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 y saber para qué es útil cada uno
- Entender de forma eficiente los datos
- Comunicar de forma concisa y clara
- Encontrar patrones/relaciones

¿Por qué es necesario graficar?

- 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

¿Por qué es necesario graficar?

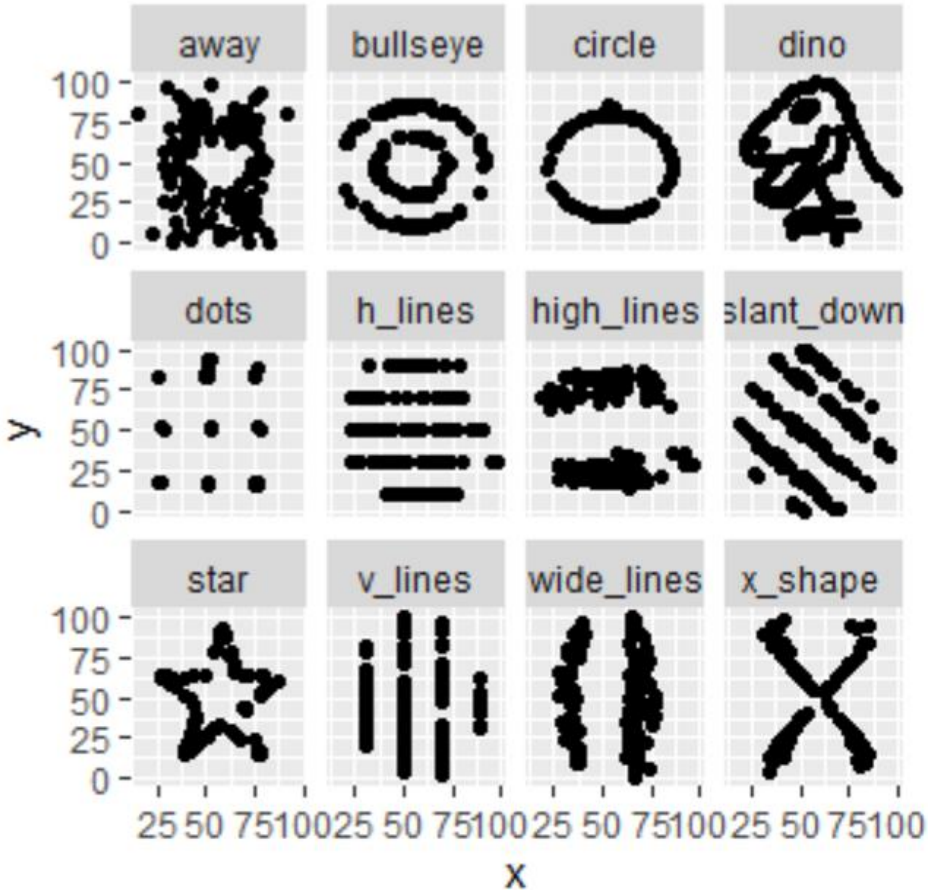
- Tratar con conjuntos de datos **multidimensionales** con más de una variable o atributo comienza a causar problemas, ya que estamos restringidos a comunicar en dos dimensiones (a lo sumo 3).
- Los gráficos no son simplemente: “imágenes bonitas”
- No toda la información importante se puede adivinar a través del análisis estadístico...

Datasaurus

- Todo estos gráficos tienen la misma media y desvío estándar

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

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



Datasaurus

- Acá el artículo de Cairo
- Acá hay más para jugar

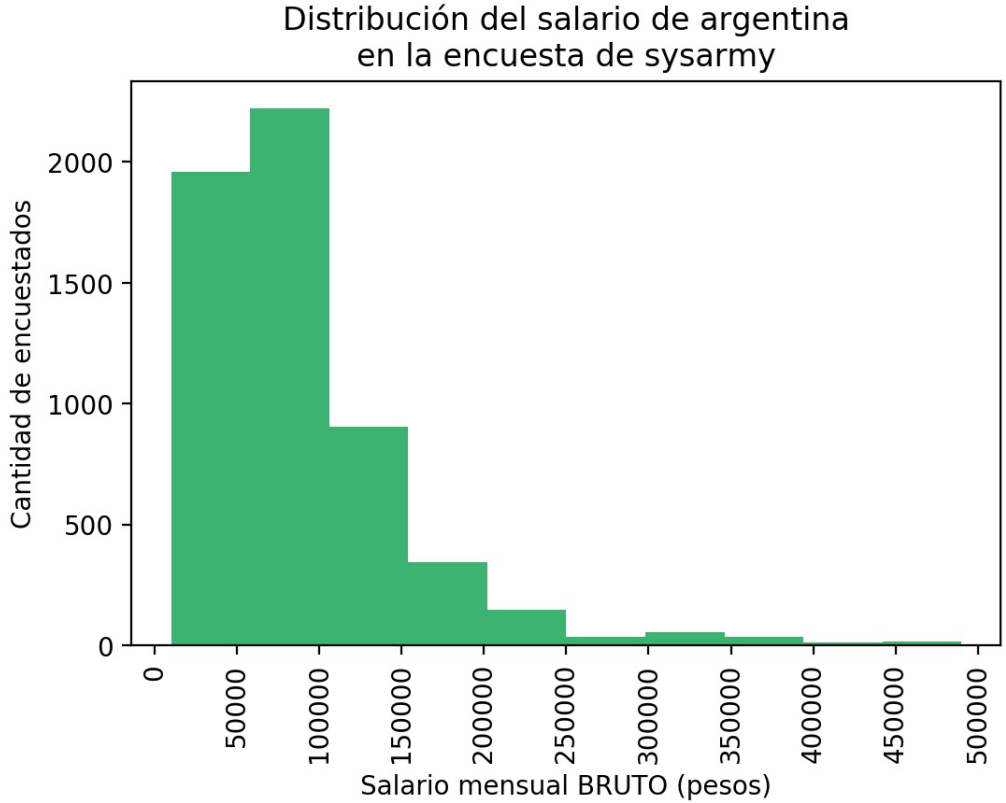
Visualización para ML

- 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 atípicos o no linealidad
- Evaluar el ajuste del modelo:
  - predicho vs observado
  - análisis de residuos

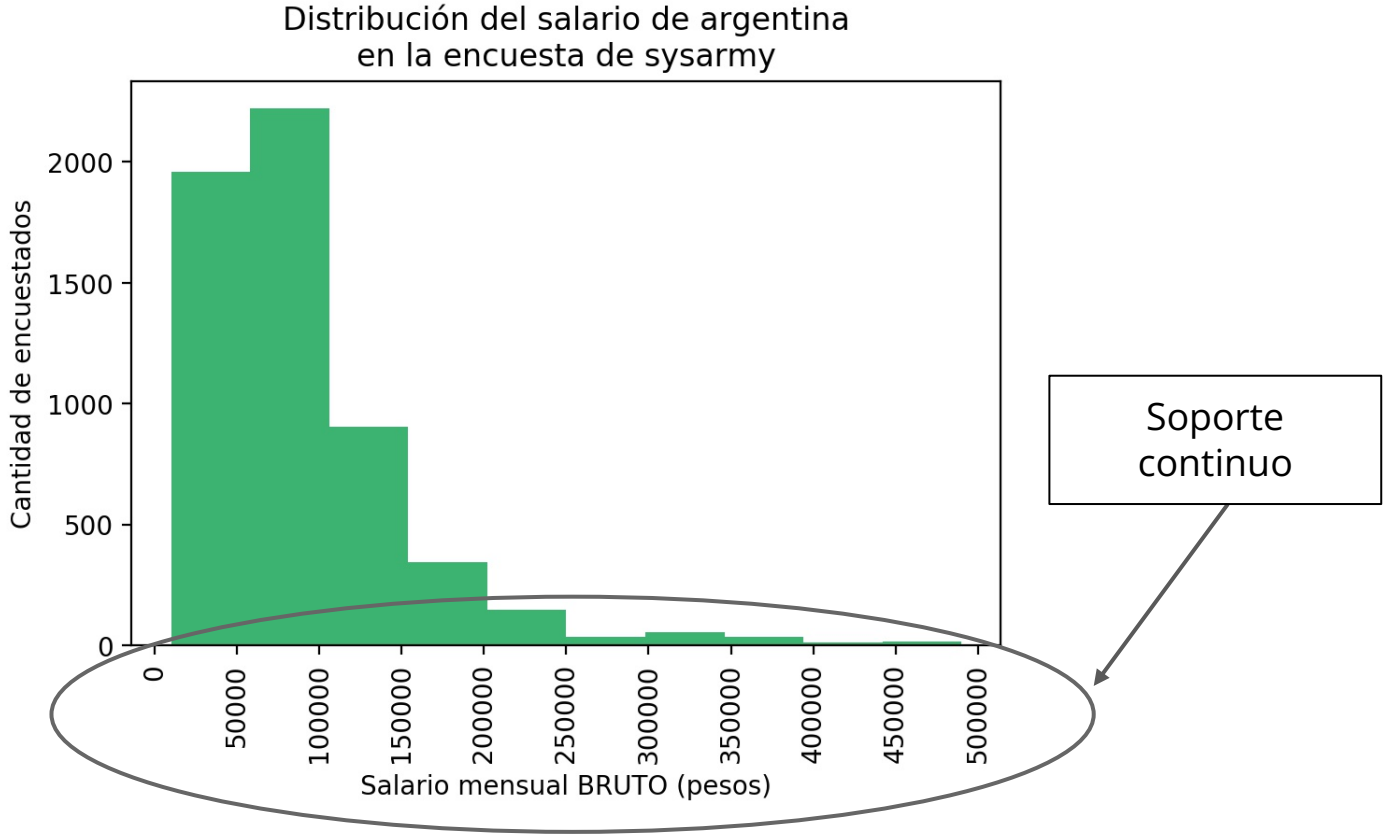
Plots

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

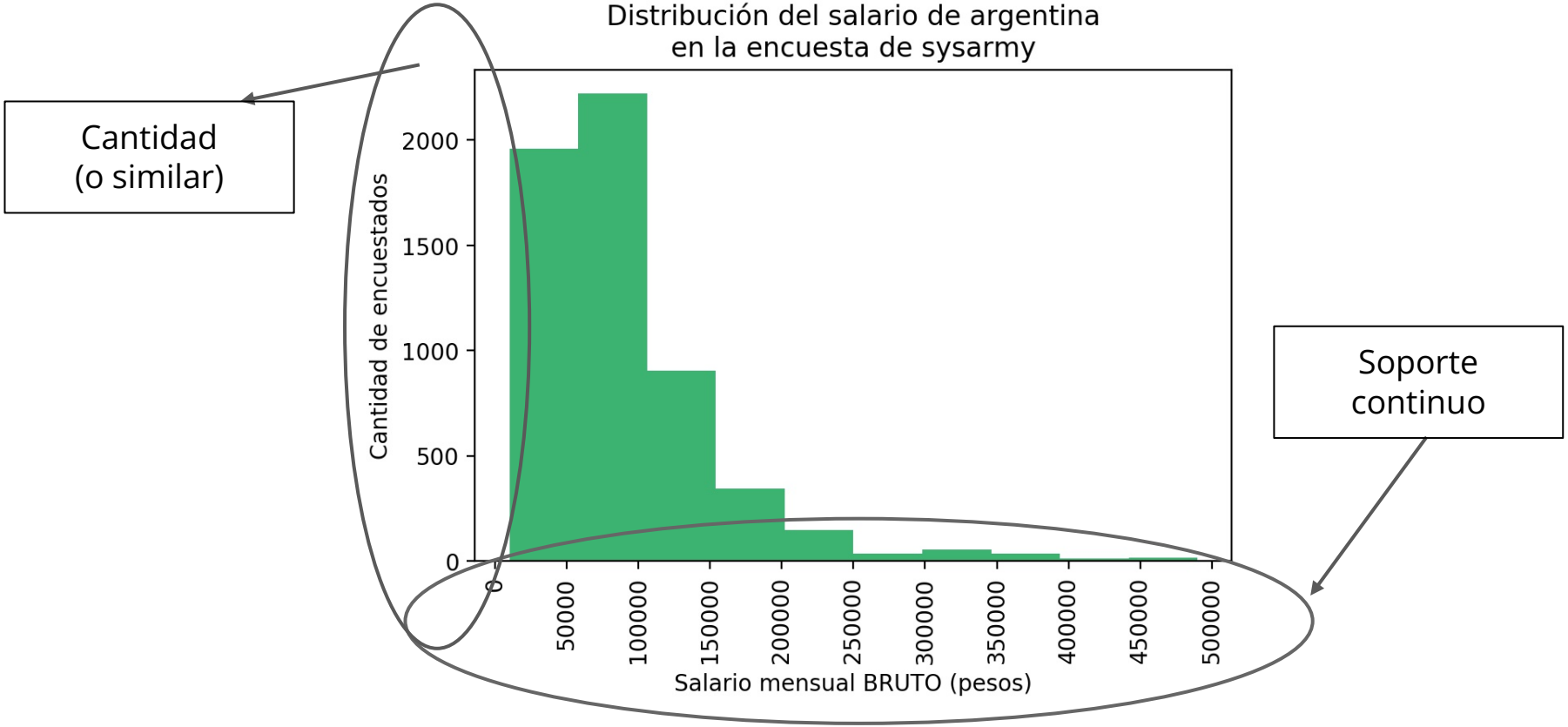
De distribución continua



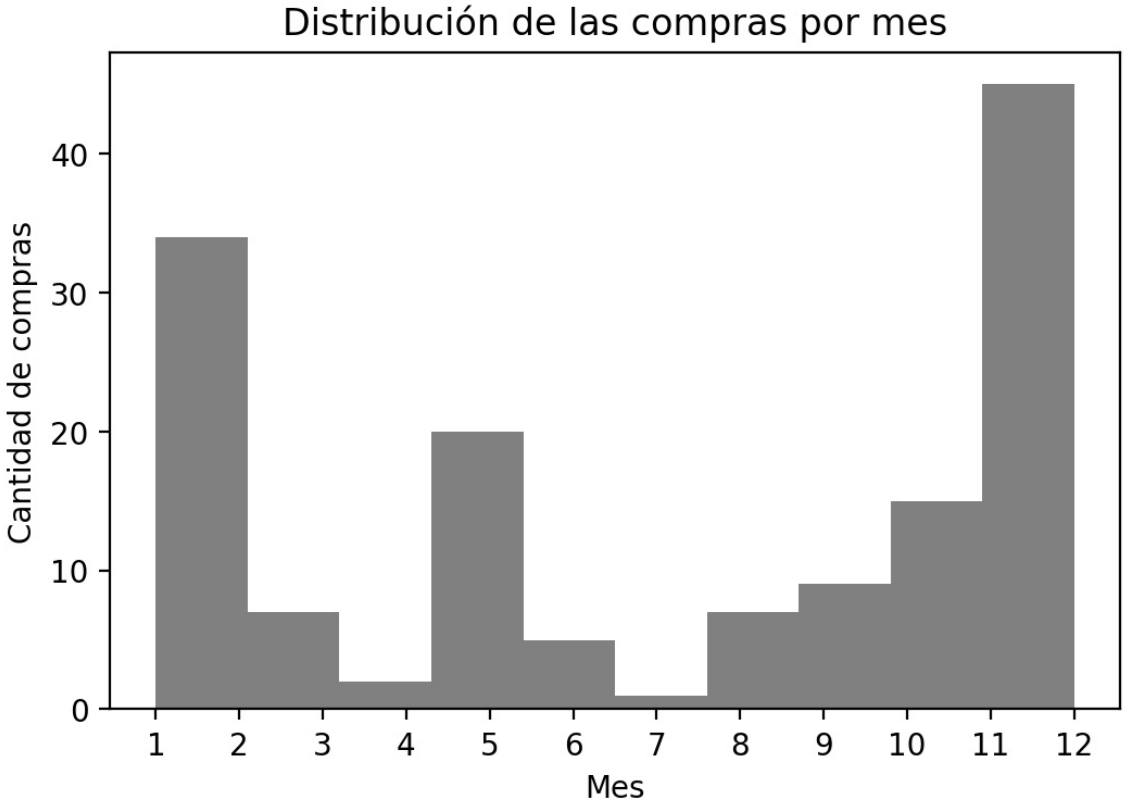
De distribución continua



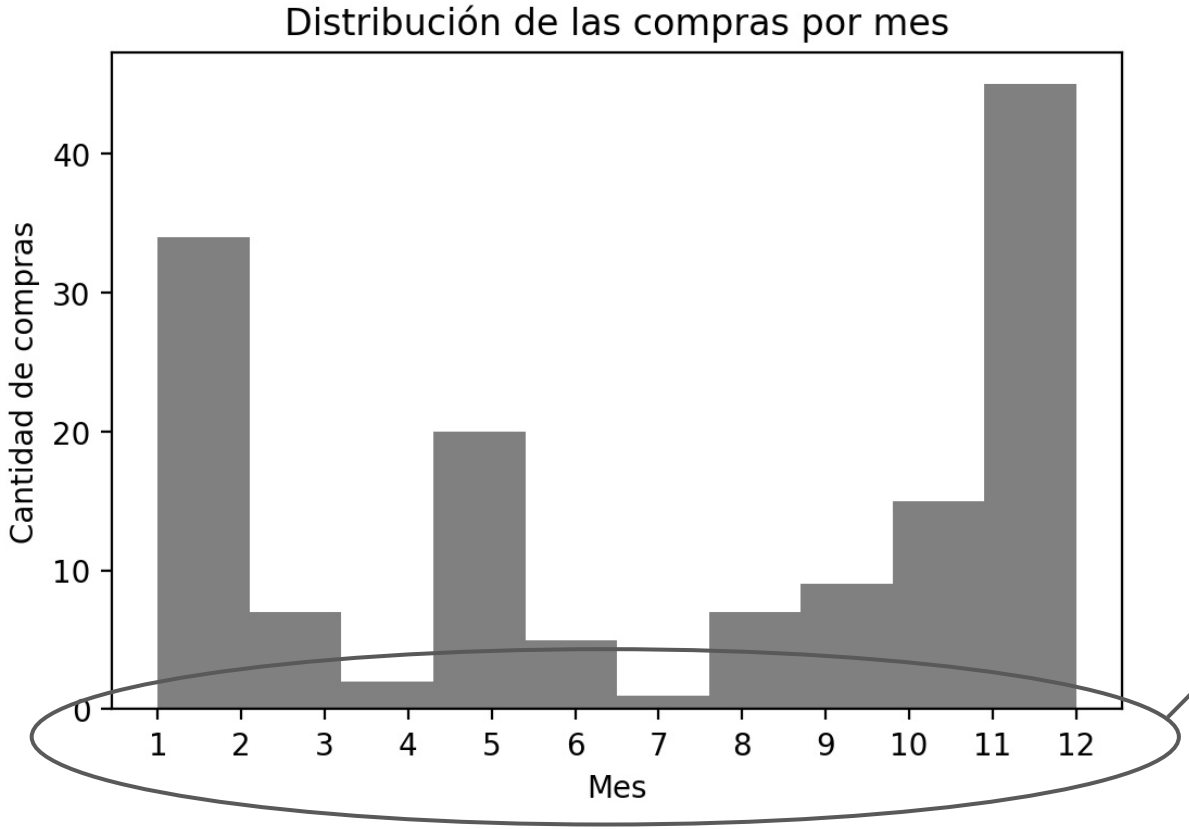
De distribución continua



De distribución continua



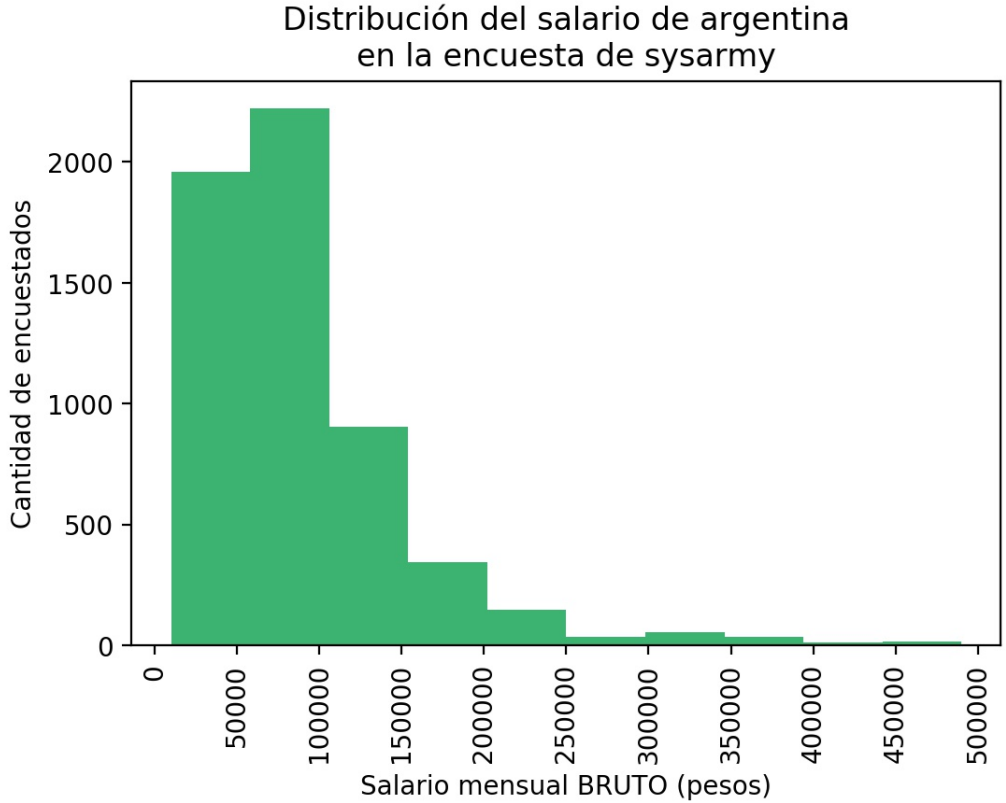
De distribución continua



Uno de los errores más comunes cuando se empieza a hacer visualizaciones es confundir cuales tienen que tipo de soporte

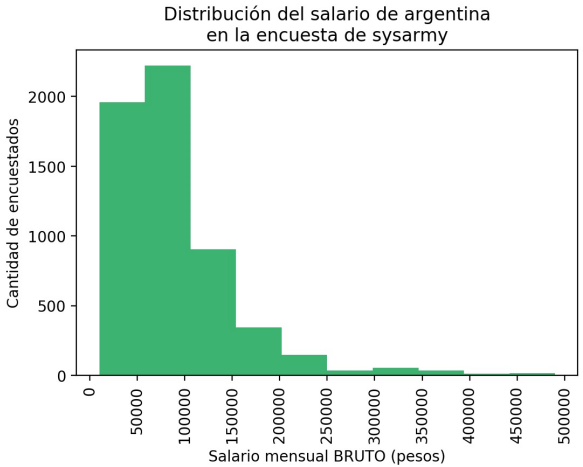
Soporte discreto

Histograma

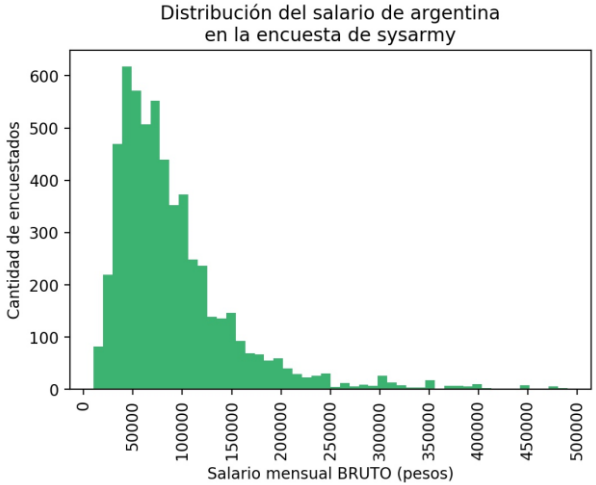

Distribución del salario de argentina
en la encuesta de sysarmy
Cantidad de encuestados
0
500
1000
1500
2000
0
50000
100000
150000
200000
250000
300000
350000
400000
450000
500000
Salario mensual BRUTO (pesos)

Histograma

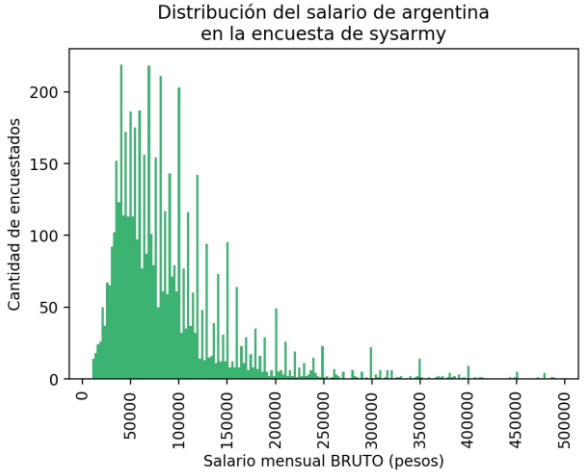
10 bins



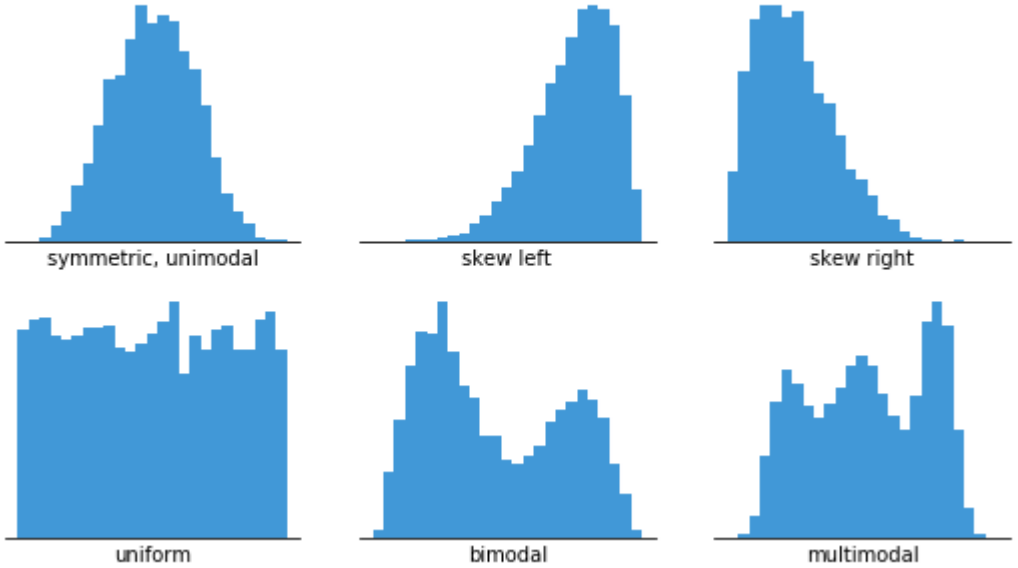
50 bins



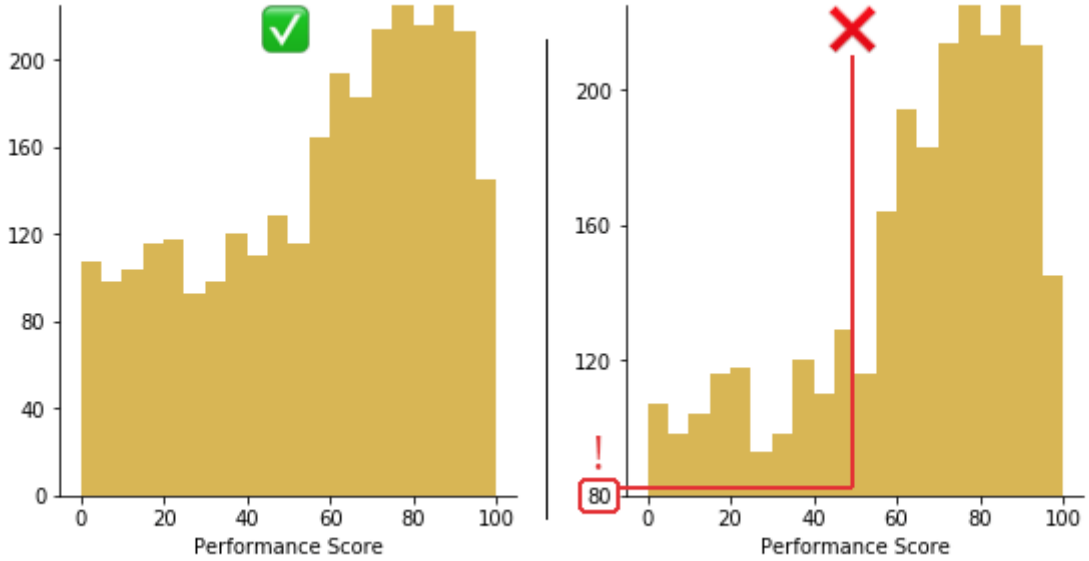
200 bins



Histograma

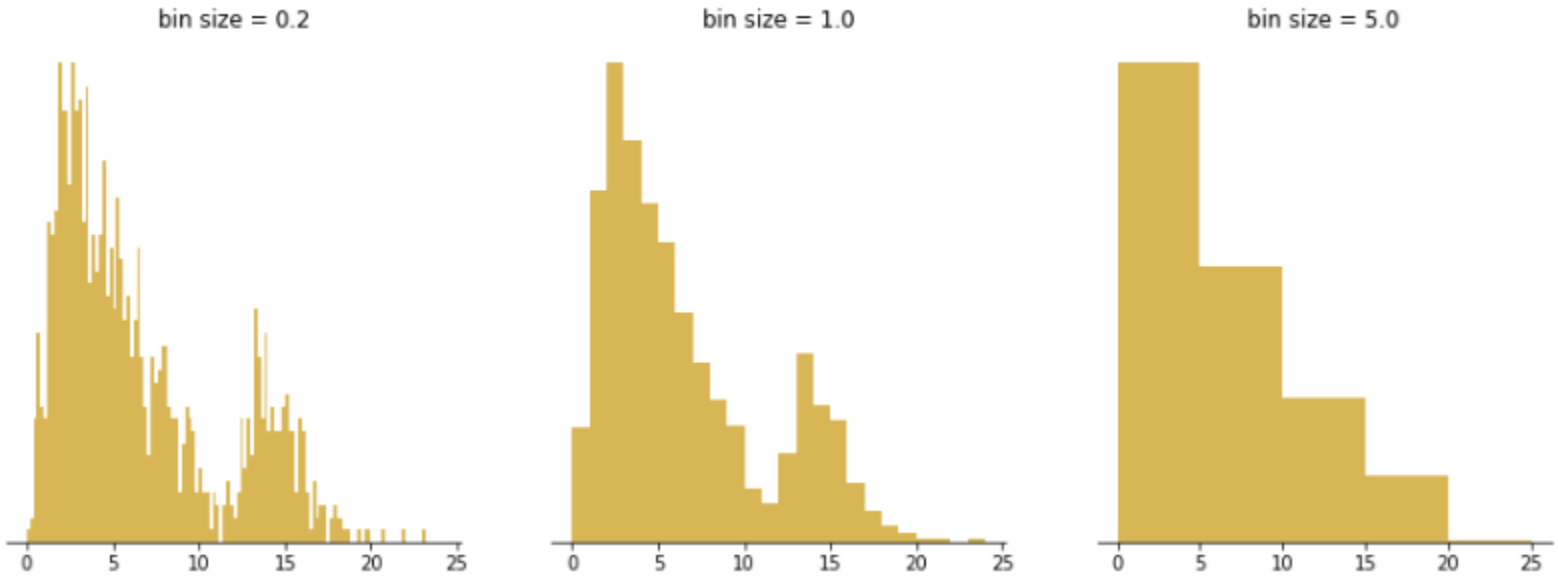

symmetric, unimodal
skew left
skew right
uniform
bimodal
multimodal

Histograma



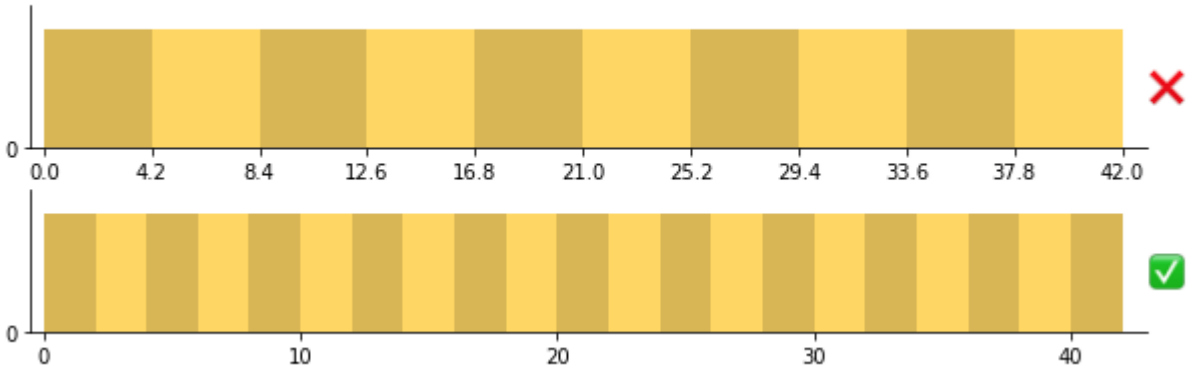
Use a zero-valued baseline

Histograma



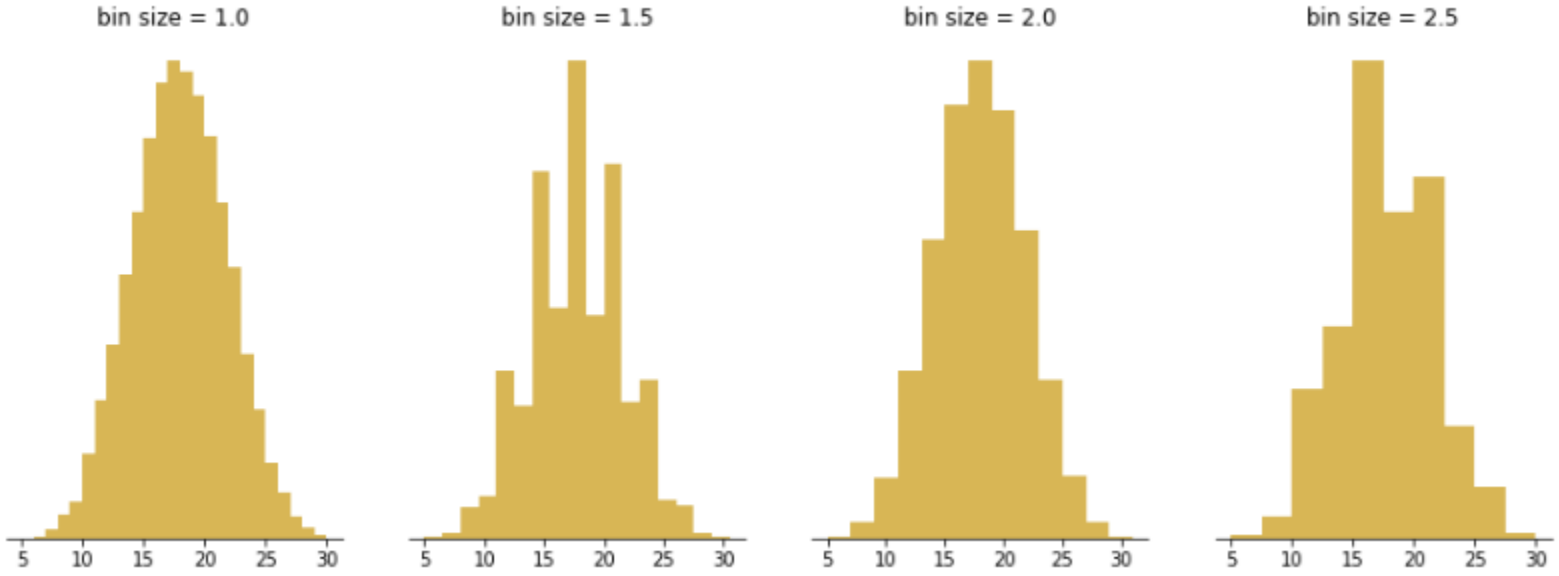
Choose an appropriate number of bins

Histograma



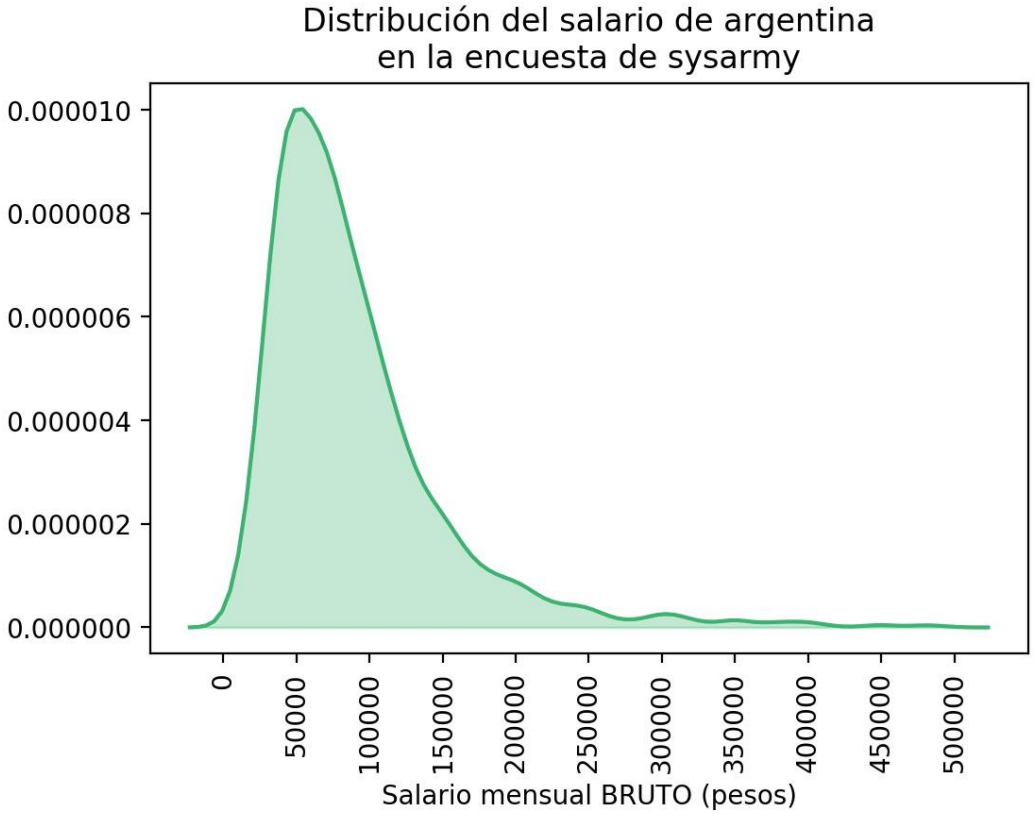
Choose interpretable bin boundaries

Histograma

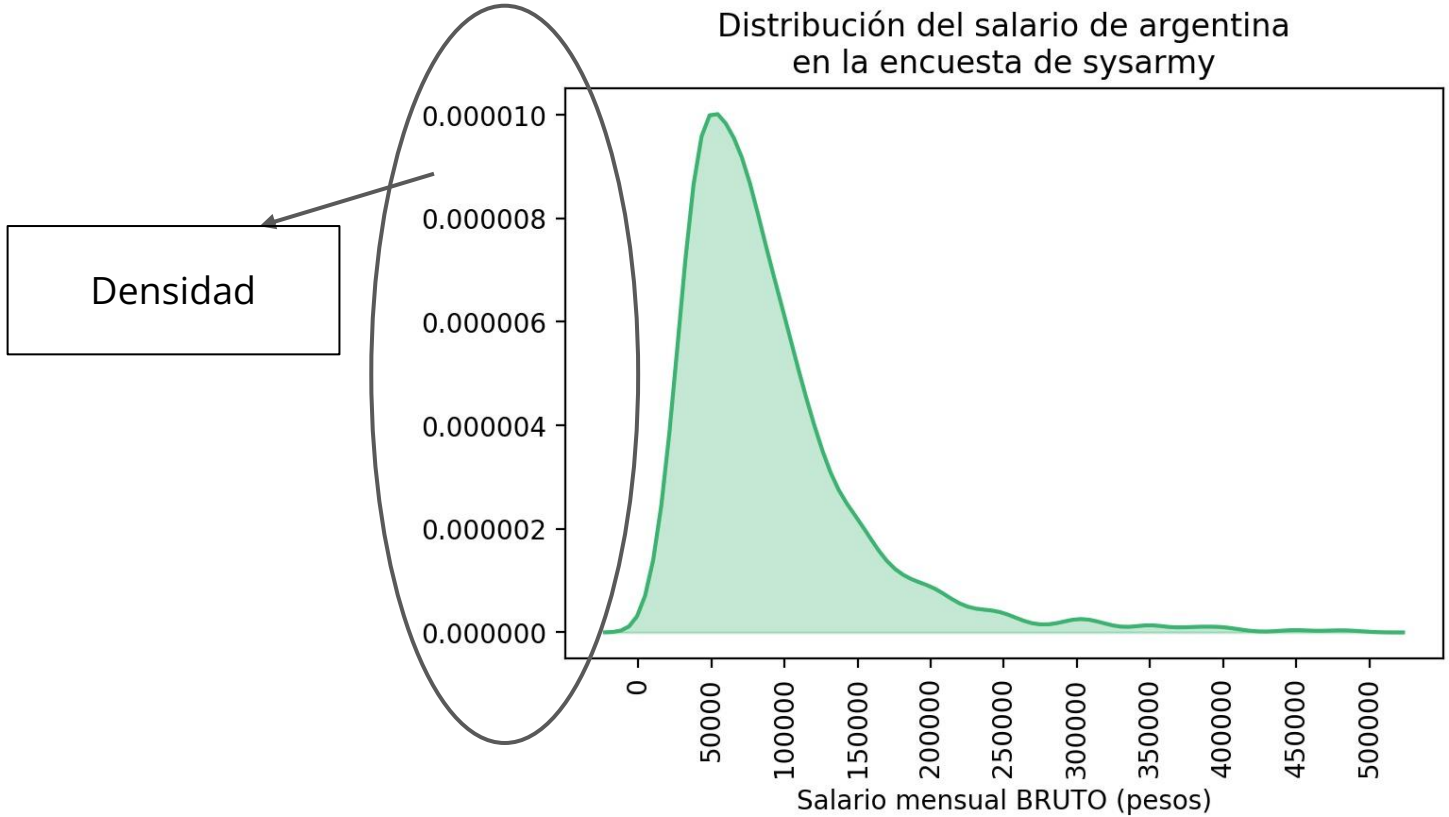


El histograma puede verse anormalmente "desigual" simplemente debido a la cantidad de valores que posiblemente podría tomar cada contenedor.

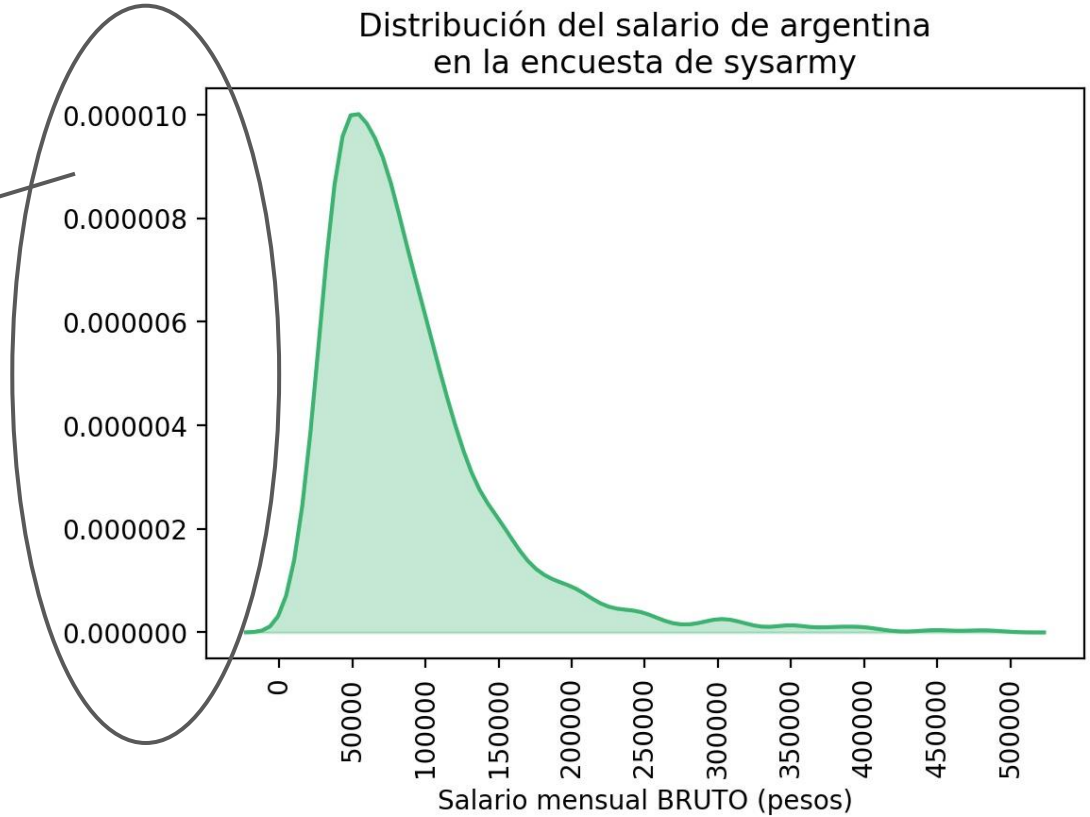
Density plot



Density plot



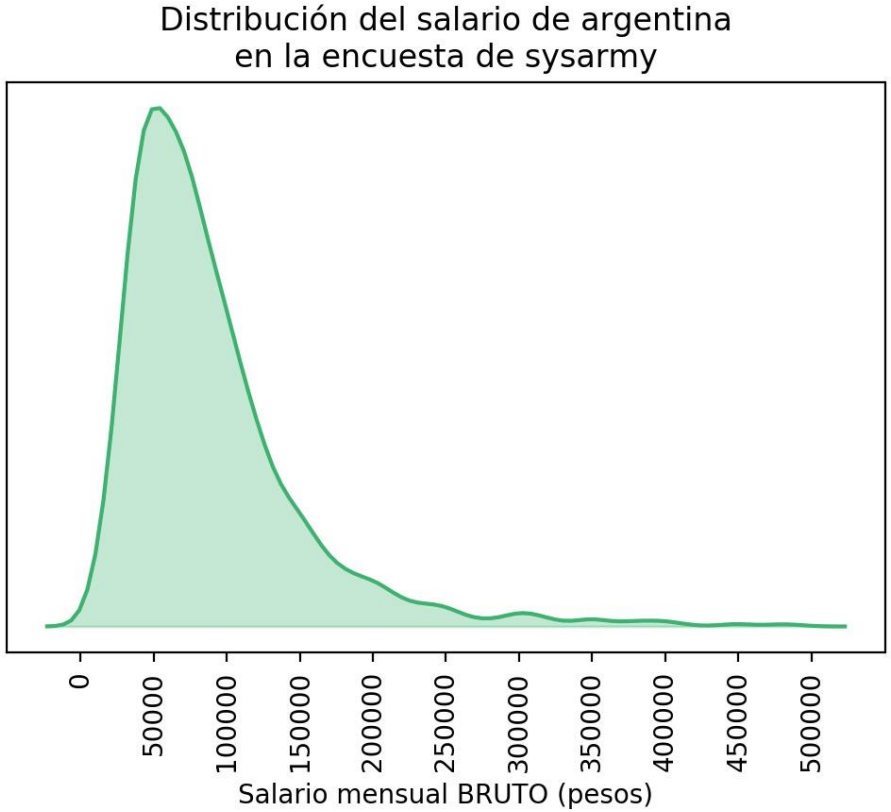
Density plot



Densidad

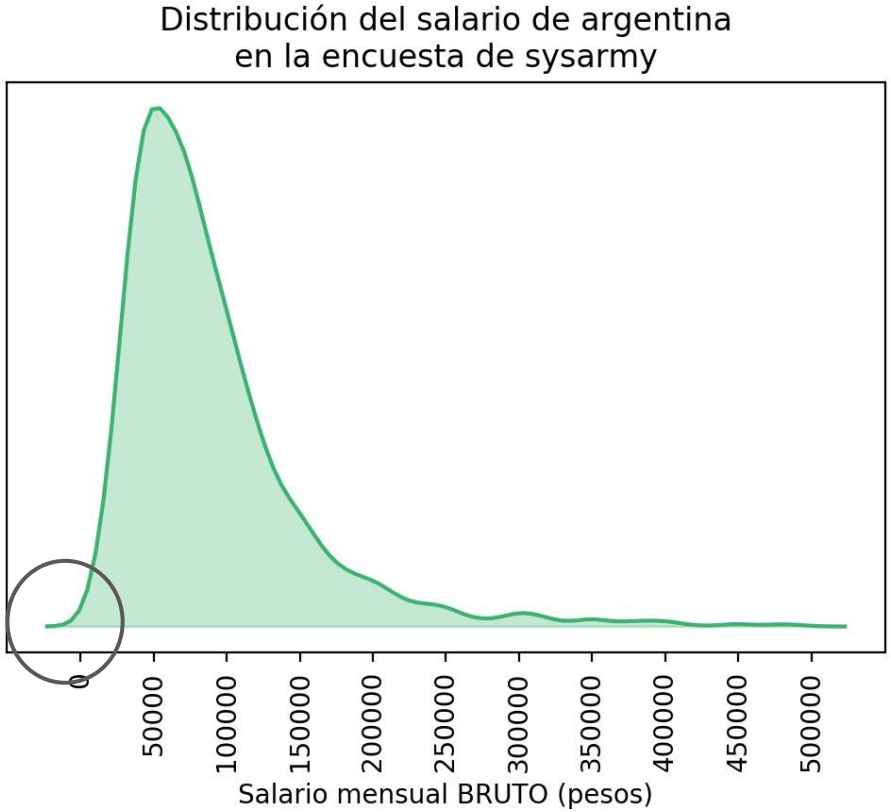
- Poco interpretable
- Ya no sabemos las cantidades

Density plot



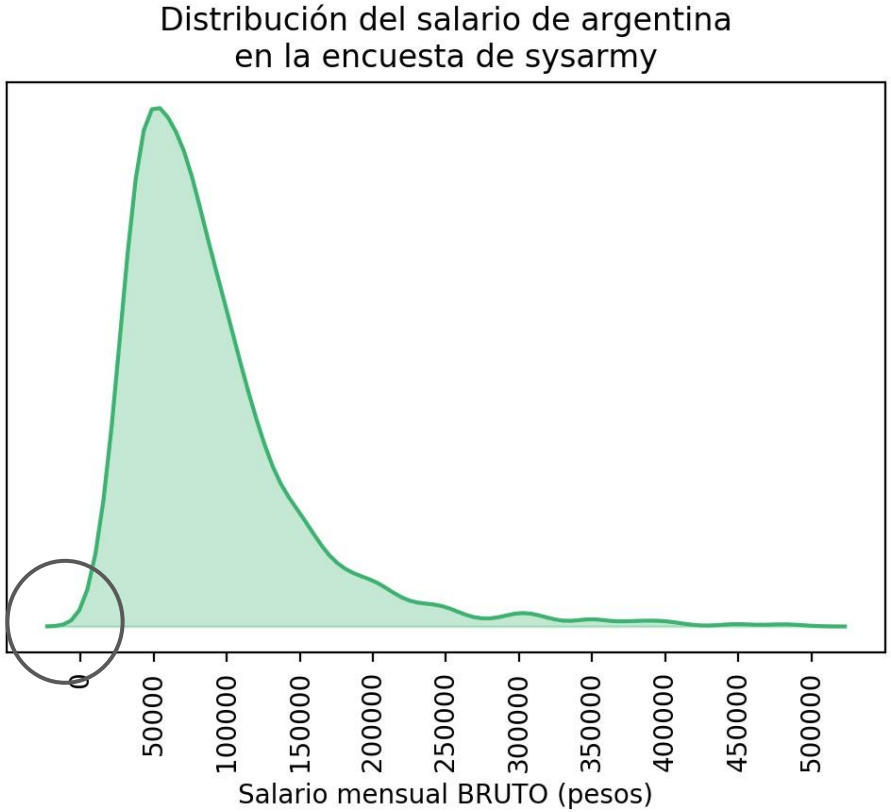
Density plot

¿Alguien cobra 0?

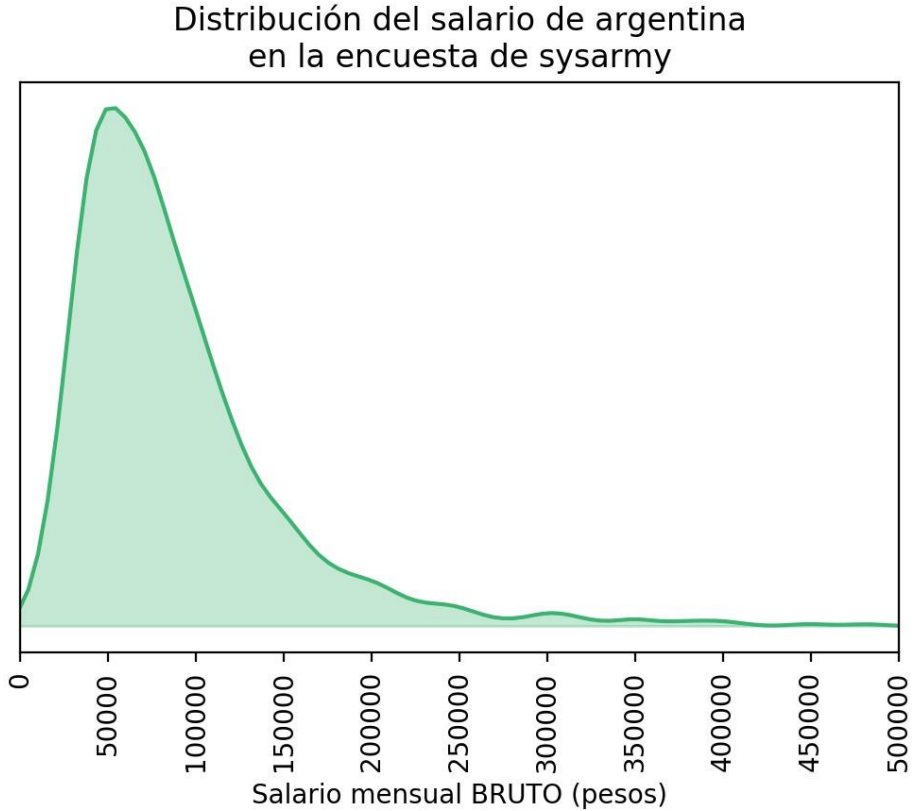


Density plot

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



Density plot



Algunos números útiles

- **Media**: Es el promedio
- **Mediana**: es el valor que está en la mitad de la población
- **Cuartil**: son los valores límite que dejan al 25% de la población entre ellos
- **Rango intercuartílico**: el rango entre el cuartil 1 y el cuartil 3

Población de salarios

$ 50000

$ 100000

$ 120000

$ 130000

$ 200000

$ 210000

$ 220000

$ 50000

Población de salarios

$ 50000

$ 100000

$ 120000

$ 130000

$ 200000

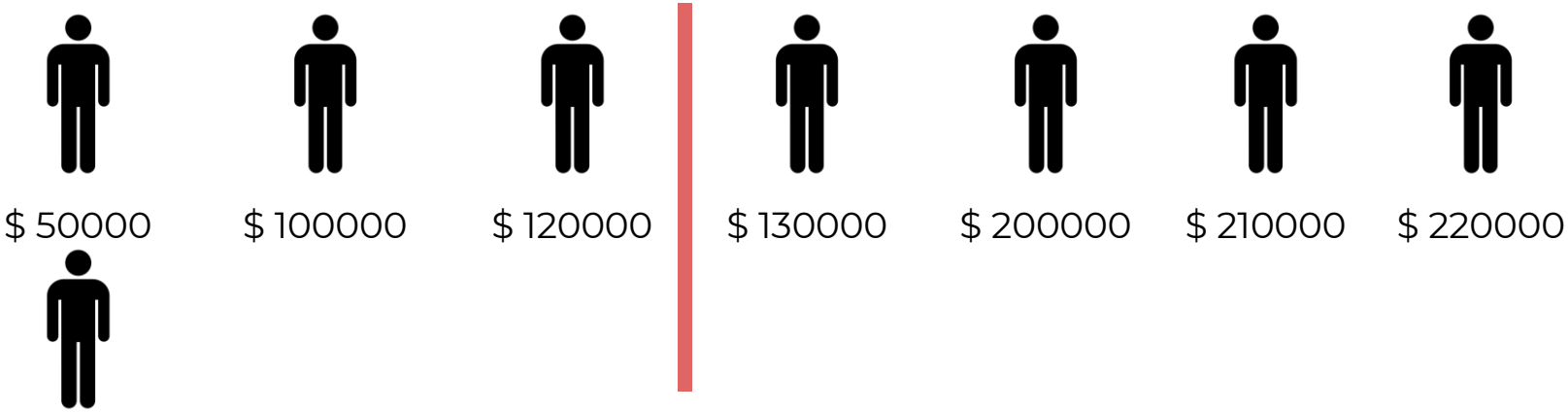
$ 210000

$ 220000

$ 50000

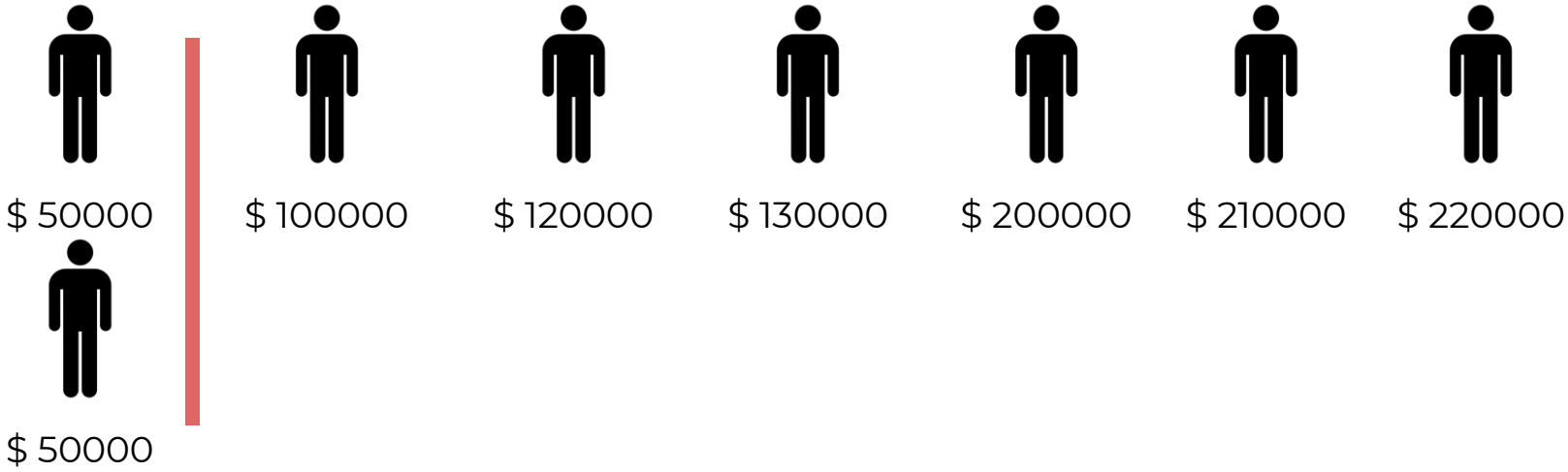
Media: 135000

Población de salarios



Media: 135000
Mediana: 125000

Población de salarios

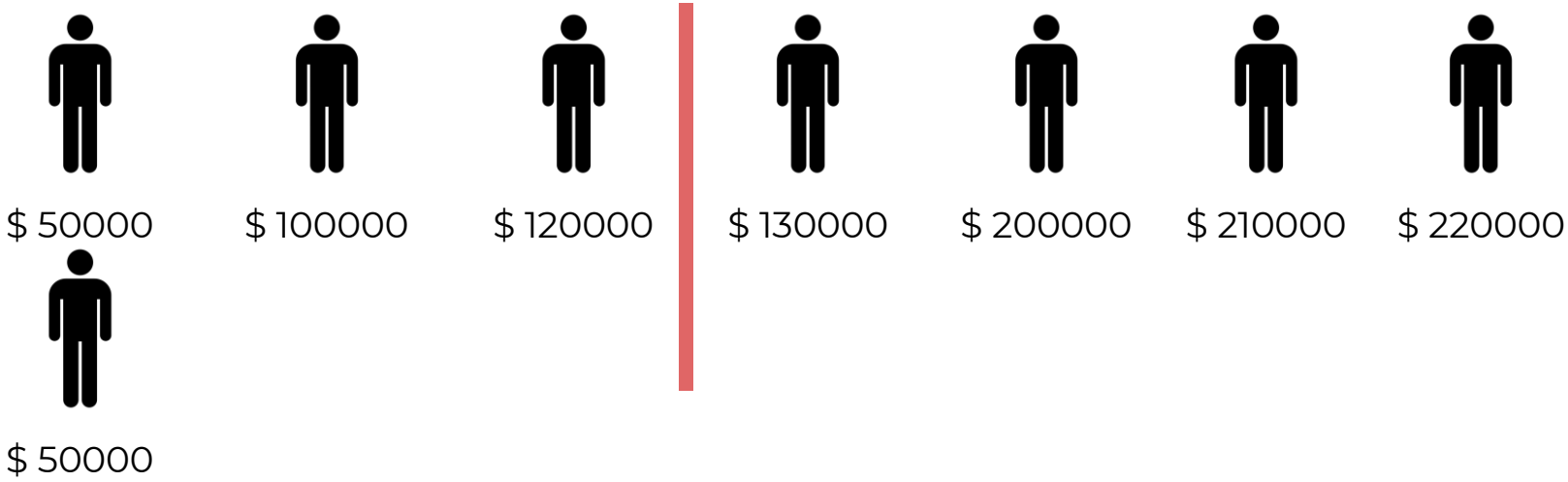


Media: 135000
Mediana: 125000
Cuartil 1: 87500

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

Población de salarios

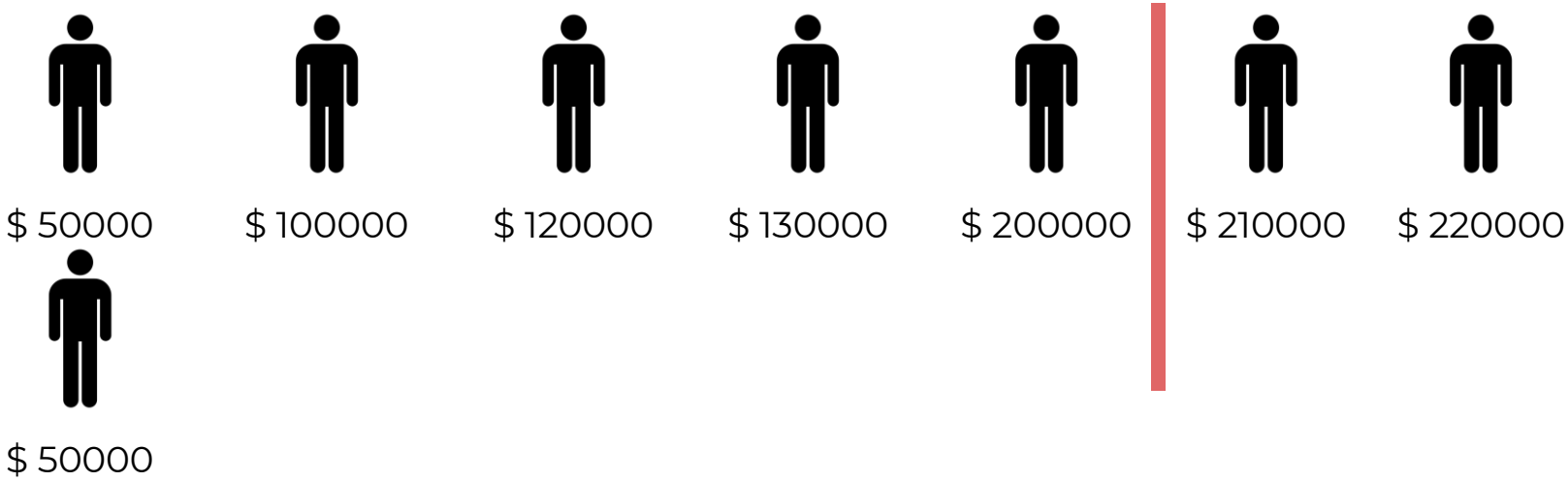


Media: 135000
Mediana: 125000
Cuartil 1: 87500
Cuartil 2: 125000

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

Población de salarios

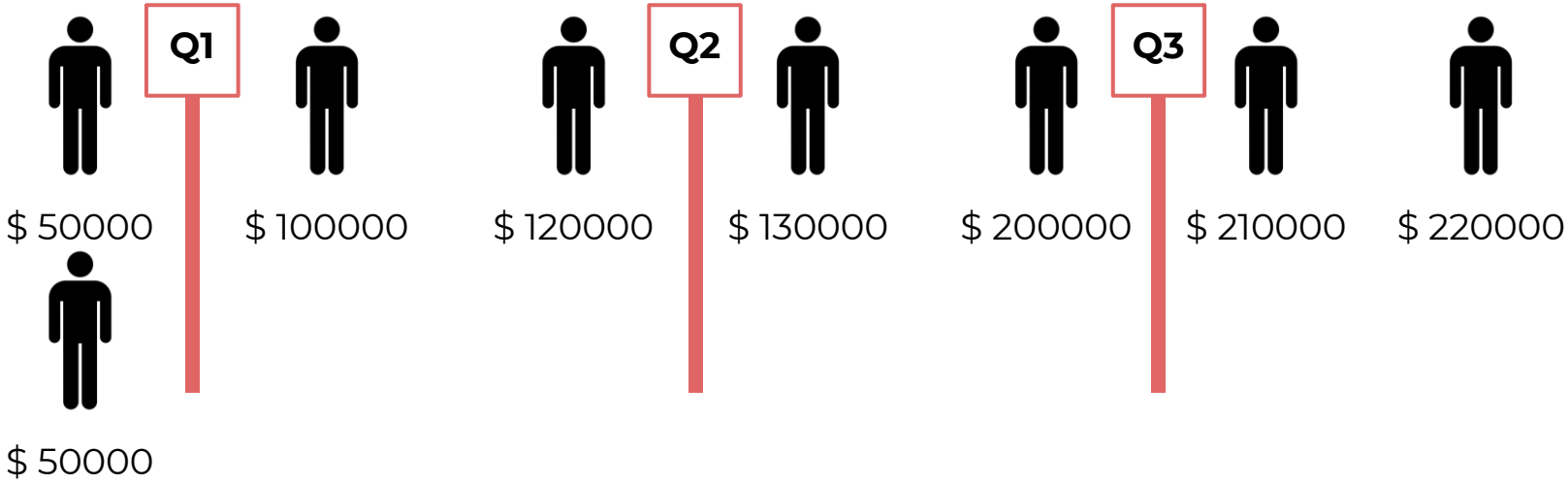


Media: 135000
Mediana: 125000
Cuartil 1: 87500
Cuartil 2: 125000
Cuartil 3: 202500

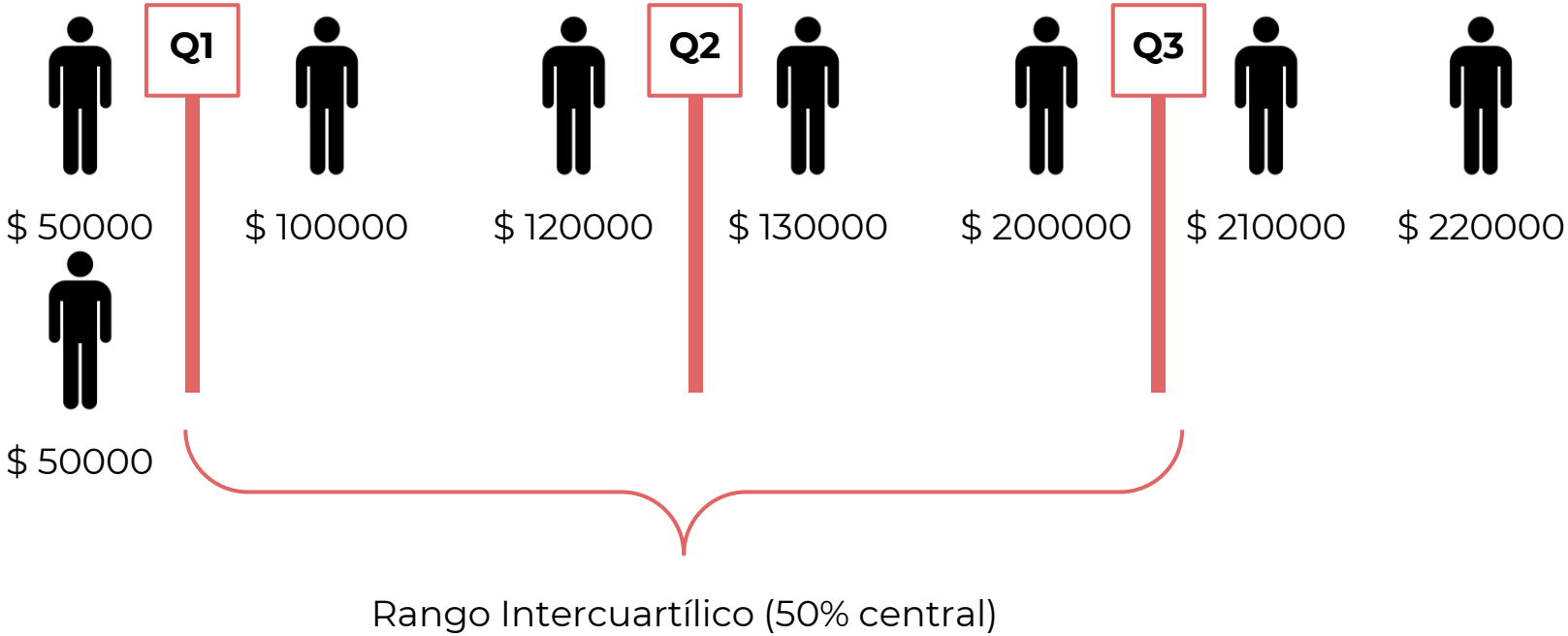
Para calcular el cuartil hay varias posibilidades, en todas se debe cumplir que se descarta la misma porción de la población.
En **numpy**, si queremos calcular el **cuartil 1** se hace la siguiente cuenta:

- (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

Población de salarios

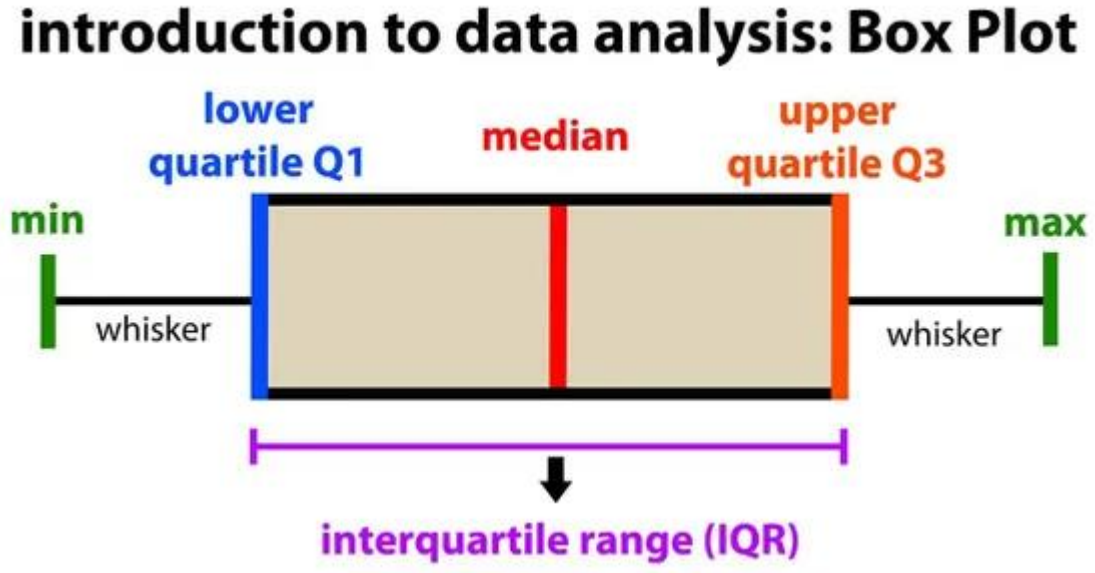


Población de salarios

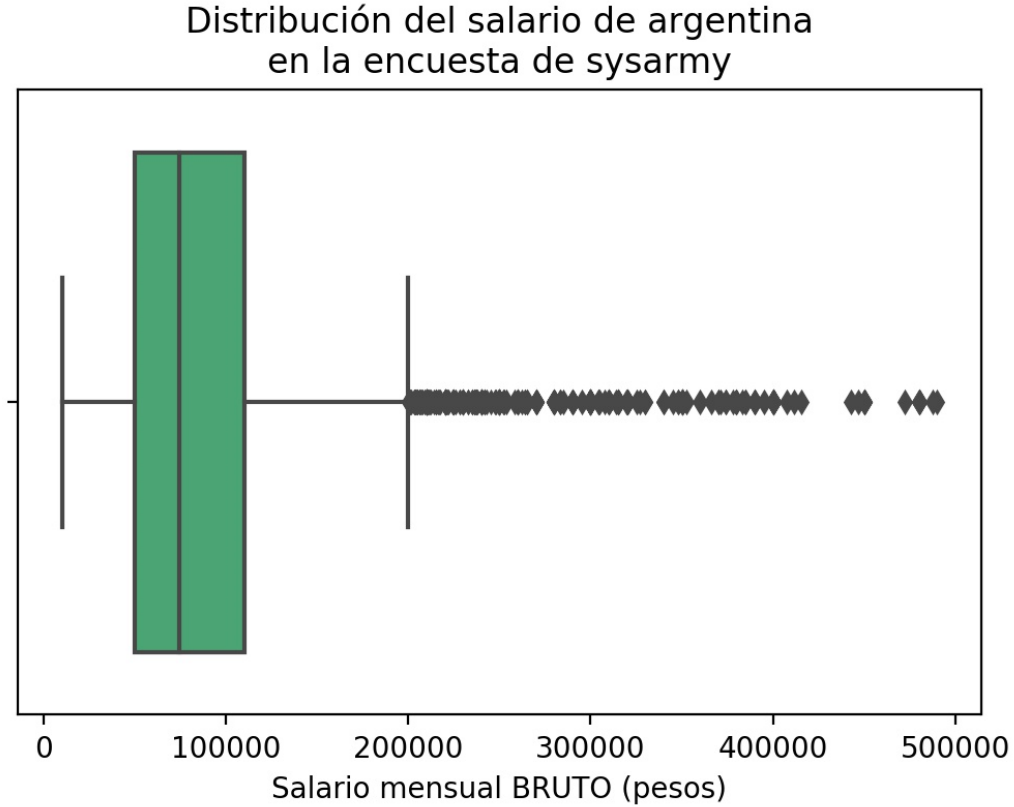


Box plot

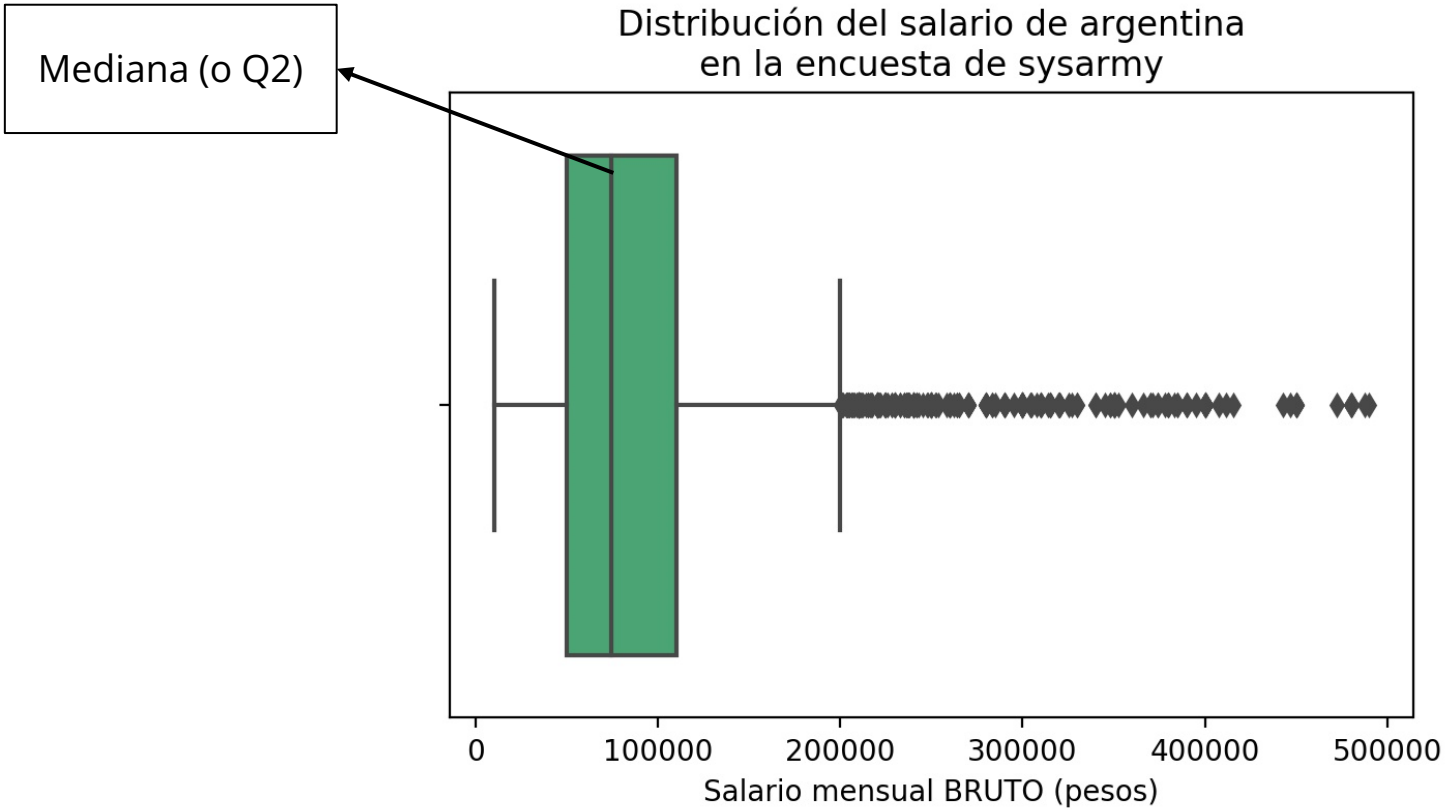
Un diagrama de caja o Box plot muestra visualmente la distribución de los datos numéricos y la asimetría mediante la visualización de los cuartiles (o percentiles) y los promedios de los datos.



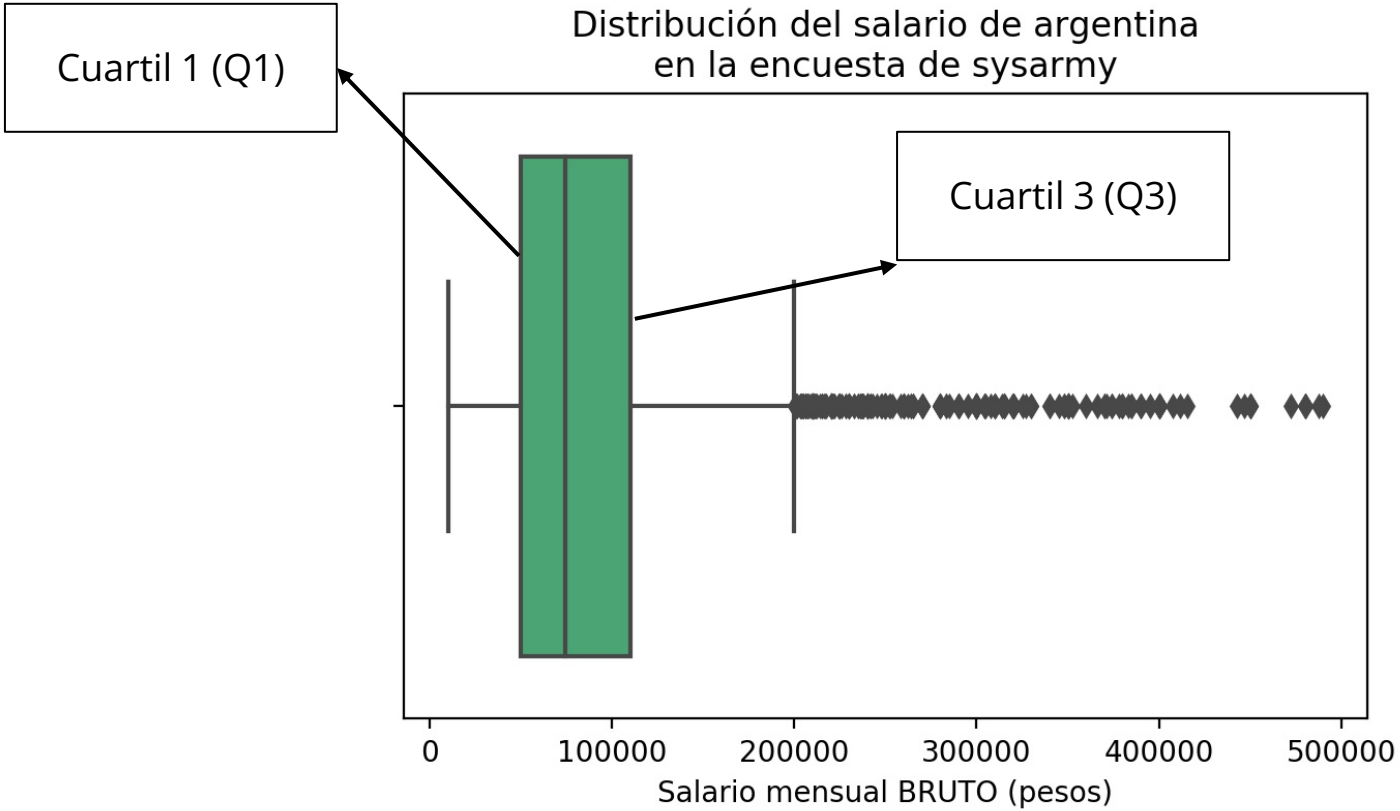
Box plot


Distribución del salario de argentina
en la encuesta de sysarmy
0
100000
200000
300000
400000
500000
Salario mensual BRUTO (pesos)

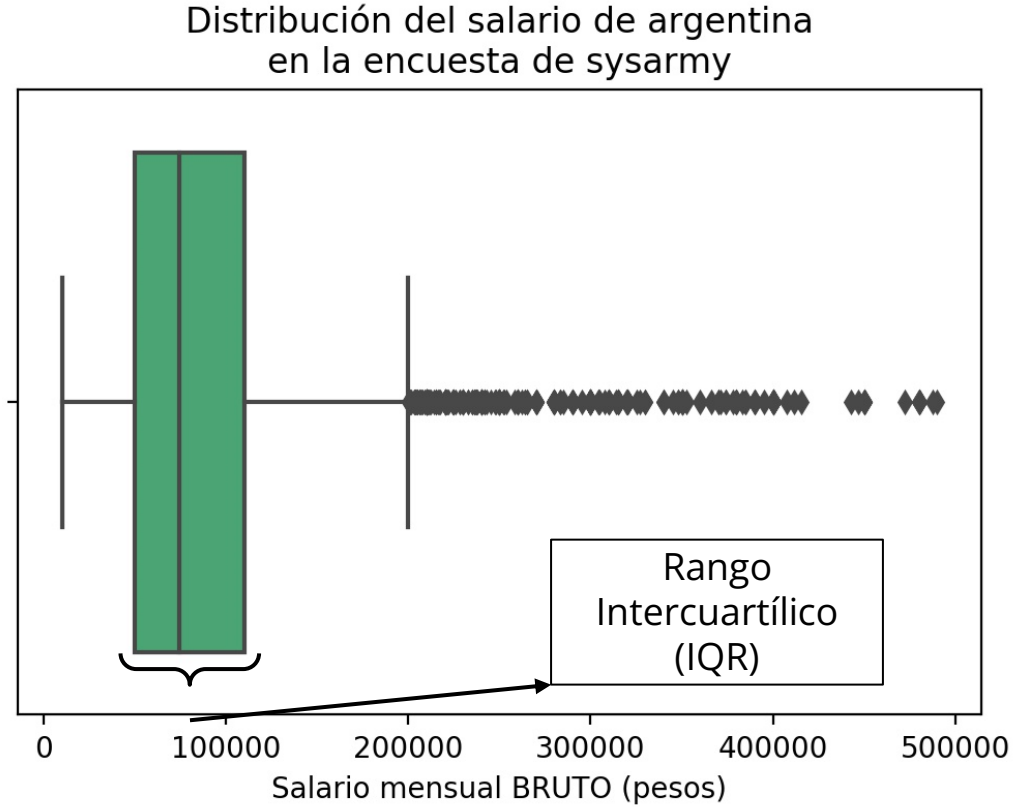
Box plot


Mediana (o Q2)
Distribución del salario de argentina
en la encuesta de sysarmy
0
100000
200000
300000
400000
500000
Salario mensual BRUTO (pesos)

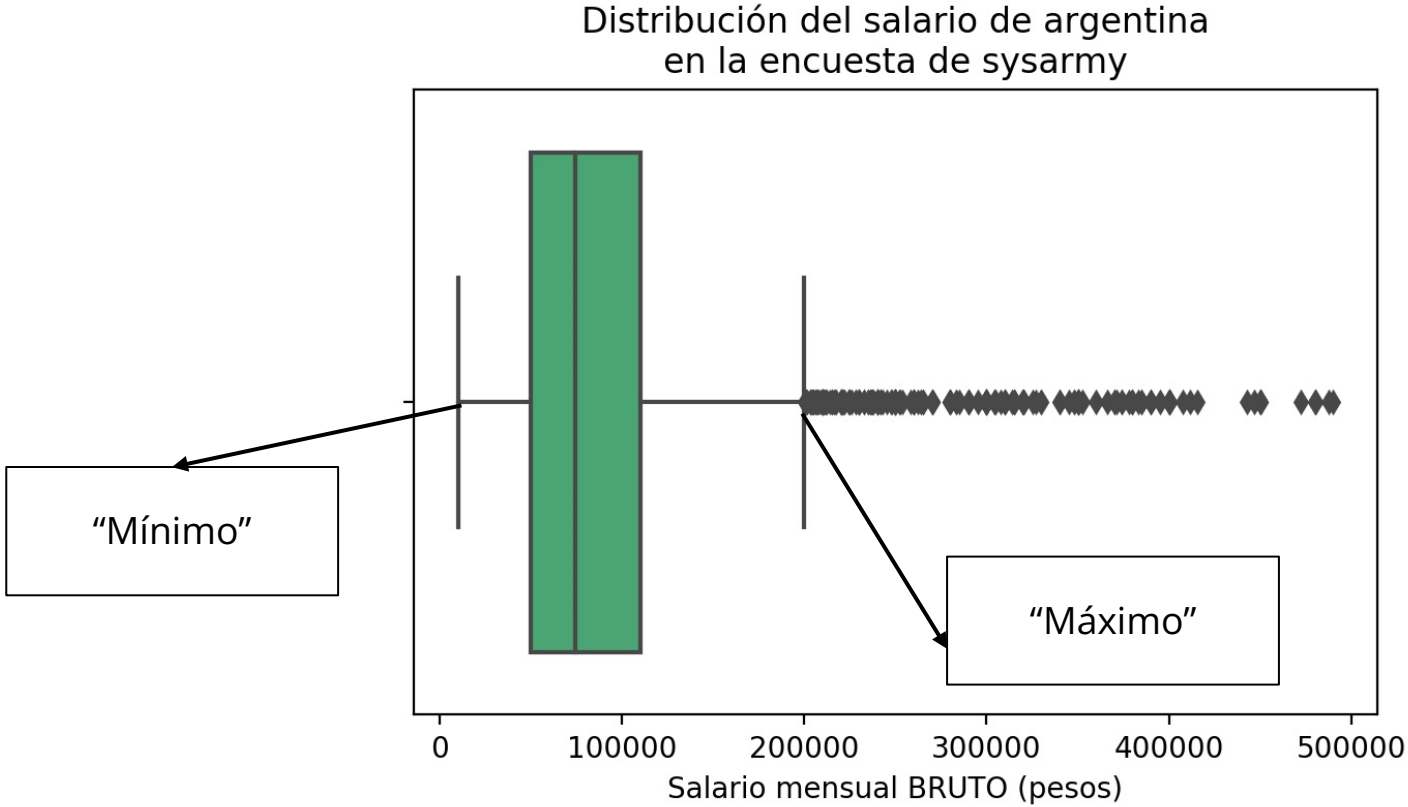
Box plot


Cuartil 1 (Q1)
Distribución del salario de argentina
en la encuesta de sysarmy
Cuartil 3 (Q3)
0
100000
200000
300000
400000
500000
Salario mensual BRUTO (pesos)

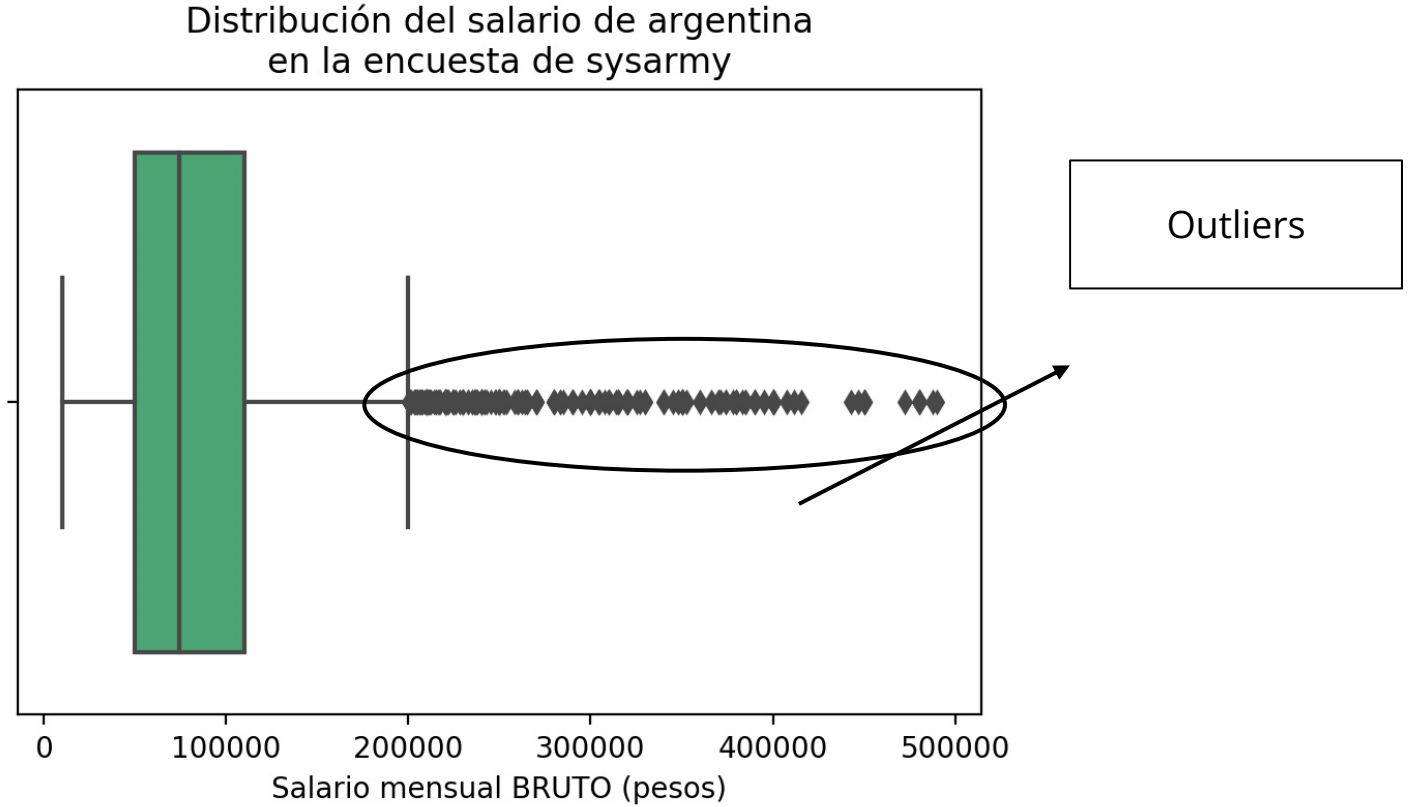
Box plot


Distribución del salario de argentina
en la encuesta de sysarmy
Rango
Intercuartílico
(IQR)
0
100000
200000
300000
400000
500000
Salario mensual BRUTO (pesos)

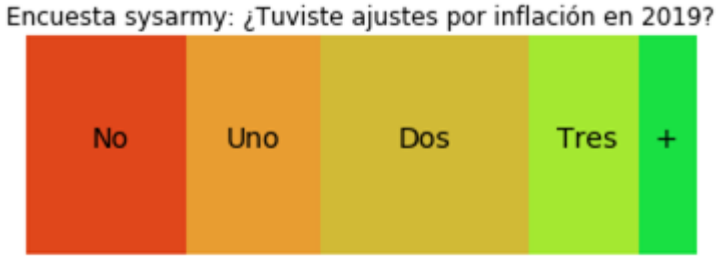
Box plot

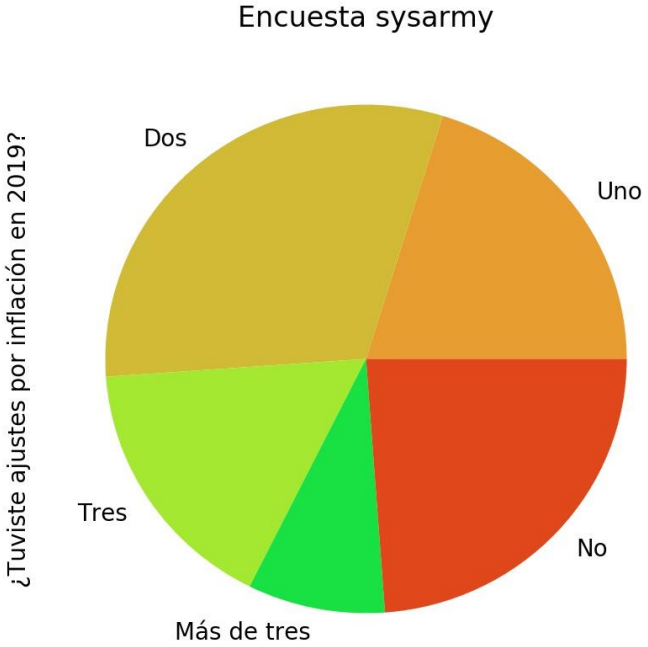

Distribución del salario de argentina
en la encuesta de sysarmy
“Mínimo”
“Máximo”
0
100000
200000
300000
400000
500000
Salario mensual BRUTO (pesos)

Box plot

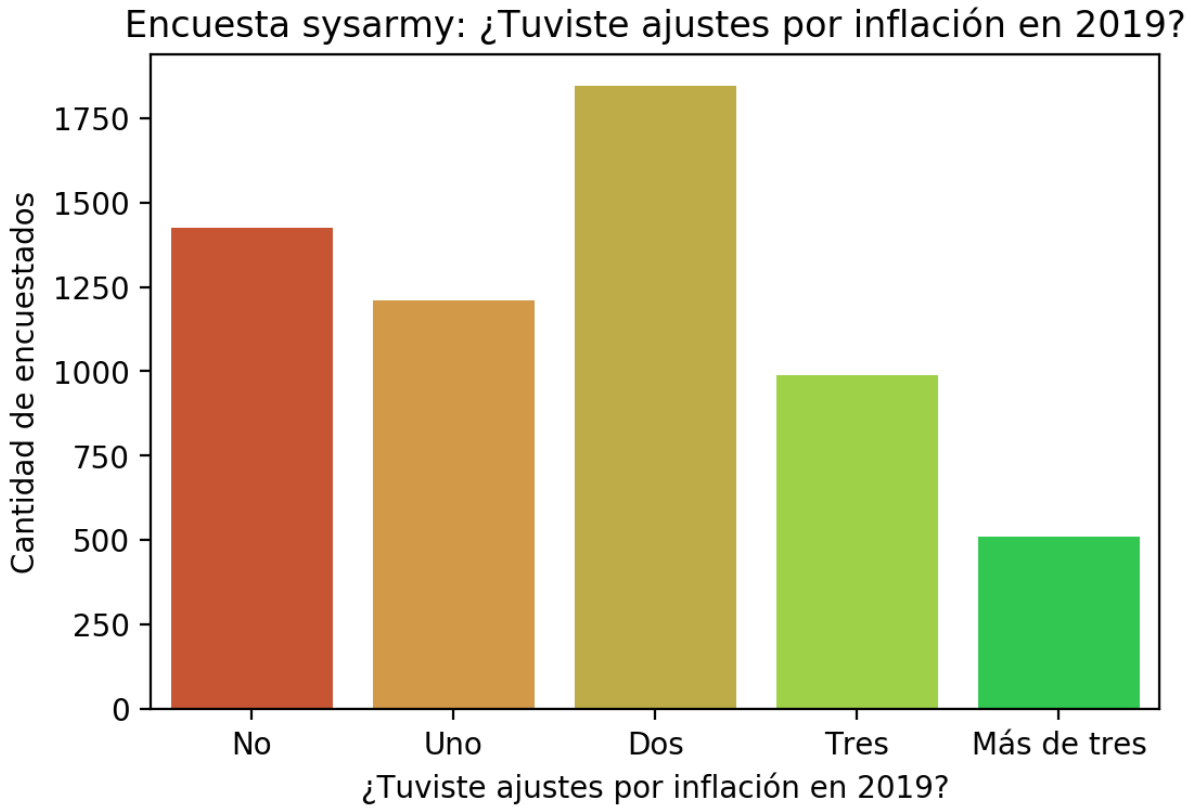

Distribución del salario de argentina
en la encuesta de sysarmy
Outliers
0
100000
200000
300000
400000
500000
Salario mensual BRUTO (pesos)

De distribución discreta

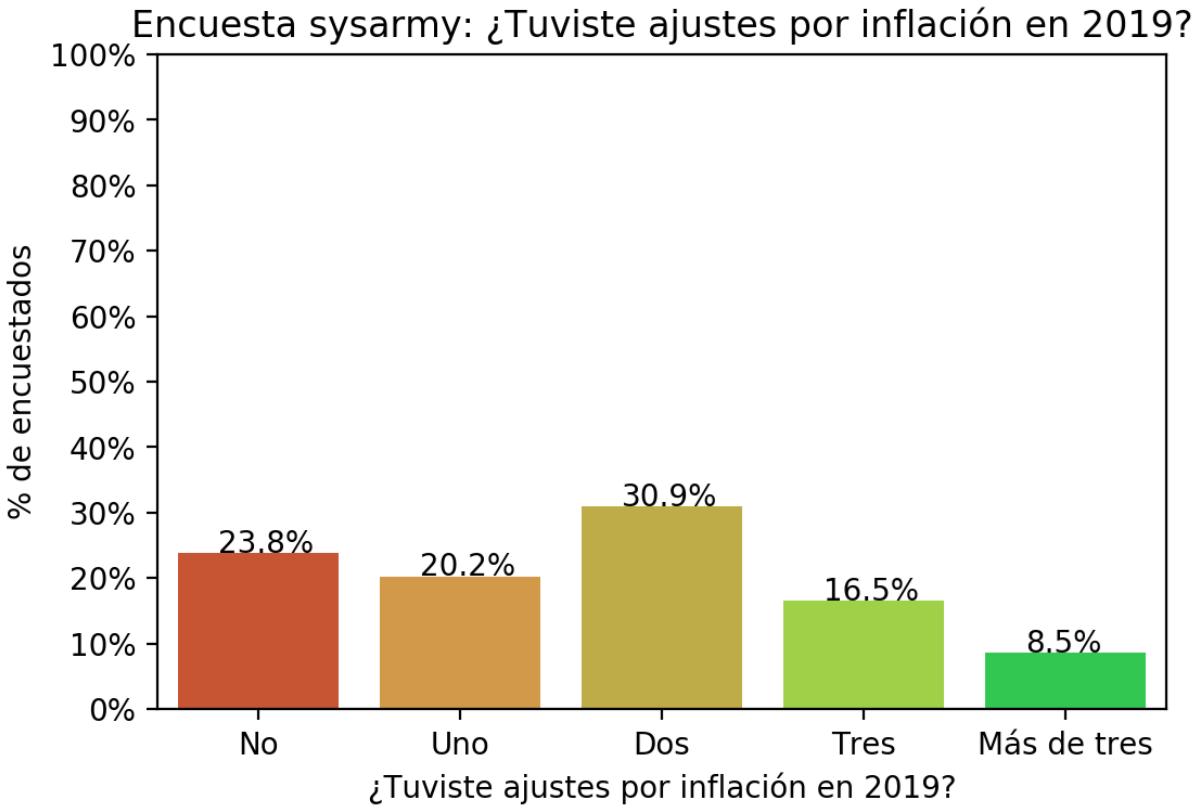




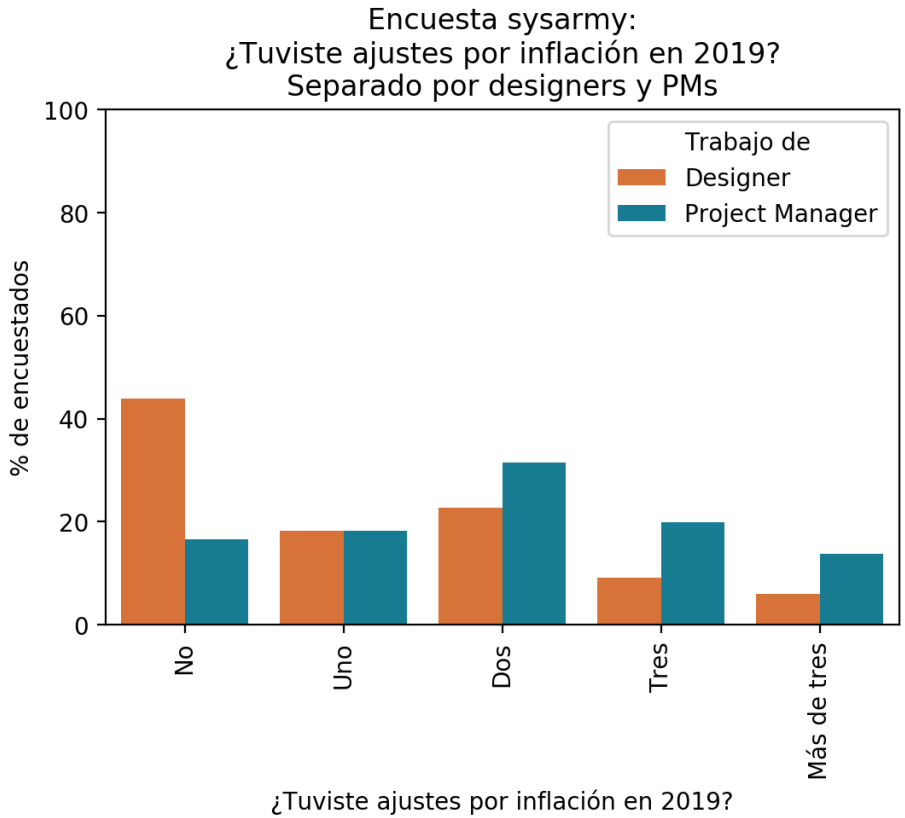
Bar plot


Encuesta sysarmy: ¿Tuviste ajustes por inflación en 2019?
Cantidad de encuestados
0
250
500
750
1000
1250
1500
1750
No
Uno
Dos
Tres
Más de tres
¿Tuviste ajustes por inflación en 2019?

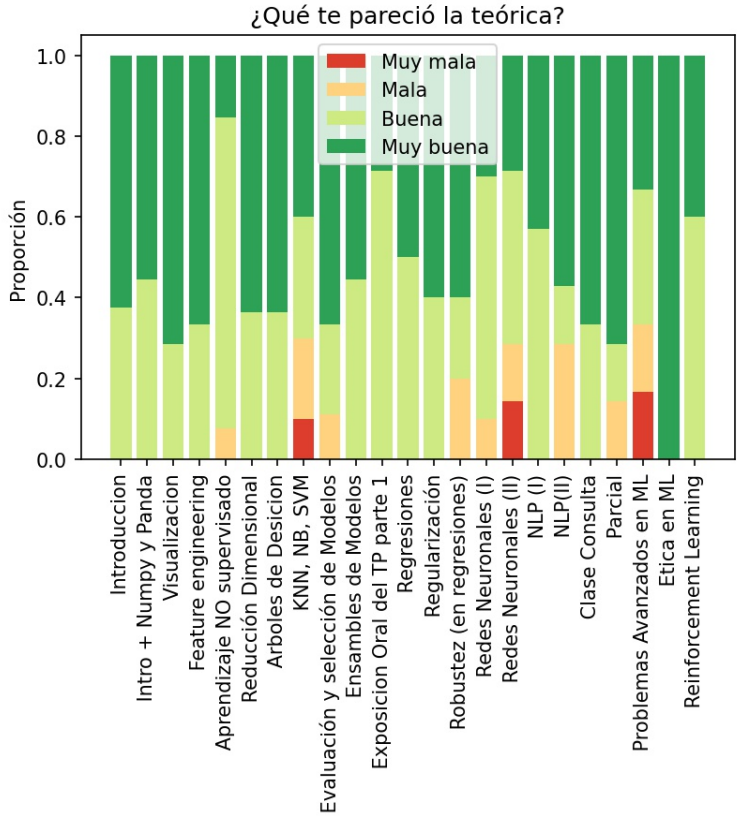
Bar plot


Encuesta sysarmy: ¿Tuviste ajustes por inflación en 2019?
% de encuestados
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
23.8%
20.2%
30.9%
16.5%
8.5%
No
Uno
Dos
Tres
Más de tres
¿Tuviste ajustes por inflación en 2019?

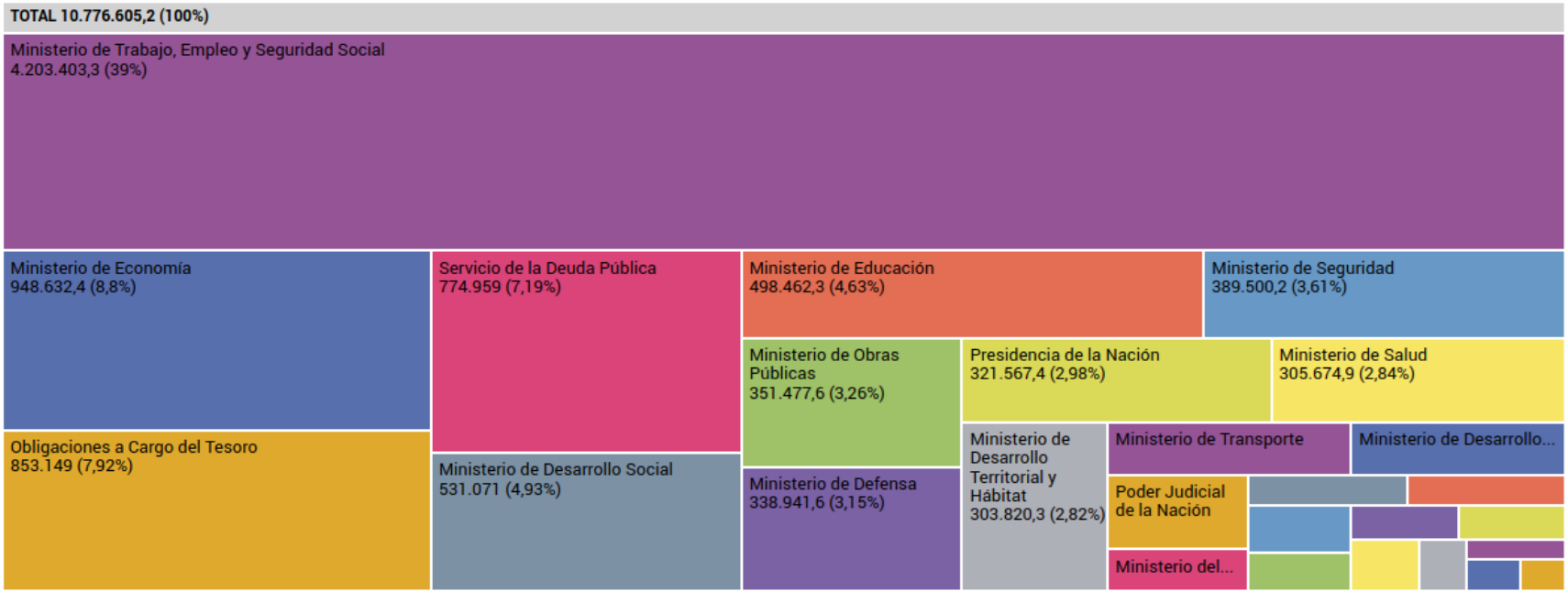
Bar plot


Encuesta sysarmy:
¿Tuviste ajustes por inflación en 2019?
Separado por designers y PMs
Trabajo de
Designer
Project Manager
% de encuestados
100
80
60
40
20
0
No
Uno
Dos
Tres
Más de tres
¿Tuviste ajustes por inflación en 2019?

Stacked bar plot



Treemap

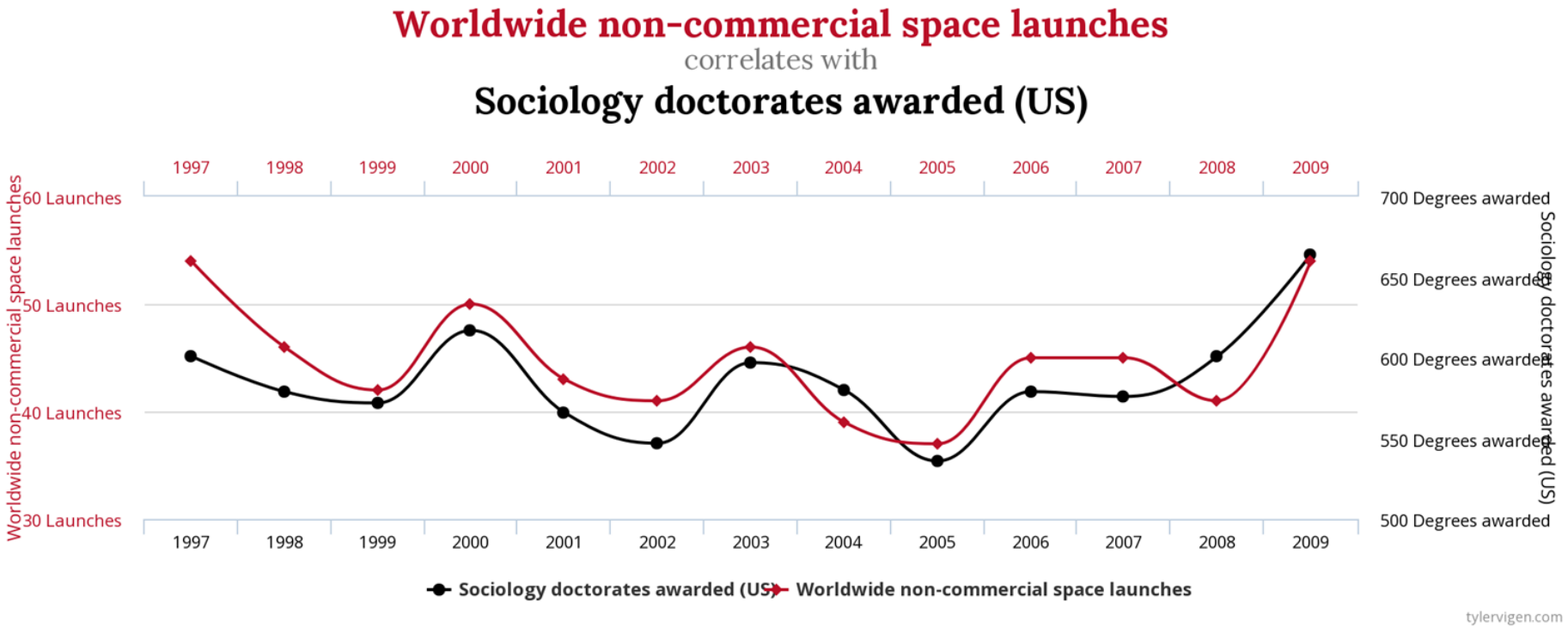


Los importes están expresados en millones de pesos. Sin Aplicaciones ni Fuentes Financieras ni Contribuciones y Gastos Figurativos. | Fuente: eSidif.

Última actualización del ejercicio 2022: 04 de Marzo de 2022

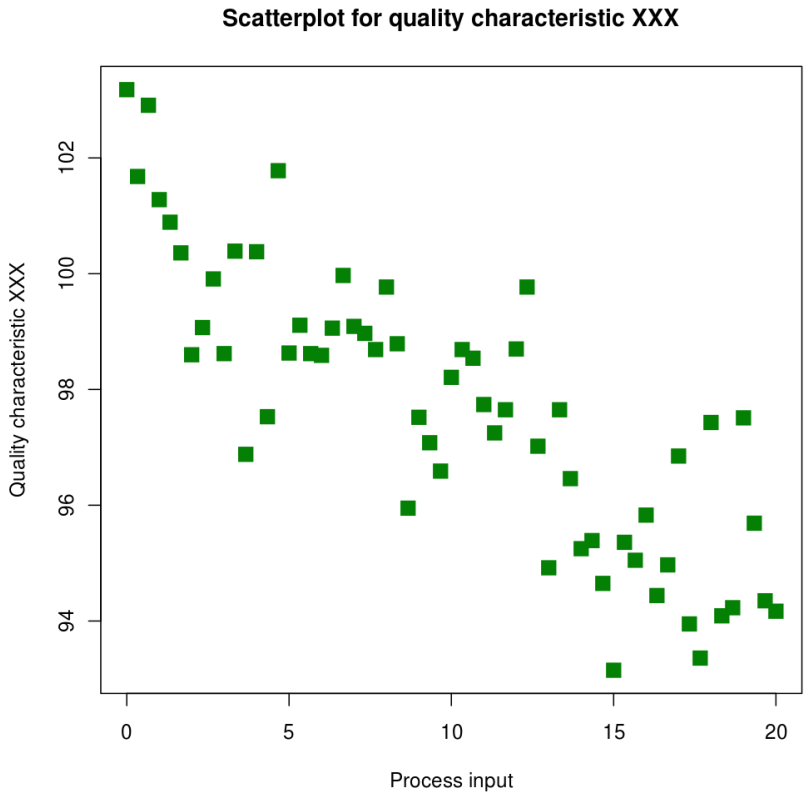
https://www.presupuestoabierto.gob.ar/sici/destacado-quien-gasta

De relación



https://www.tylervigen.com/spurious-correlations

Scatter plot (de dispersión)



Utiliza las coordenadas cartesianas para mostrar los valores de dos variables

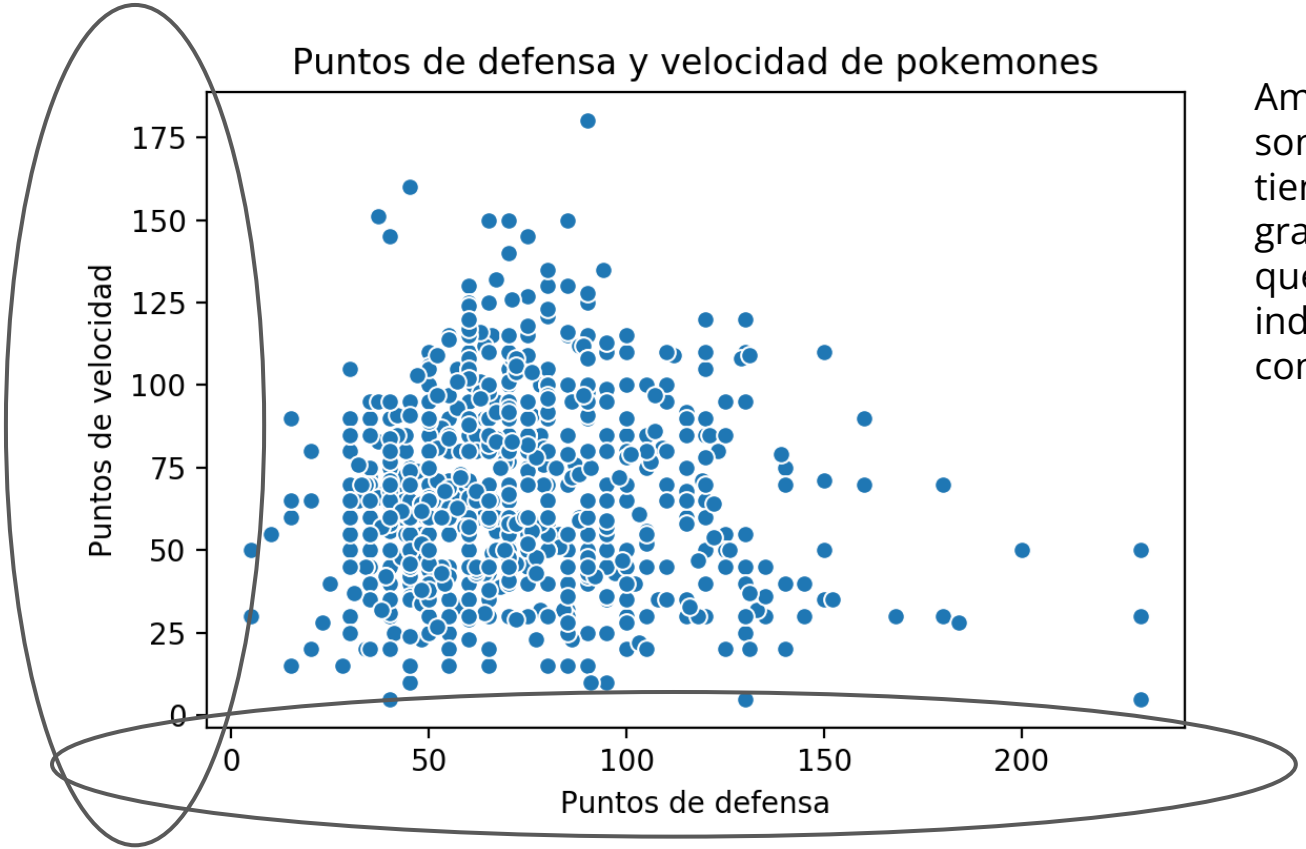
Correlación de Pearson

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

Para 2 variables podemos medir su correlación lineal con el coeficiente de correlación r (Pearson). Este coeficiente, es una función que mide cuán relacionada estan 2 variables de forma lineal.

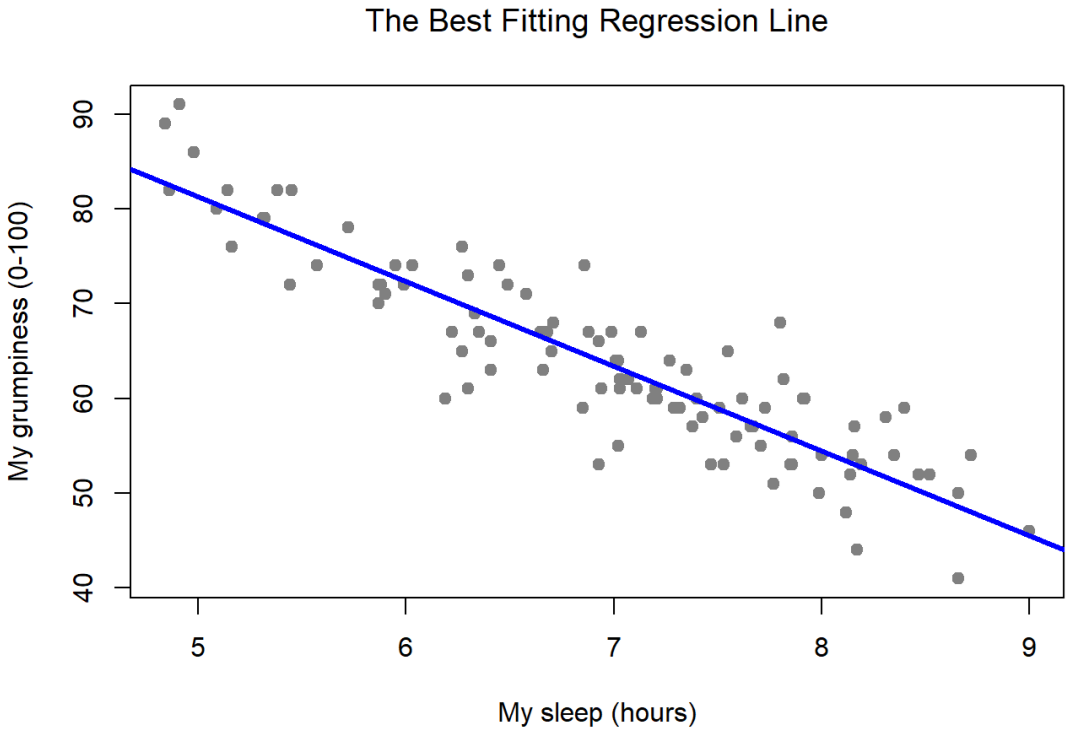
- Si da 0 NO existe correlación
- Si da 1 Están relacionadas linealmente de forma perfecta (todos los puntos están en una línea)
- Si da -1 Existe una correlación negativa perfecta.

Scatter plot (de dispersión)



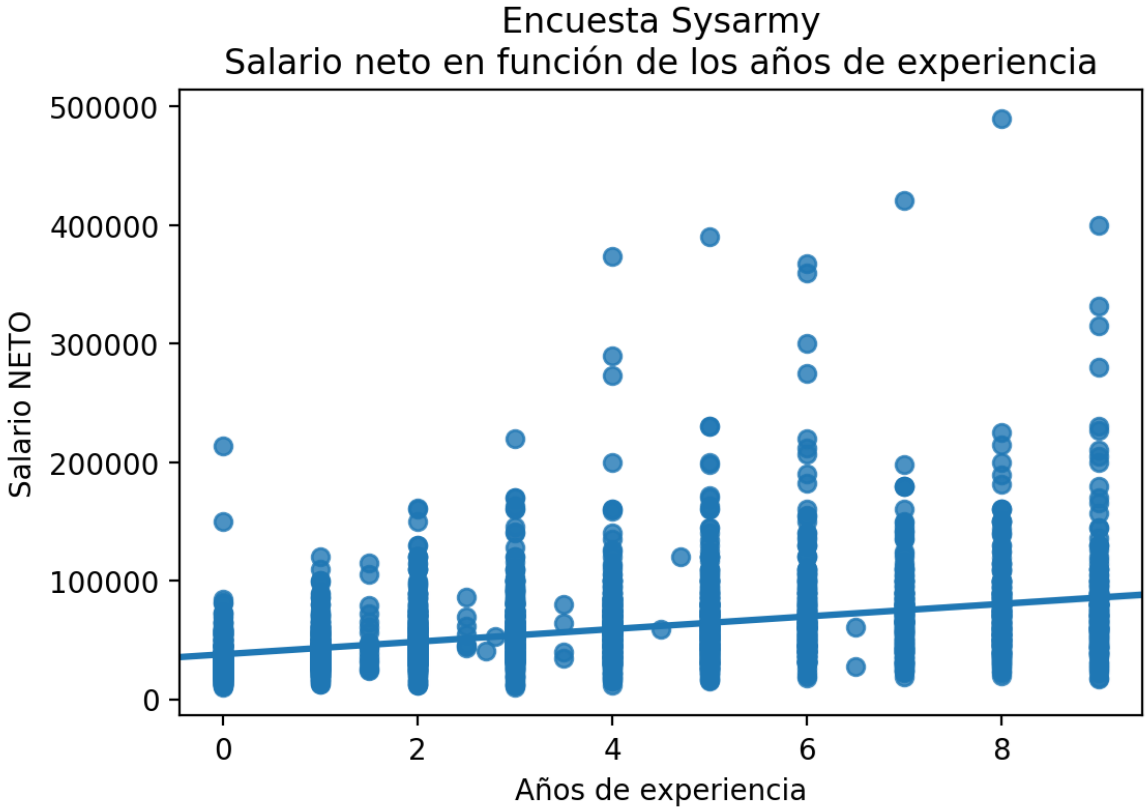
Ambos ejes o bien son continuos o tienen una buena granularidad (lo que los vuelve indistintos del continuo)

Regression plot

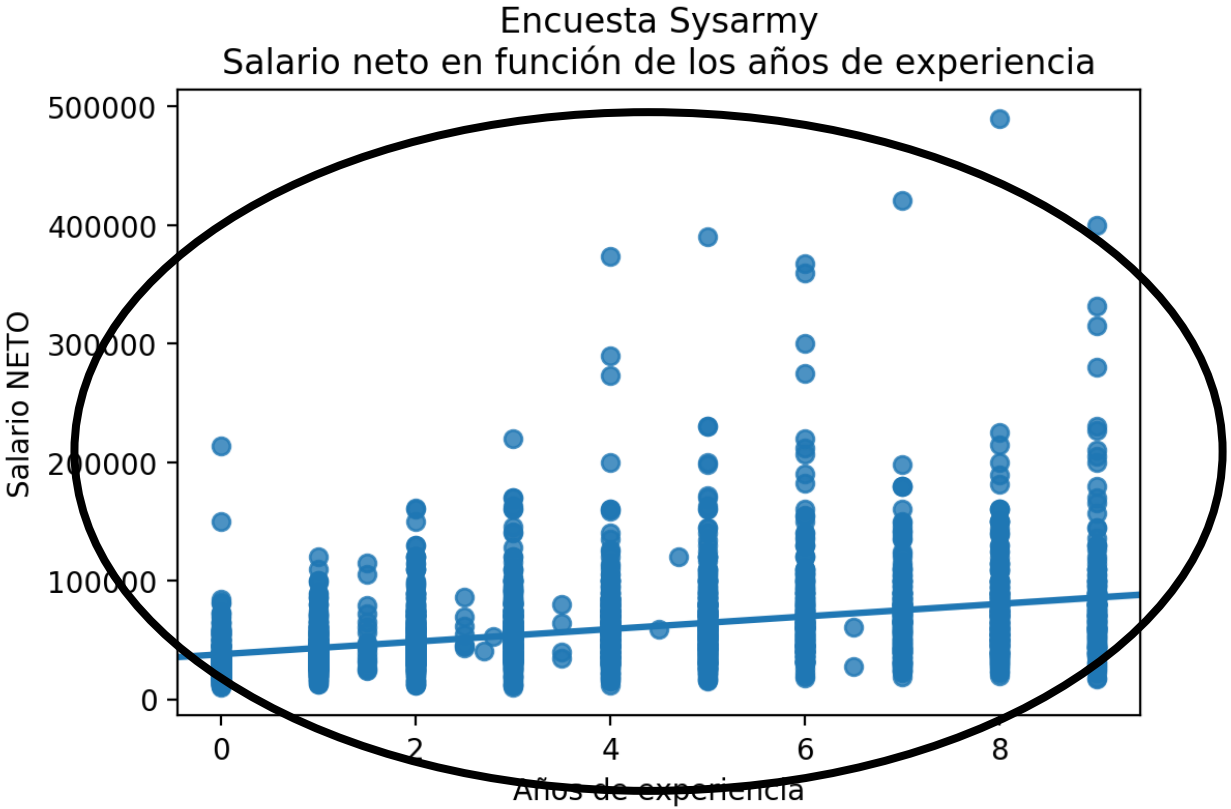


Se incluye una guía visual que muestra la relación entre las variables

Ejemplo confuso



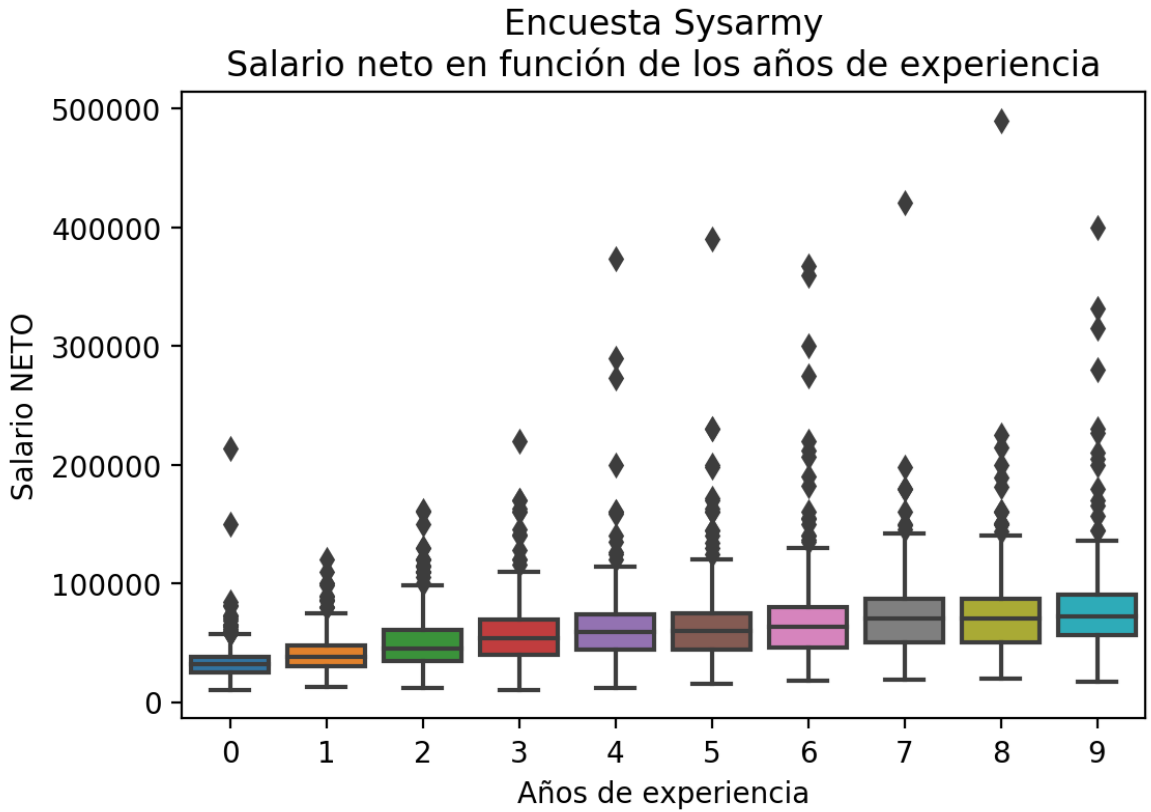
Ejemplo confuso



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.

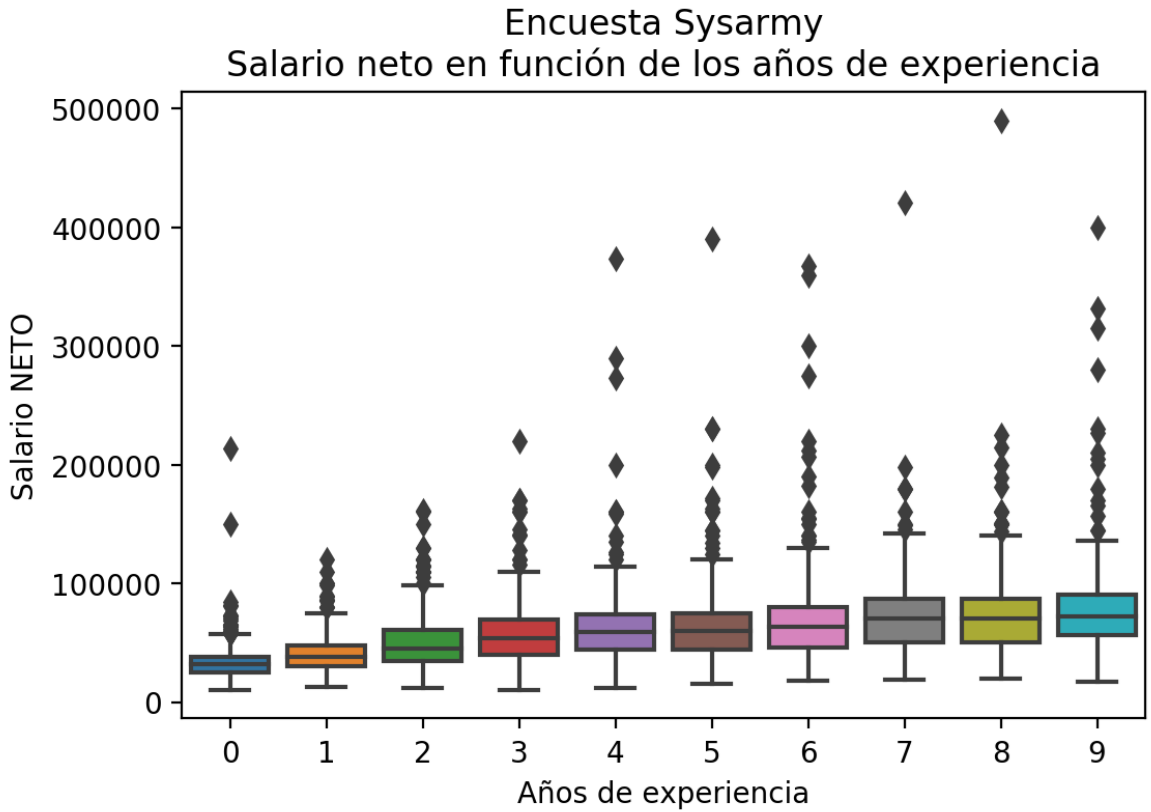
Podemos pensarlo como comparar distribuciones continuas.

Una mejora

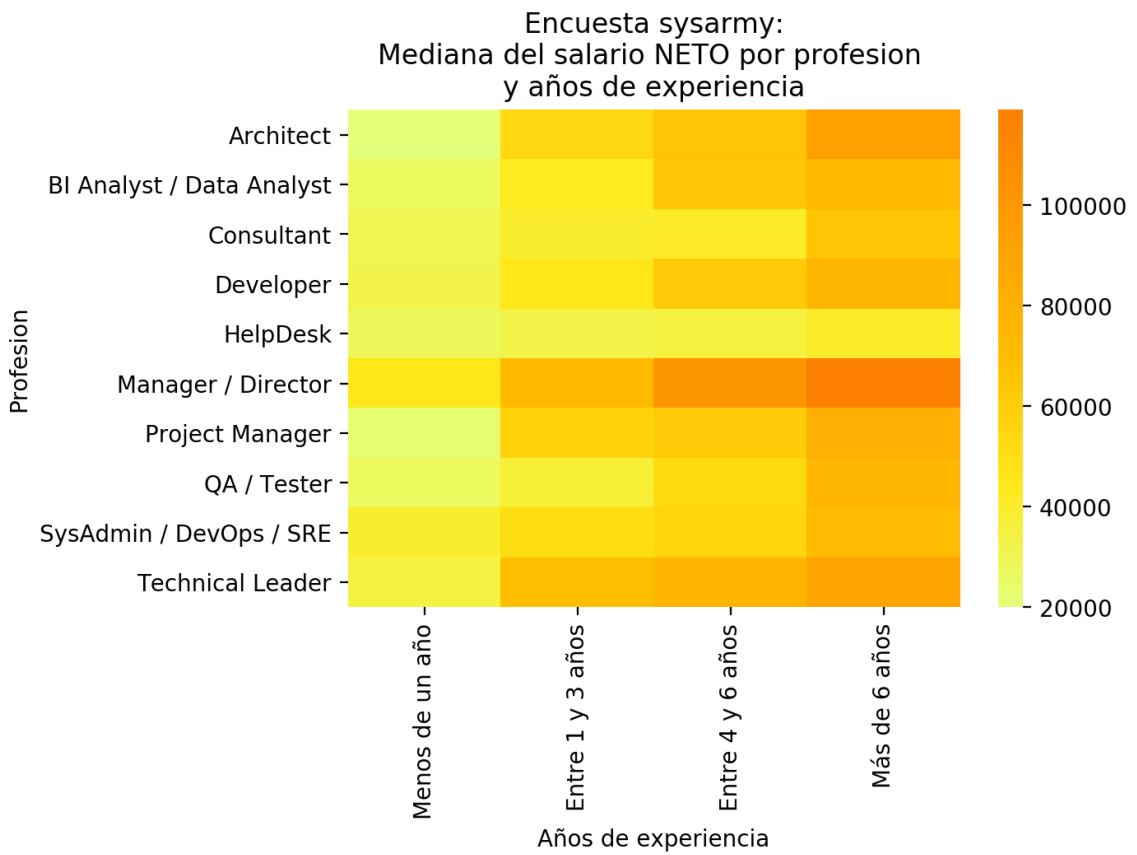

Encuesta Sysarmy
Salario neto en función de los años de experiencia
Salario NETO
500000
400000
300000
200000
100000
0
0
1
2
3
4
5
6
7
8
9
Años de experiencia

Una mejora

Y si queremos también verlo por profesión?



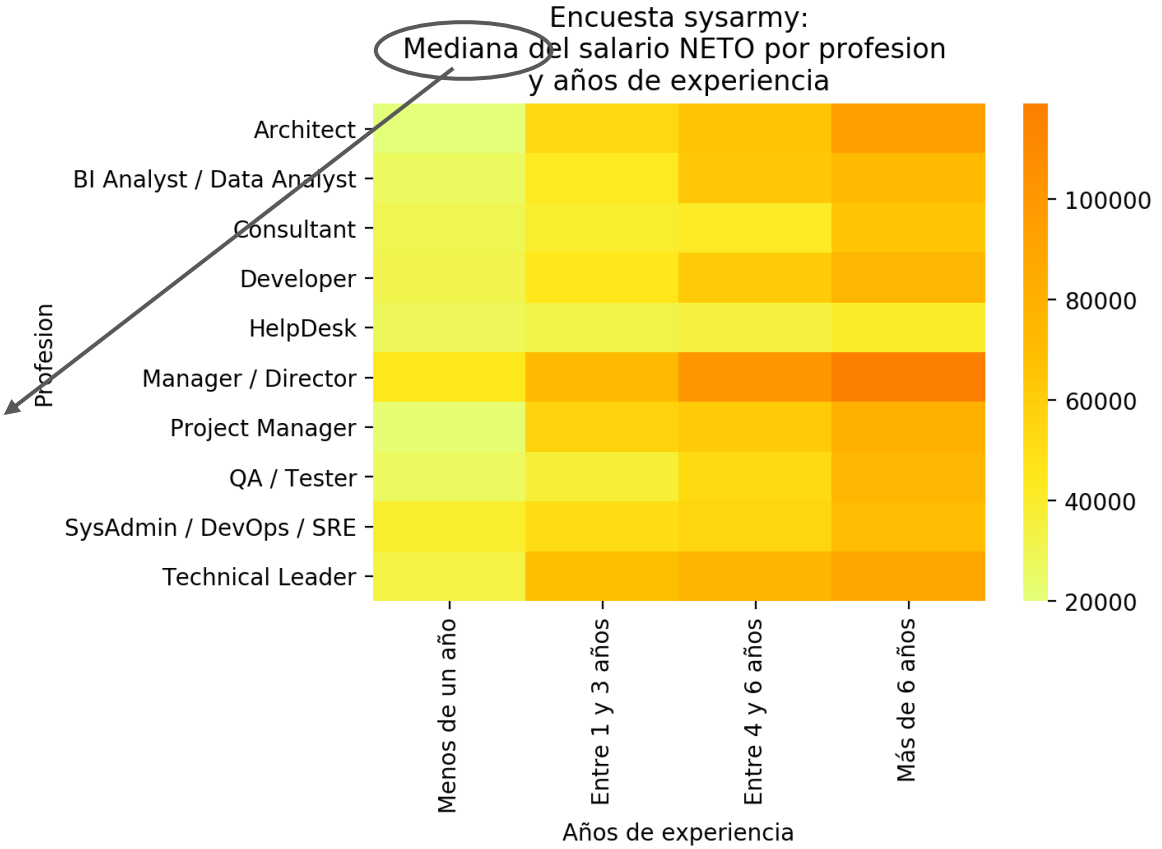
Heatmap


Encuesta sysarmy:
Mediana del salario NETO por profesion
y años de experiencia
Profesion
Architect
BI Analyst / Data Analyst
Consultant
Developer
HelpDesk
Manager / Director
Project Manager
QA / Tester
SysAdmin / DevOps / SRE
Technical Leader
Menos de un año
Entre 1 y 3 años
Entre 4 y 6 años
Más de 6 años
Años de experiencia
100000
80000
60000
40000
20000

Heatmap

Sirve para comprar 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.



Series de tiempo

Lineplot

cuarentena
Término de búsqueda

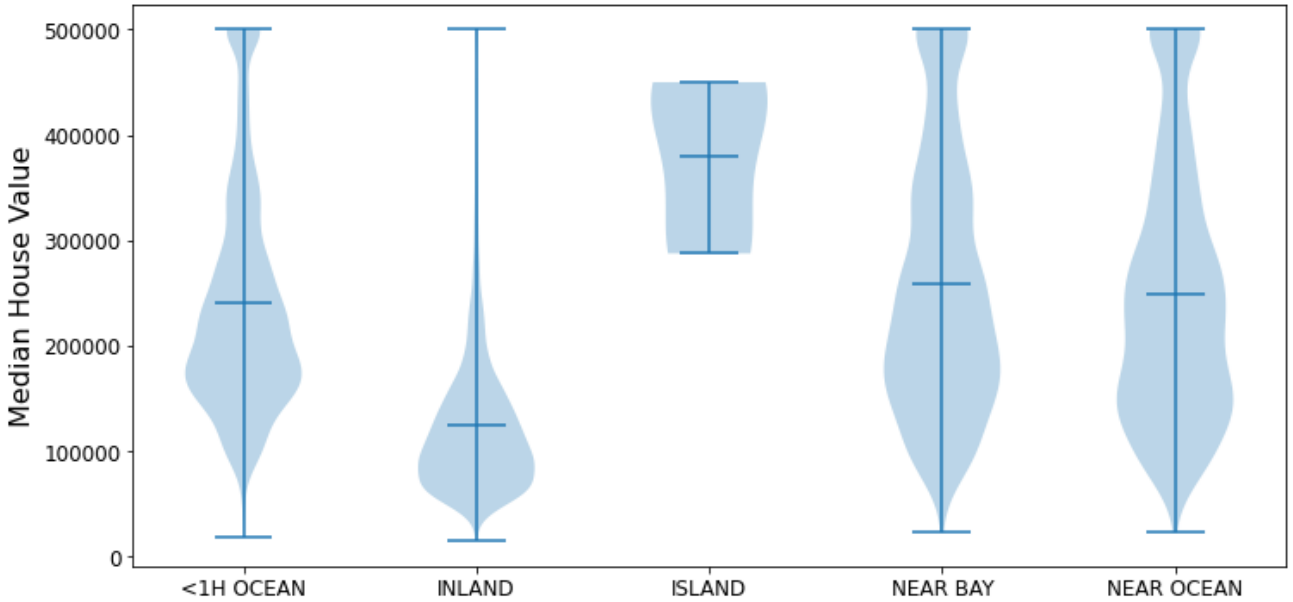
+ Comparar

Argentina | 1/3/20 - 26/8/20 | Todas las categorías | Búsqueda web

Interés a lo largo del tiempo

https://trends.google.com.ar/trends/explore?date=2020-03-01%202020-08-26&geo=AR&q=cuarentena

Violin 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.

Otros plots

https://python-graph-gallery.com/

https://datavizproject.com/

Más material

- “Everything we know about how humans interpret graphics”, Kennedy Elliot. https://www.youtube.com/watch?v=s0J6EDvlN30

- Tufte, Edward R., 1942-. (2001). The visual display of quantitative information.