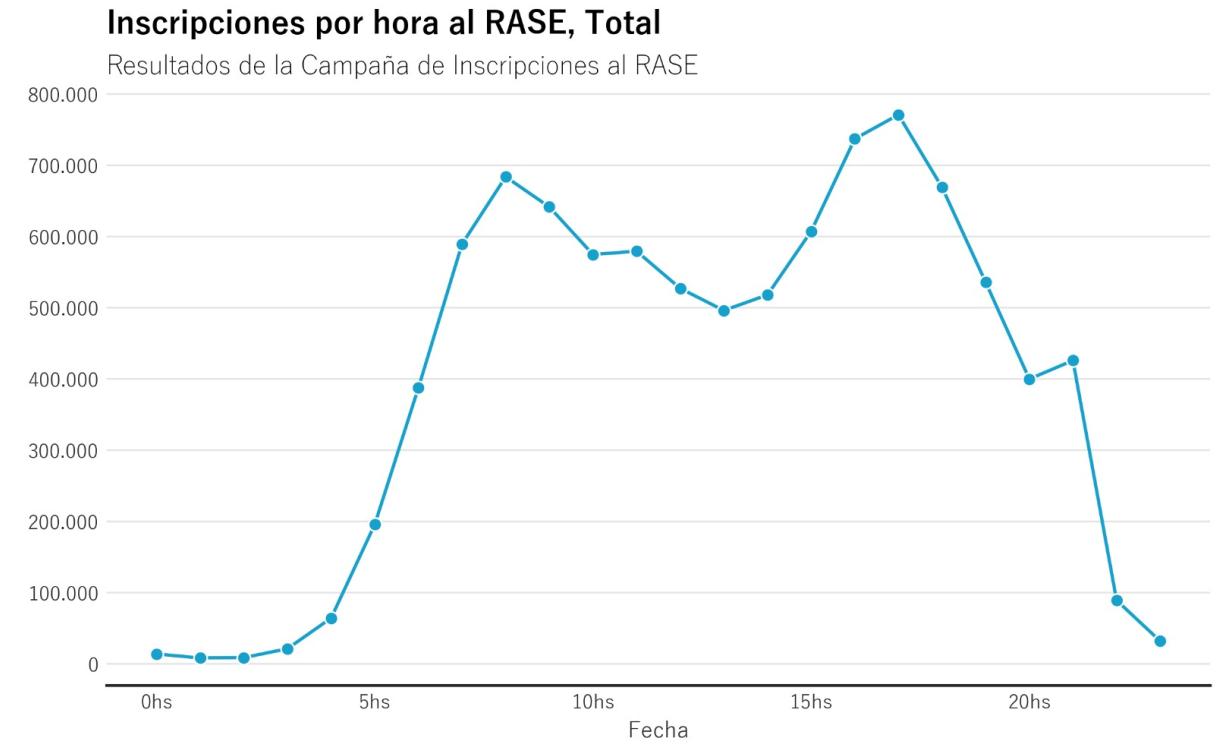
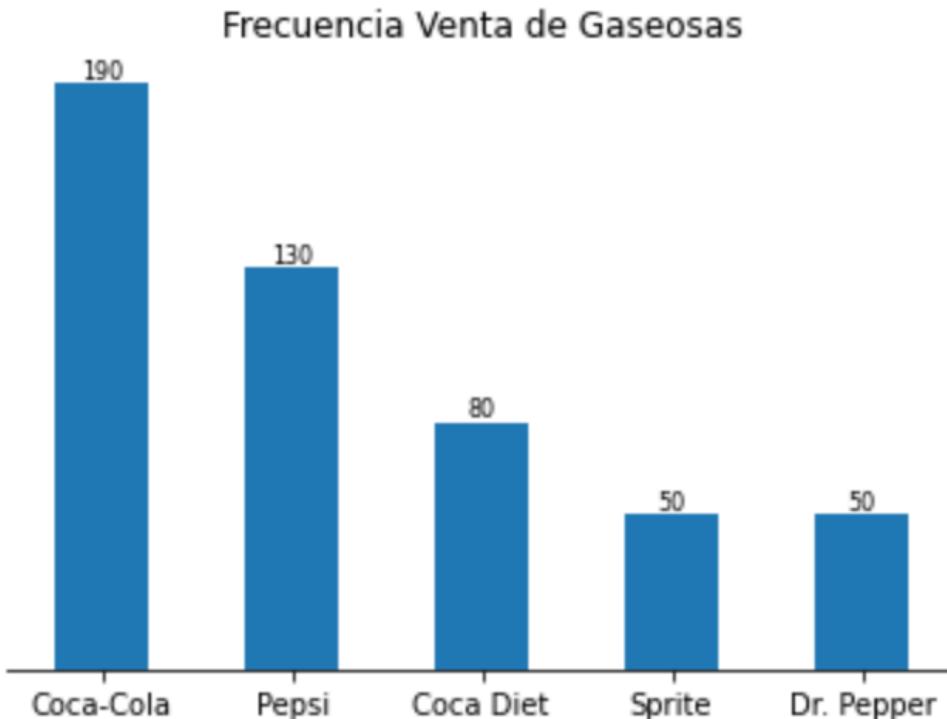
Oct-Dic, 2020



¿Qué conclusiones sacan de este gráfico? ¿Es un problema para la empresa?

Histogramas

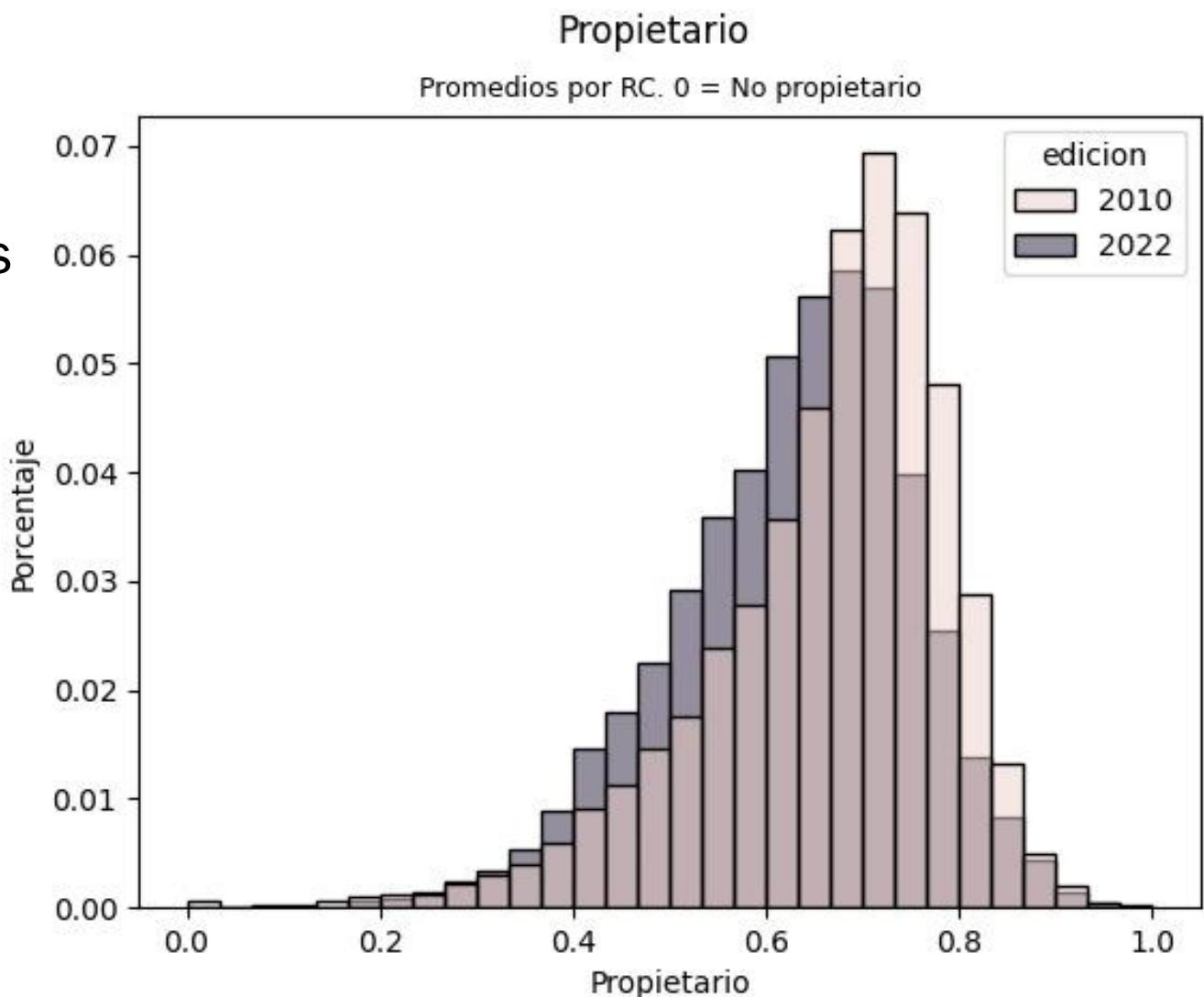
- Muestra cuántas veces una variable toma ciertos valores
 - Puede ser de variables categóricas o continuas
 - Si es continua, se discretiza separando en bins (barras)
- La frecuencia se puede mostrar de manera absoluta o relativa



(en realidad, si el eje x es categórico, se le llama gráfico de frecuencias)

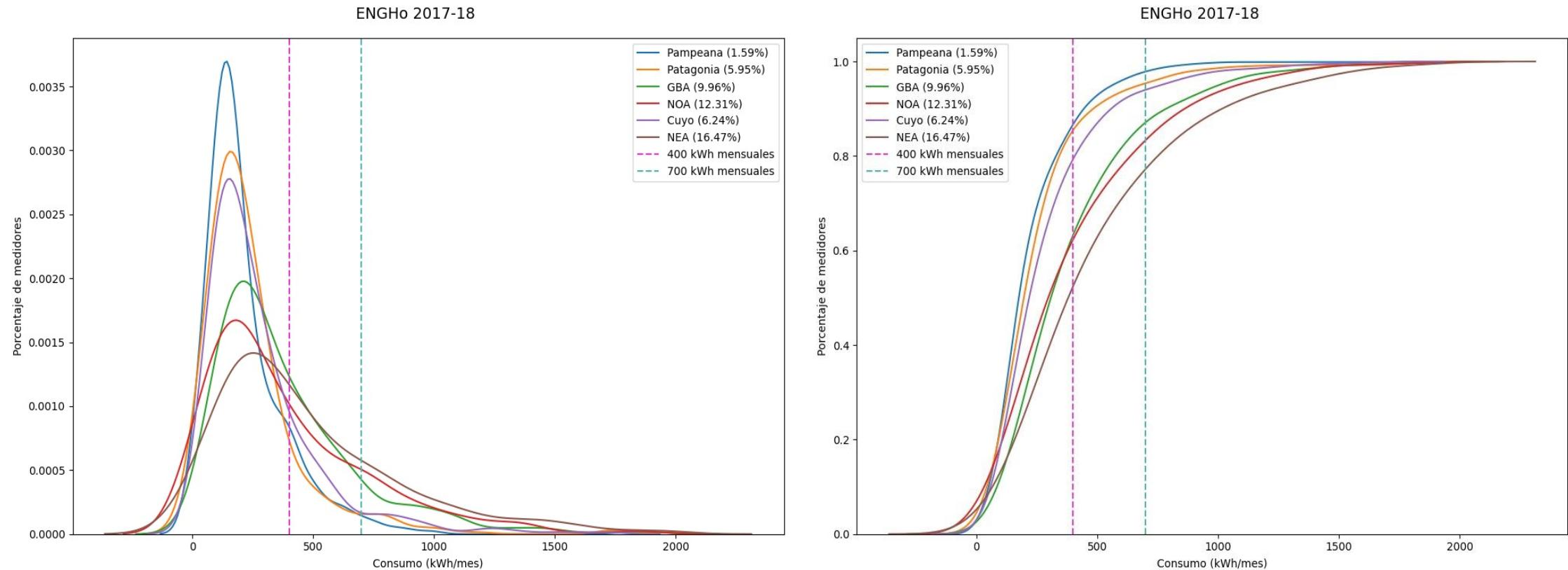
Histogramas

- Se pueden juntar histogramas



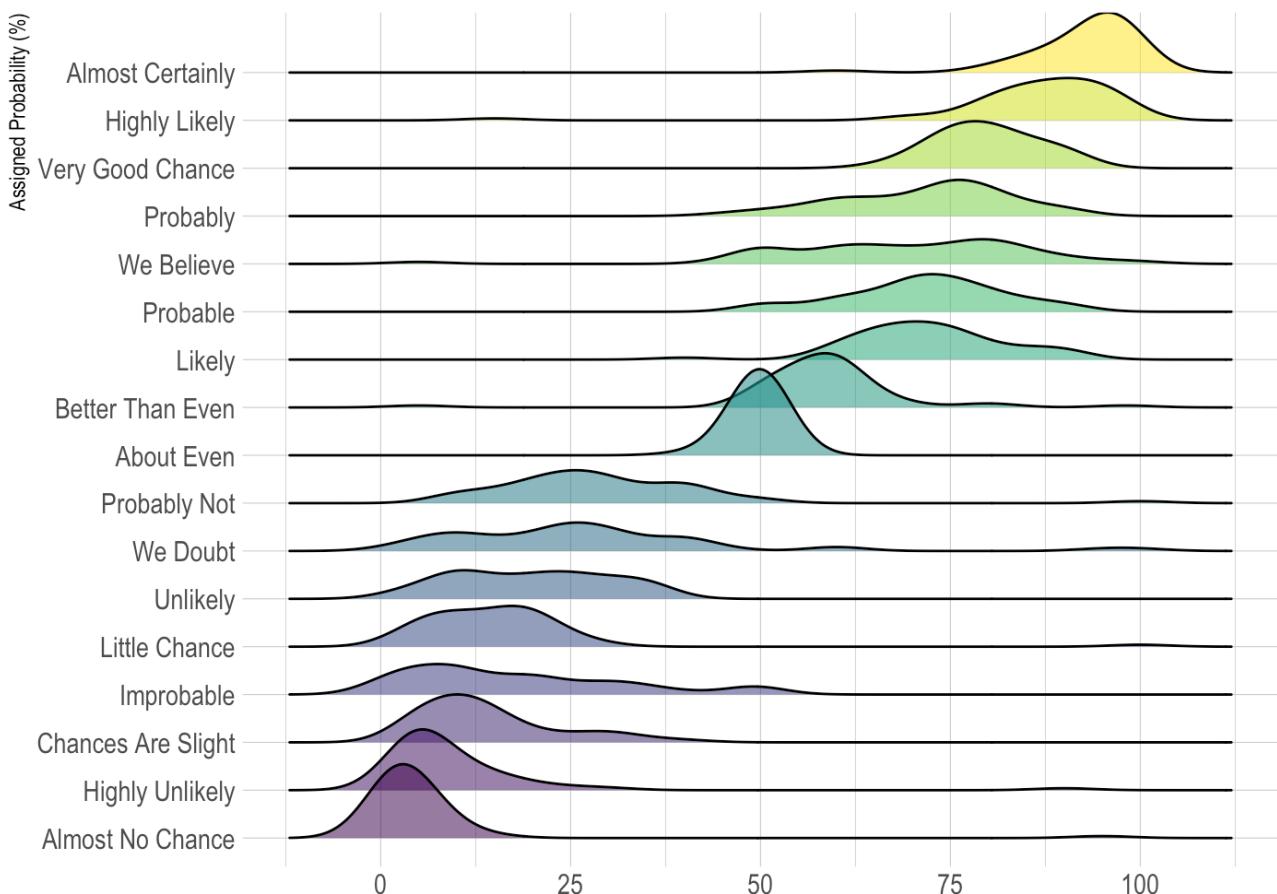
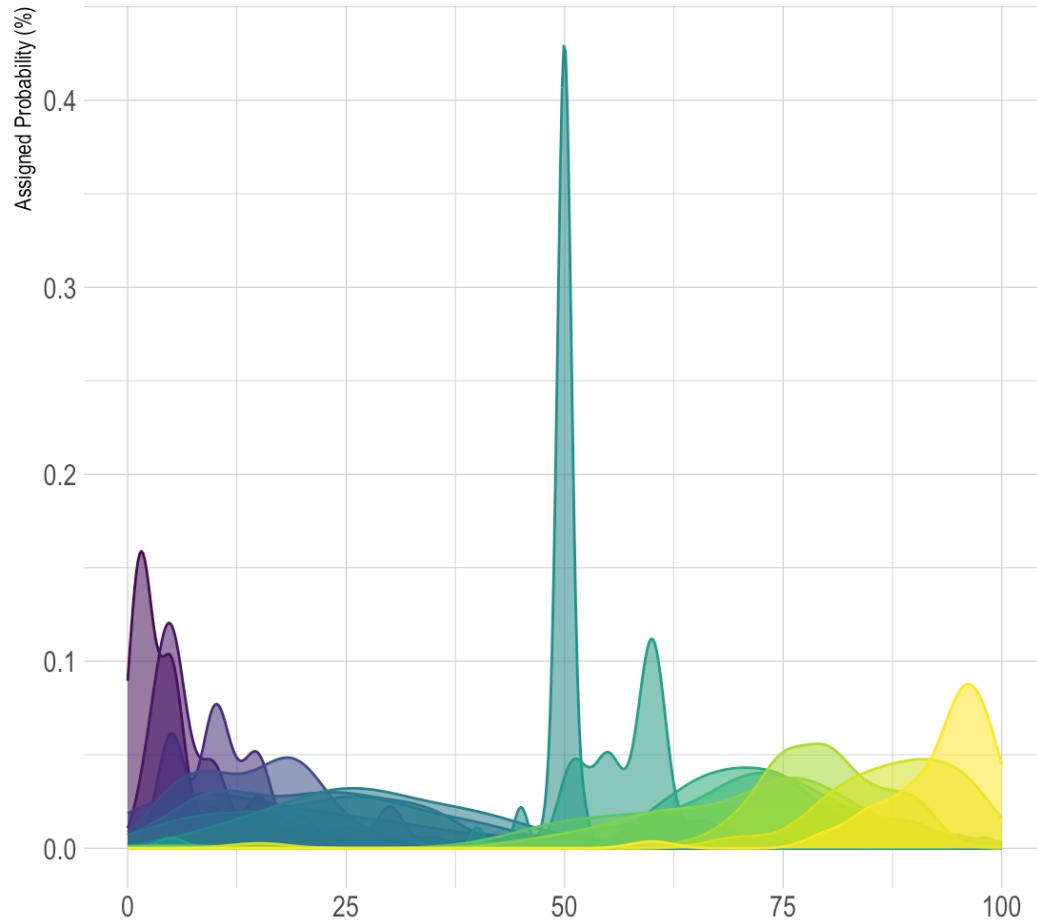
KDE: histograma suavizado

- Si el n es suficiente, directamente se puede suavizar el gráfico
 - Ya no se llama histograma, ahora es un KDE (kernel density estimate)
- También se puede graficar el acumulado (en $X=5$, la altura es la cantidad de veces que el valor es ≤ 5)



Ambos representan los mismos datos

Muchos KDE: Ridgeplot



Estadística descriptiva

Ejemplo: Hallar un valor que describa esta distribución de ingresos

€ 15,670.00
€ 15,680.00
€ 16,680.00
€ 17,500.00
€ 18,800.00
€ 20,000.00
€ 21,000.00
€ 21,000.00
€ 25,000.00
€ 25,120.00
€ 26,000.00
€ 33,000.00
€ 35,000.00
€ 83,000.00
€ 101,500.00

Seguramente pensaron en una medida de tendencia central

Medidas de tendencia central

- Valor único que resume la magnitud central (el “medio”) de una variable
- Las más conocidas:

Media (mean):

$$\text{mean}(x) = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{101500 + 83000 + \dots + 15670}{15} = 31663$$

Suma de valores sobre cantidad

Mediana (median):

Valor que acumula la mitad de las observaciones (o sea, la mitad es menor, y la mitad mayor).

Ordenar y agarrar el del medio. Si son pares, promediar los dos del medio

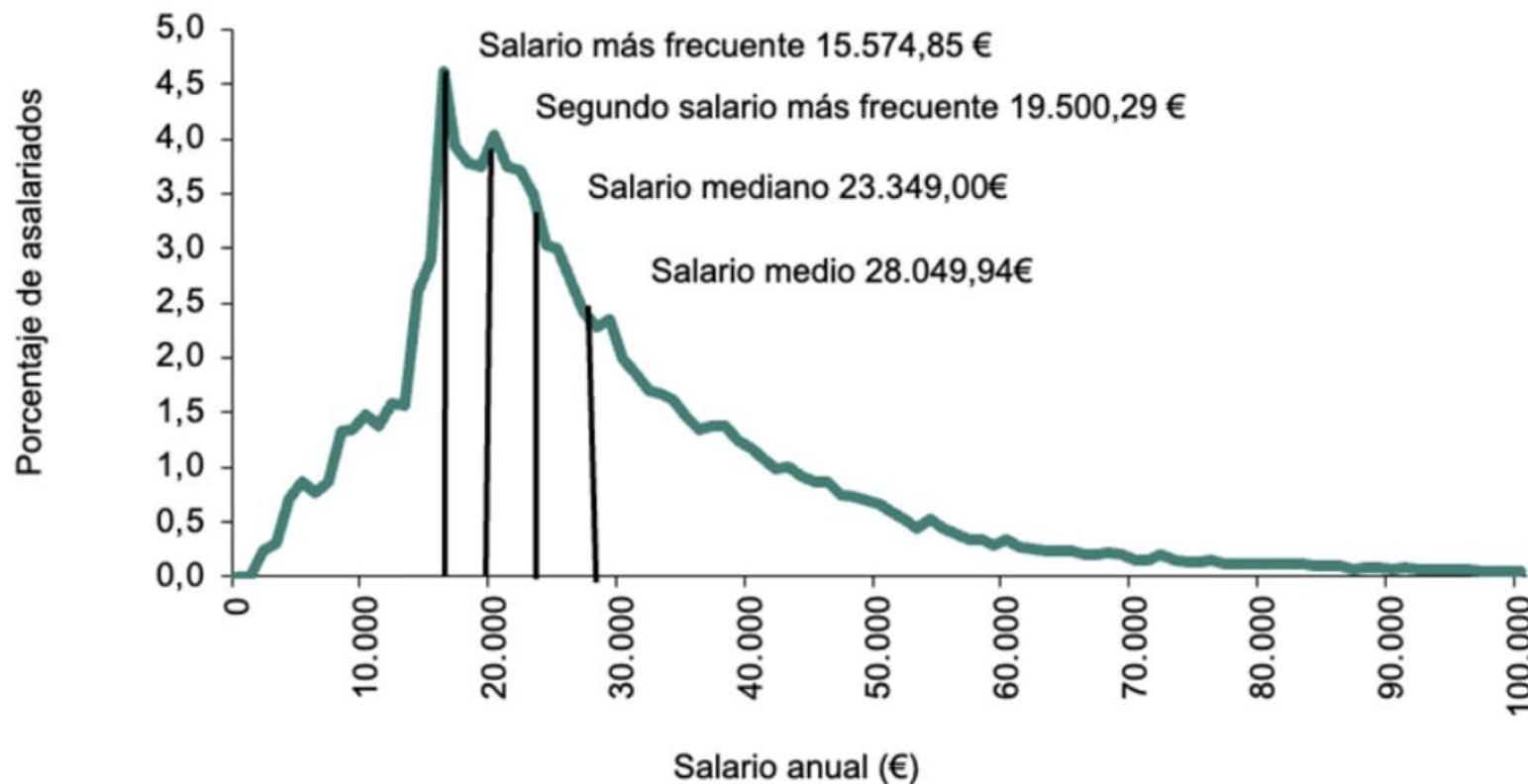
Moda (mode):

Valor que más se repite. No tiene mucho sentido si hay pocas observaciones

€ 101,500.00
€ 83,000.00
€ 35,000.00
€ 33,000.00
€ 26,000.00
€ 25,120.00
€ 25,000.00
€ 21,000.00
€ 21,000.00
€ 20,000.00
€ 18,800.00
€ 17,500.00
€ 16,680.00
€ 15,680.00
€ 15,670.00

Medidas de tendencia central

¿Cuál hay que usar? Depende:

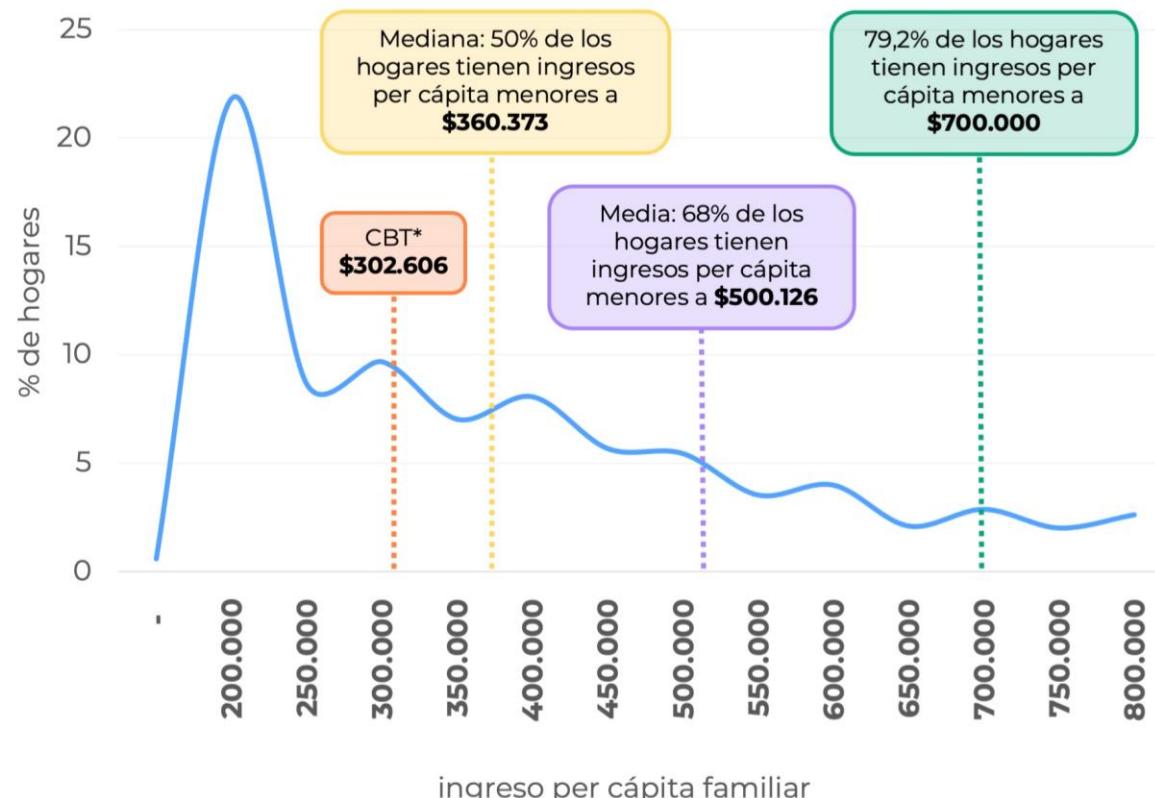


Fuente: INE. Encuesta de Estructura Salarial 2023

Medidas de tendencia central

La mala distribución del ingreso en Argentina

Ingreso per cápita familiar medio y mediano. Argentina, 3er trimestre 2024.



Actualizando por inflación a Agosto 2025 (asumiendo que la distribución de ingresos mantuvo la misma forma):

- CBT (canasta básica total): \$ 383.428
- Mediana: \$ 456.624
- Media: \$ 633.703

¿Pueden ver en este gráfico el porcentaje de hogares por debajo de la CBT (hogares bajo la línea de pobreza)? ¿Cómo lo editarían para poder hacerlo?

¿Por qué creen que oscila después de la moda? ¿Ven un patrón?

Medidas de dispersión

- Valor único que resume la dispersión de la variable
- Antes queríamos el valor que represente al “centro”, ahora queremos uno que muestre “la variación”

Rango:

$$\max(x) - \min(x) = 101500 - 15670 = 85830$$

Diferencia entre máximo y mínimo. Se basa sólo en las puntas, por lo que si tengo casos extremos, pesan mucho

Desvío estándar:

$$\sigma(x) = \text{sd}(x) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(15670 - 31663)^2 + \dots + (101500 - 31663)^2}{14}} = 25509$$

Representa la desviación promedio de la media de la variable (“en promedio, una observación está a 25509 de la media”)
Como es la “distancia promedio al promedio”, tiene los mismos problemas que el promedio (susceptible a distribuciones chuecas)

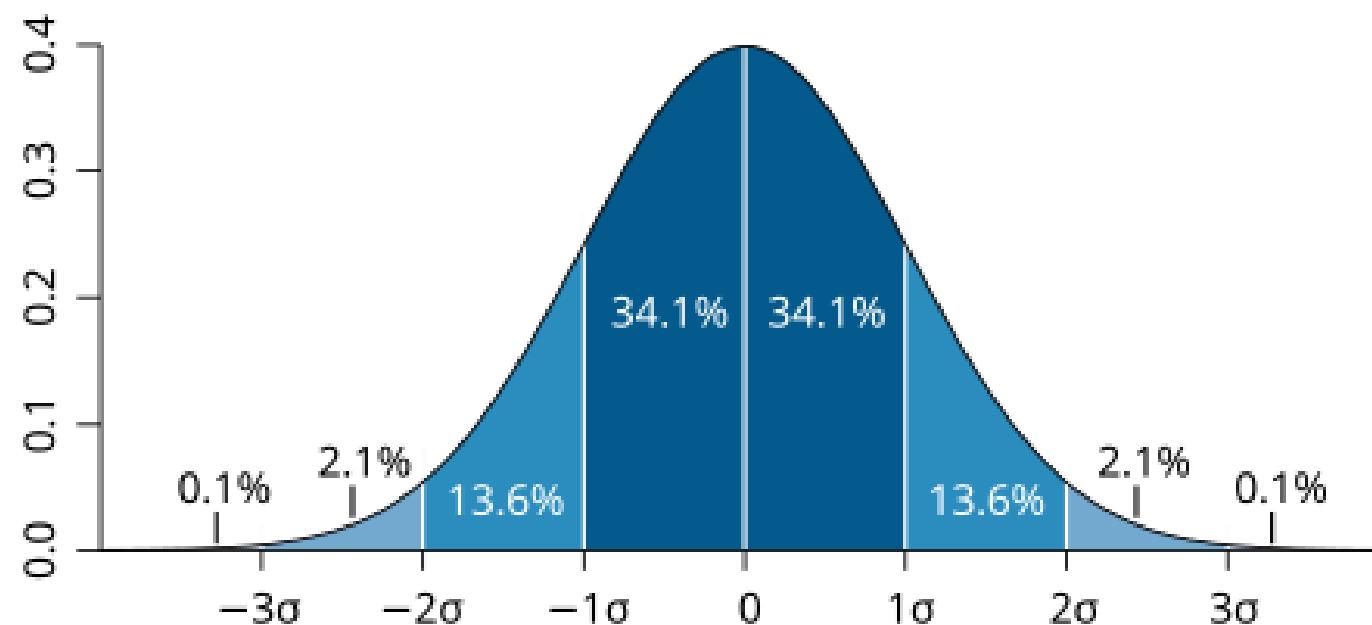
€ 101,500.00
€ 83,000.00
€ 35,000.00
€ 33,000.00
€ 26,000.00
€ 25,120.00
€ 25,000.00
€ 21,000.00
€ 21,000.00
€ 20,000.00
€ 18,800.00
€ 17,500.00
€ 16,680.00
€ 15,680.00
€ 15,670.00

Medidas de dispersión

Desvío estándar:

$$\sigma(x) = \text{sd}(x) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

El desvío permite agrupar los datos en intervalos



“Si la distribución es normal, el 68% de los datos están a +/- un desvío de la media”

Cuantiles

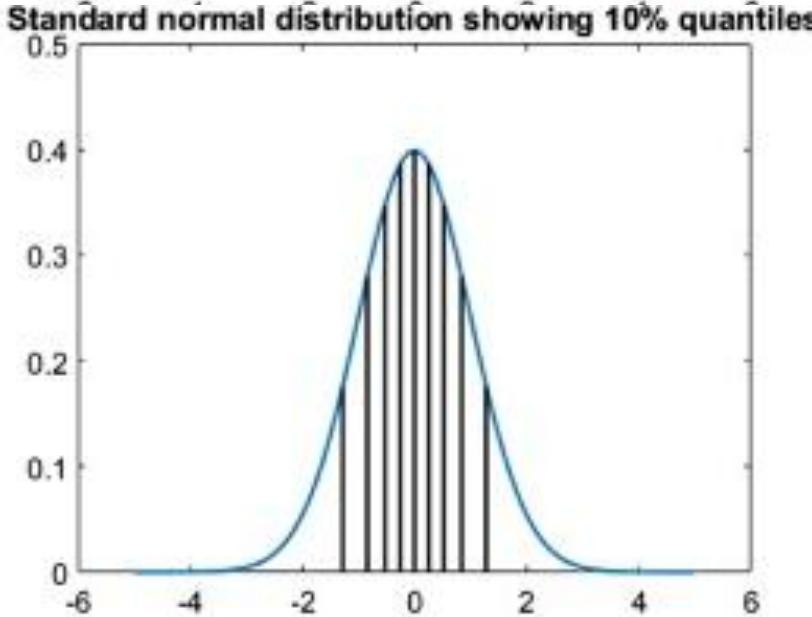
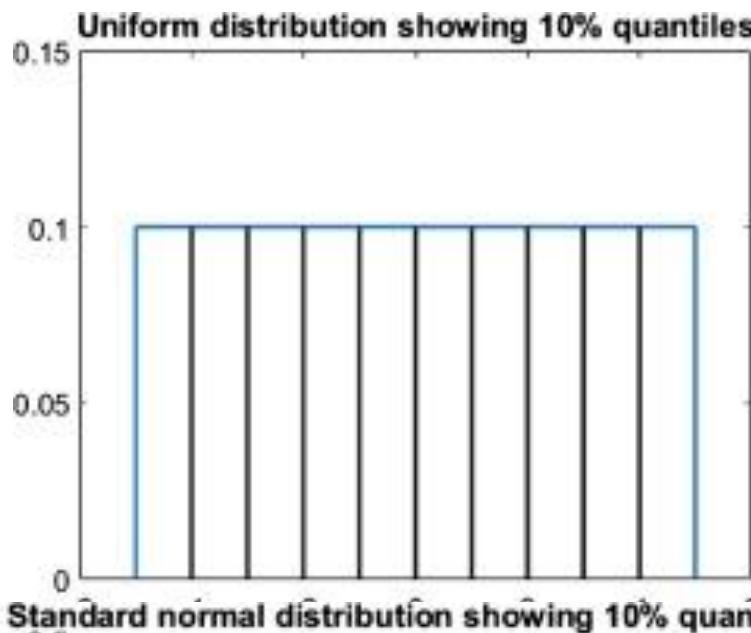
Percentiles, deciles, quintiles, cuartiles, q-tiles

- Deciles: valores que dividen al conjunto en 10 intervalos (de distinta longitud), cada uno teniendo el 10% de los datos.
 - “El primer decil es el valor que deja al 10% de los datos por debajo”
- Percentiles: valores que dividen al conjunto en 100 partes, cada una teniendo el 1% de los datos
 - “El percentil 47 es el valor que deja al 47% de los datos por debajo”
 - El percentil 50 es el valor que tiene al 50% de los datos por debajo, ¿les suena esta definición?
 - Percentil 50 = Mediana

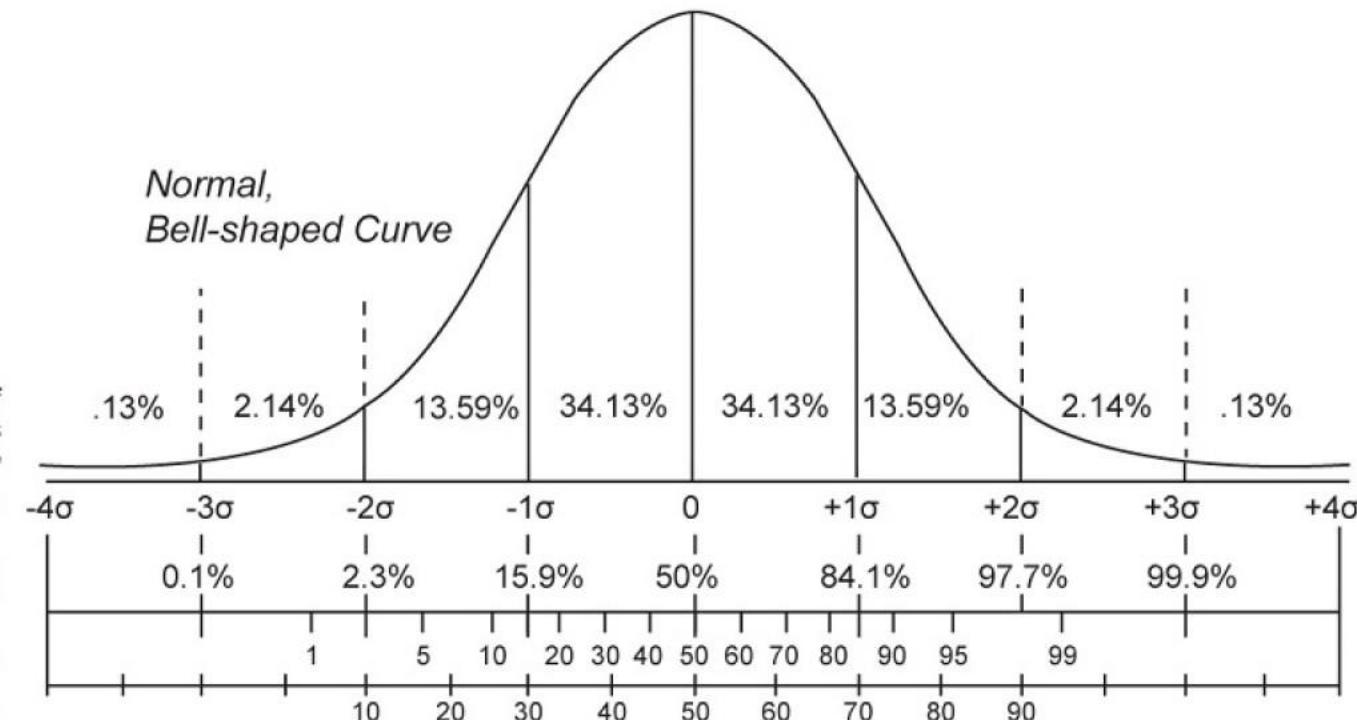
Además de los anteriores, existen quintiles (separan en 5, c/u 20%), cuartiles (separan en 4, c/u 25%), veintiles (separan en 20, c/u 5%)

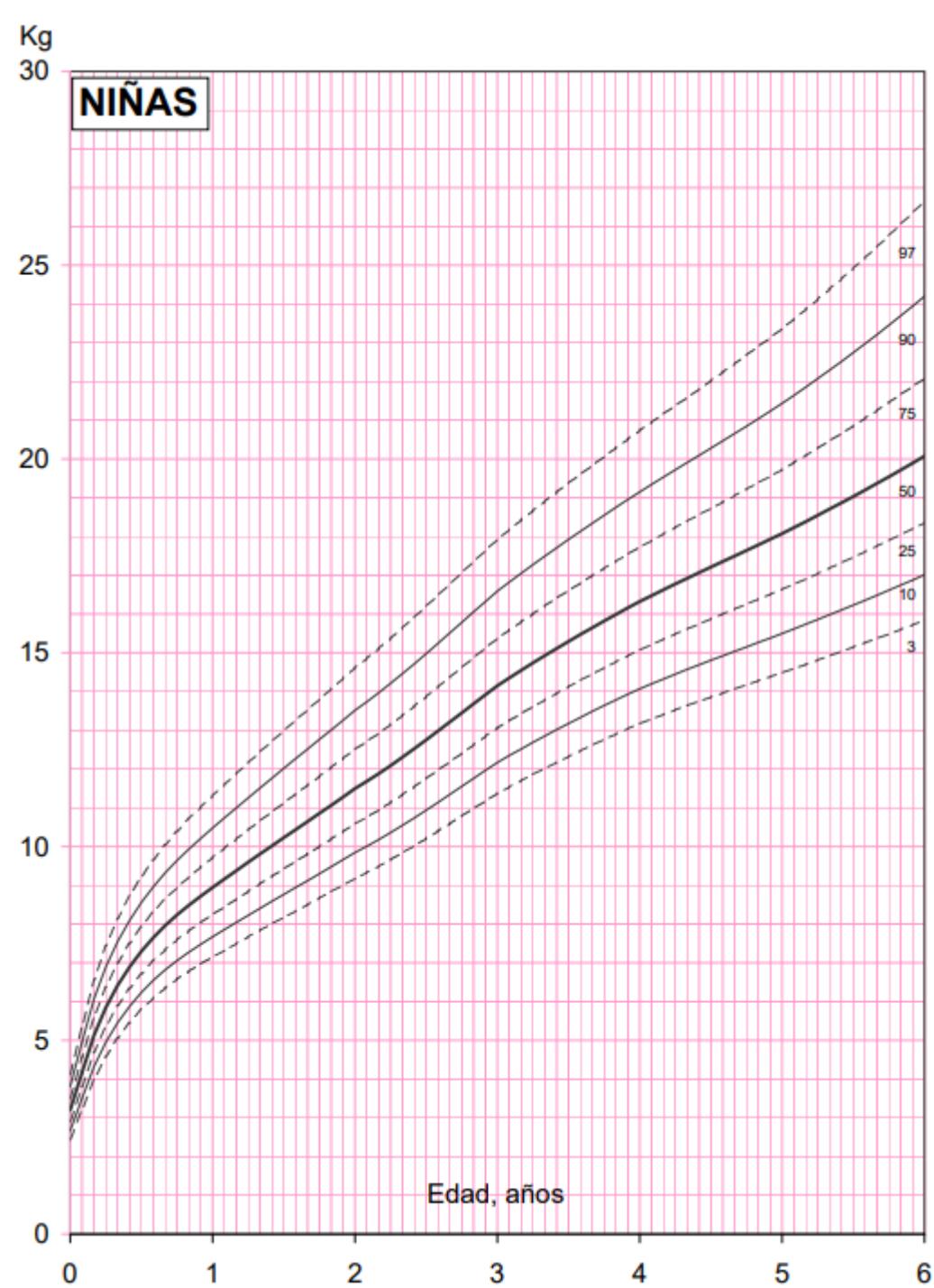
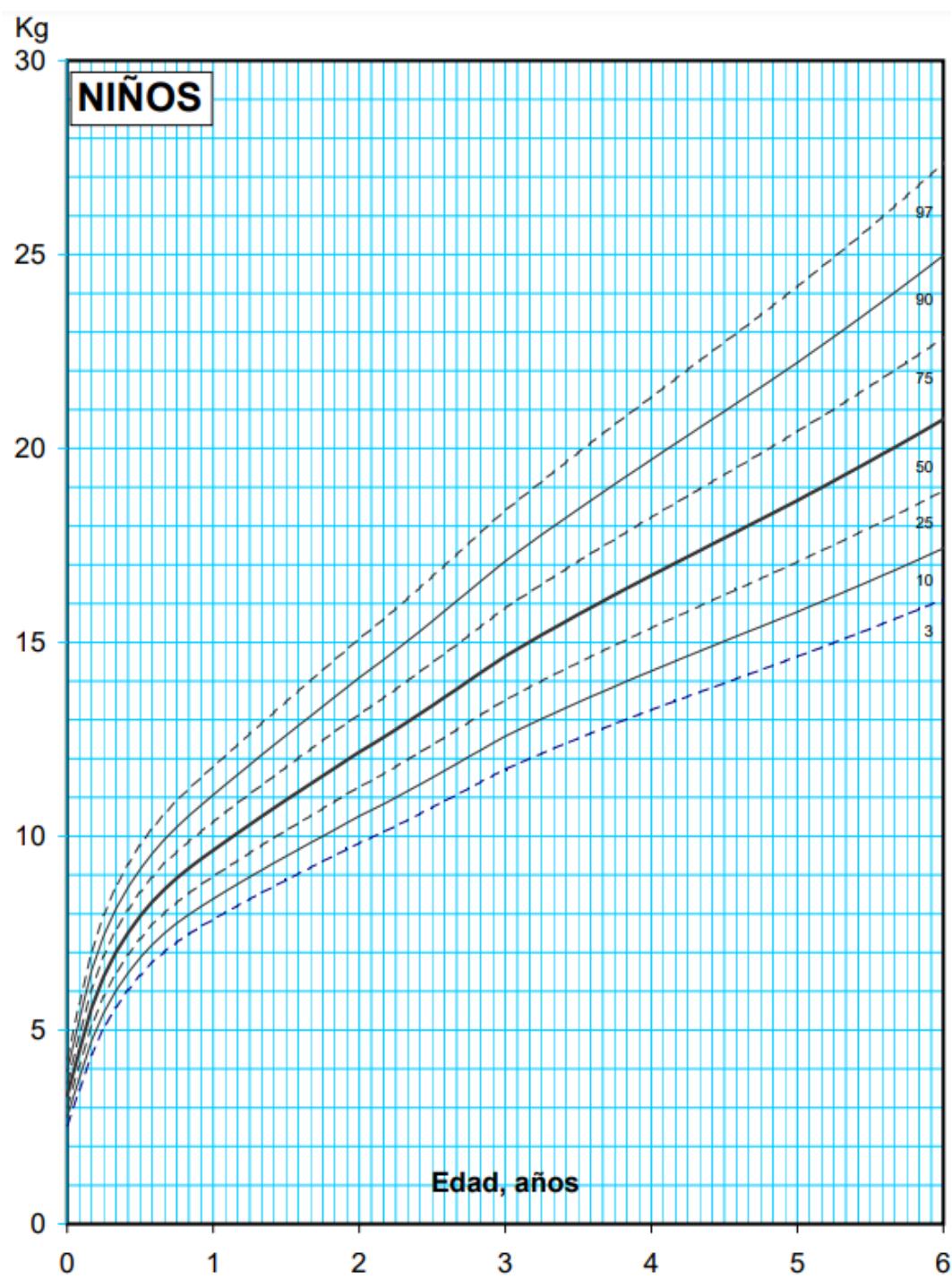
Mediana = Percentil 50 = Cuartil 2 = Decil 5 = Veintil 10

Cuantiles



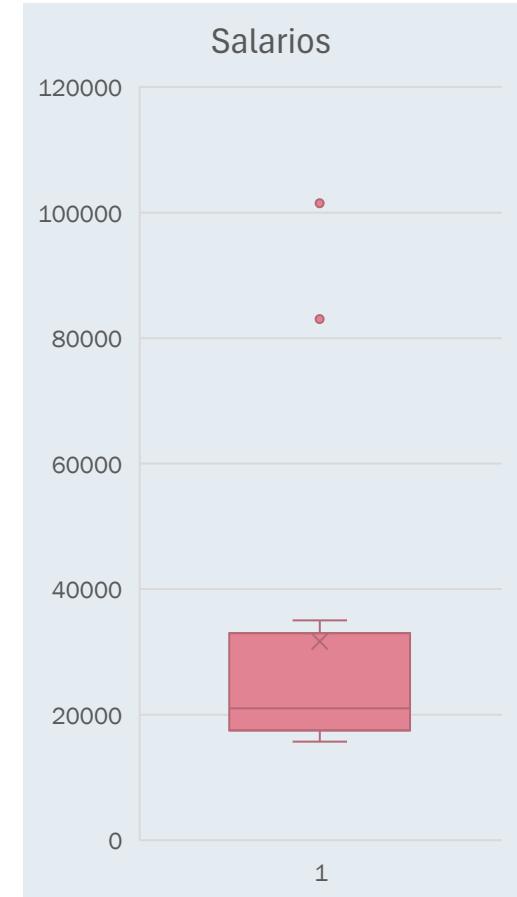
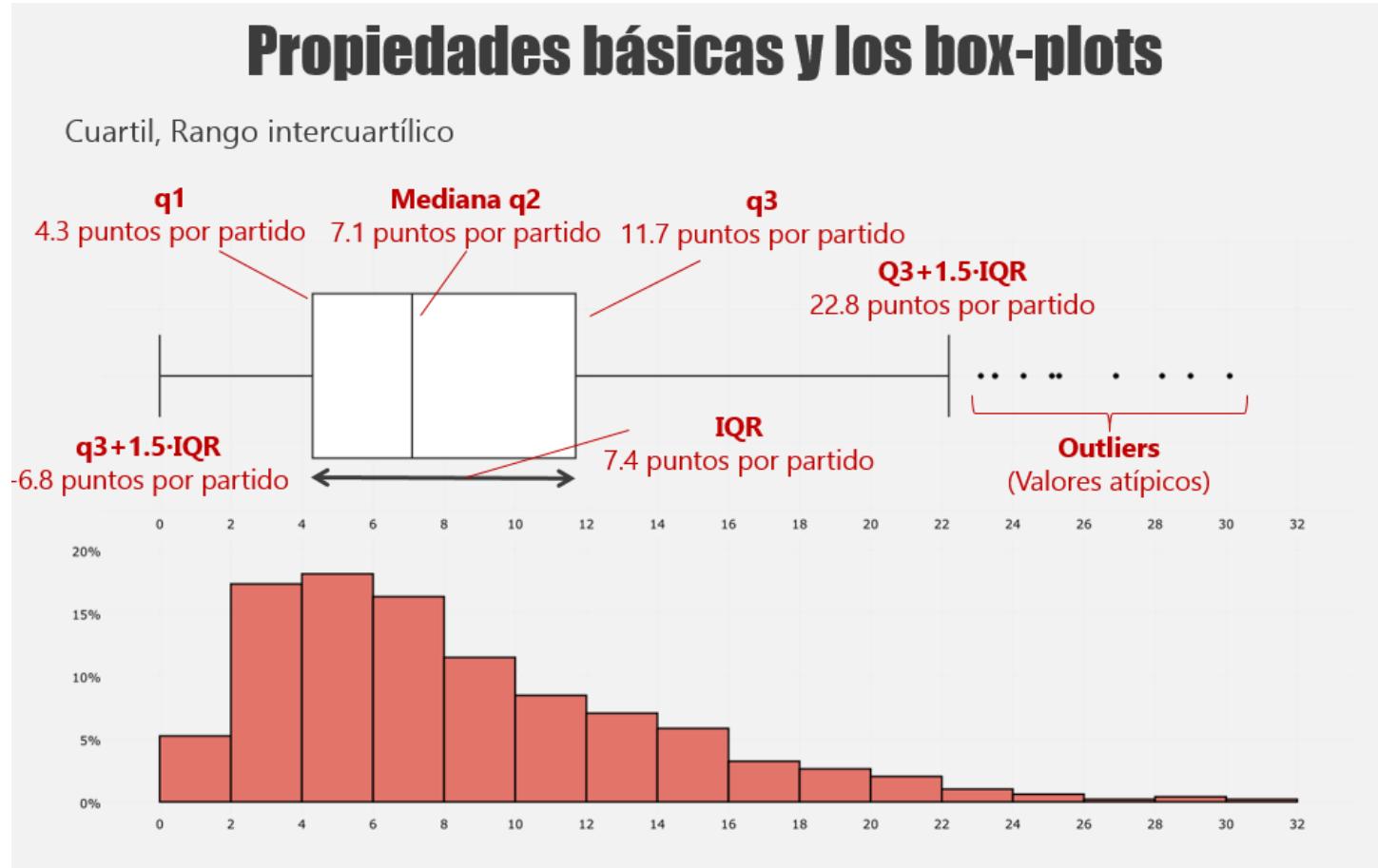
Percentage of
is in 8 portions
of the curve
ard Deviations
Cumulative
Percentages
Percentiles
Normal Curve
Equivalents





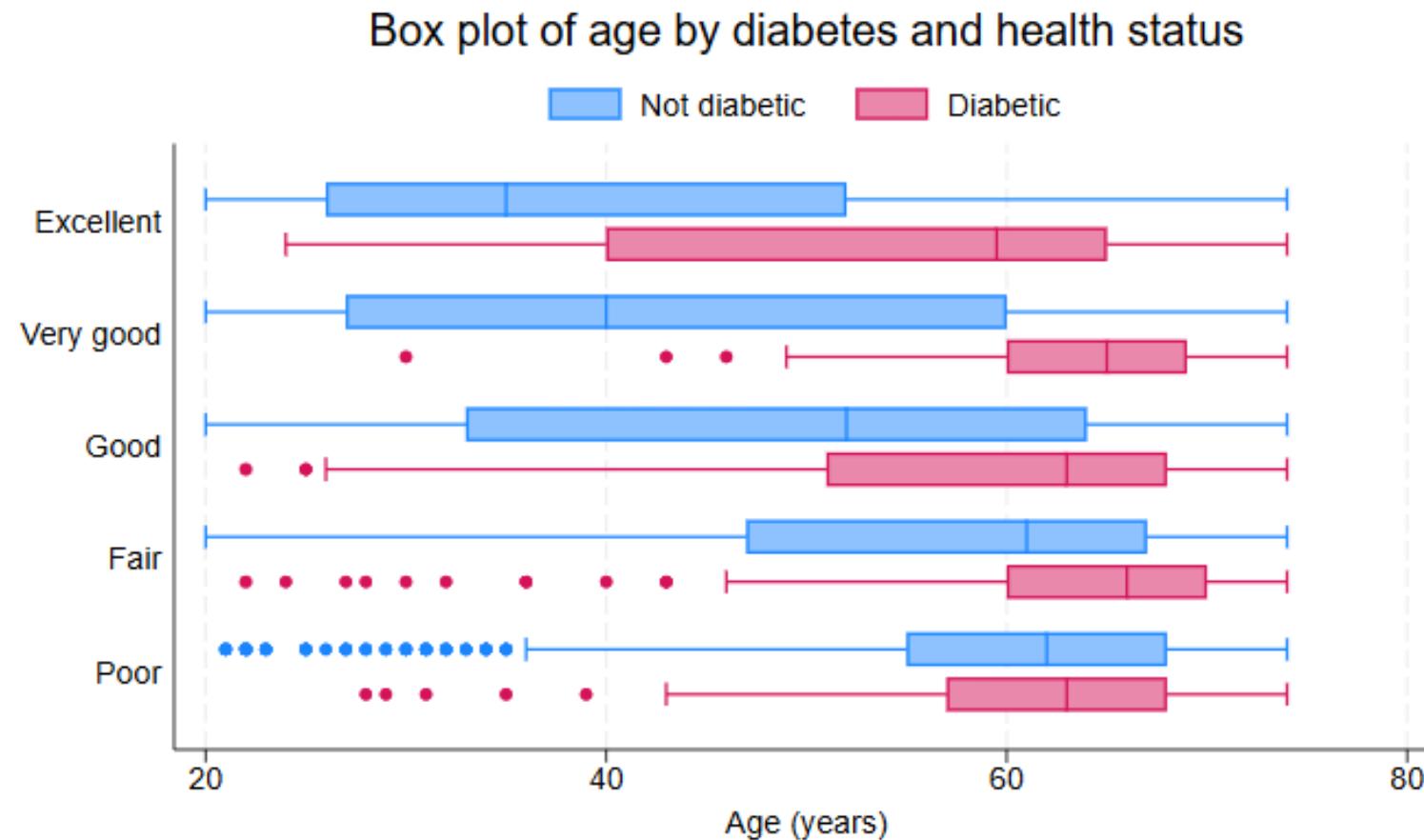
Boxplot

Gráfico que resume una distribución

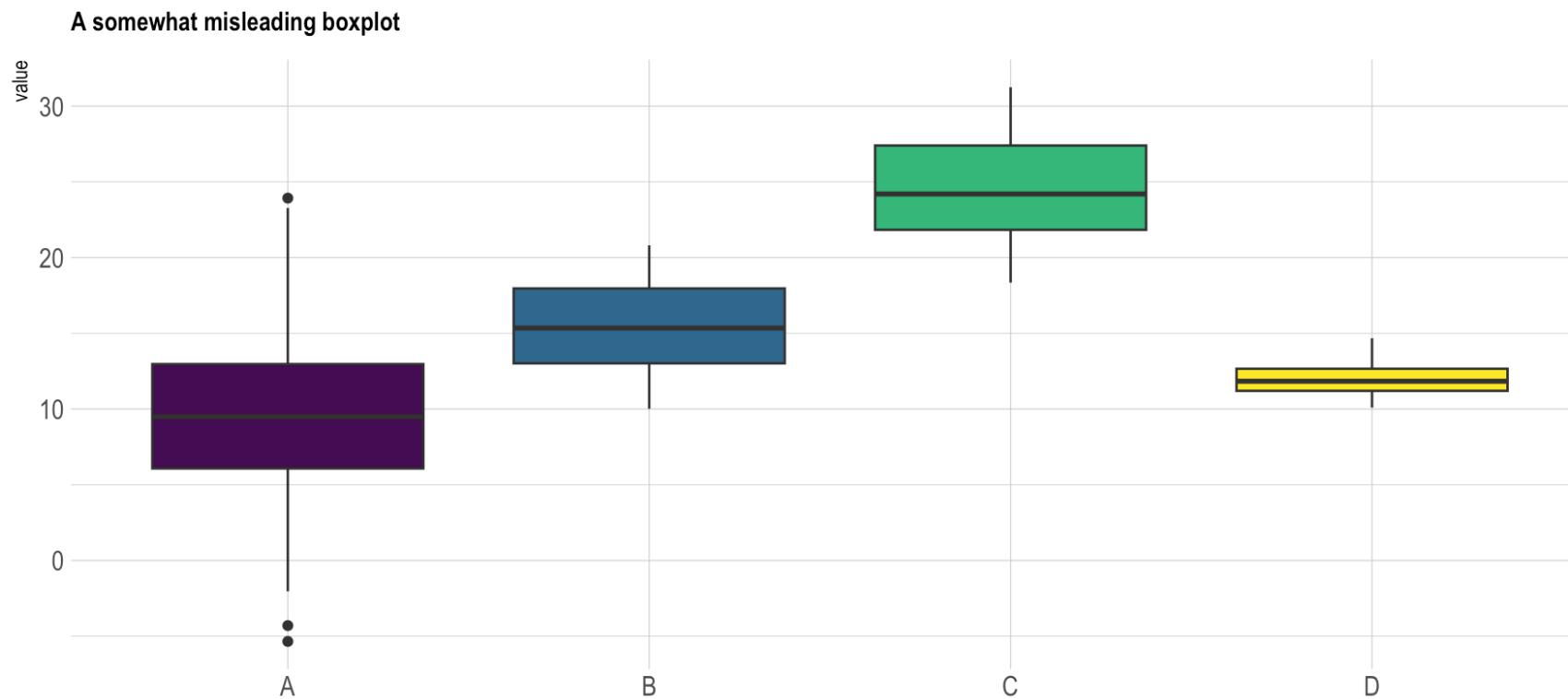


Boxplot

Gráfico que resume una distribución



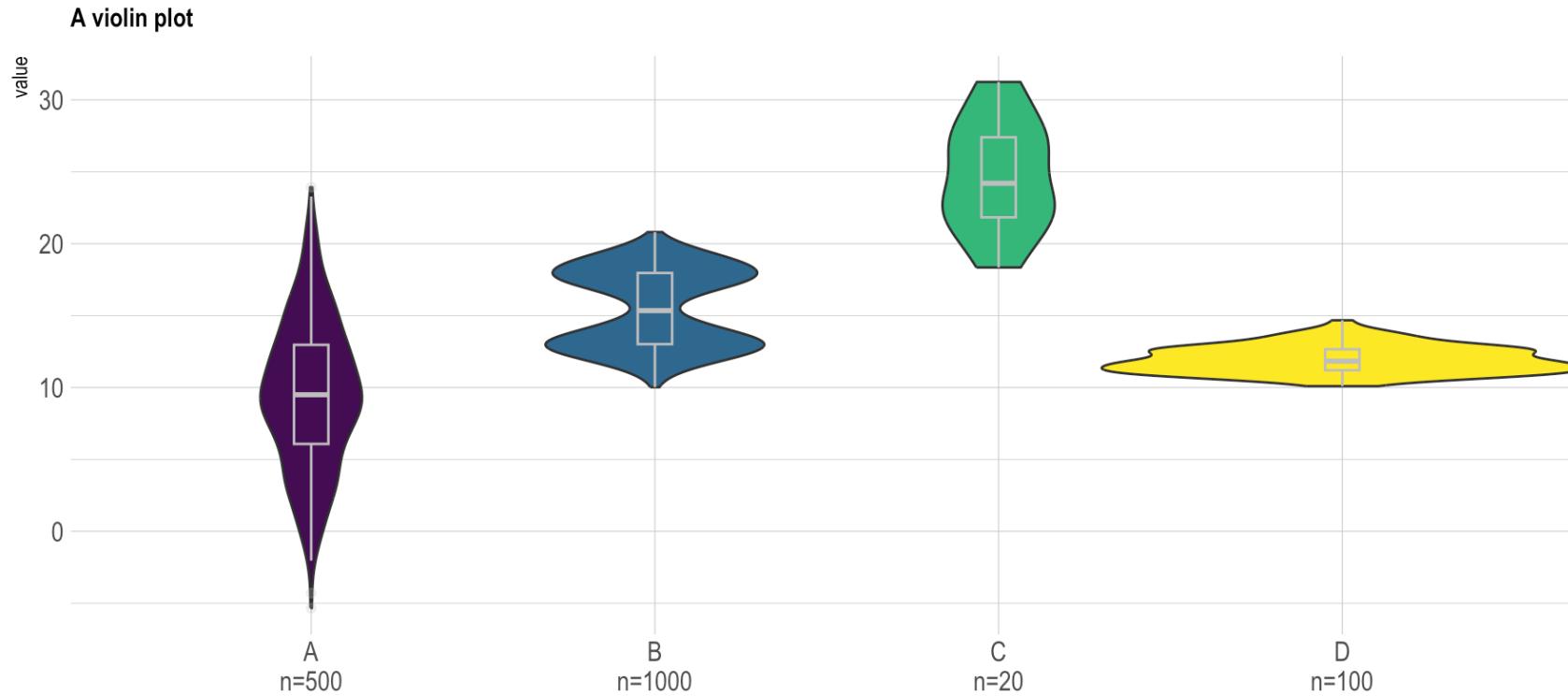
Boxplot



Viendo este gráfico, B y C parecieran tener distribuciones parecidas, centradas en un valor distinto.

¿Cómo lo puedo asegurar? Con un Violin plot

Violinplot



Resulta que B y C no tenían mucho que ver

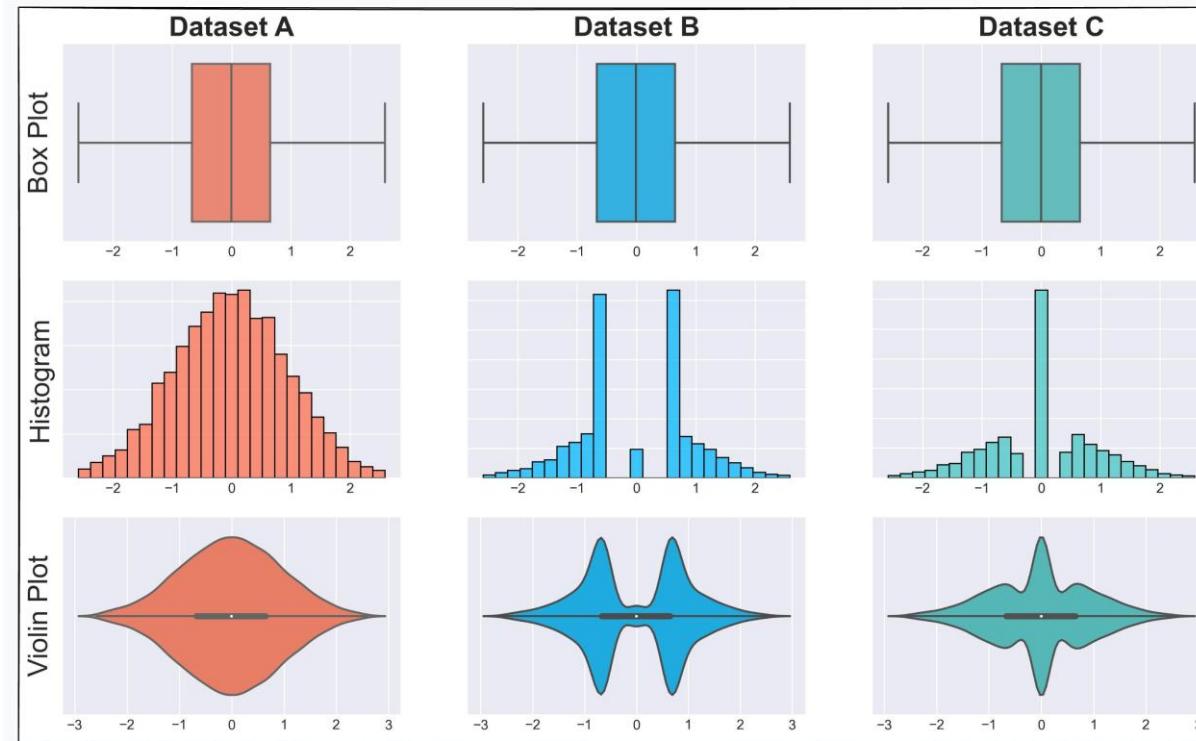
Violinplot

Box plots are Misleading



blog.DailyDoseofDS.com

Different datasets → Same Box Plots



Trabajo en equipo



A dibujar!

Dataset: [IMDB Movies Dataset](#).

10.000 películas, desde 1920 hasta 2025.

# Title	# Year	# Certificate	# Duration (min)	# Rating	# Metascore	# Votes	# Review Count
The Idea of You	2023.0	R	115.0	6.4	67.0	28,744	166
Kingdom of the Planet of	2023.0	PG-13	145.0	7.3	66.0	22,248	183
Unfrosted	2023.0	PG-13	97.0	5.5	42.0	18,401	333
The Fall Guy	2023.0	PG-13	126.0	7.3	73.0	38,953	384
Challengers	2023.0	R	131.0	7.7	82.0	32,517	194
Abigail	2023.0	R	109.0	6.8	62.0	27,284	168
Civil War	2023.0	R	109.0	7.5	75.0	64,014	610
Twisters	2023.0	Missing value	Missing value	Missing value	Missing value	0	0
Anyone But You	2023.0	R	103.0	6.1	52.0	82,215	373
The Ministry of Ungentlen	2023.0	R	120.0	7.0	57.0	21,084	117

A dibujar!

Siguiendo visualizaciones_2.py:

1. Abrir el dataset con Pandas
2. Generar quintiles de duración, deciles de rating, quintiles de metascore y cuartiles de reviews.
3. Generar columna de polémica (definimos como peli polémica si está en el top 25% de reviews escritas por usuarios)
4. Graficar histograma de Metascore, de duración y de rating
 - a) Elegir el número de bins para que el gráfico quede suave. Probar después con la siguiente instrucción: `int(np.ceil(np.log2(len(df)) + 1))`
 - b) Ponerle título al gráfico y a los ejes

A dibujar!

Siguiendo visualizaciones_2.py:

5. Hacer en un mismo gráfico, KDEplots de rating y metascore
 - a) ¿Quiénes son más exigentes, los usuarios (rating) o los críticos (metascore)
 - La variable rating es el promedio del rating de los usuarios. ¿Cambiará el gráfico si esa variable fuese la mediana?
 - b) ¿Se puede ver la correlación entre ambas con este tipo de gráfico? ¿Por qué? Si no se puede, hacer un gráfico que permita compararlas
6. Elegir una entre duración, rating o metascore. Calcular media, mediana, percentil 16 y desvío estándar.
 - a) Agregar media y mediana al gráfico como una línea. ¿Se comportan como las de una distribución simétrica?
 - b) Media, mediana, percentil 16 y desvío, ¿se comportan como los de una distribución normal? Agregarlos al gráfico
7. Graficar la distribución de años
 - a) Supongamos que llegamos a alguna conclusión importante con este dataset. ¿Les parece representativo para toda la historia del cine?

A dibujar!

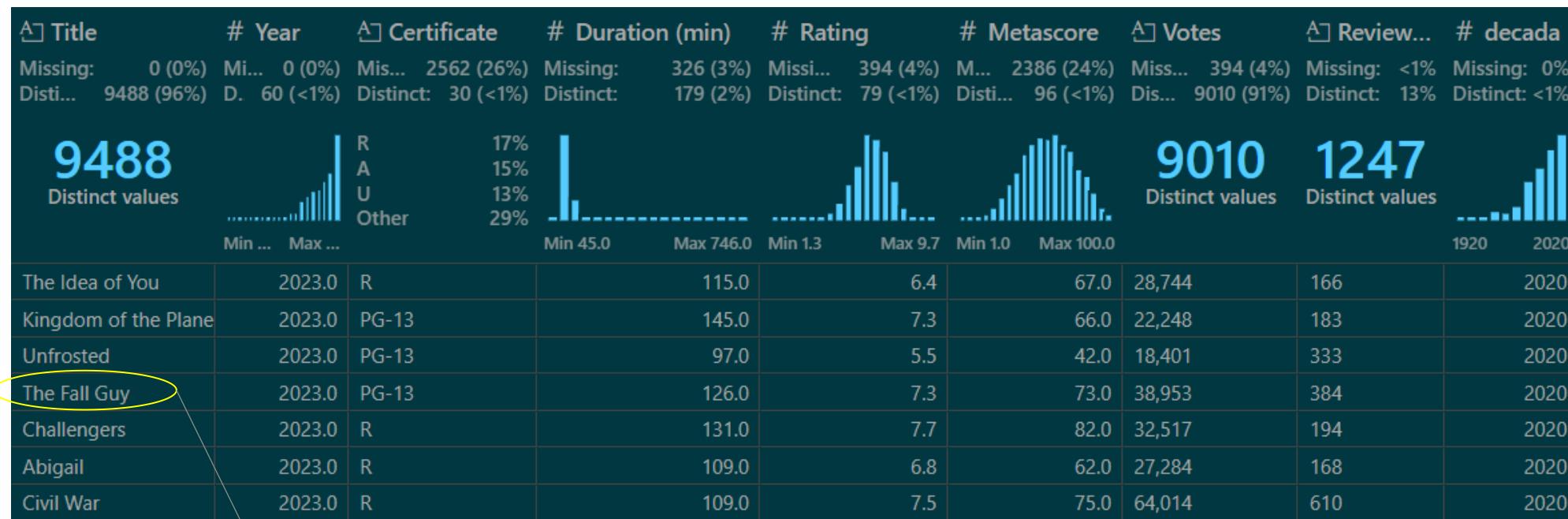
Siguiendo visualizaciones_2.py:

8. Hacer un boxplot de la duración de las pelis para los distintos deciles de rating.
 - a) Repetirlo sacando outliers del gráfico. ¿Ven algún patrón?
9. Hacer un violinplot del metascore por década
 - a) ¿Puede tener valores negativos? Hacer un boxplot y comparar
10. Graficar kdeplot de rating para quintiles de duración
 - a) ¿Hay relación? Verificar con otro gráfico
11. Hacer KDEplot de duración separado por quintiles de duración. ¿Tiene sentido?
12. Hacer boxplot (con y sin outliers) de reviews para distintos quintiles de duración. ¿Tienen relación?
13. Hacer boxplot (sin outliers) de duración para quintiles de metascore, separando además las pelis polémicas
 - a) ¿Duran lo mismo las pelis polémicas?
14. Hacer kdeplot de rating separando películas polémicas
 - a) ¿Por qué una mide tanto menos? Arreglar con **common_norm=False**

Análisis exploratorio

Todo lo que acaban de hacer, forma parte de lo que se llama análisis exploratorio.

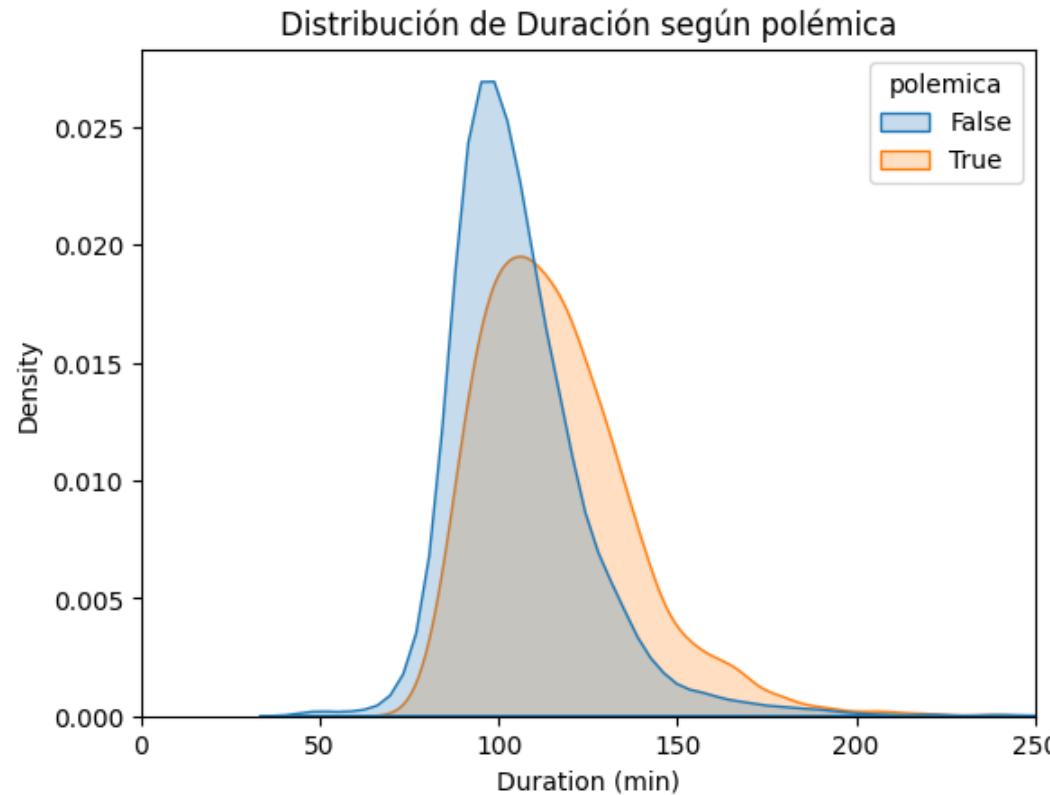
Mediante visualizaciones y metiéndole mano a la base, vemos las relaciones entre variables, si cumplen patrones, etc.



Peliculón

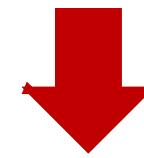
Análisis exploratorio

Sin embargo, hay cosas para las que se necesita más que sólo gráficos



¿Puedo asegurar que las pelis polémicas duran más?

¿Puedo cuantificar la probabilidad de que de verdad sea así, y no sea que mi muestra es especial?



Estadística inferencial:

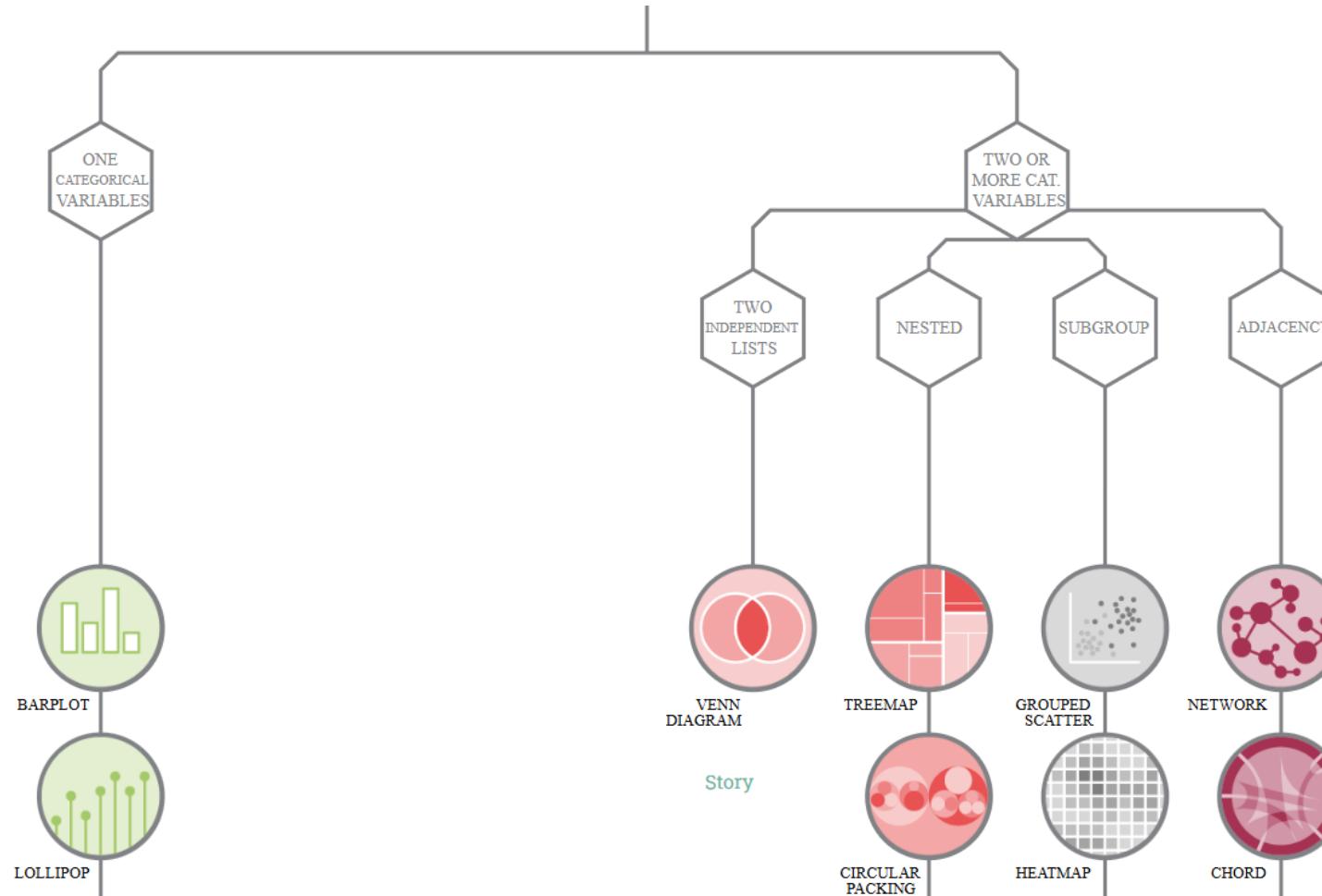
- Intervalos de confianza
- P-valor
- Potencia, significancia
- ...

(no se ve en esta materia)

Herramientas

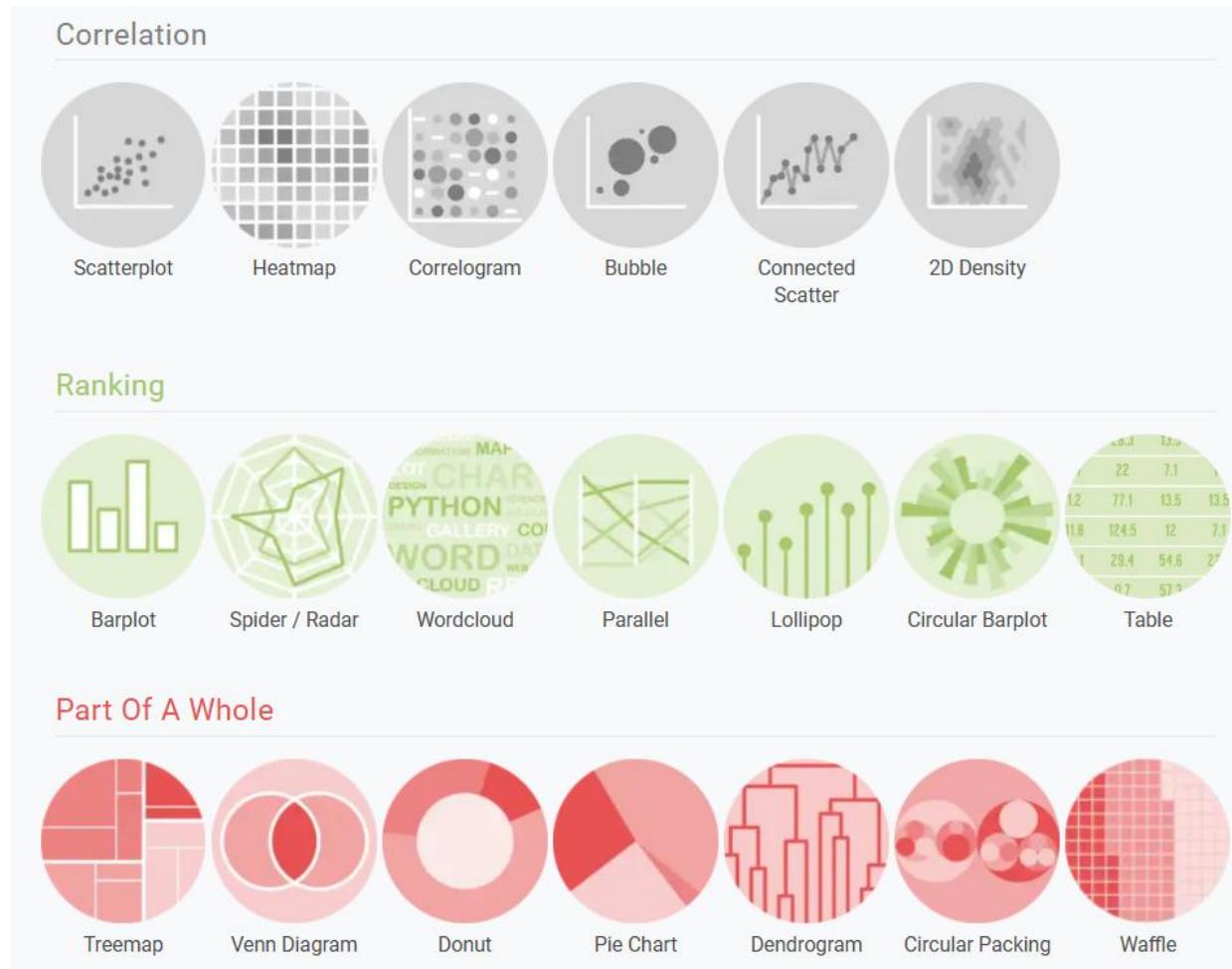
Herramientas

Data to viz: Según el tipo de datos que tengas, muestra qué gráficos puedes usar



Herramientas

[Python Graph Gallery](#): ejemplos de gráficos, con el código para hacerlos



Herramientas

[Viz Palette](#): Cargas tu paleta de colores y muestra cómo se ve con distintos tipos de daltonismo (o en blanco y negro)

Color Population:

No Color Deficiency - 96% Deuteranomaly - 2.7% Protanomaly - 0.66% Protanopia - 0.59% Deutanopia - 0.56%

Greyscale

Sample font

Randomize Data

Stroke: Dark None

