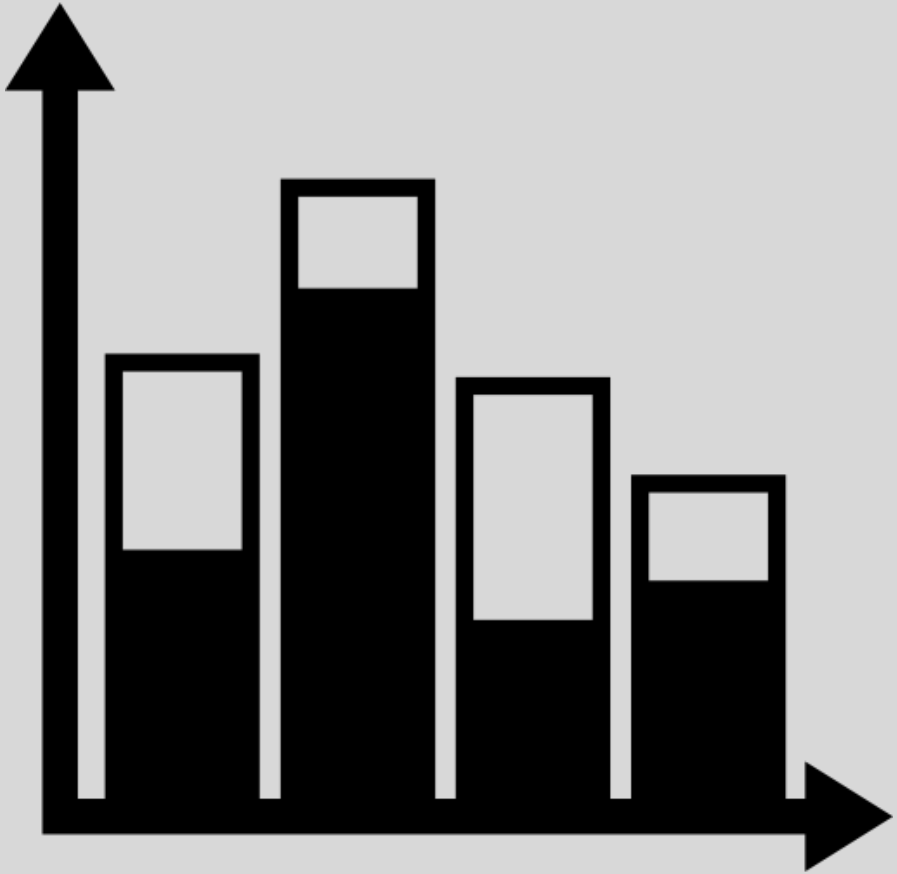


DigitalHouse > Coding School

DATA SCIENCE

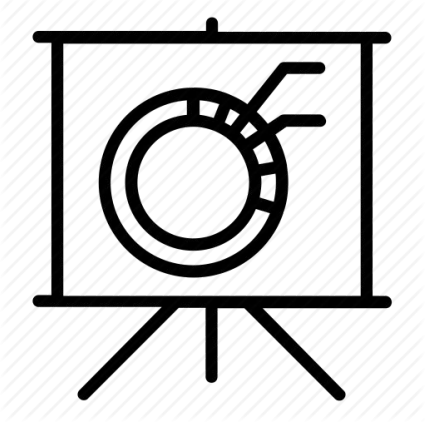
MÓDULO 2

Visualización

OBJETIVOS DE LA CLASE

1. Adquirir los principios de visualización de datos

2. Conocer y aplicar herramientas de visualización en Python



¿Cuál afirmación sobre visualización es válida?

A. Debemos siempre poder obtener una intuición sobre los datos.
B. La visualización debería siempre ser hecha antes de empezar a modelar.
C. Si modelamos sin haber visto los datos primero, seguramente estaremos yendo hacia un problema más adelante.
D. Todas.
E. Ninguna.

Tips sobre visualización de datos en Designing data visualization

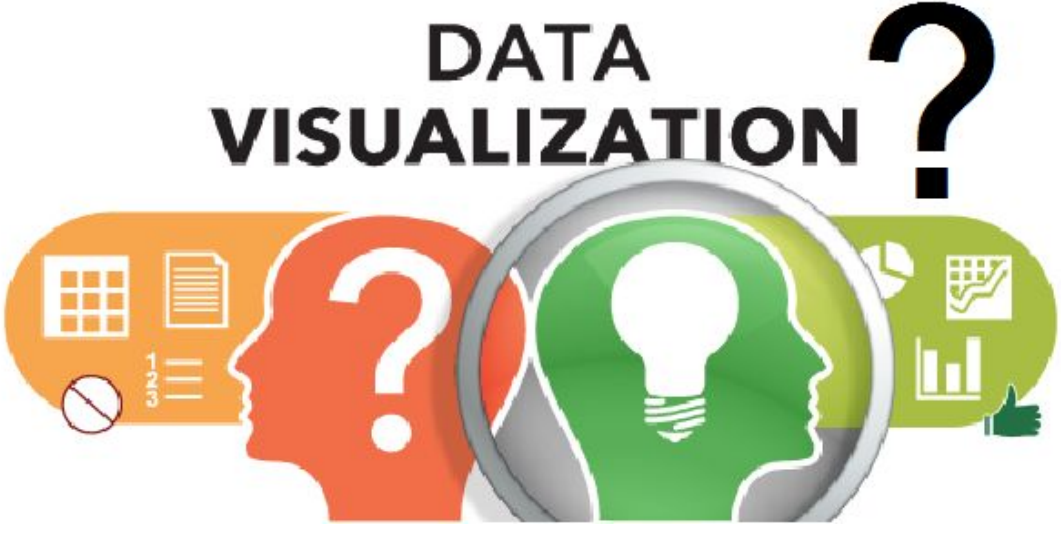
¿Cuál afirmación sobre visualización es válida?

A. Debemos siempre poder obtener una intuición sobre los datos.
B. La visualización debería siempre ser hecha antes de empezar a modelar.
C. Si modelamos sin haber visto los datos primero, seguramente estaremos yendo hacia un problema más adelante.
D. Todas.
E. Ninguna.

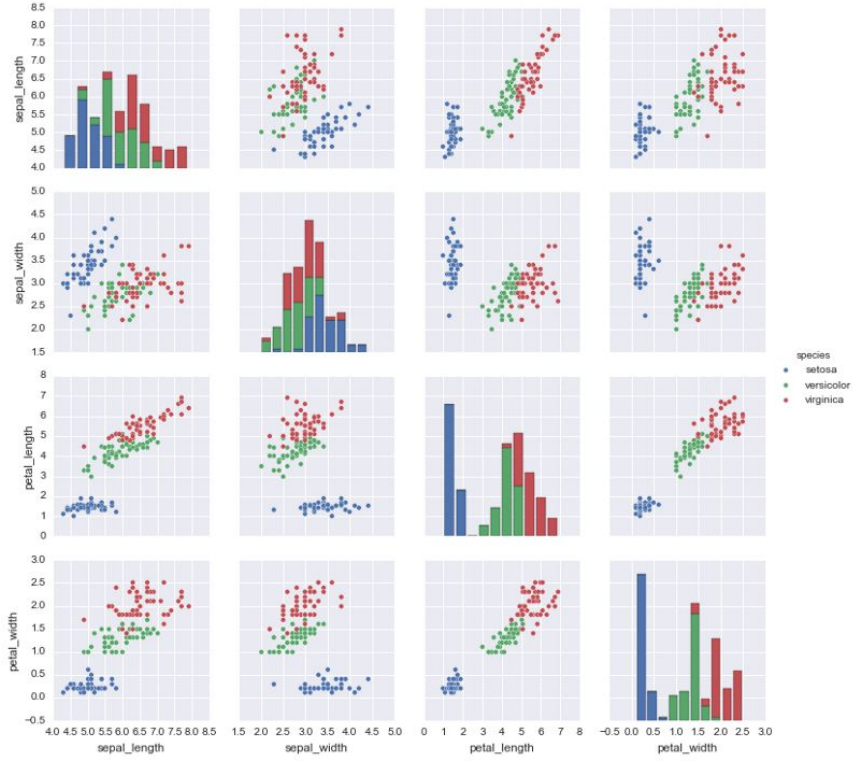
Tips sobre visualización de datos en Designing data visualization

¿PARA QUÉ VISUALIZAR DATOS?

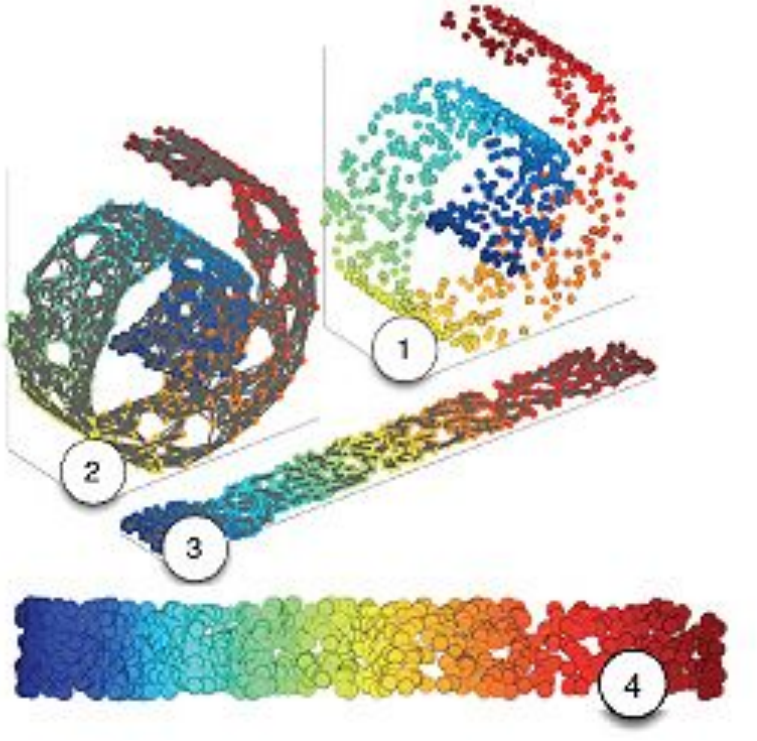
Dada la manera que el **cerebro humano procesa la información**, utilizar cuadros o gráficos para **visualizar** grandes volúmenes de datos complejos es mucho más fácil que sumergirse planillas o informes.



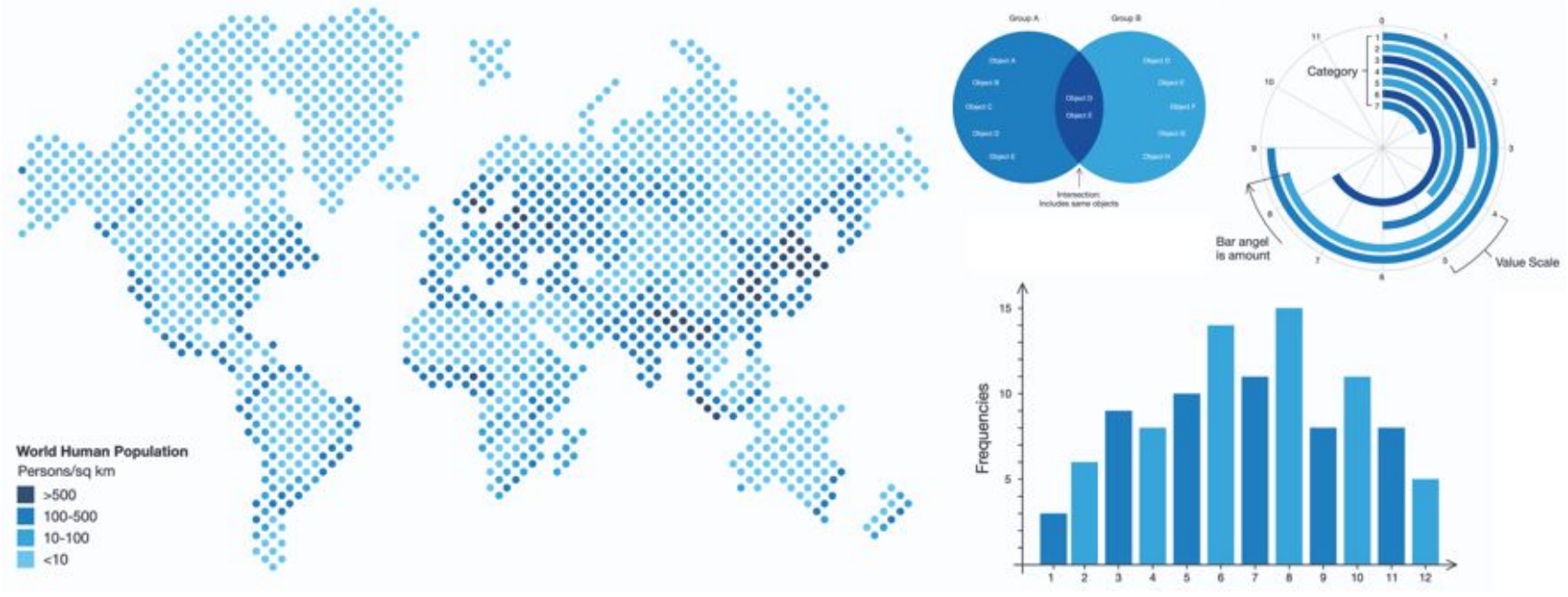
La visualización de datos permite: **Explorar los datos**



La visualización de datos permite: **Expresar relaciones complejas**



La visualización de datos permite: **Condensar información y comunicar de manera más potente**



¿POR QUÉ VISUALIZAR LOS DATOS?

Consideremos 4 datasets que contienen dos variables o columnas (x,y).
La siguiente información estadística resume las características de 4 grupos:

Cuarteto de Anscombe

Plot	sum X	sum Y	avg X	avg Y	stdev X	stdev Y
I	99.0	82.5	9.00	7.50	3.32	2.03
II	99.0	82.5	9.00	7.50	3.32	2.03
III	99.0	82.5	9.00	7.50	3.32	2.03
IV	99.0	82.5	9.00	7.50	3.32	2.03

¿Podemos concluir que los datasets son iguales? ¿o son diferentes?

¿POR QUÉ VISUALIZAR LOS DATOS?

Ahora observemos los 4 datasets y grafiquemos cada uno:

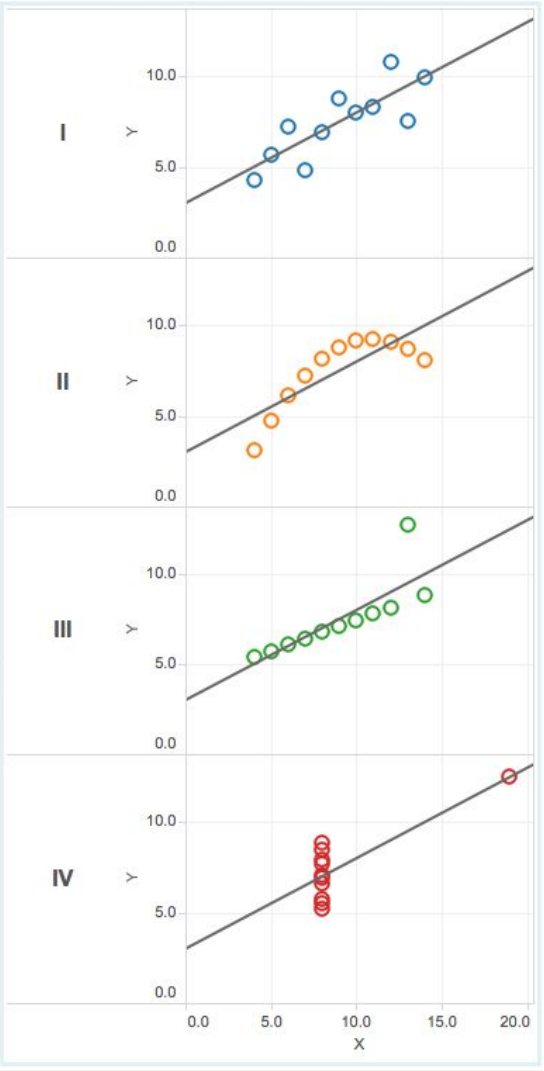
Este ejemplo nos recuerda que la información sintética tiene que ser complementada con mayor conocimiento del dominio.

Visualizar los datos puede evitar hacer supuestos incorrectos.

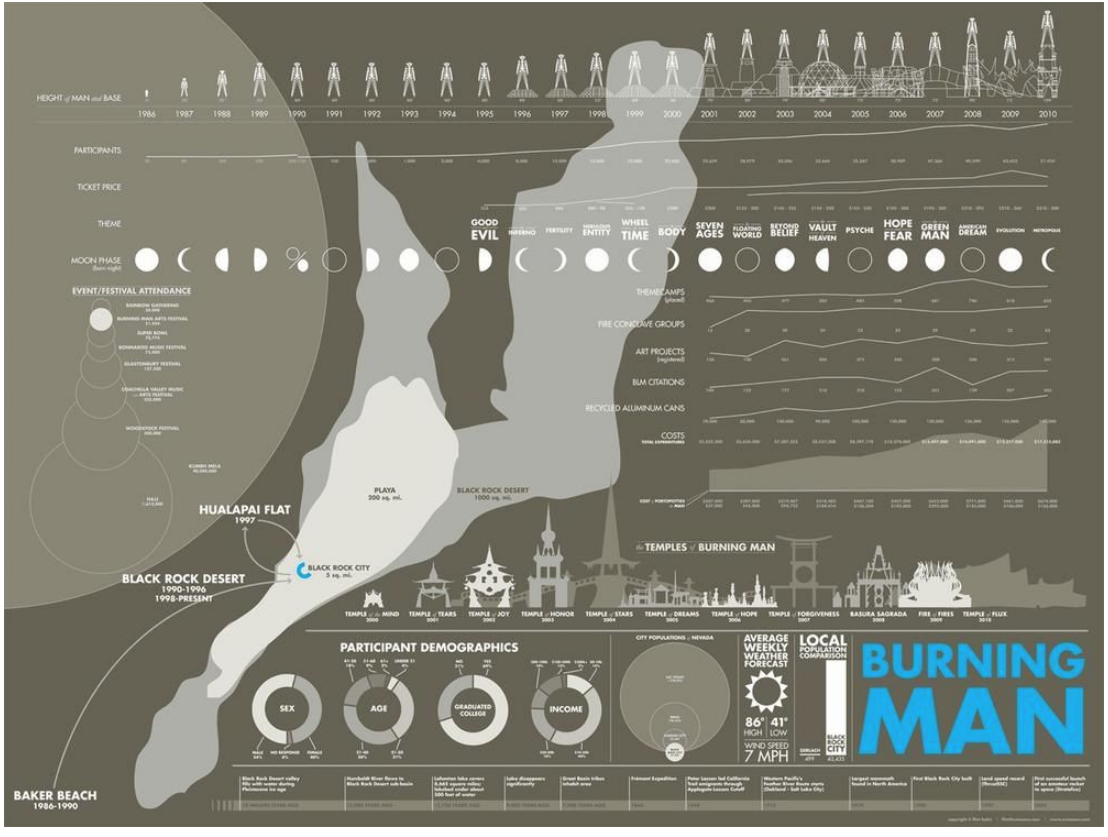
Cuarteto de Anscombe

I	II	III	IV
(4, 4.3)	(4, 3.1)	(4, 5.4)	(8, 5.3)
(7, 4.8)	(5, 4.7)	(5, 5.7)	(8, 5.6)
(5, 5.7)	(6, 6.1)	(6, 6.1)	(8, 5.8)
(8, 7.0)	(7, 7.3)	(7, 6.4)	(8, 6.6)
(6, 7.2)	(14, 8.1)	(8, 6.8)	(8, 6.9)
(13, 7.6)	(8, 8.1)	(9, 7.1)	(8, 7.0)
(10, 8.0)	(13, 8.7)	(10, 7.5)	(8, 7.7)
(11, 8.3)	(9, 8.8)	(11, 7.8)	(8, 7.9)
(9, 8.8)	(12, 9.1)	(12, 8.2)	(8, 8.5)
(14, 10)	(10, 9.1)	(14, 8.8)	(8, 8.8)
(12, 10.8)	(11, 9.3)	(13, 12.7)	(19, 12.5)

Plot	sum X	sum Y	avg X	avg Y	stdev X	stdev Y
I	99.0	82.5	9.00	7.50	3.32	2.03
II	99.0	82.5	9.00	7.50	3.32	2.03
III	99.0	82.5	9.00	7.50	3.32	2.03
IV	99.0	82.5	9.00	7.50	3.32	2.03



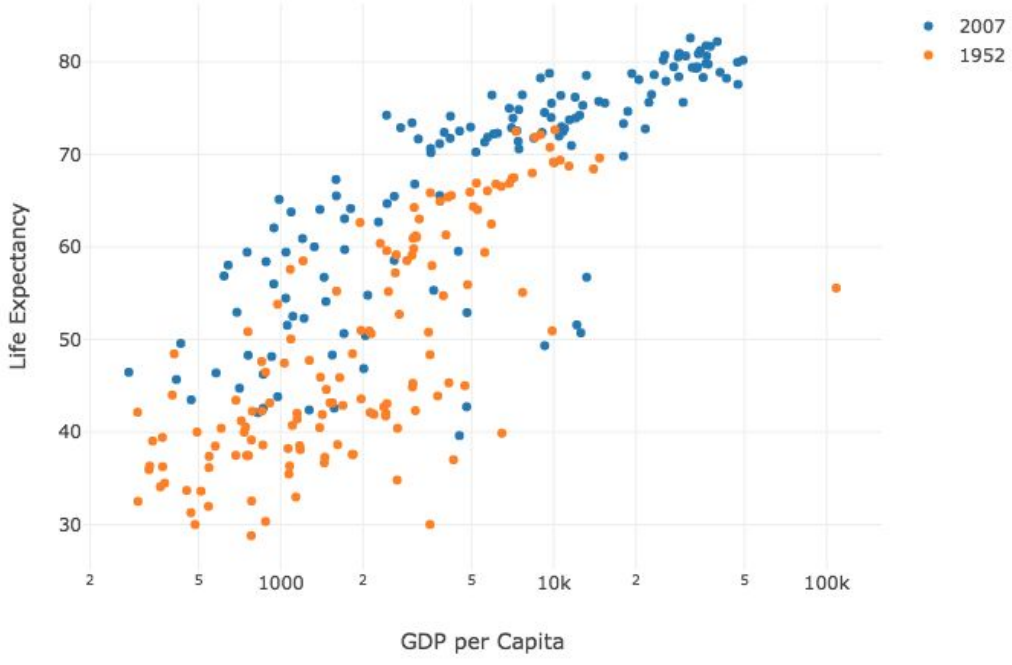
INFOGRAFÍAS



1. Tratamiento customizado de la información (dibujo manual).
2. Específico para el dataset.
3. Foco en aspectos estéticos.
4. Relativamente pobre de datos.

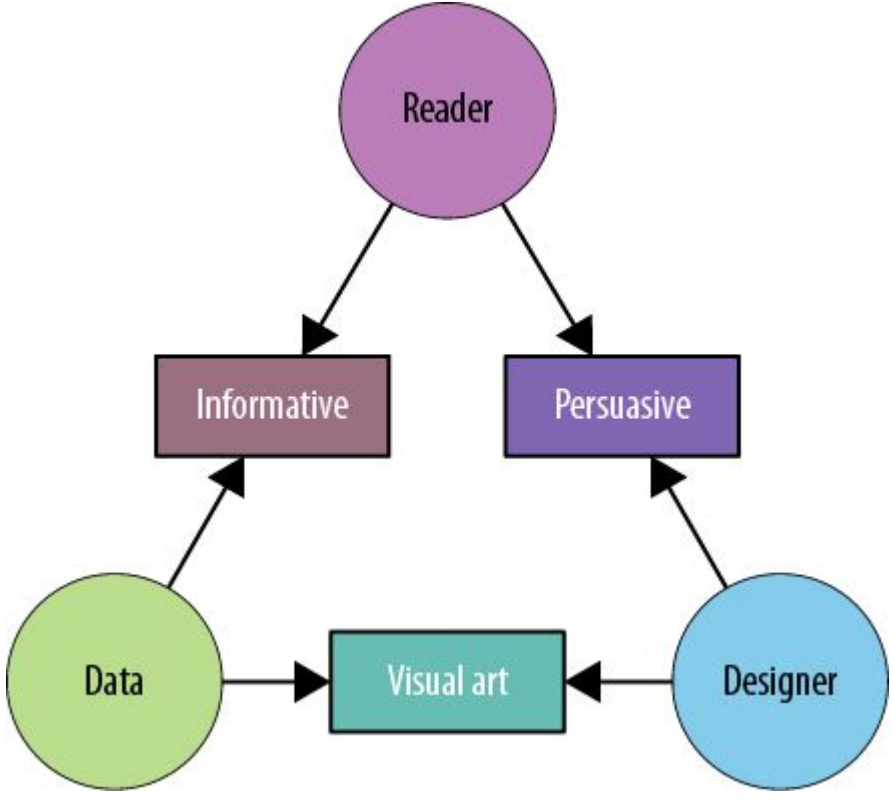
DigitalHouse > Coding School

VISUALIZACIÓN DE DATOS

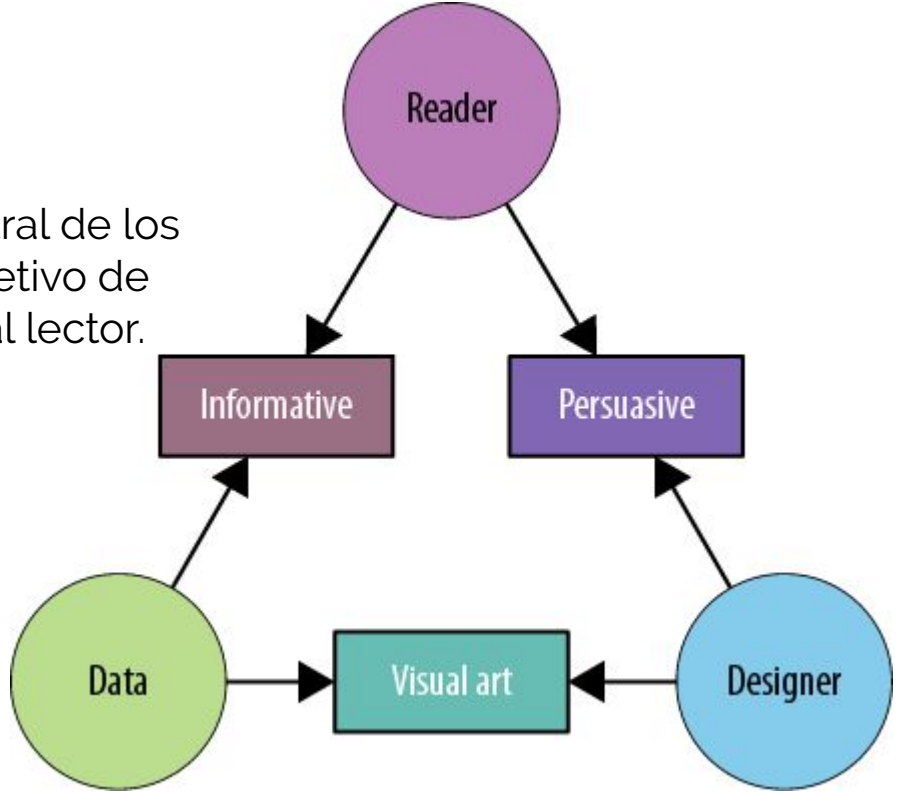


1. Generado por algoritmos.
2. Aplicables a datasets diferentes (el mismo tipo de gráfico se puede usar para diferentes bases de datos).
3. Generalmente son austeros estéticamente.
4. Rico en cantidad de datos.

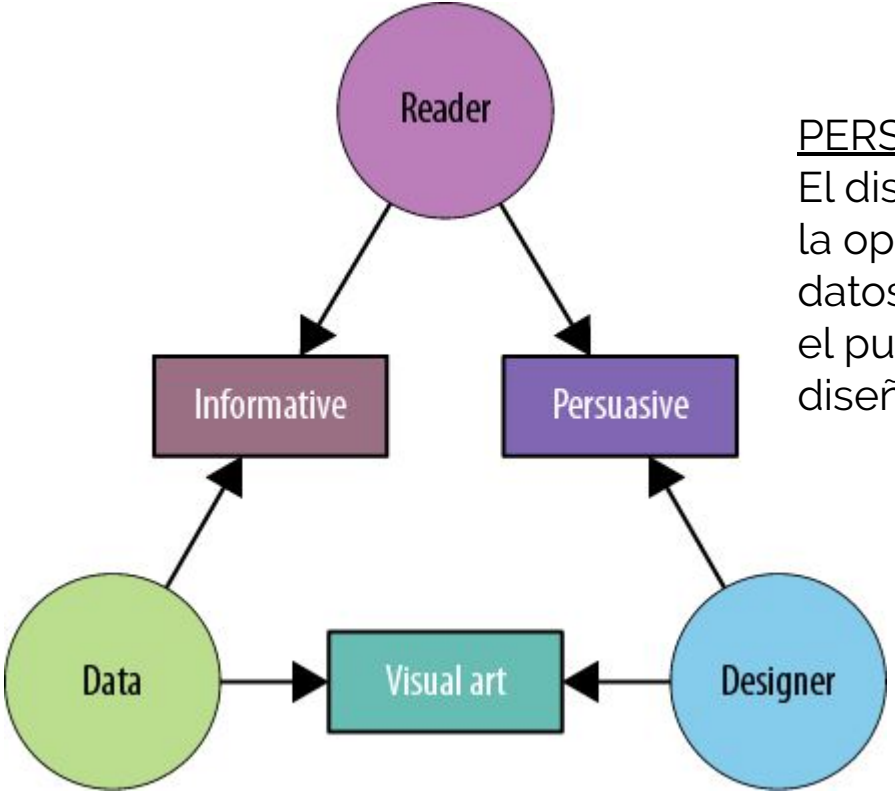
DISEÑO BASADO EN LAS RELACIONES DE LA TRÍADA DESIGNER-READER-DATA



INFORMATIVA
Presentación neutral de los hechos con el objetivo de educar/informar al lector.



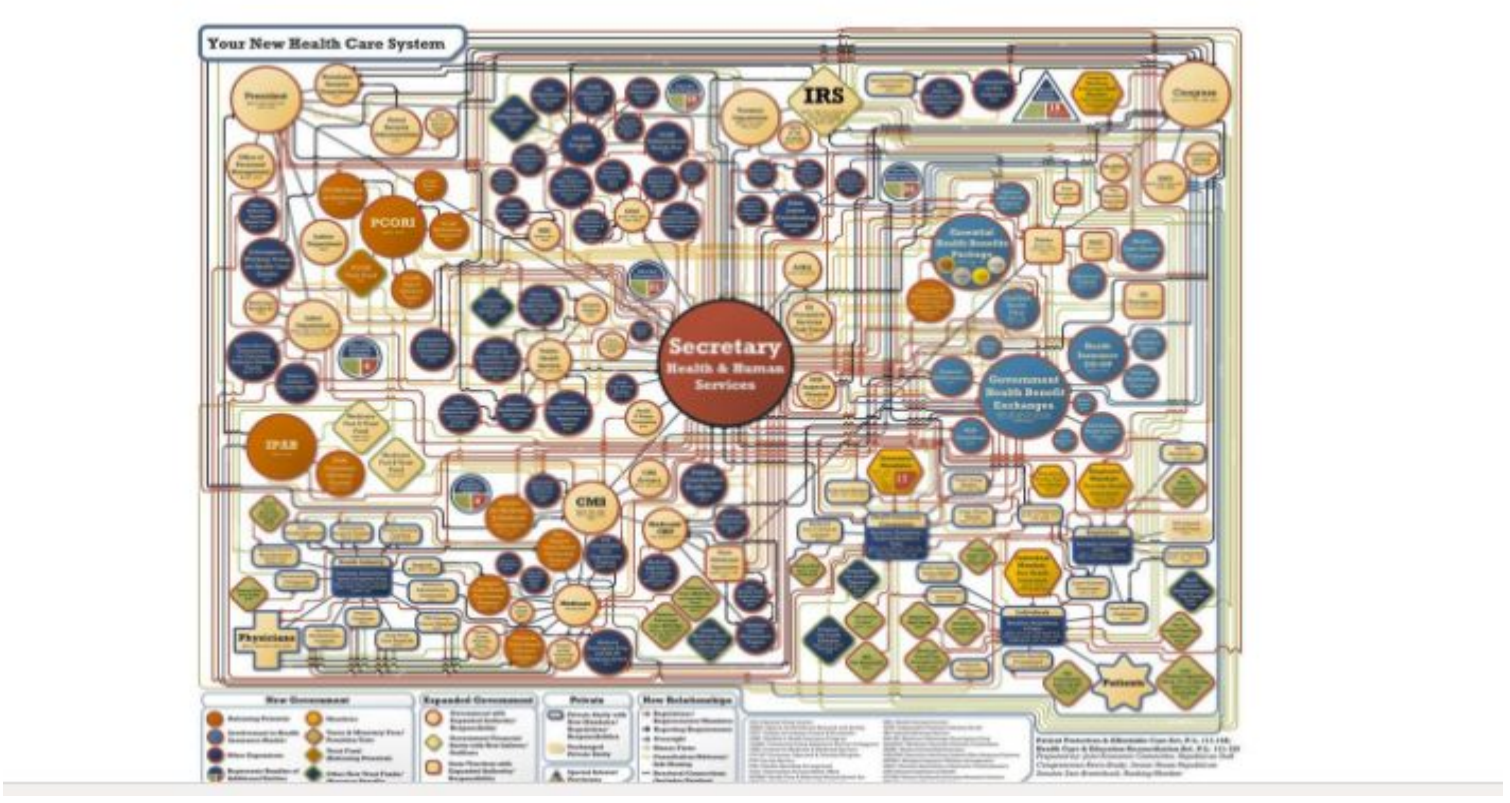
DISEÑO BASADO EN LAS RELACIONES DE LA TRÍADA DESIGNER-READER-DATA



PERSUASIVA

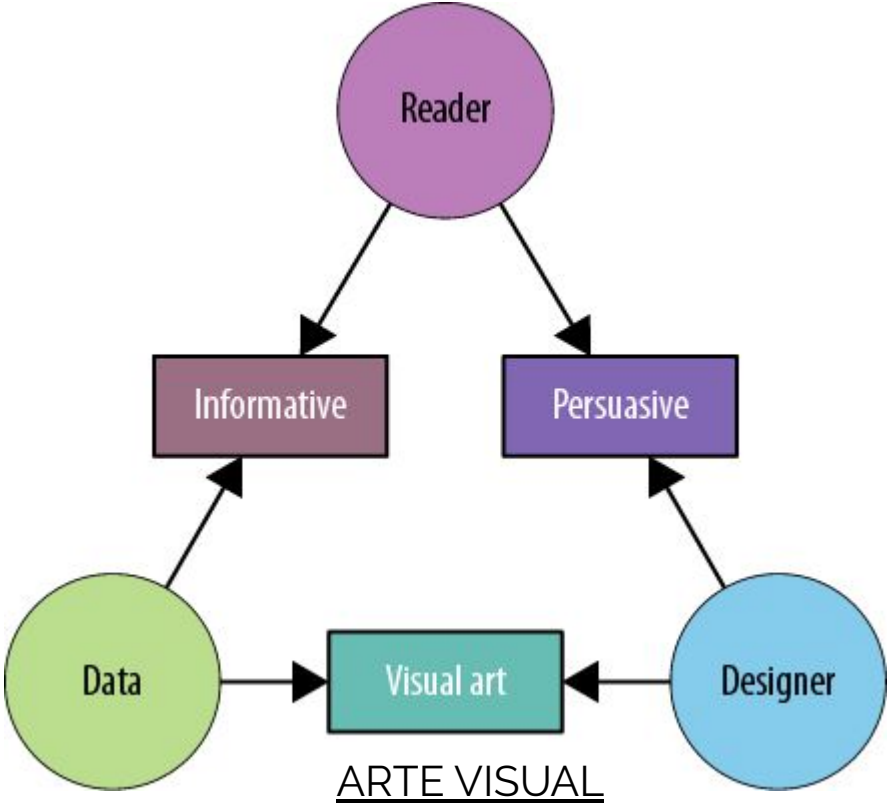
El diseñador busca cambiar la opinión del lector. Los datos sirven para sustentar el punto de vista del diseñador.

Persuasiva



Representación hecha por los Republicanos del plan de salud propuesto por los Demócratas en 2010. Claramente pretende exagerar la complejidad del sistema.

DISEÑO BASADO EN LAS RELACIONES DE LA TRÍADA DESIGNER-READER-DATA



ARTE VISUAL

Expresa el vínculo entre el diseñador y los datos.
El lector puede no ser tenido en cuenta.

DATA VIZ Y PERCEPCIÓN VISUAL

Algunos atributos generan un impacto mayor en nuestro cerebro.

PERCEPCIÓN VISUAL

¿Cuántos cuadrados hay? ¿Cuántos círculos?

PERCEPCIÓN VISUAL

¿Cuántos cuadrados hay? ¿Cuántos círculos?

PERCEPCIÓN VISUAL

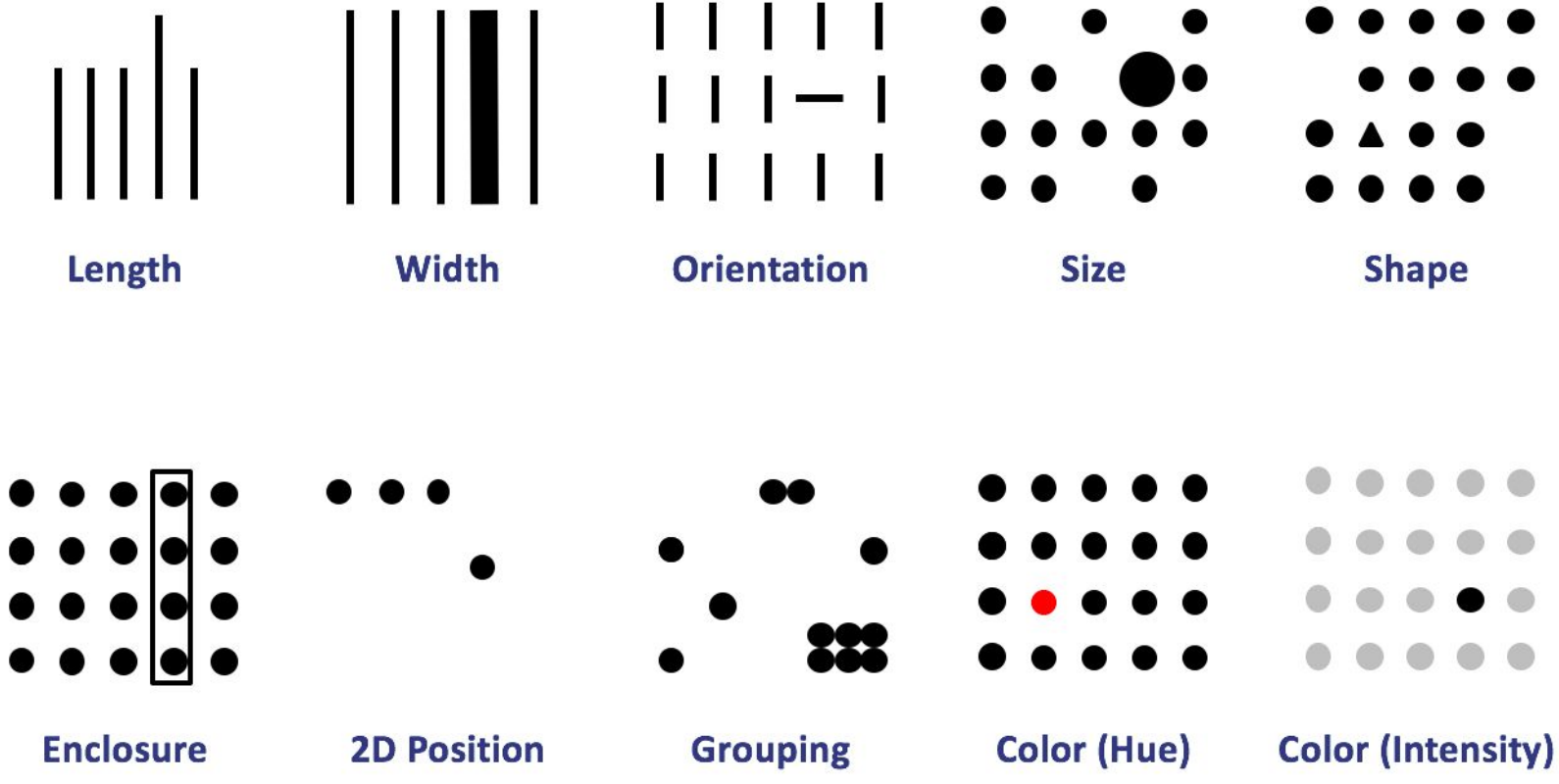
¿Cuántos cuadrados hay? ¿Cuántos círculos?

PERCEPCIÓN VISUAL

¿Cuántos cuadrados hay? ¿Cuántos círculos?
¿Qué imagen transmite mejor la información?

ALGUNOS RECURSOS...

Contamos con diferentes recursos visuales para transmitir información:



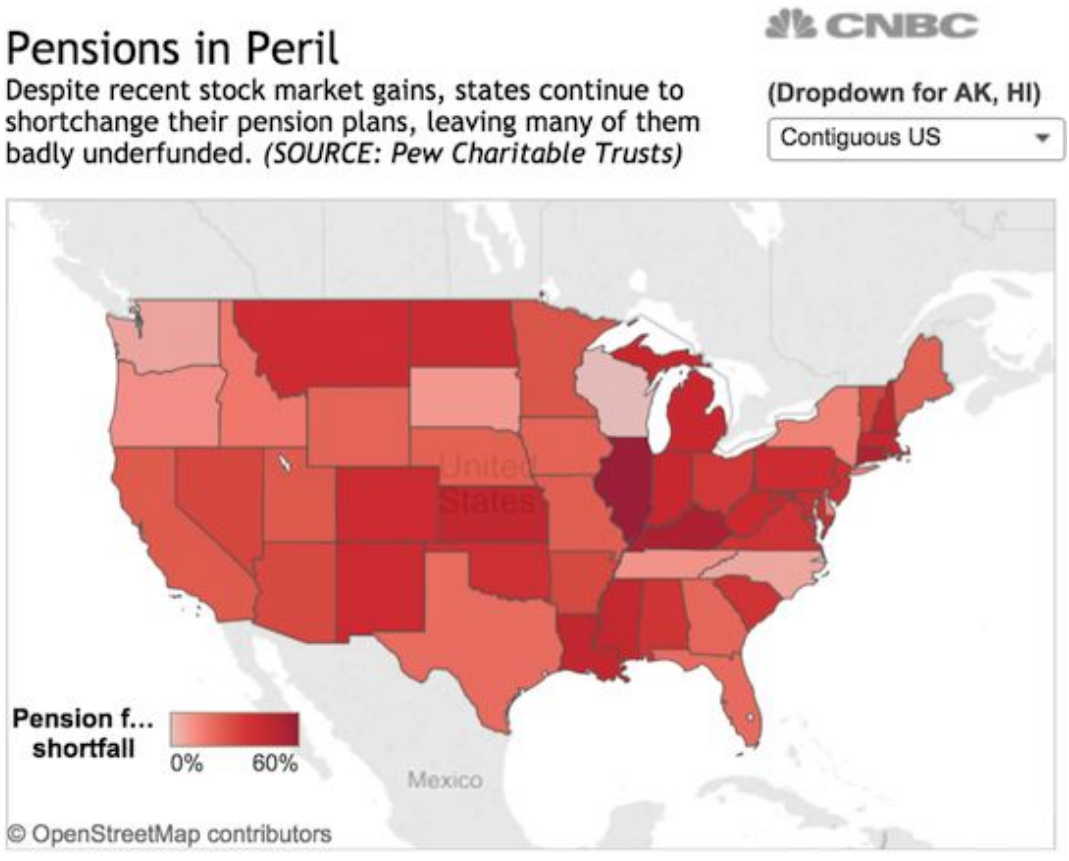
COLOR

Los usos del color en visualización de datos permiten indicar:

- **Secuencia**
- **Divergencia**
- **Categoría**

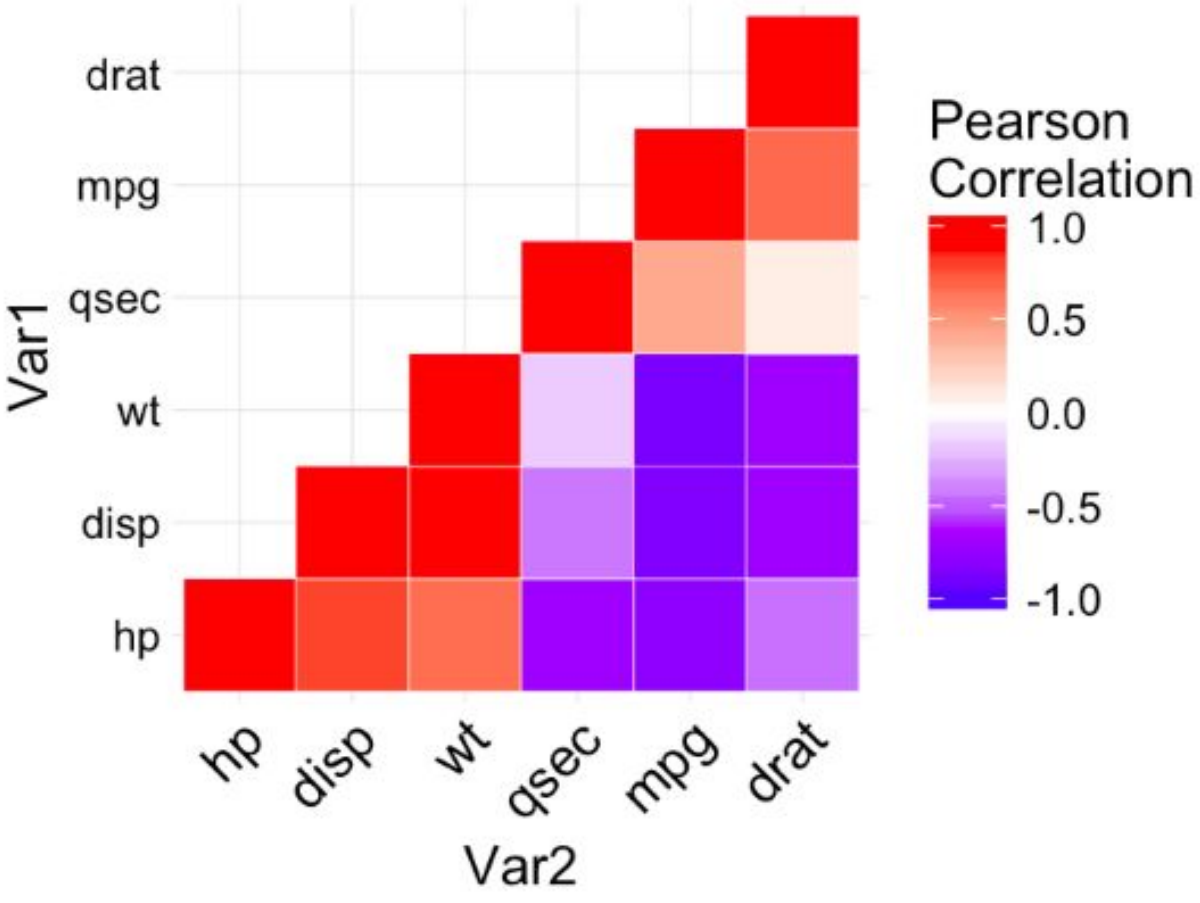
COLOR: SECUENCIA

Los colores ***secuenciales*** se utilizan para mostrar valores ordenados de menor a mayor:



COLOR: DIVERGENCIA

Los colores ***divergentes*** se utilizan para mostrar valores ordenados que tienen un valor crítico, tales como un promedio o cero:



COLOR: CATEGORÍA

Los colores ***categóricos*** se utilizan para distinguir datos pertenecientes a diferentes grupos.

En particular, se asocia a la representación de ***variables categóricas***.

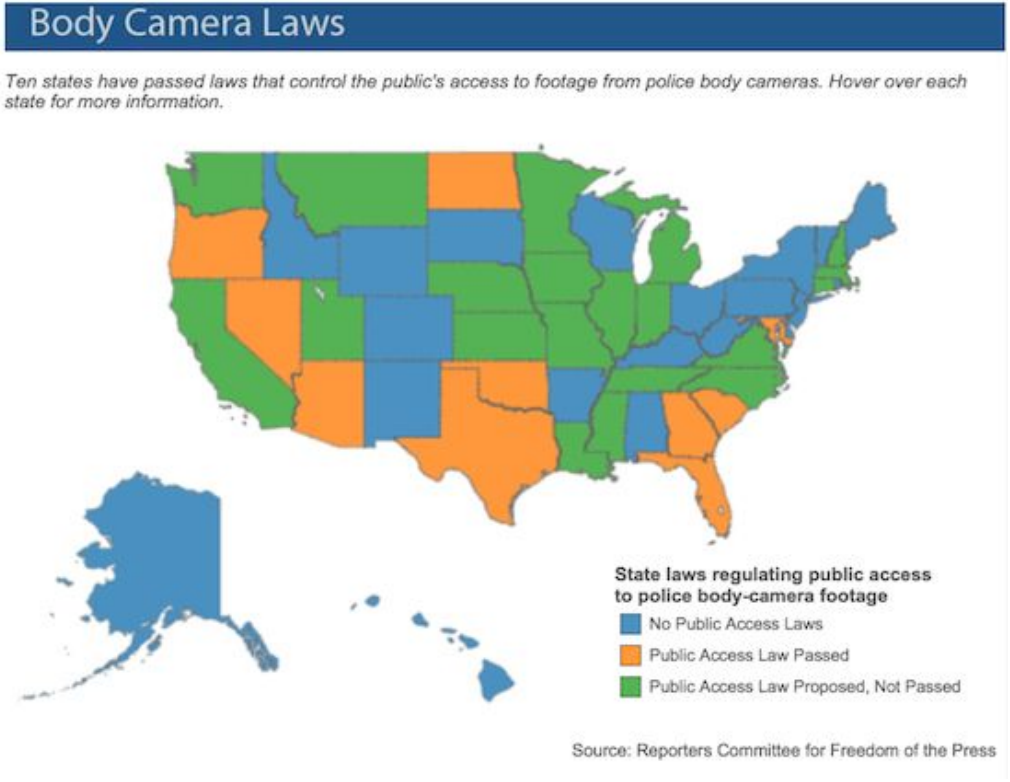


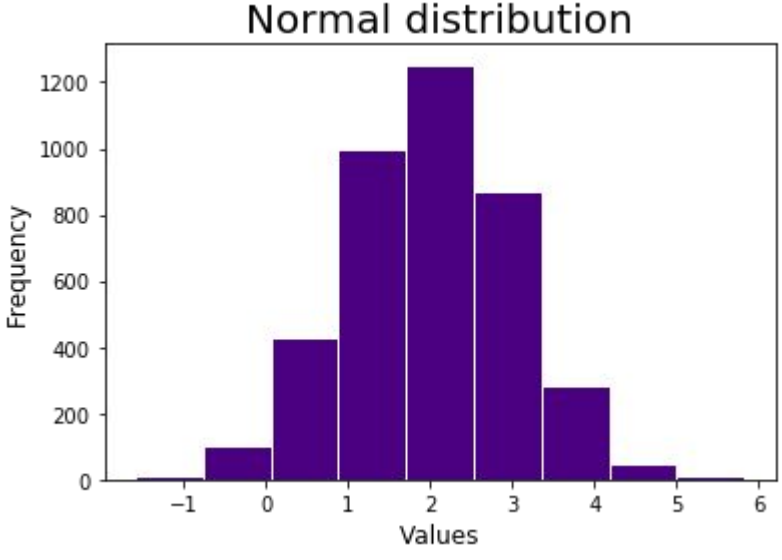
DIAGRAMA Y GRÁFICOS

Además de los atributos de visualización, podemos considerar qué tipo de diagrama o gráfico usar. Veamos algunos de los diagramas y gráficos más utilizados:

- **Histogramas**
- **Diagrama de caja** (box plot)
- **Dispersión** (scatter plot)
- **De series (líneas)** (plot)
- **Barras** (bar chart)
- **Tortas** (pie chart)

HISTOGRAMAS

Los histogramas nos indican qué forma toma la **distribución de frecuencias de una variable.** En otras palabras, muestran cómo y en qué valores se concentran los datos. Cuando sea posible identificar la distribución, podremos discernir, por ejemplo, si es válido suponer 'normalidad' o 'uniformidad', por ejemplo, para trabajar con determinados métodos.



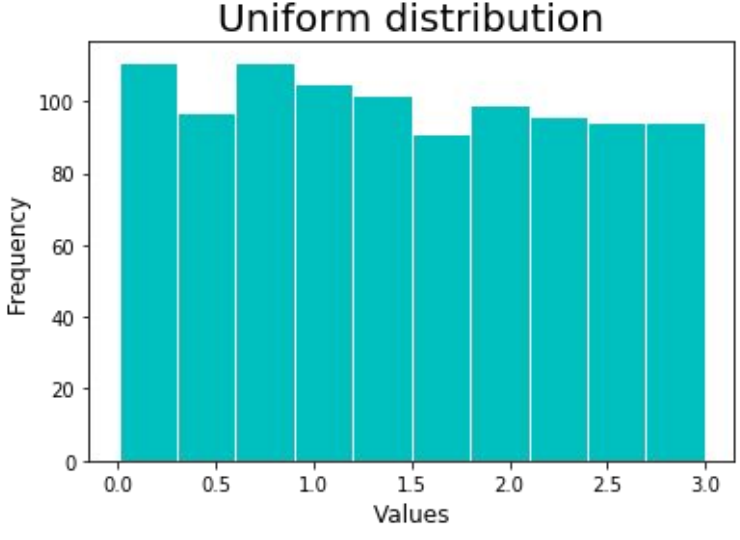


DIAGRAMA DE CAJA (BOX PLOT)

Un boxplot muestra la distribución de los valores de una variable, destacando los valores críticos que sirven de límite de los rangos intecuartílicos (RIC). Hay distintos tipos:

- Box plots que representan el rango completo de valores que toma la variable, segmentando su distribución en cuartiles .

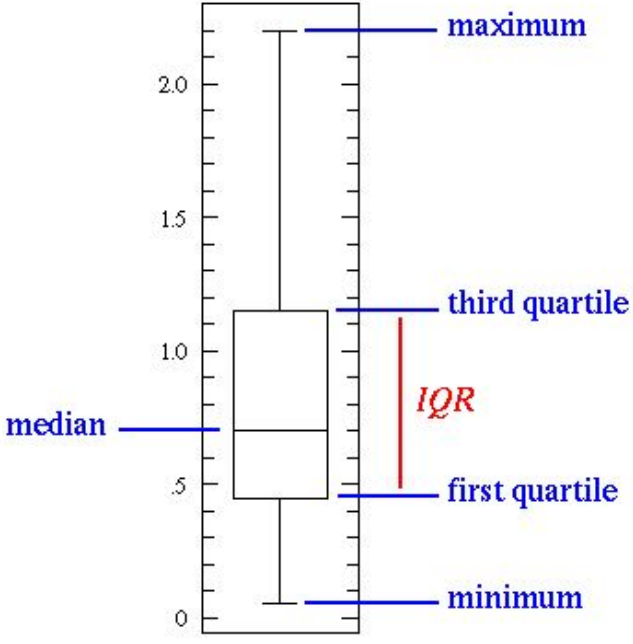


DIAGRAMA DE CAJA (BOX PLOT)

- Los box plots que excluyen los extremos de la distribución a partir de
  - 1) considerar la distribución del RIC (+/-1,5) o
  - 2) excluir percentiles extremos de forma simétrica.
- En estos casos, los outliers deben ser ploteados (círculos, puntos, estrellas).

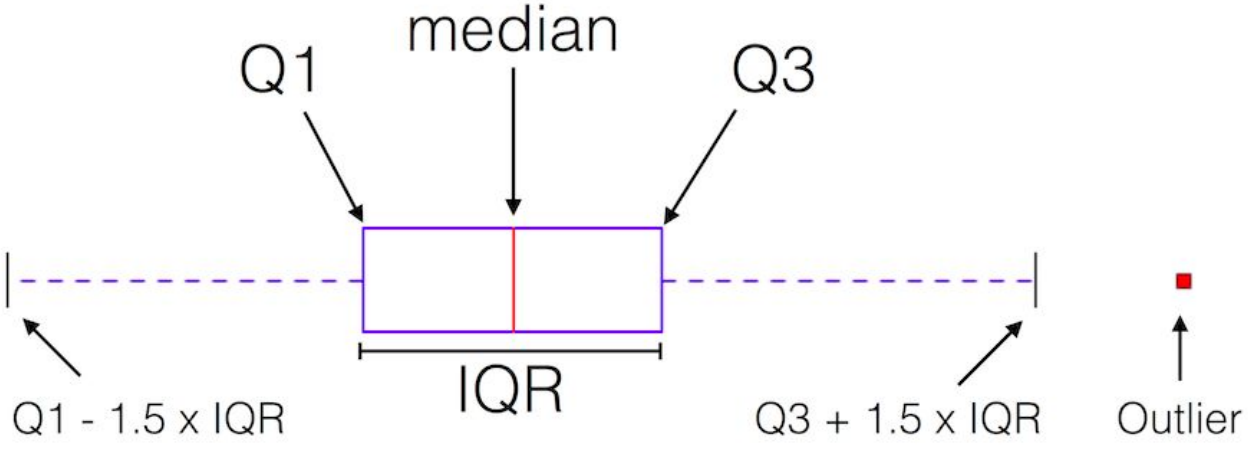


GRÁFICO DE DISPERSIÓN (SCATTER PLOT)

Los gráficos de dispersión son una buena manera para conocer principales tendencias, concentraciones y outliers.

Esta información puede orientar hacia dónde profundizar
la investigación.

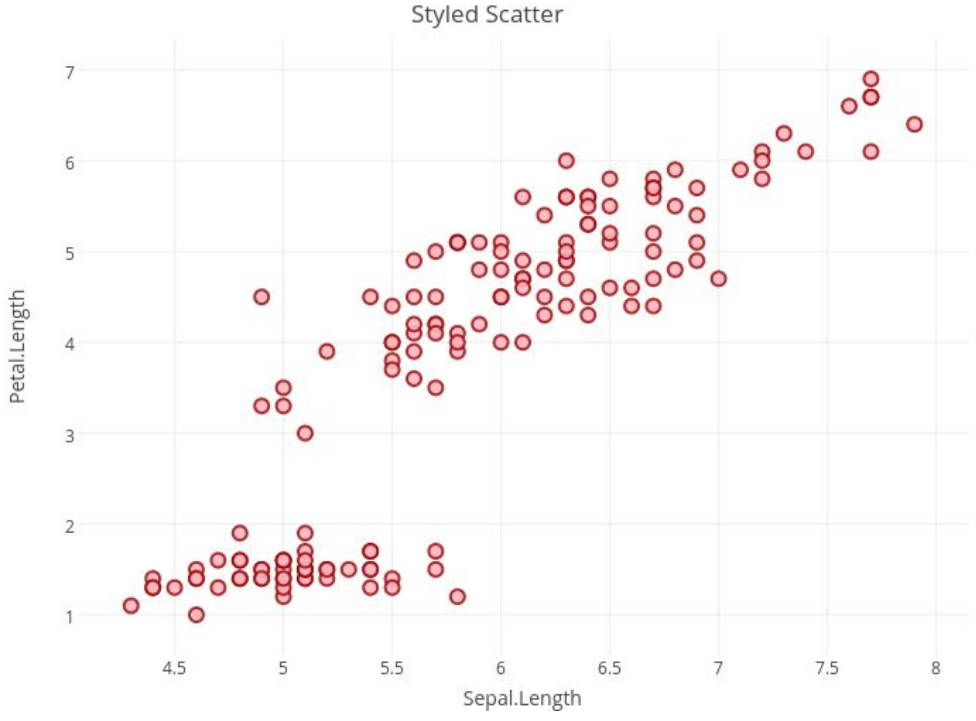


GRÁFICO DE LÍNEAS (PLOT)

Los gráficos de líneas permiten observar cómo es la relación existente entre dos variables continuas. En general, se utilizan para graficar la evolución temporal de una variable. La unión de los puntos presenta una idea sobre su recorrido, mostrando picos y valles de la serie.

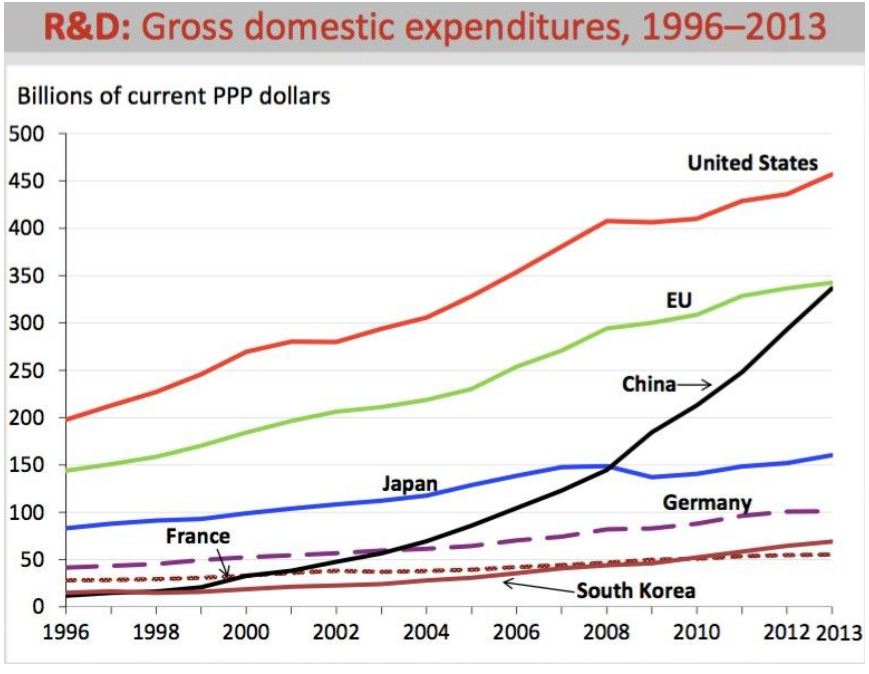
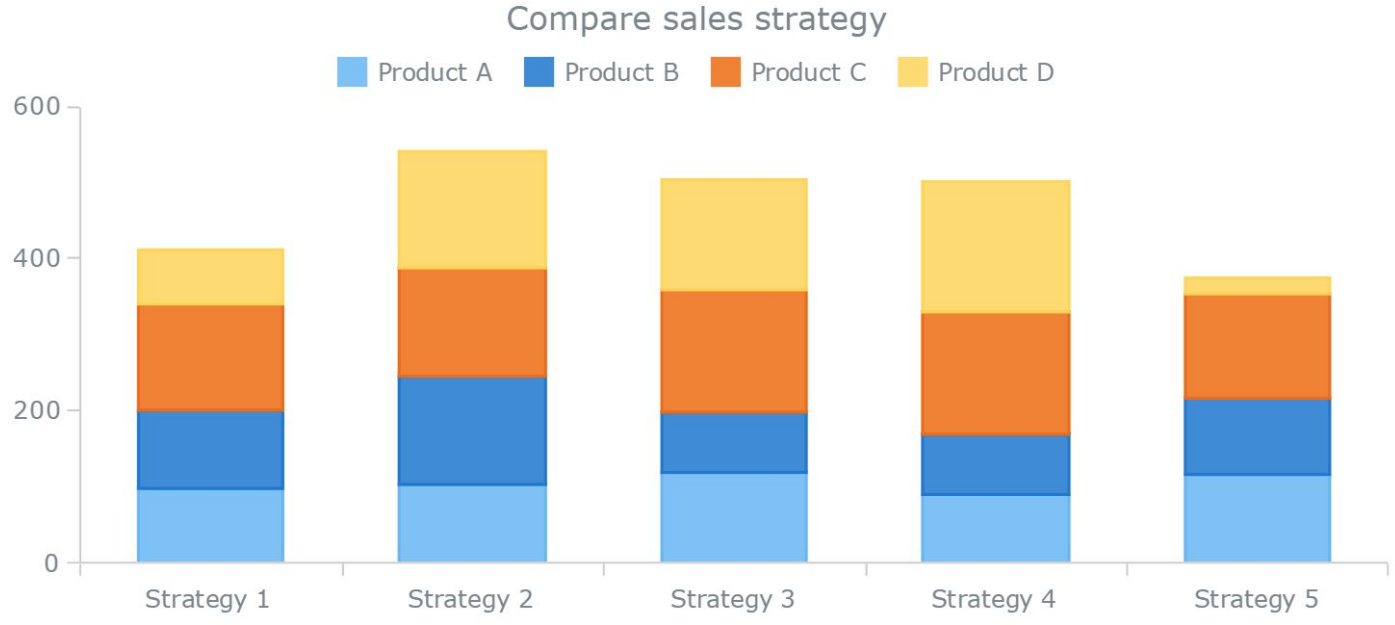


GRÁFICO DE BARRAS

Es una de las formas más utilizadas para visualizar datos. ¿Por qué? Es fácil comparar, mostrando rápidamente máximos y mínimos. Es efectivo para mostrar datos numéricos que son separables en diferentes categorías.

GRÁFICOS DE BARRAS

Los gráficos de barras apiladas también son útiles para comparar distribuciones de distintas poblaciones o series de elementos.

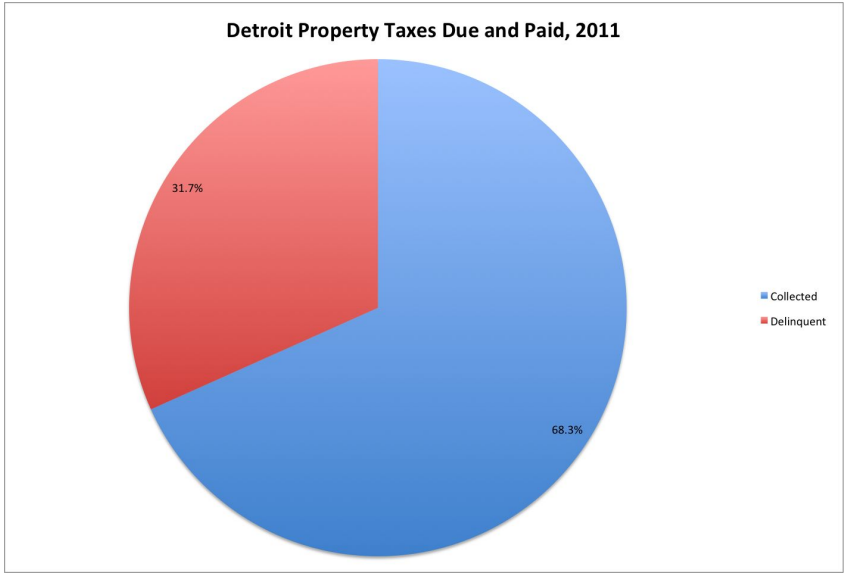


GRÁFICOS DE BARRAS

En este caso, el ploteo simultáneo e independiente de dos variables permite visualizar niveles y distribución relativa de cada variable y, a la vez, realizar comparaciones entre ellas.

GRÁFICO DE TORTAS

Se pueden usar para mostrar proporciones relativas o porcentajes (y "pocas porciones"); para varios datos o desagregaciones, suelen ser reemplazados por gráficos de barras.



Escenario de utilización:

- 2 o 3 "porciones" a mostrar
- Tamaño de "porciones" significativamente diferentes

Crítica a los gráficos de torta:
The Worst Chart In The World

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data

Examples of use

Income distribution, population (age/sex) distribution

Chart types

histogram

The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

boxplot

Summarise multiple distributions by showing the median (centre) and range of the data

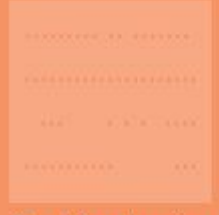
violin

Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

population-pyramis

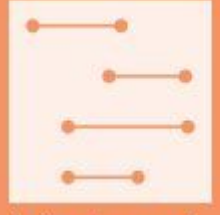
A standard way for showing the age and sex breakdown of a population distribution; effectively, back to back histograms

dot-plot-strip



Good for showing individual values in a distribution, can be a problem when too many dots have the same value

dot-plot



A simple way of showing the range (min/max) of data across multiple categories.

DigitalHouse > Coding School

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other)

Examples of use

Inflation & unemployment, income & life expectancy

Chart types

scatterplot

The standard way to show the relationship between two variables, each of which has its own axis

line-column

A good way of showing the relationship between an amount (columns) and a rate (line)

scatterplot-connected

Usually used to show how the relationship between 2 variables has changed over time

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable

XY-heatmap

A good way of showing the patterns between 2 categories of data, less good at showing fine differences in amounts

Change v Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries: Choosing the correct time period is important to provide suitable context for the reader

Examples of use

Share price movements, economic time series

Chart types

line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points

column-timeline

Columns work well for showing change over time - but usually best with only one series of data at a time

column-line-timeline

A good way of showing the relationship over time between an amount (columns) and a rate (line)

stock-price

Usually focused on day-to-day activity, these charts show opening/closing and hi/low points of each day

slope

Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story

area

Use with care. These are good at showing changes to total, but seeing change in components can be very difficult.

DigitalHouse > Coding School

Part to whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead

Examples of use

Fiscal budgets, company structures, national election results

Chart types

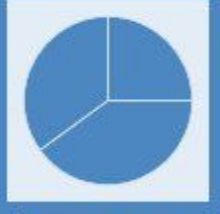
column-stacked

A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

bar-stacked-proportional

A good way of showing the size and proportion of data at the same time, as long as the data are not too complicated.

pie



A common way of showing part-to-whole data - but be aware that it's difficult to accurately compare the size of the segments.

doughnut

Similar to a pie chart - but the centre can be a good way of making space to include more information about the data (eg. total)

treemap

Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments

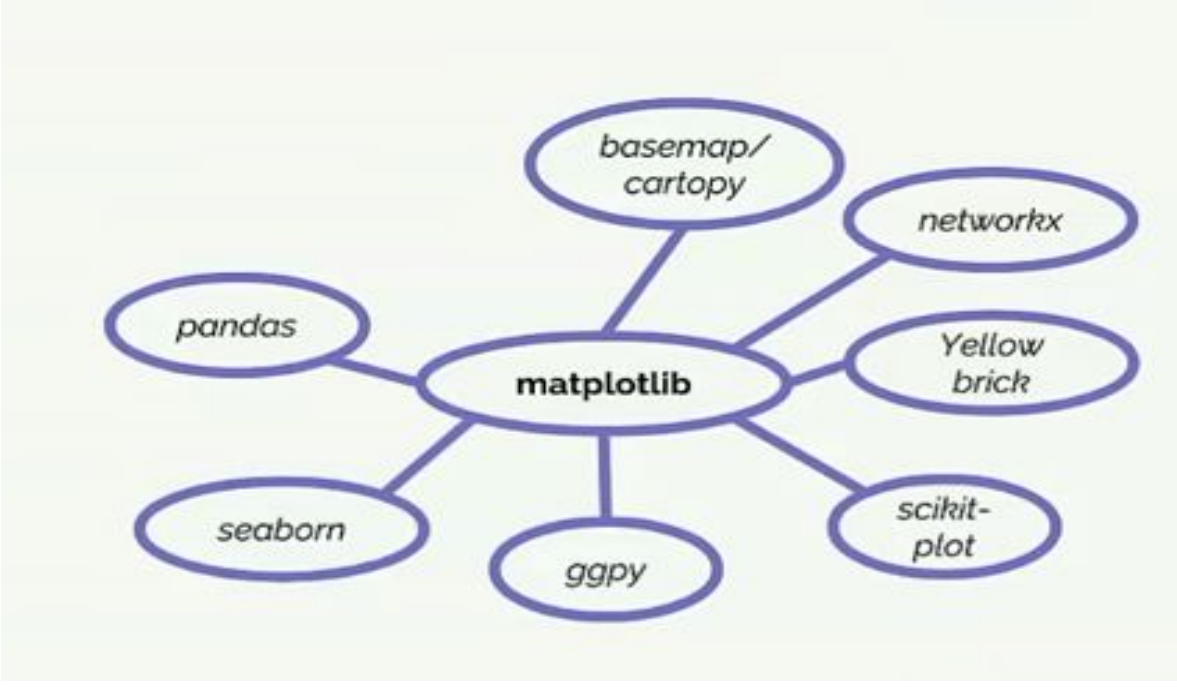
Voronoi

A way of turning points into areas - any point within the area is closer to the central point than any other point

ECOSISTEMA DE LIBRERÍAS DE VISUALIZACIÓN EN PYTHON

Librerías de Visualización en Python

Construyendo sobre las fortalezas de Matplotlib



Mantienen Matplotlib como backend versátil y probado, agregan nueva API específica al dominio

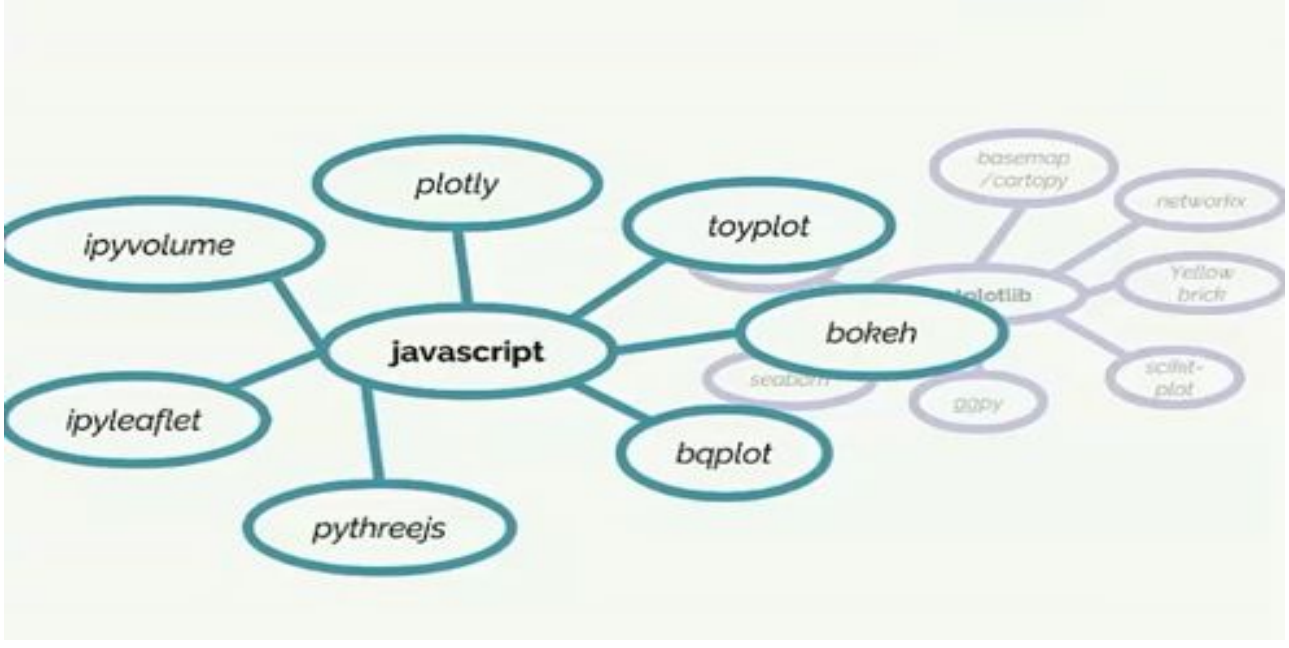
Ventajas de Matplotlib

- Interfaz similar a Matlab, fácil pasaje
- Soporta diferentes graphical backends y sistemas operativos, permite diferentes formatos de output.
- Puede reproducir casi cualquier gráfico
- Standard por más de una década

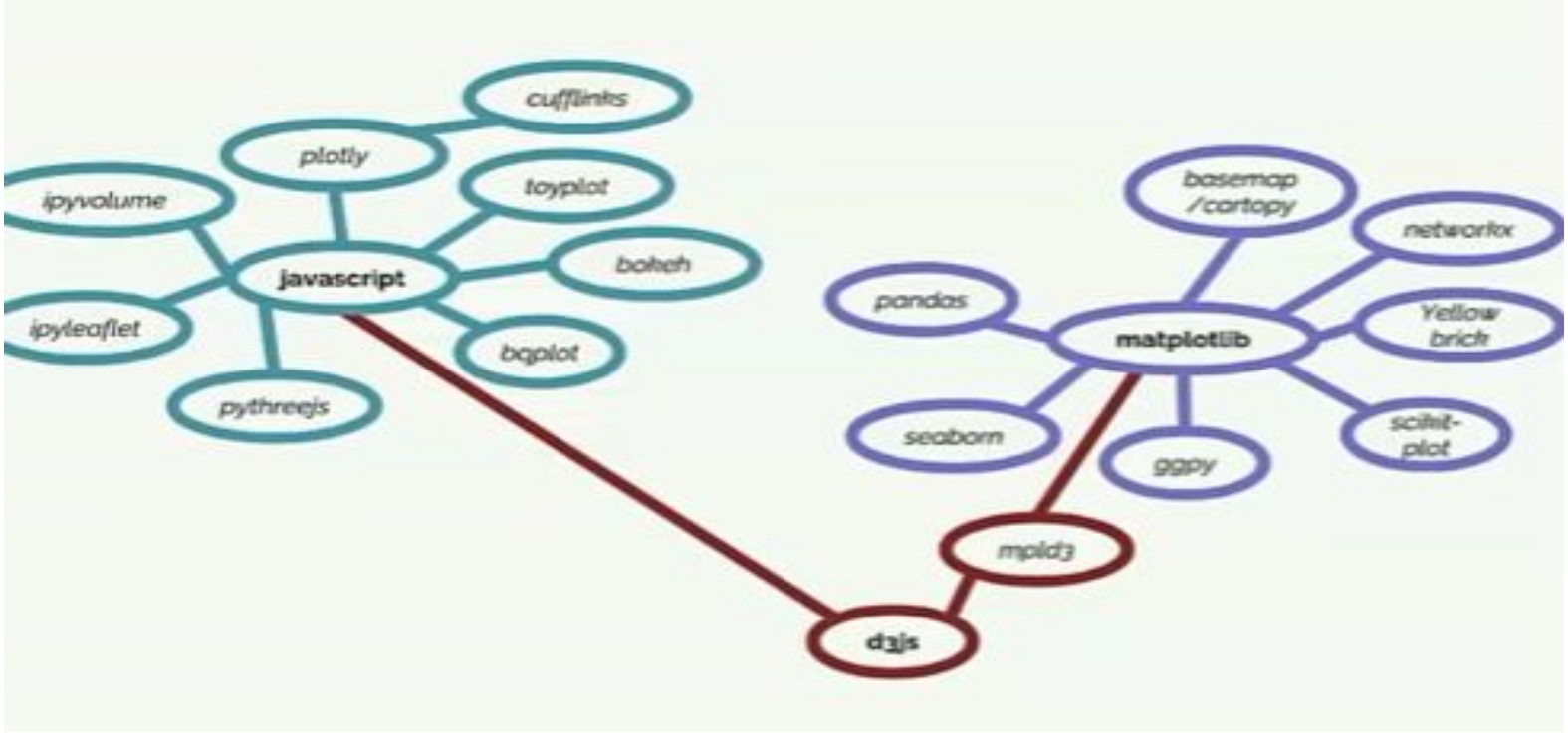
Desventajas de Matplotlib

- Frecuentemente lenta para datasets grandes/complejos
- Pobre soporte para gráficos interactivos/web
- Algunos defaults de estilo poco estéticos

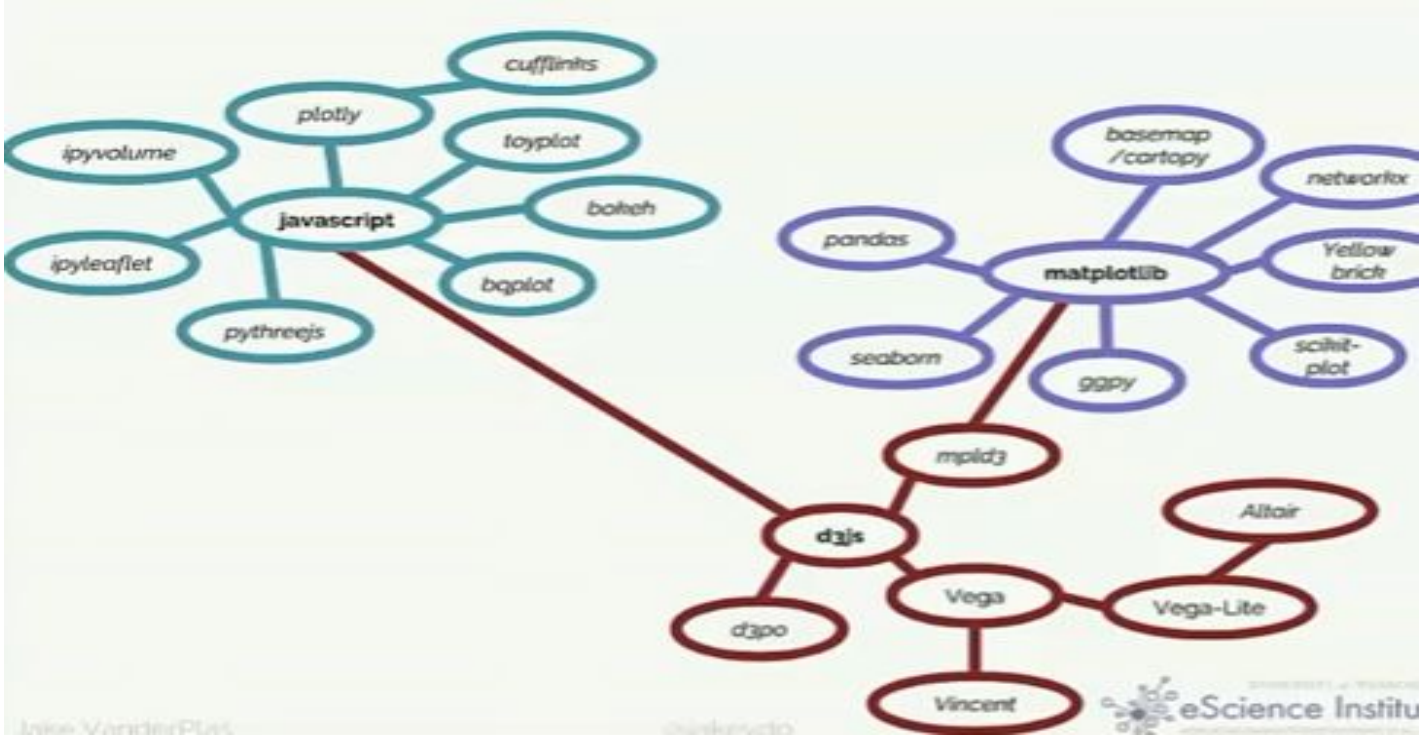
Librerías de Visualización en Python



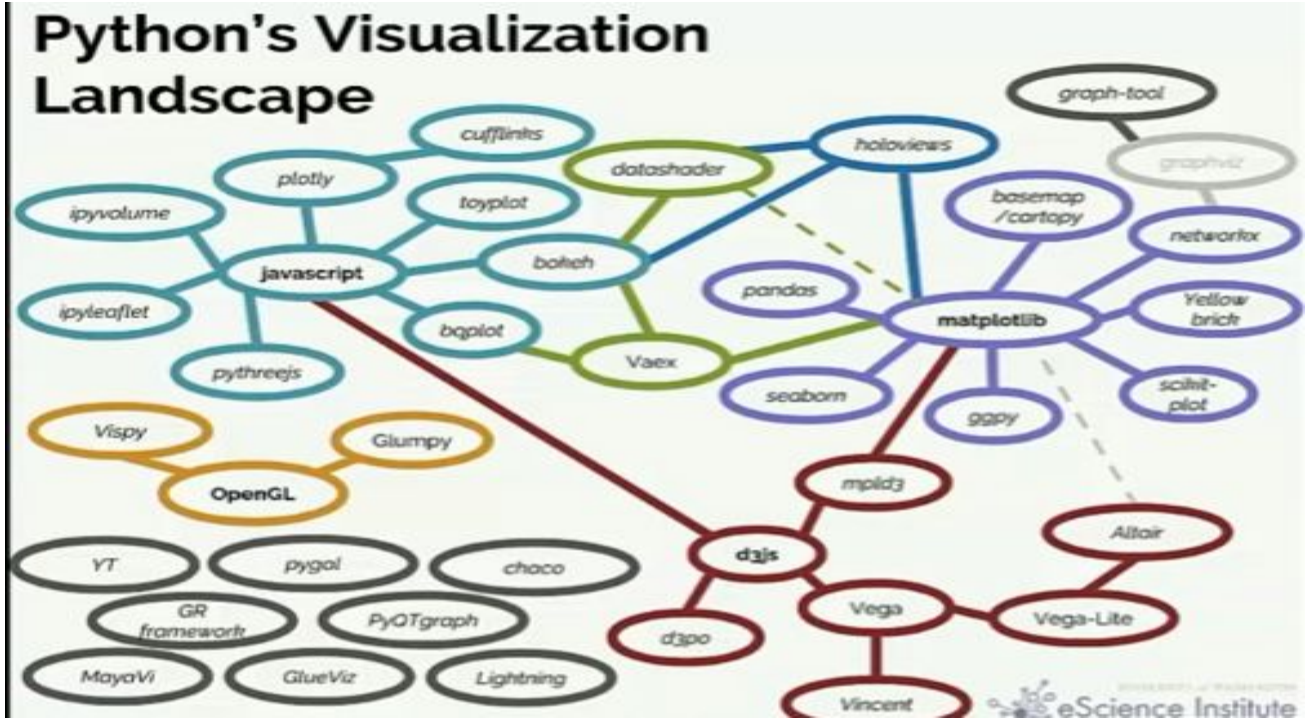
Librerías de Visualización en Python



Librerías de Visualización en Python

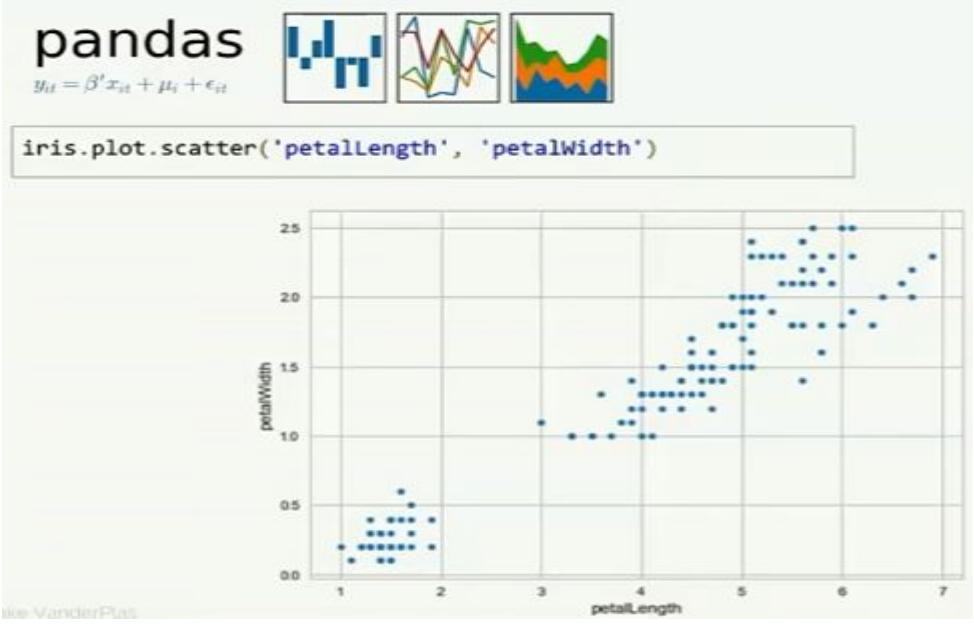


Librerías de Visualización en Python



Librerías de Visualización en Python

Pandas

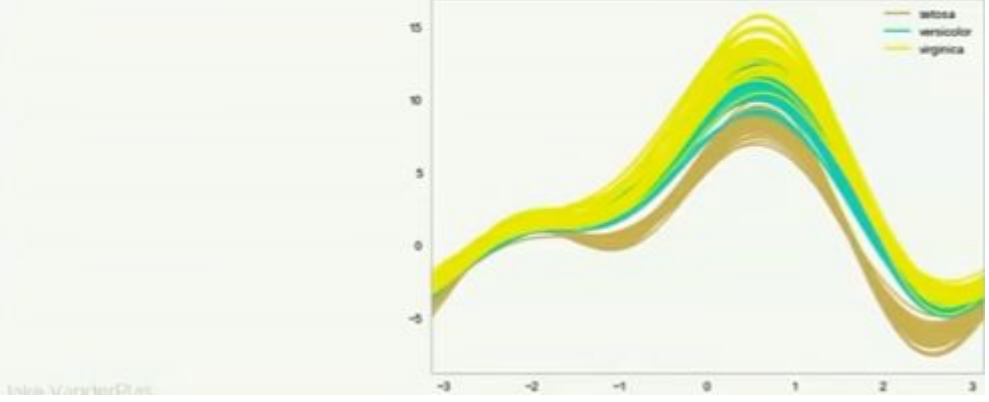


Provee un objeto de la clase DataFrame y una API sencilla para plotear DataFrames

Librerías de Visualización en Python

Pandas

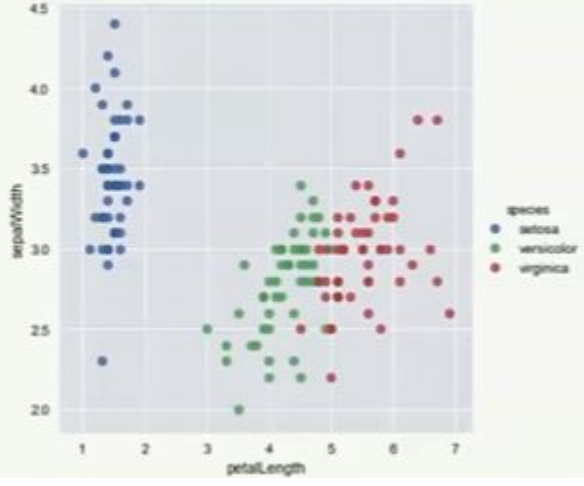
```
from pandas.tools.plotting import andrews_curves
andrews_curves(iris, 'species')
```



Recientemente se han agregado herramientas de visualización estadística más avanzadas

Seaborn

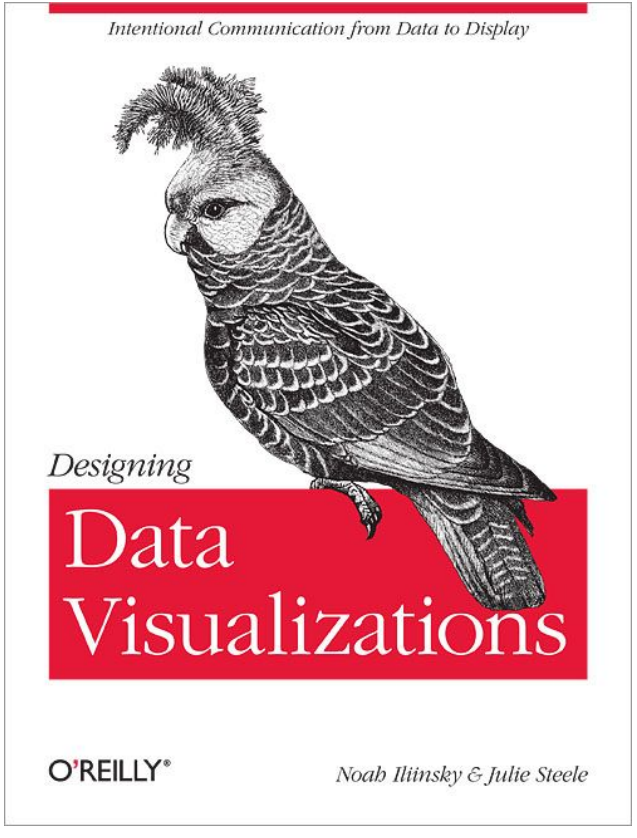
```
import seaborn as sns
sns.lmplot('petalLength', 'sepalWidth', iris,
           hue='species', fit_reg=False)
```

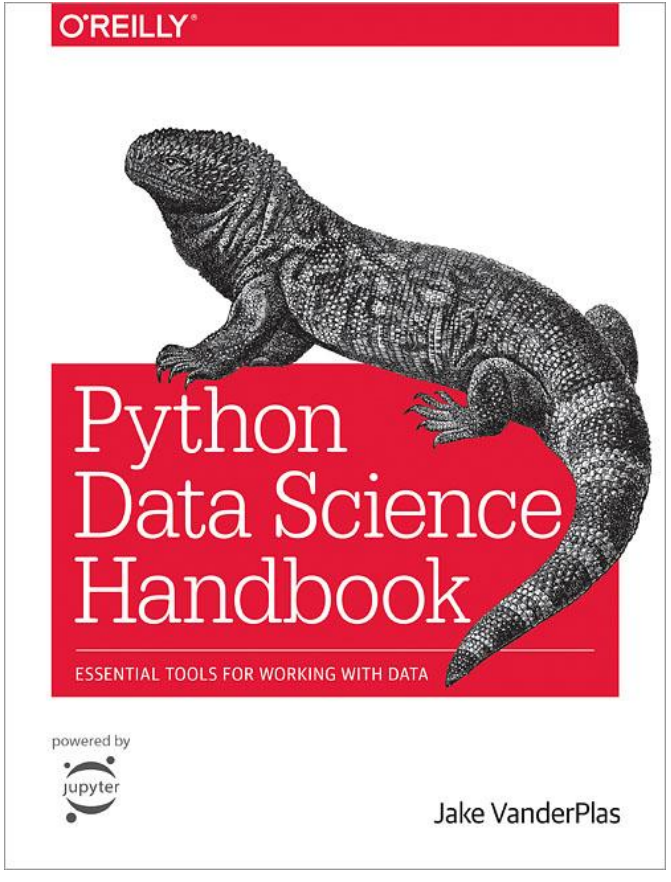


Como Pandas, envuelve (wraps) Matplotlib, atractivo set de paletas de colores y estilos de gráfico, foco en la visualización estadística y modelización

Referencias

Bibliografía adicional





Ver pdf

Visitar