Visualización de datos

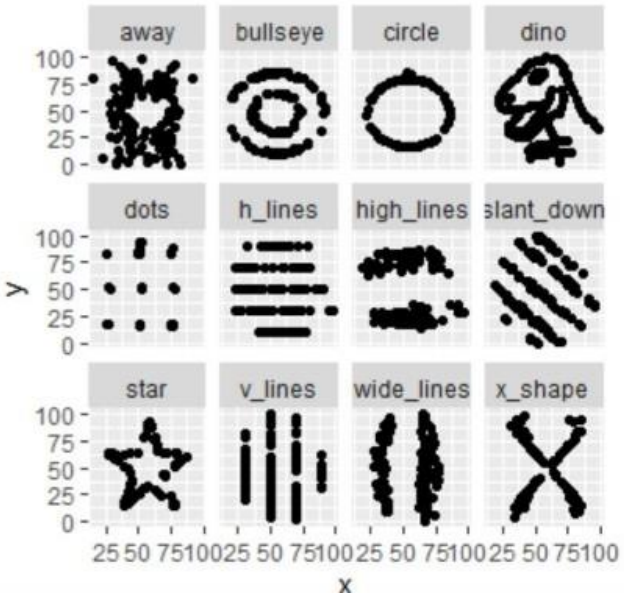
¿Por qué es necesario graficar?

- Las técnicas de visualización de datos son muy importantes tanto para nuestro trabajo como para comunicarlo
- La cantidad de tipos de gráficos disponibles es enorme y es importante entenderlos ysaber para qué es útil cada uno
- Entender de forma eficiente los datos
- **Comunicar de forma concisa y clara**
- Encontrar patrones/relaciones
- El análisis descriptivo es uno de las partes principales de cualquier análisis
- relacionado con un proyecto de ciencia de datos o de una investigación específica
- La agregación de datos, el resumen y la visualización son algunos de los pilares principales que respaldan este área
- La visualización de datos es una herramienta poderosa y ampliamente adoptada debido a su efectividad para extraer la información correcta, comprender e interpretar los resultados de manera clara y fácil
- Tratar con conjuntos de datos multidimensionales con más de una variable o atributo comienza a causar problemas, ya que estamos restringidos a comunicar en dos dimensiones (a lo sumo 3).
- Los gráficos no son simplemente: “imágenes bonitas”
- No toda la información importante se puede adivinar a través del análisis estadístico...

- Todo estos gráficos tienen la misma media y desvío estándar

$$\hat{\mu}_x = 54.3 \quad \hat{\mu}_y = 47.8$$

$$\hat{s}_x = 16.8 \quad \hat{s}_y = 26.9 \quad \hat{\rho}_{xy} = -0.1$$



Visualización para ML

En aprendizaje automático, la visualización se utiliza para:

- Análisis inicial de los datos:
  - para examinar si los datos satisfacen los supuestos requeridos para el método
  - tienen complicaciones inesperadas como valores atipicos o no linealidad
- Evaluar el ajuste del modelo:
  - predicho vs observado
  - análisis de residuos → lo que se aparta de la prediccion de un numero

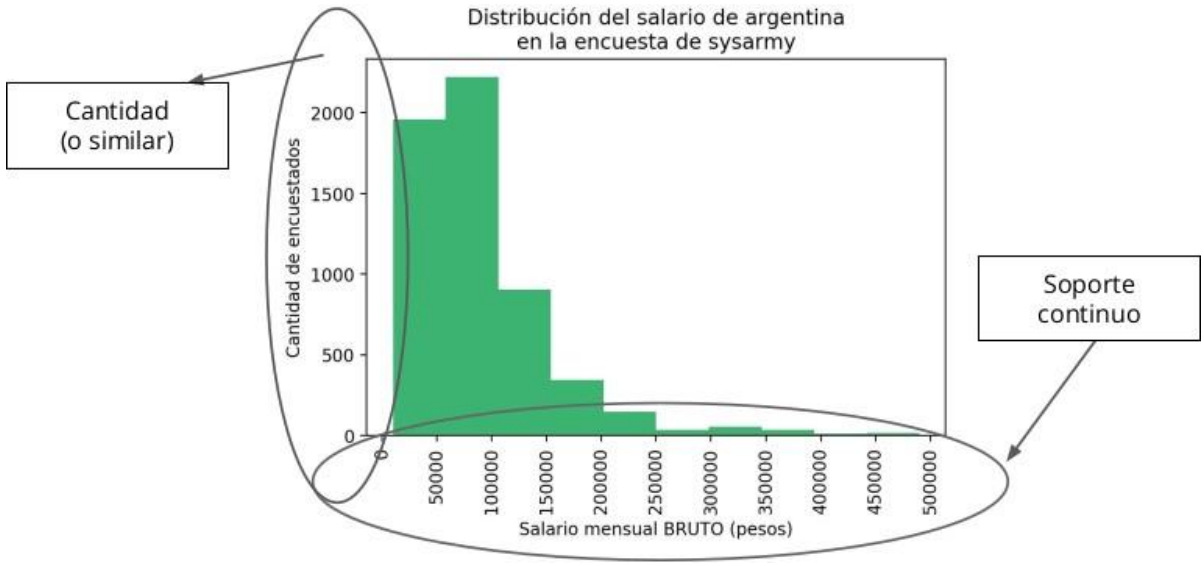
Tipos de graficos - Plots

- De distribución continua
- De distribución discreta
- De relación
- Series de tiempo
- Otros

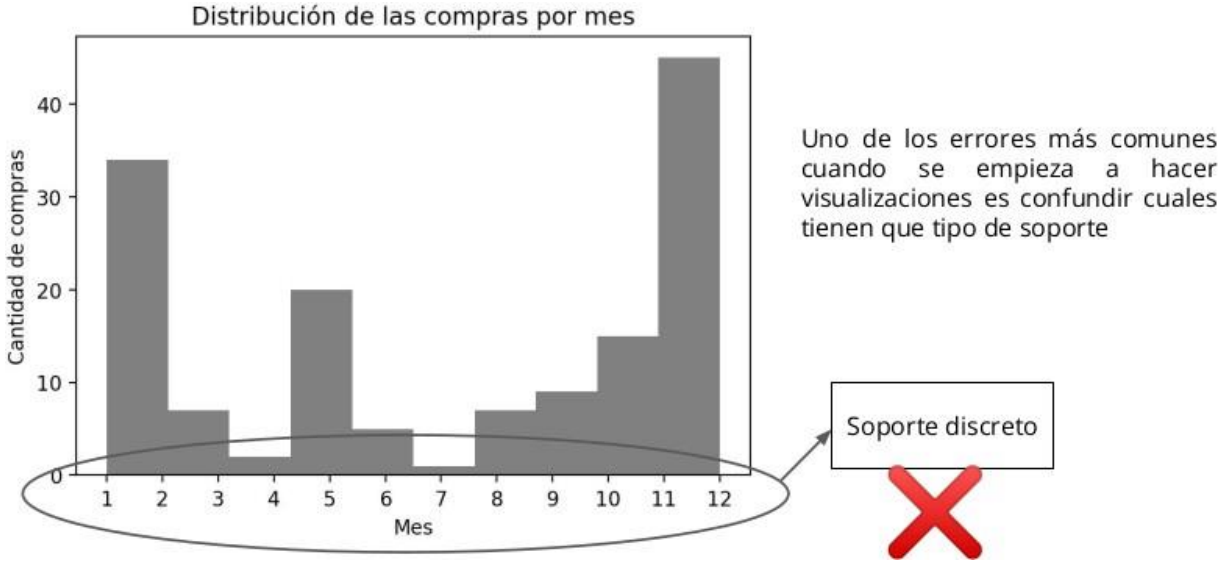
Distribucion continua

Queremos ver una variable continua en funcion de otras cosas

Histograma



Podemos ver que la mayor cantidad de gente estan cobrando entre 0 y 150.000



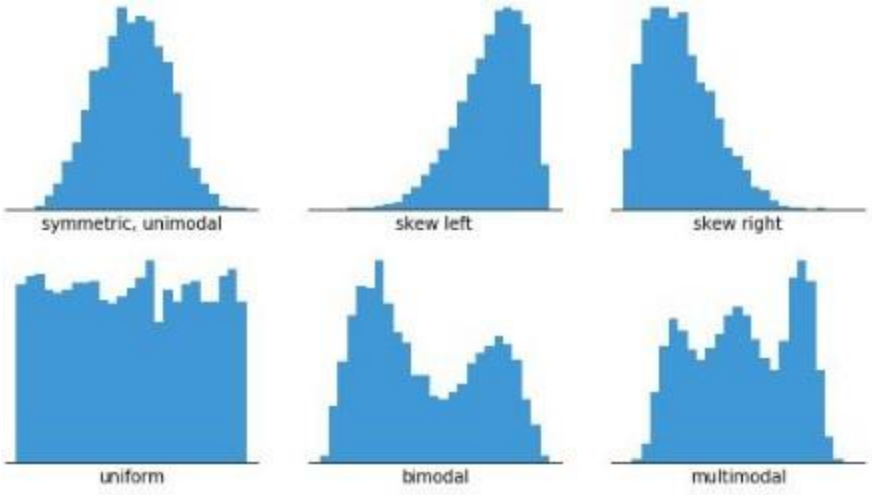
Mes es una variable discreta

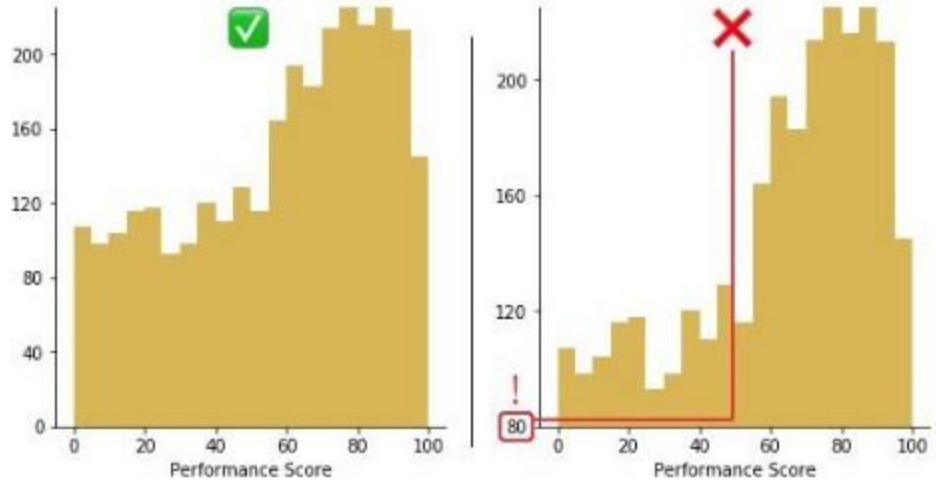
Este soporte no es continuo

Variar los bins → cantidad de barras que se van a mostrar

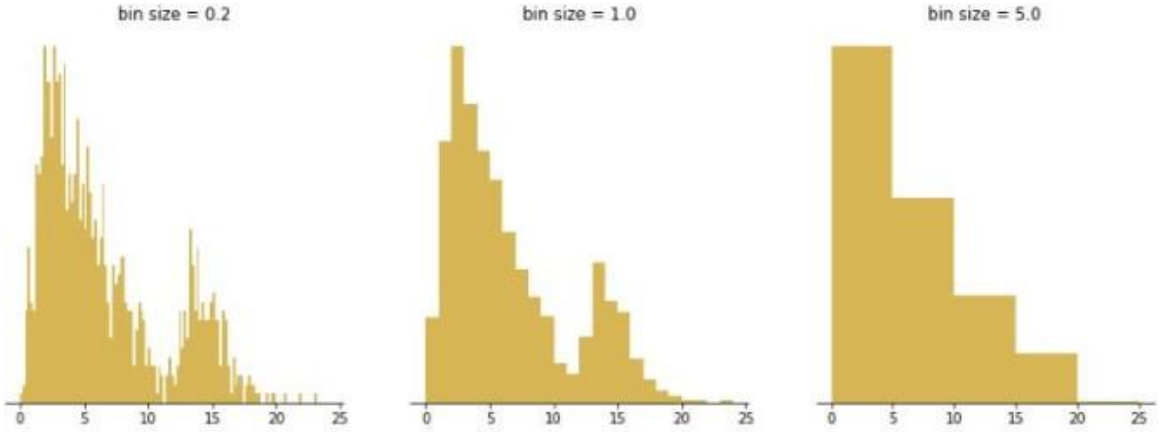
Hay un punto en que el detalle nos juega en contra

Esto se puede hacer con soporte continuo donde se va haciendo mas personalizado, pero esto no se podria hacer en el discreto.



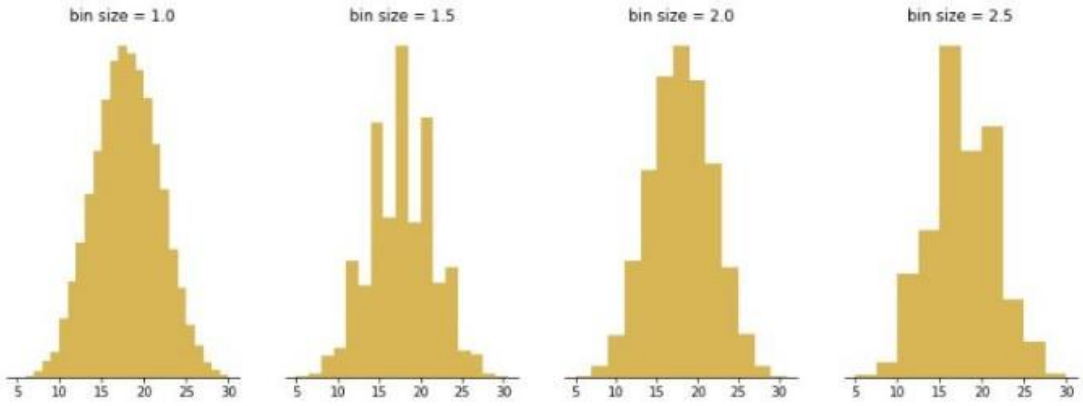


Use a zero-valued baseline

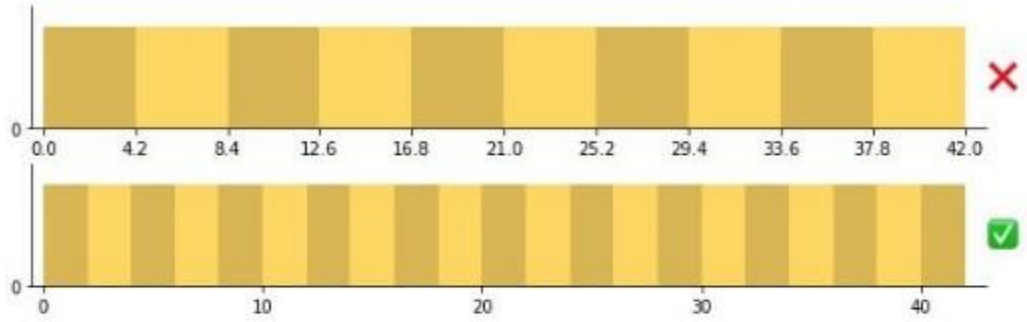


Choose an appropriate number of bins

En los dos primeros vemos que es una bimodal, pero en el tercero se pierde esa caracteristica.



El histograma puede verse anormalmente "desigual" simplemente debido a la cantidad de valores que posiblemente podría tomar cada contenedor.

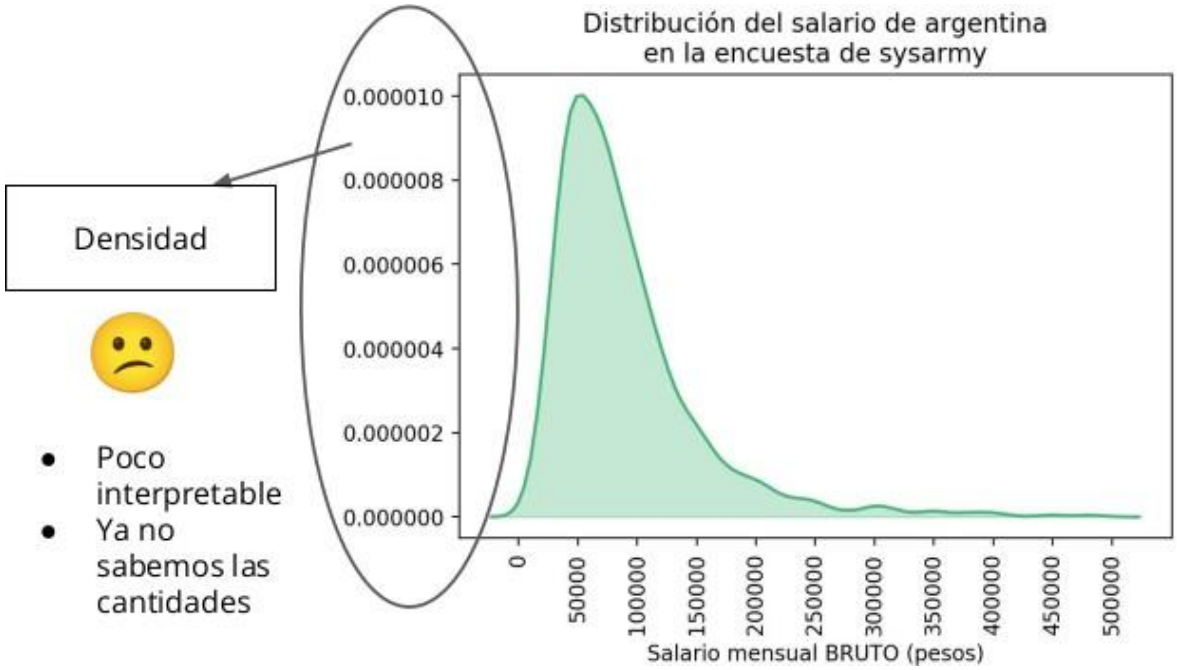


Choose interpretable bin boundaries

Que los valores sean faciles de entender

Density Plot

Nos deja ver una version continua y suavizada con mucho detalle de todo el espectro.



Lo que no se entiende ⇒ lo sacamos del grafico

La altura indica que tan comun es algo

- Suaviza los bordes para lograr la densidad, no sabe que no tiene sentido < 0

Alguien cobra 0? No tiene sentido eso

Aca el grafico correcto

Números útiles

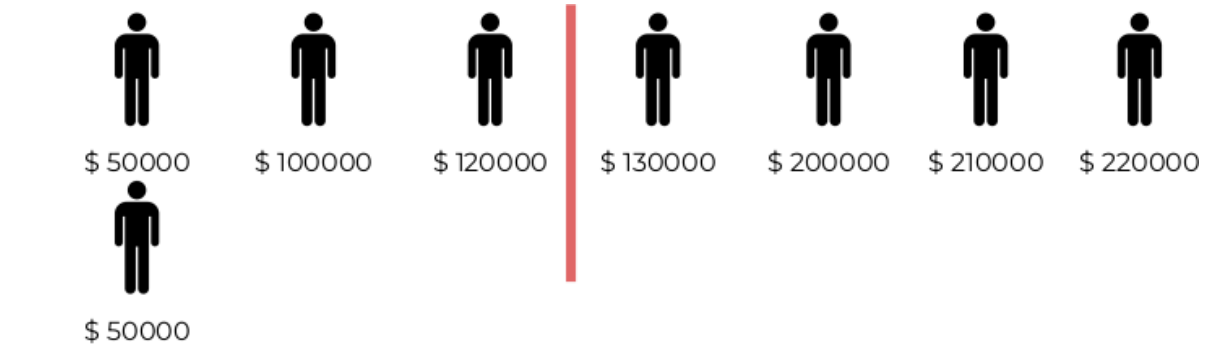
- Media: Es el promedio
- Mediana: es el valor que está en la mitad de la población

Puede ser que el valor exista y sea una observación

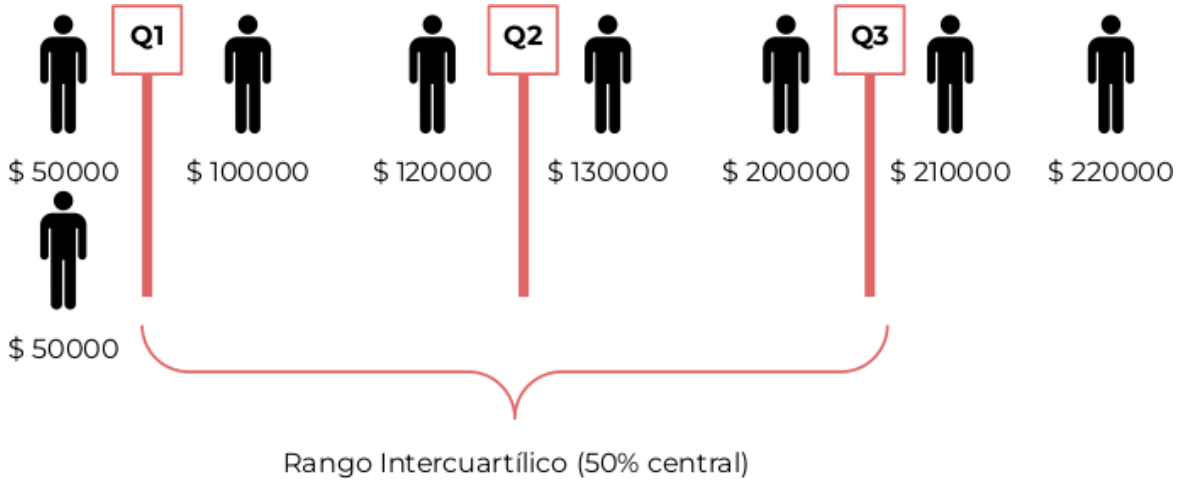
O que no exista y será el promedio de las dos observaciones centrales

- Cuartil: son los valores límite que dejan al 25% de la población entre ellos Lo partimos en 4
- Rango intercuartílico: el rango entre el cuartil 1 y el cuartil 3

Ejemplo: Población de salarios



Media: 135000
Mediana: 125000



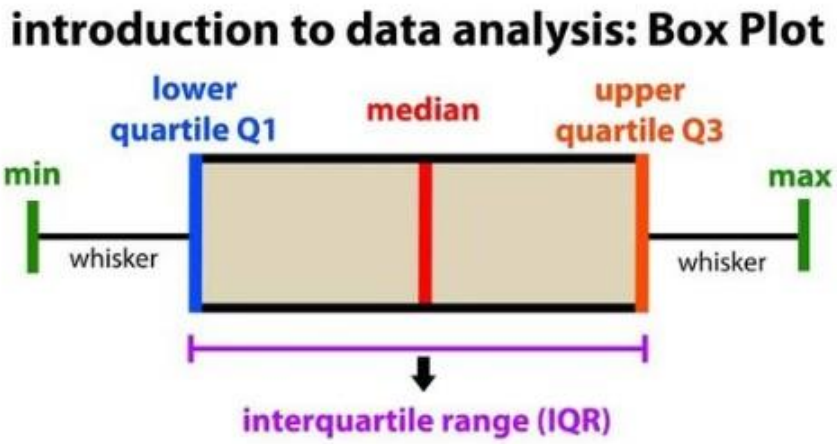
Media: 135000
Mediana: 125000
Cuartil 1: 87500
Cuartil 2: 125000
Cuartil 3: 202500

Para calcular el cuartil hay varias posibilidades, en todas se debe cumplir que se descarta la misma porción de la población.
En **numpy**, si queremos calcular el **cuartil 1** se hace la siguiente cuenta:

- (N-1) * 0.25 ⇒ en este caso (8-1)*0.25 = **1.75**
- Luego se devolvería: array[1] + (array[2]-array[1])*0.75
  - En este caso: 50000+(100000-50000)*0.75=87500

Box Plot

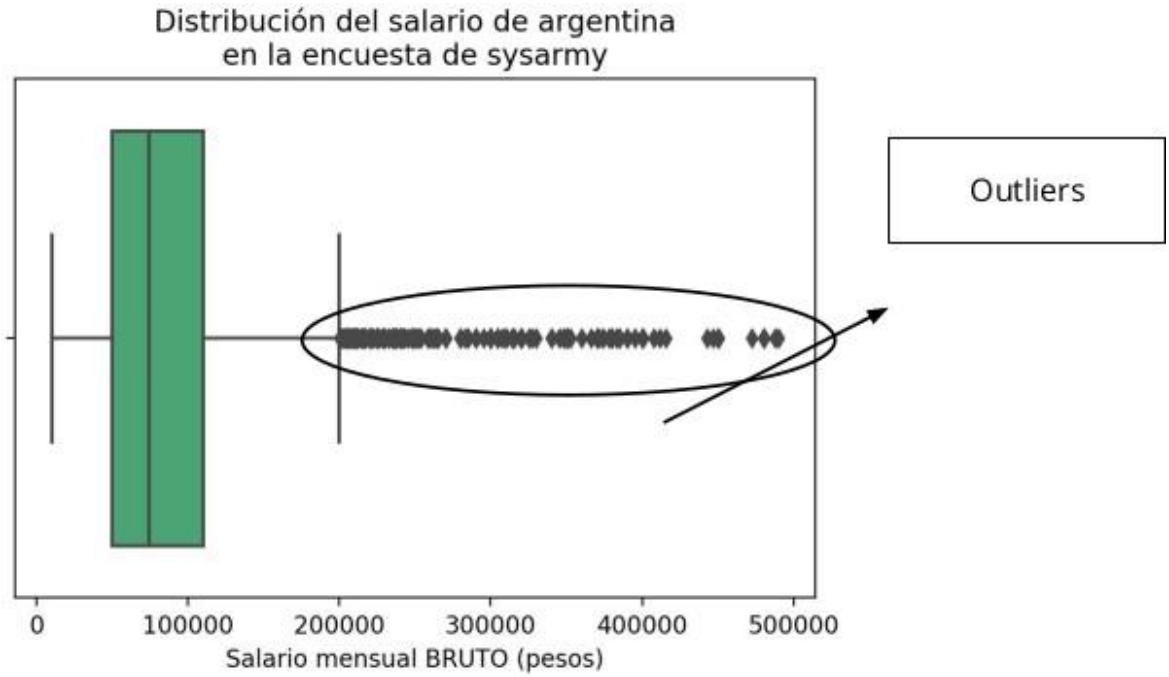
Un diagrama de caja o Box plot muestra visualmente la distribución de los datos numéricos y la asimetría mediante la visualización de los cuartiles (o percentiles) y los promedios de los datos.



Ordenar los valores que tenemos en una recta donde tenemos a los extremos el min y max.

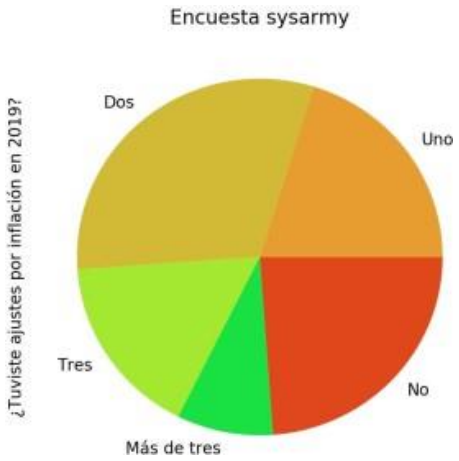
Nos dibuja la mediana y los cuartiles 1 y 3.

Es una forma de ver las variables continuas de forma mas simple



Distribucion Discreta

Pie Chart

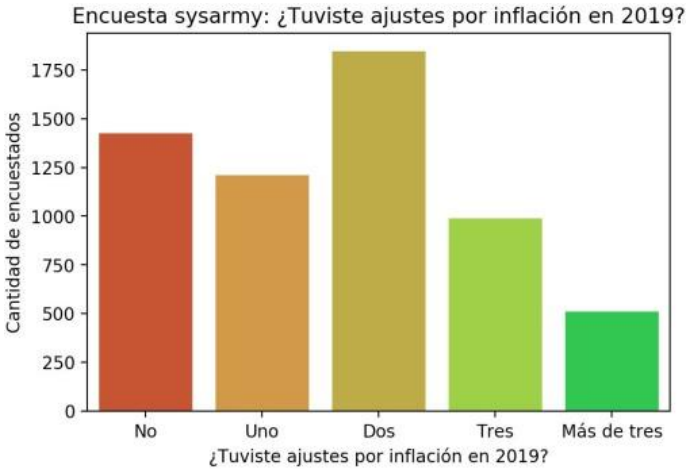


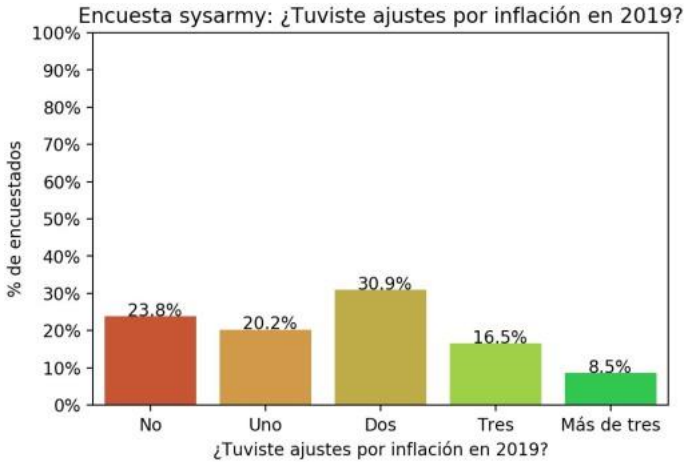
Stacked Bar Plot

Cada porcentaje de la barra muestra una proporcion que tiene esa categoria

Usar el pie chart cuando la diferencia es grande, tipo 90% y 10%, asi las diferencias se ven mejor.

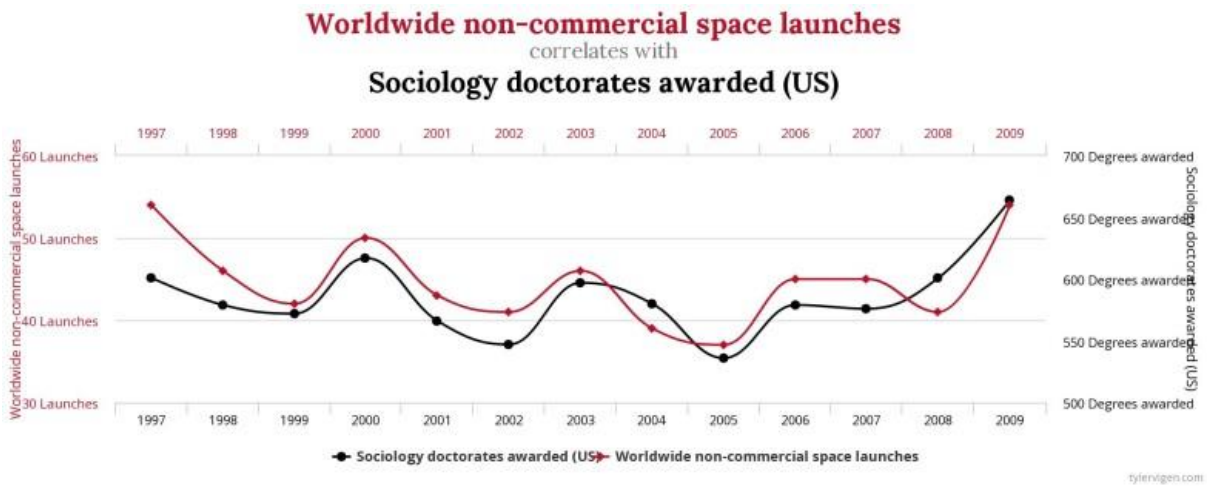
Bar plot





En el primero se ve claramente la diferencia de magnitud → puede ser debido a que se usa la cantidad y no porcentajes

De relación



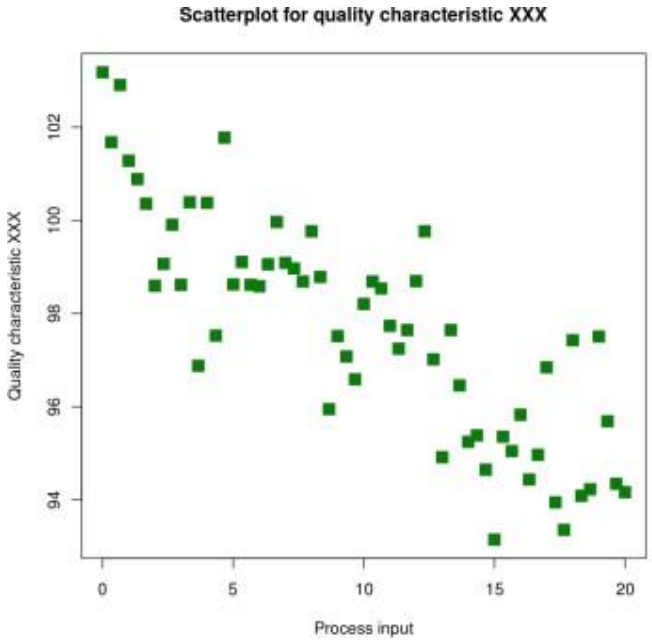
Hay una correlacion entre las curvas, pero en este caso es de casualidad → correlaciones espurias

De dispersión

Los diagramas de dispersión son útiles para ver si dos variables están correlacionadas

Ponemos una variable en el eje x y otra en el y.
Si los puntos se alinean ⇒ decimos que estan correlacionados

Scatter plot



Utiliza las coordenadas cartesianas para mostrar los valores de dos variables

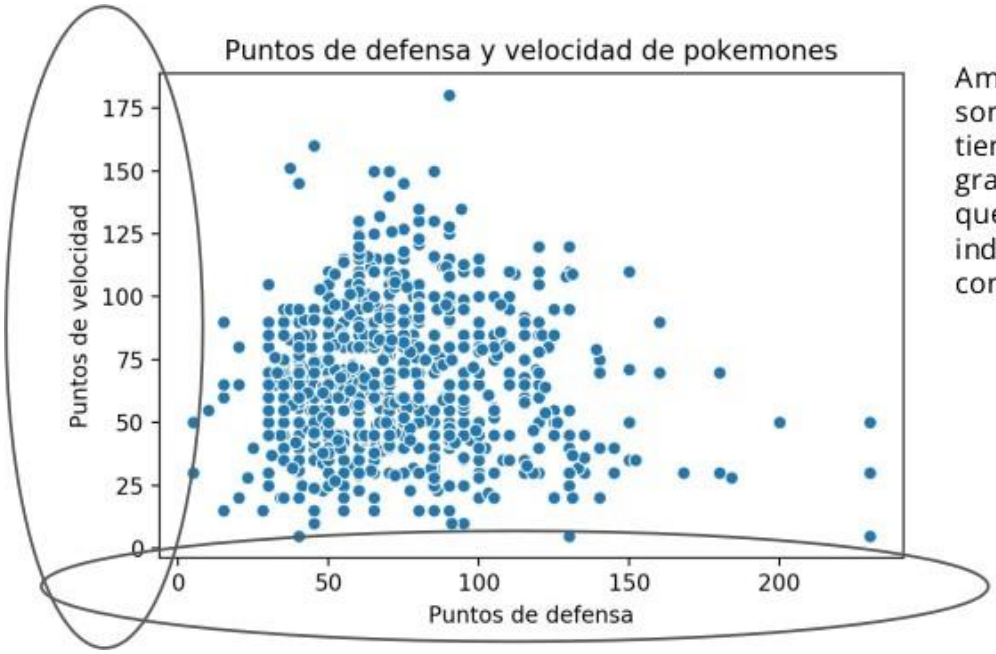
Aca vemos una correlacion negativa (va hacia abajo)

Correlación de Pearson

Para 2 variables podemos medir su correlación lineal con el **coeficiente de correlación r**.

Este coeficiente, es una función que mide cuán relacionada estan 2 variables de forma lineal.

- Si da 0 ⇒ NO existe correlación
- Si da 1 ⇒ Están relacionadas linealmente de forma perfecta (todos los puntos están en una línea)
- Si da -1 ⇒ Existe una correlación negativa perfecta.



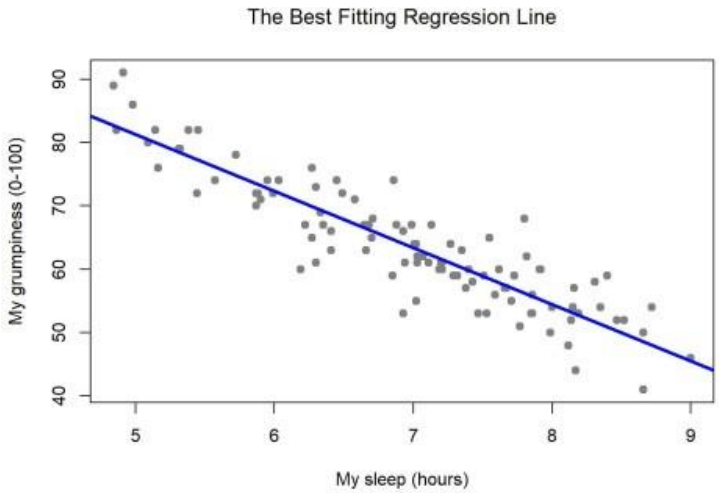
Ambos ejes o bien son continuos o tienen una buena granularidad (lo que los vuelve indistintos del continuo)

No hay ningun tipo de correlacion, esta totalmente disperso

Regression plot

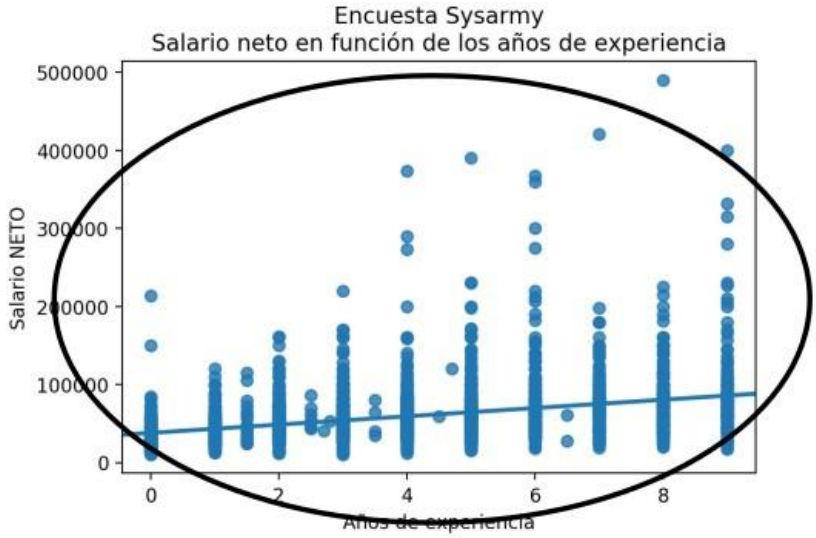
Es un Scatter Plot con la regresion de cuadrados minimos dibujada encima.

Aca si hay una correlacion



Se incluye una guía visual que muestra la relación entre las variables

Ejemplo confuso

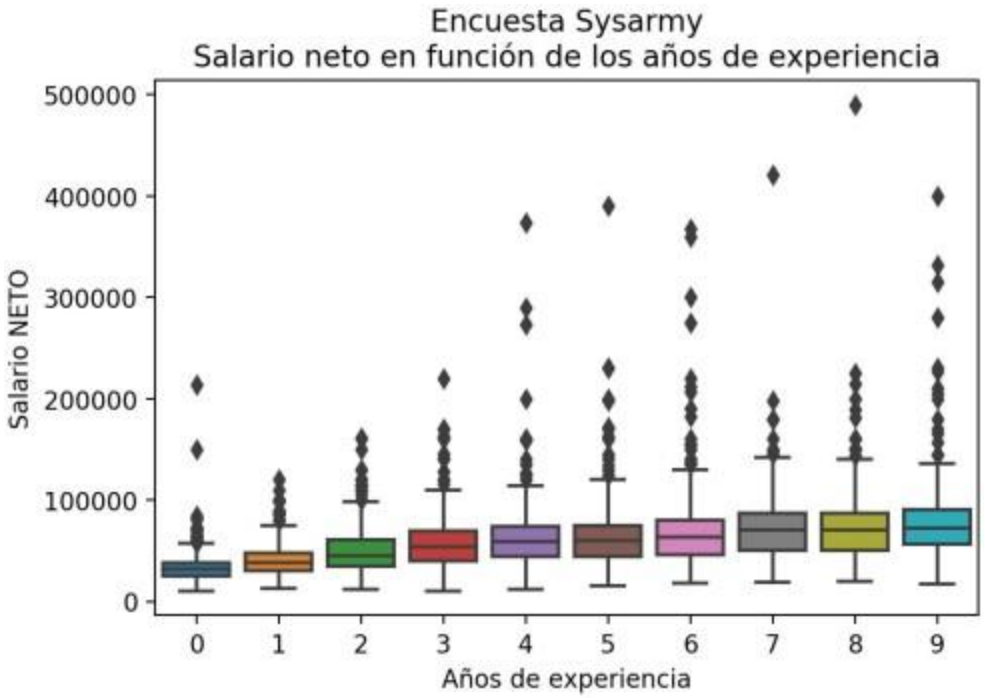


Si bien los años de experiencia son una variable que puede ser real y continua, nadie va a contestar que tiene 3.253 años de experiencia, no hay buena granularidad, los valores más comunes son enteros.

Podemos pensarlo como comparar distribuciones continuas.

La variable no es lo suficientemente continua

Una mejora



Heat Map

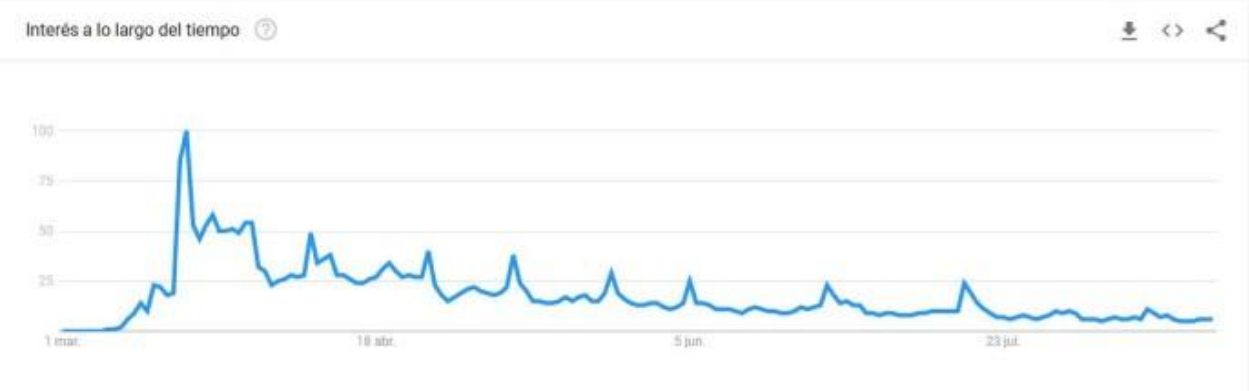
Sirve para comparar distribuciones en donde **ambos ejes son discretos y un tercero de "profundidad" numérico**.

Generalmente surge de calcular algún agregado del grupo al que corresponde el rectángulo.

Aca los salarios estan indicados por el color.

Lineplot

→ Graficos en funcion del tiempo



Aca es la cantidad de veces que se busco la palabra cuarentena a lo largo del tiempo

Violin plots

Similares a los box plots

Es un diagrama de caja con un diagrama de densidad kernel rotado en cada lado.
El diagrama de violín es similar a los diagramas de caja, excepto que también muestran la densidad de probabilidad de los datos en diferentes valores.

En el eje y se muestra la distribucion de los datos