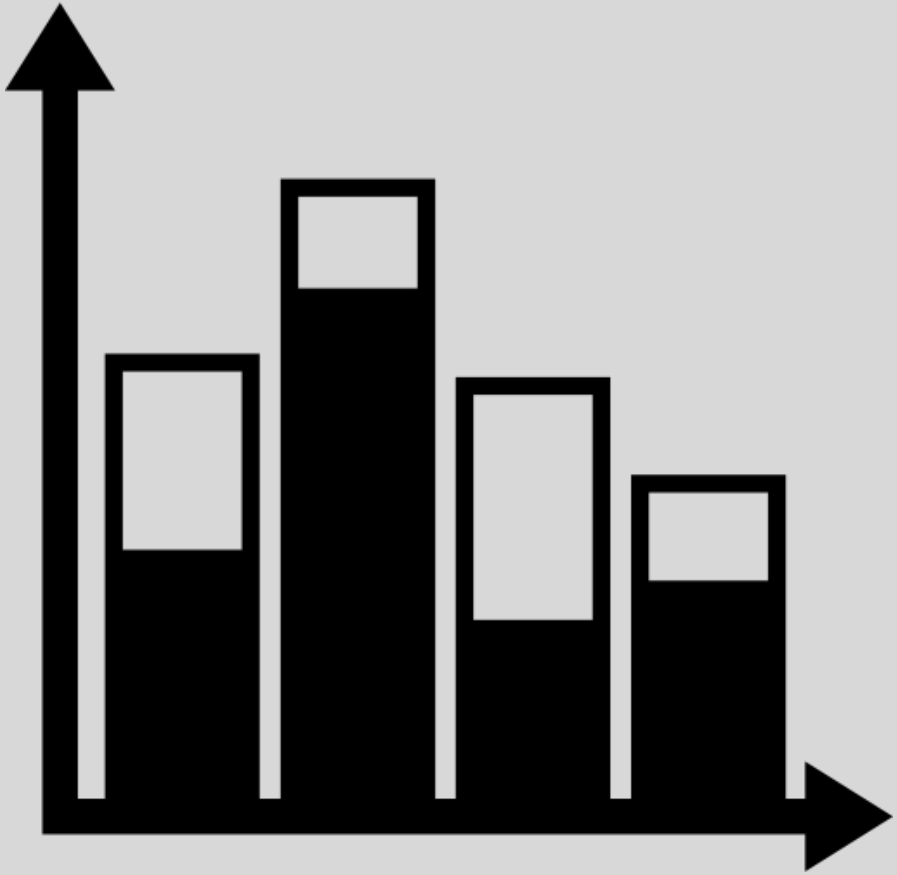


DigitalHouse > Coding School

DATA SCIENCE

MÓDULO 1

Visualización

Agosto de 2017

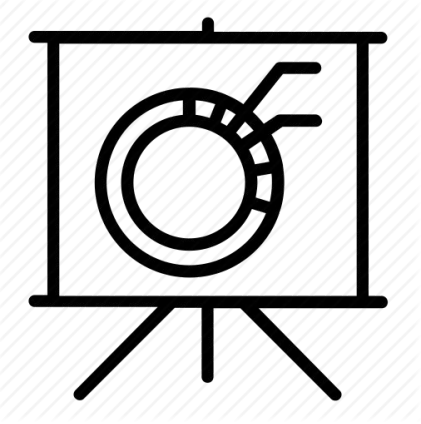
OBJETIVOS DE LA CLASE

1. Recorrer los principios de visualización de datos

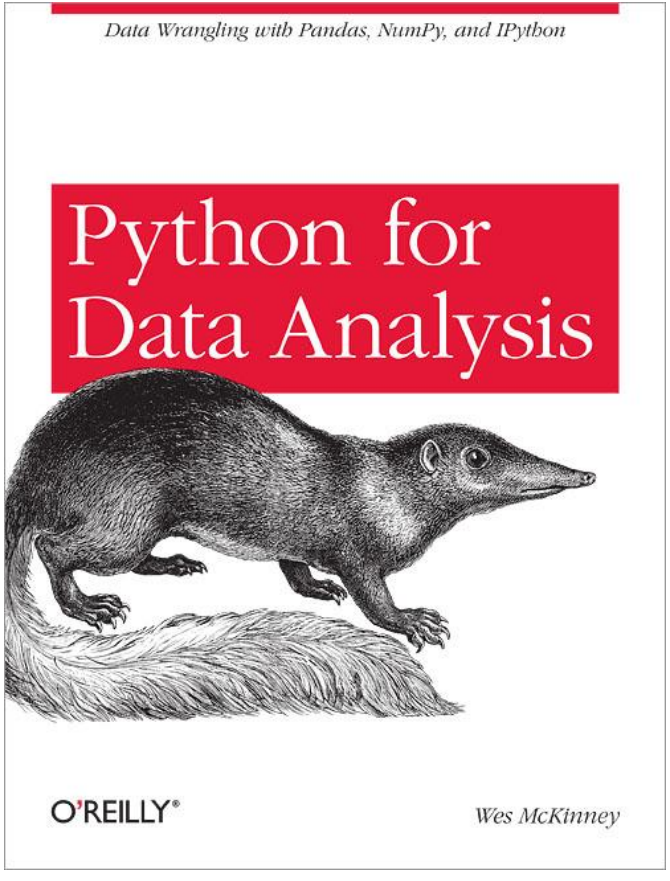
2. Introducir y aplicar herramientas de visualización en Python

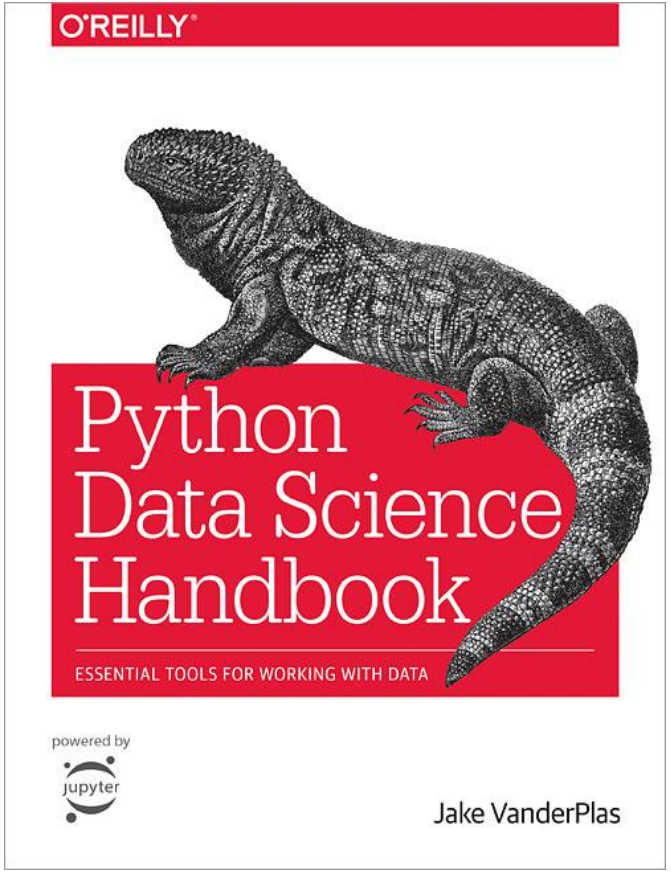
   - PRÁCTICA_GUIADA_Visualizacion.ipynb
   - PRÁCTICA_INDEPENDIENTE_Solution_Visualizacion.ipynb
   - LAB_Visualizacion.ipynb

3. Repasar e integrar conceptos



Bibliografía adicional





INTRODUCCIÓN
¿DE QUÉ TRATA *DATA VIZ*?

La visualización de datos se refiere a toda representación de datos en tanto que:

- Se grafica a través de algoritmos,
- Es fácil de reproducir con dataset diferentes (aunque de similar dimensión o características),
- Puede carecer de refinamiento estético (datos crudos, sin decoración),
- Es dato-intensiva: grandes volúmenes de datos son bienvenidos y viables (contraste con infografías).

Tips sobre visualización de datos en Designing data visualization

INTRODUCCIÓN
¿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
- Expresar relaciones complejas de manera fácil
- Condensar información y comunicar de manera mucho más potente
- Comunicar de manera universal

Según sea el escenario hay técnicas más adecuadas que otras para visualizar los datos.

¿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:

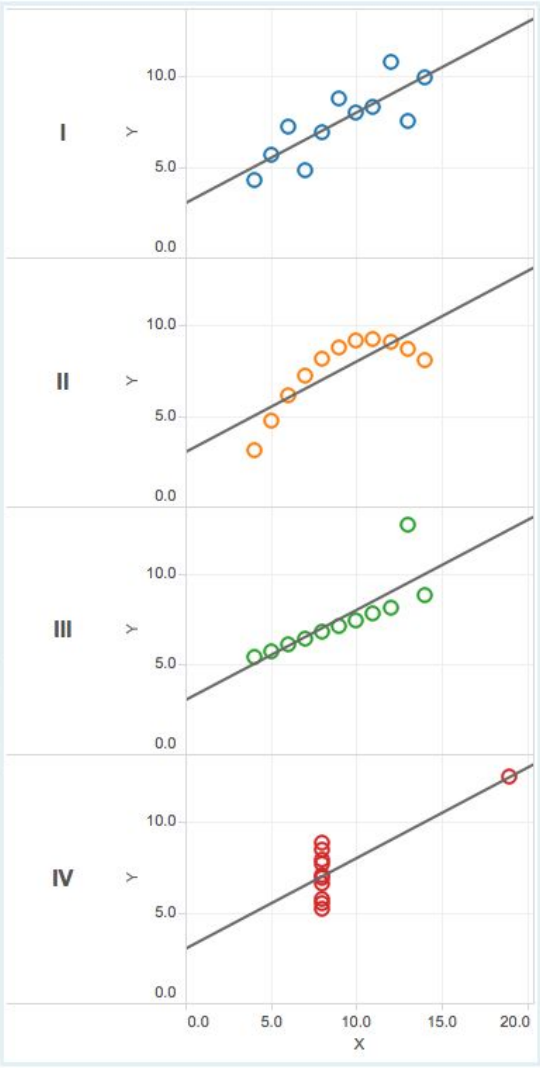
Este ejemplo nos recuerda que la información sintética tiene que ser complementada con mayor conocimiento del dominio.

Visualizar los datos puede evitar hacer supuestos incorrectos.

Cuarteto de Anscombe

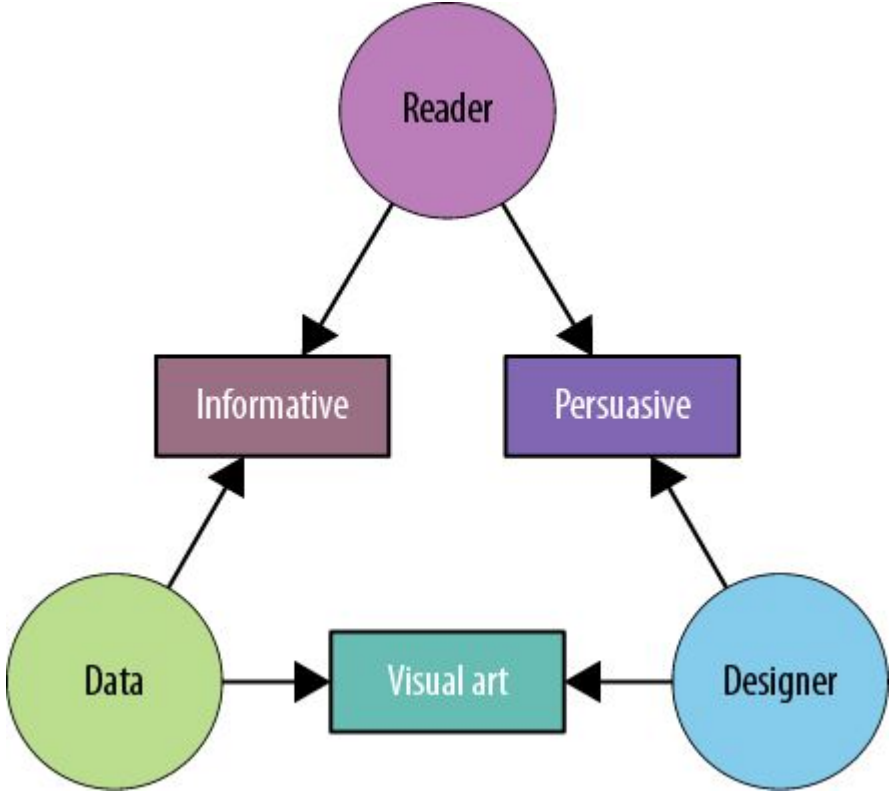
I	II	III	IV
(4, 4.3)	(4, 3.1)	(4, 5.4)	(8, 5.3)
(7, 4.8)	(5, 4.7)	(5, 5.7)	(8, 5.6)
(5, 5.7)	(6, 6.1)	(6, 6.1)	(8, 5.8)
(8, 7.0)	(7, 7.3)	(7, 6.4)	(8, 6.6)
(6, 7.2)	(14, 8.1)	(8, 6.8)	(8, 6.9)
(13, 7.6)	(8, 8.1)	(9, 7.1)	(8, 7.0)
(10, 8.0)	(13, 8.7)	(10, 7.5)	(8, 7.7)
(11, 8.3)	(9, 8.8)	(11, 7.8)	(8,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



PRESENTACIÓN DE RESULTADOS

Diseño basado en las relaciones predominantes de la triada RDD



DATA VIZ Y PERCEPCIÓN VISUAL

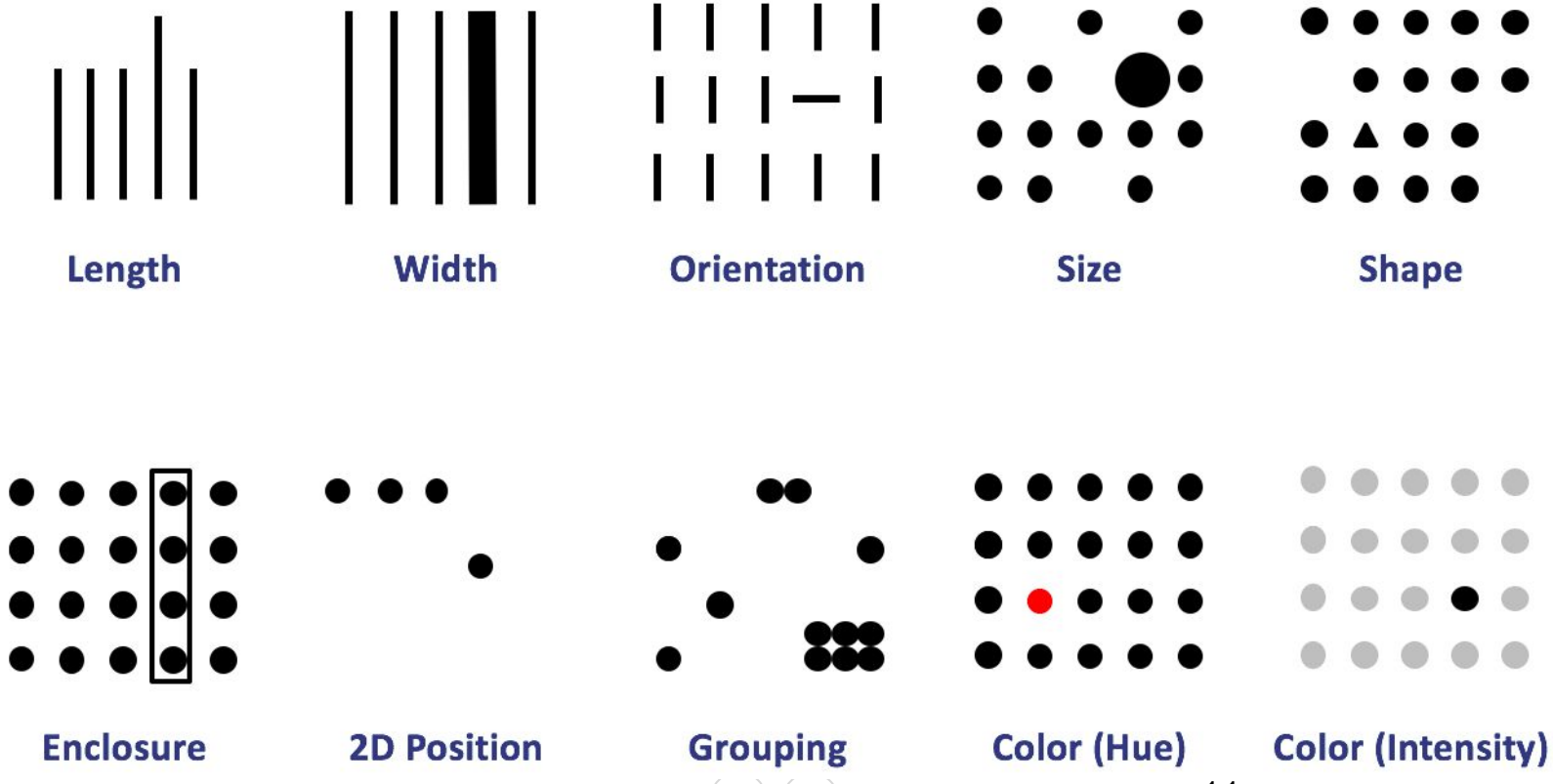
Algunos atributos generan un impacto mayor en nuestro cerebro.

PERCEPCIÓN VISUAL

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

ALGUNOS RECURSOS...

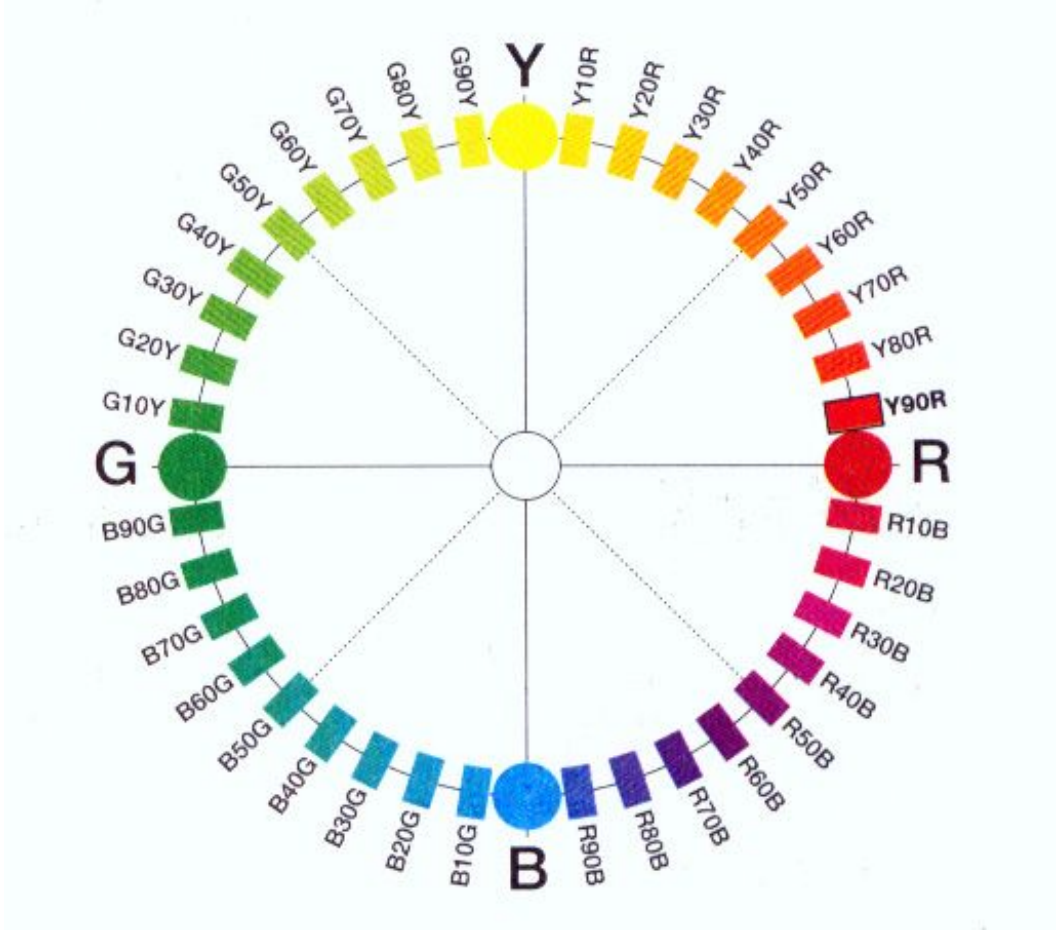
Contamos con diferentes recursos visuales para transmitir información:



COLOR

Propiedades del color:

- Tono o matiz
- Intensidad (saturación)
- Brillo o luminosidad
- Contrastes



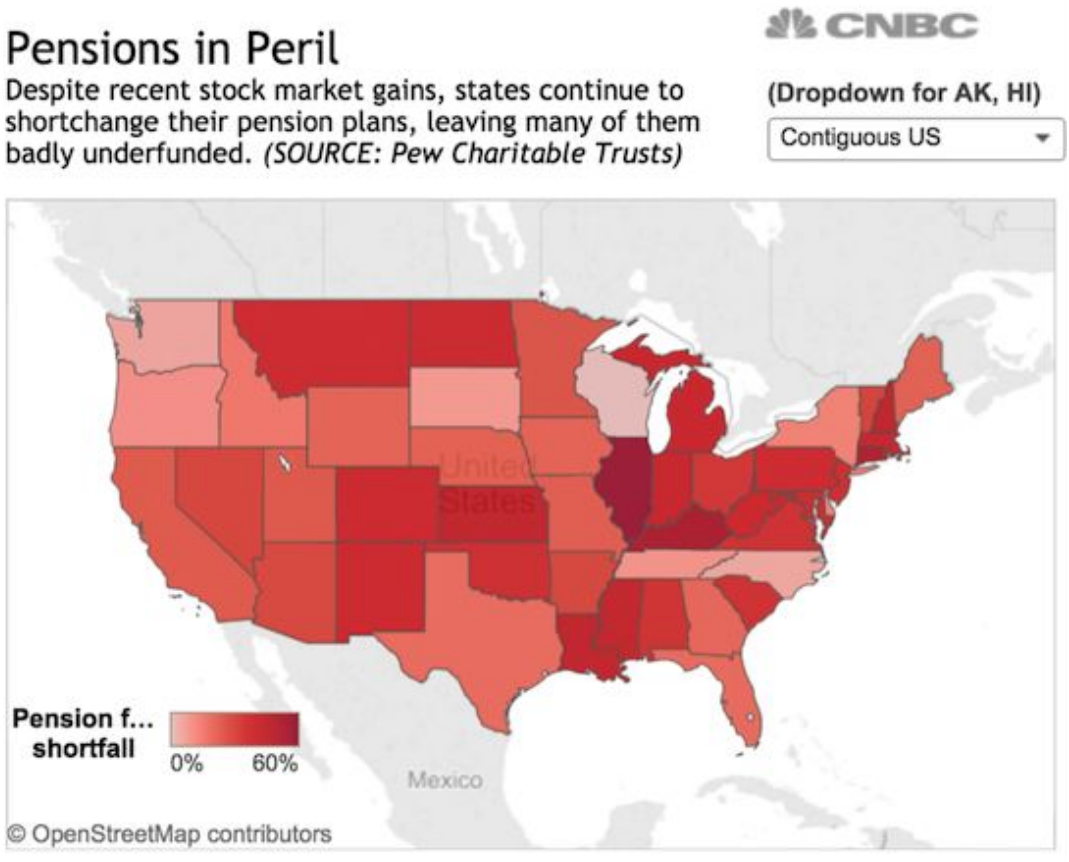
COLOR

Los usos del color en visualización de datos permiten indicar:

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

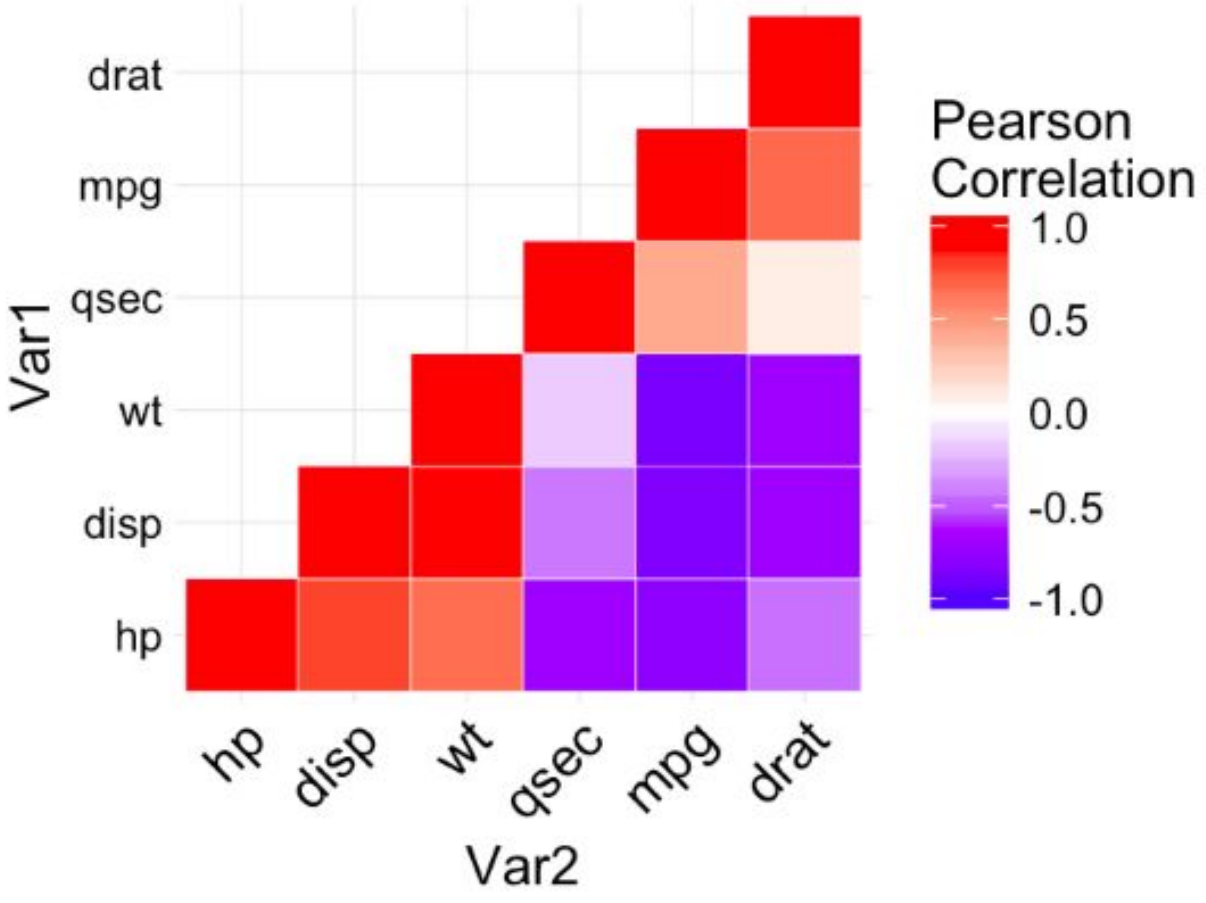
COLOR: SECUENCIA

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



COLOR: DIVERGENCIA

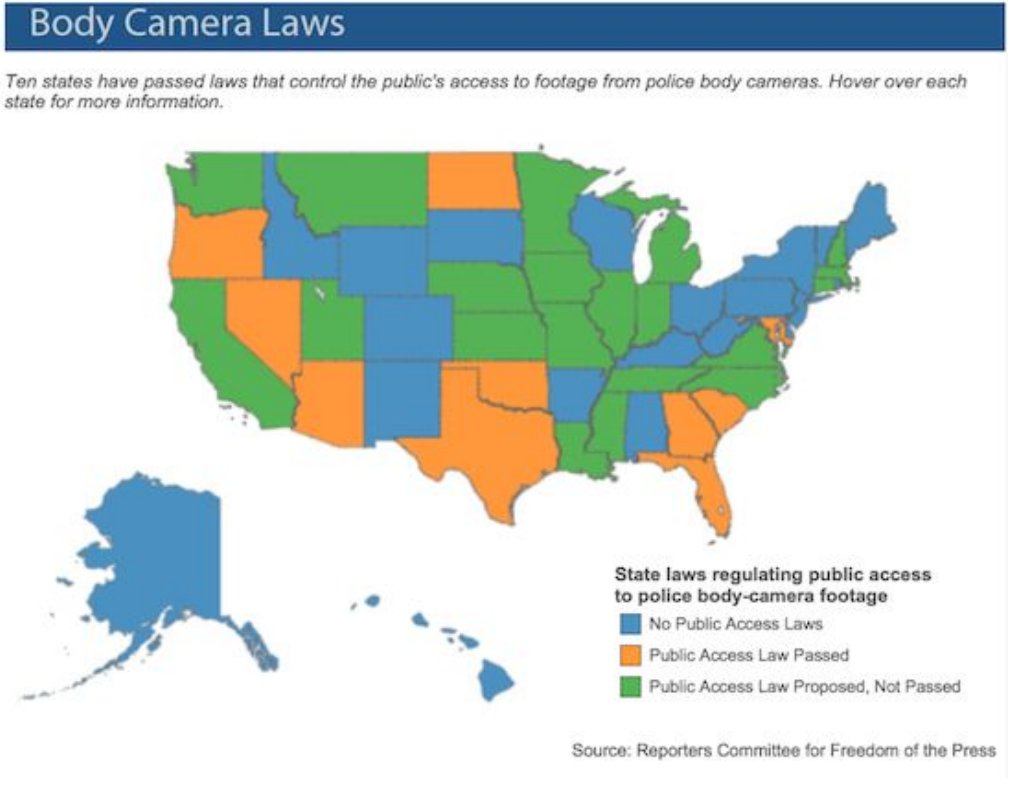
Los colores ***divergentes*** se utilizan para mostrar valores ordenados que tienen un valor crítico, tales como un promedio o cero:



COLOR: CATEGORÍA

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

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



COLOR: CATEGORÍA

Es importante identificar contrastes posibles para diferenciar distintos grupos o categorías:

- Complementarios
  - Par
  - Split
- Triada
- Pares complementarios
  - Rectangular
  - Cuadrangular

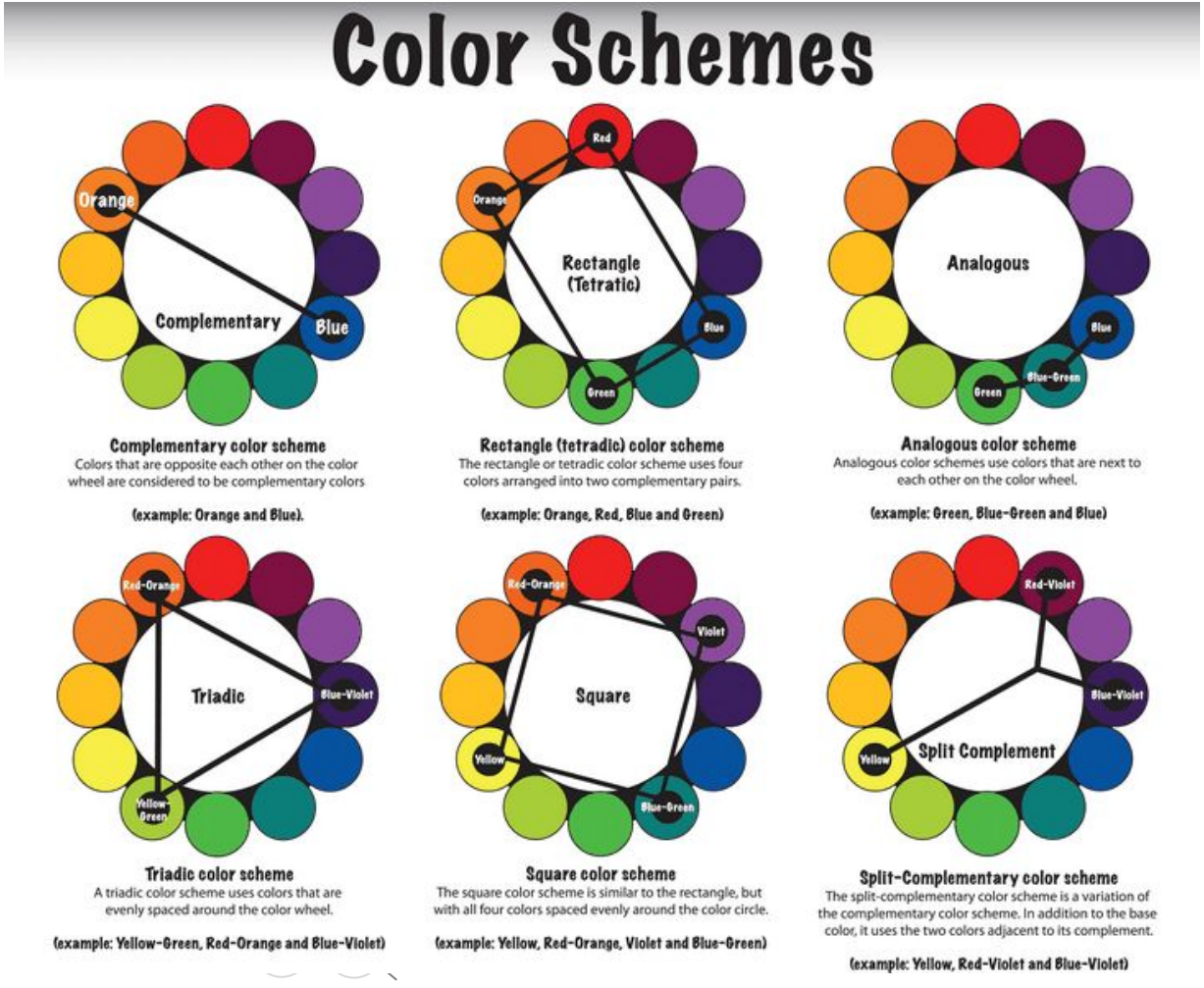


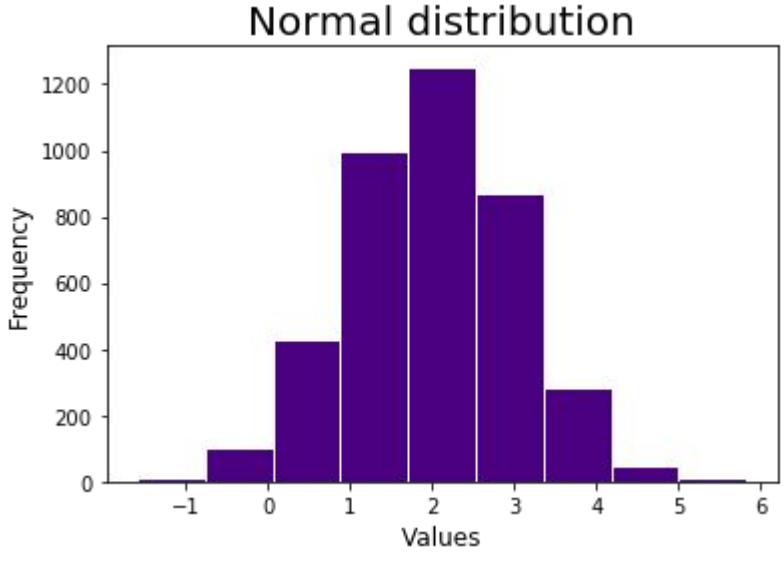
DIAGRAMA Y GRÁFICOS

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.



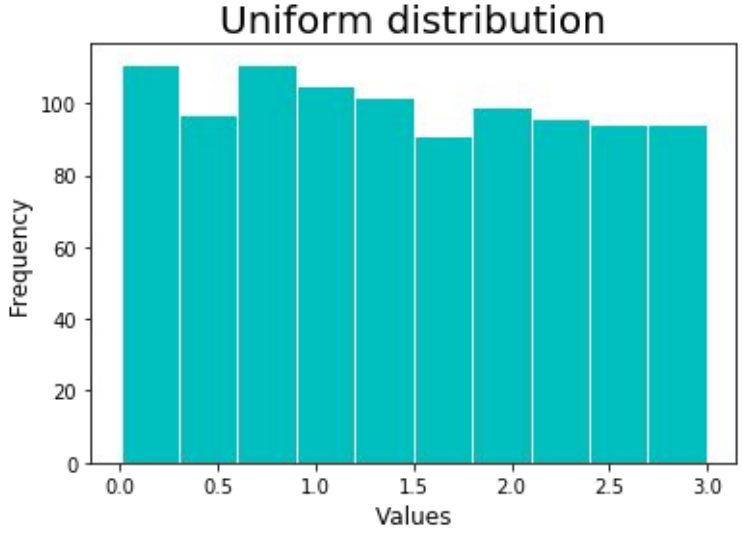


DIAGRAMA DE CAJA (BOX PLOT)

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:

- 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).

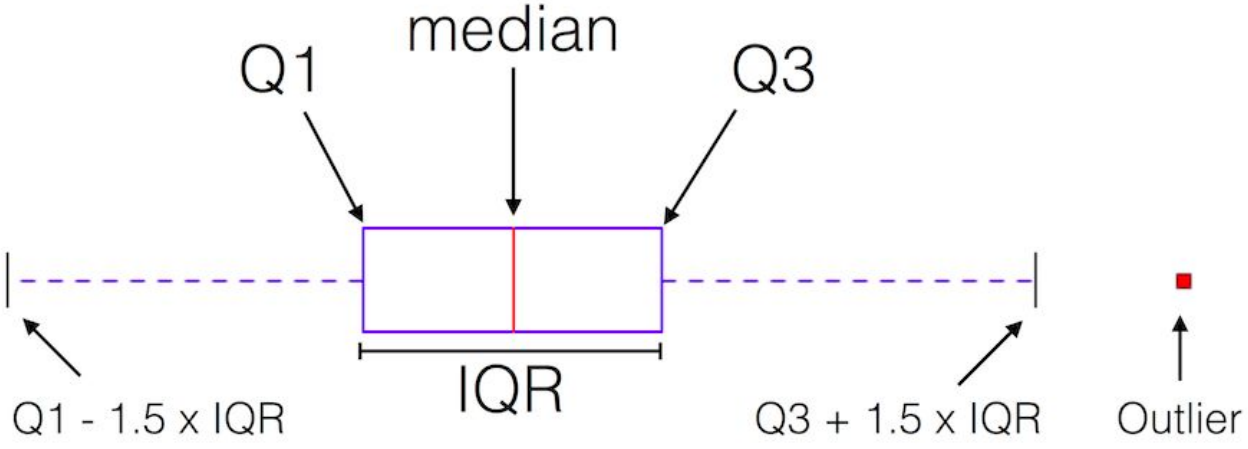


DIAGRAMA DE CAJA (BOX PLOT)

- Los box plots también pueden representar el rango completo de valores que toma la variable, segmentando su distribución en cuartiles .

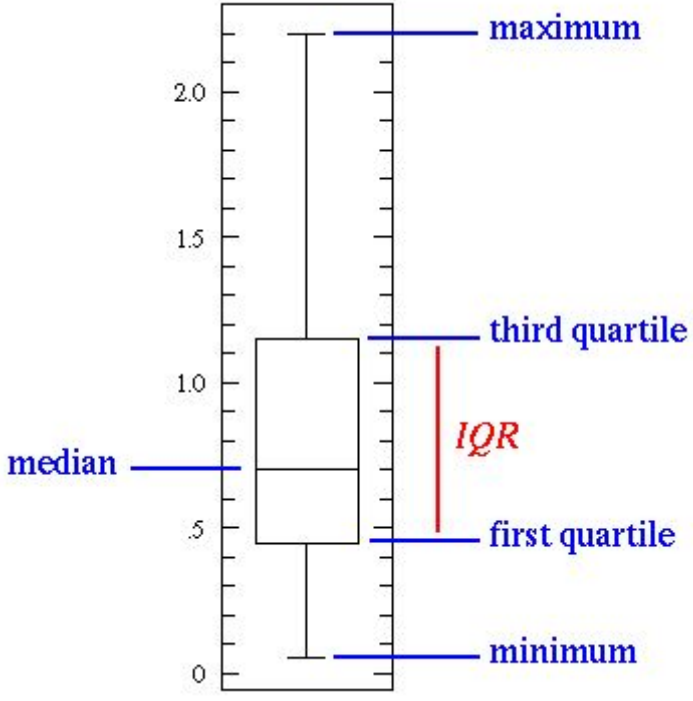


GRÁFICO DE DISPERSIÓN (SCATTER PLOT)

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

Esta información puede orientar hacia dónde profundizar la investigación.

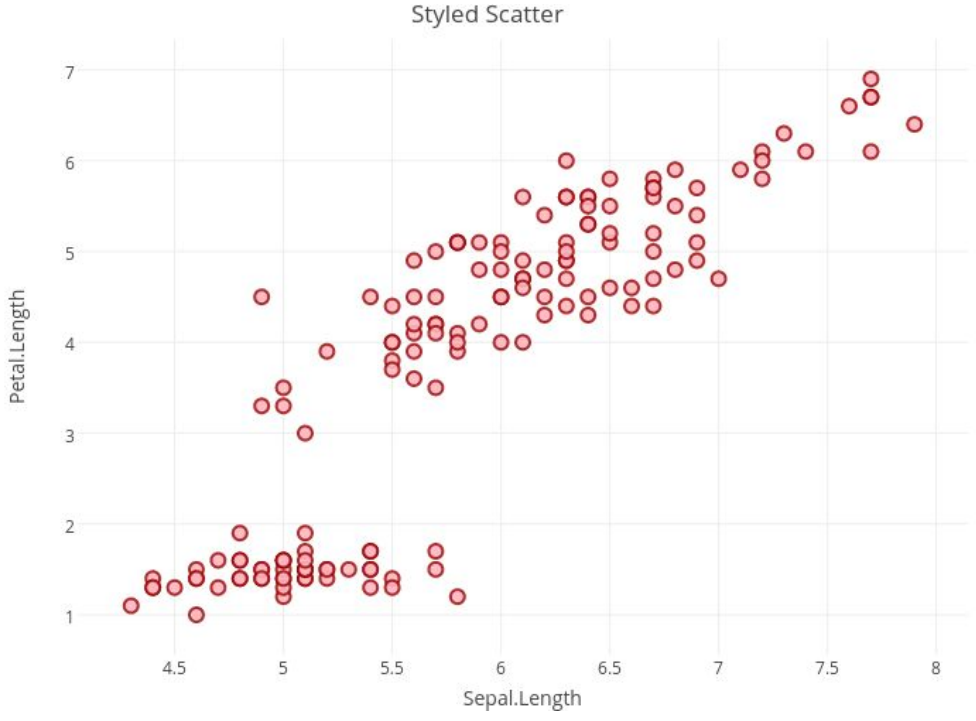


GRÁFICO DE LÍNEAS (PLOT)

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.

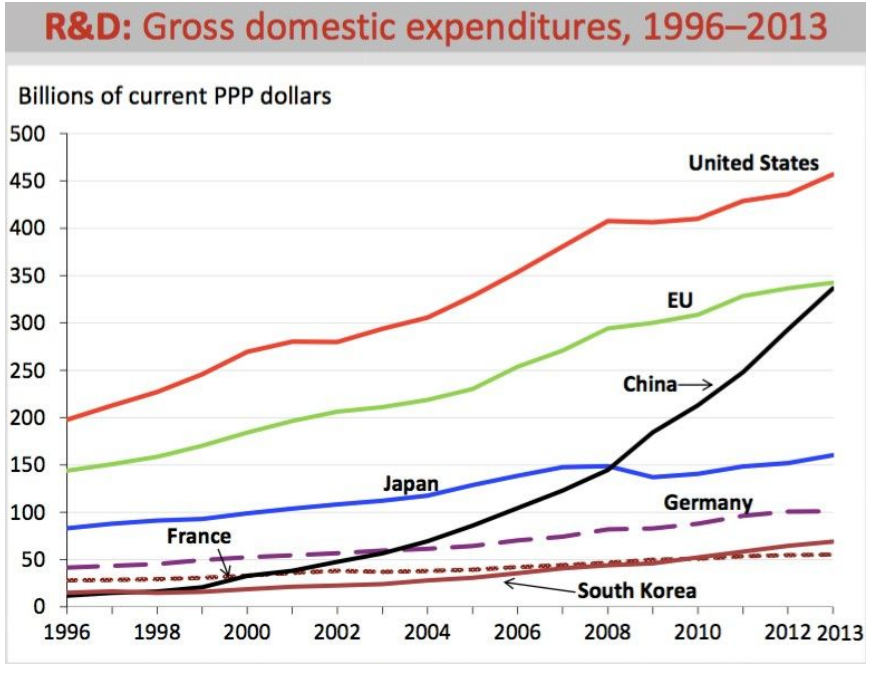
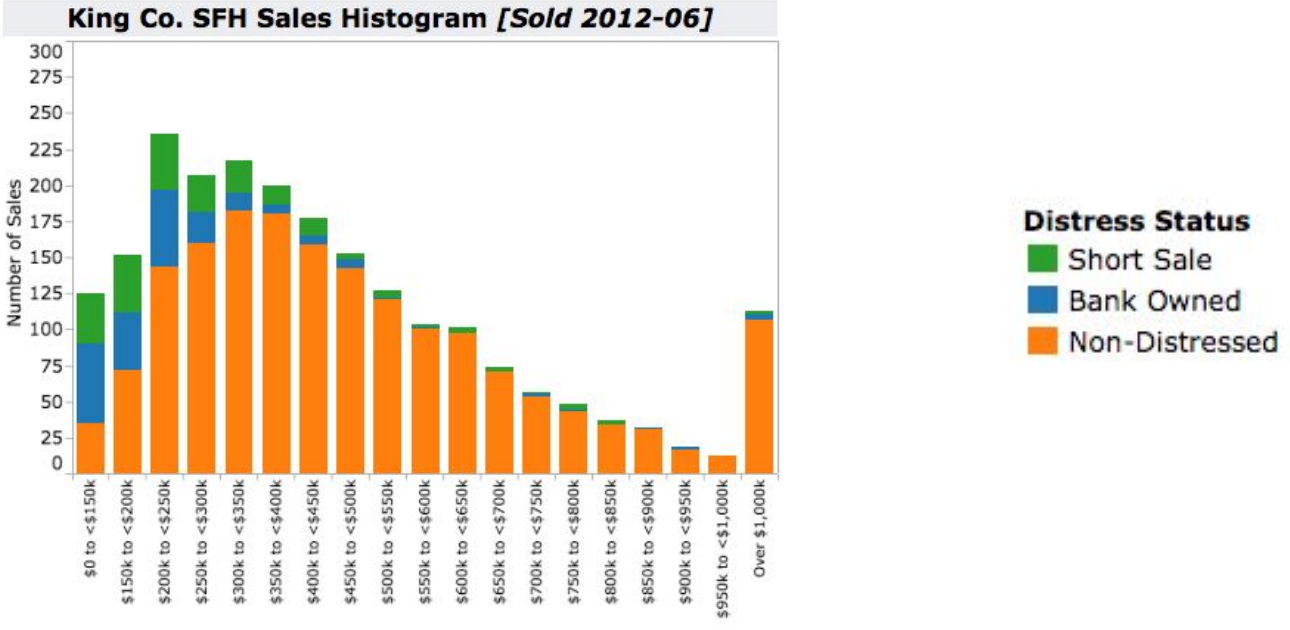


GRÁFICO DE BARRAS

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.

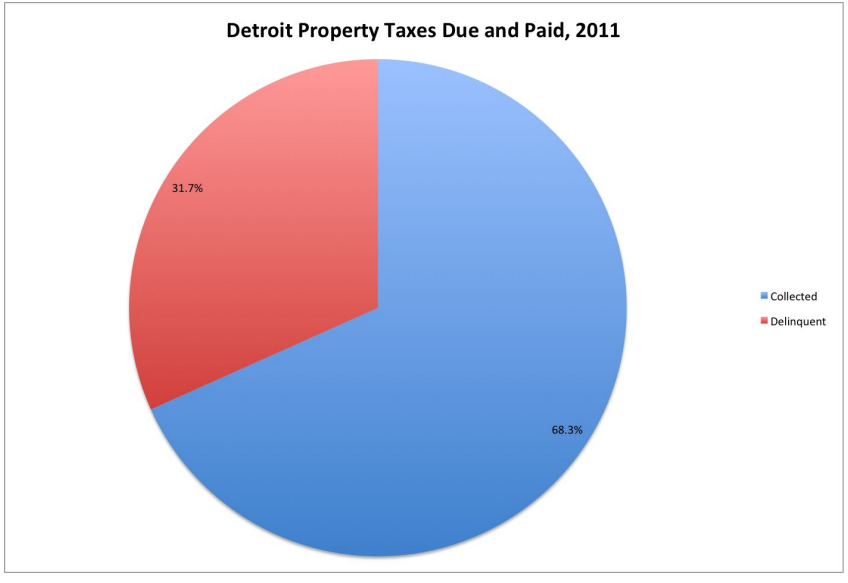


GRÁFICOS DE BARRAS

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 barra:
The Worst Chart In The World

Visualización