

ACÀMICA

¡Bienvenidos/as a Data Science!



Agenda

¿Cómo anduvieron?

Repaso: Glosario Estadístico

Explicación: Outliers

Hands-on training

Break

¿Sabías que...?

Actividad: Explorando mis datos

Cierre



¿Cómo anduvieron?



Proyecto 1: Análisis Exploratorio de Datos (EDA)



Análisis Exploratorio de Datos (EDA)

fase	ADQUISICIÓN Y EXPLORACIÓN		MODELADO				DEPLOY
	Exploración de datos	Feature Engineering	Regresión	Optimización de parámetros	Procesam. del lenguaje natural	Sistema de recomendación	Publicación de modelos
tiempo	SEM 1	SEM 5	SEM 8	SEM 11	SEM 13	SEM 18	SEM 22
	SEM 2	SEM 6	SEM 9	SEM 12	SEM 14	SEM 19	SEM 23
	SEM 3	SEM 7	SEM 10		SEM 15	SEM 20	SEM 24
	SEM 4				SEM 16	SEM 21	
					SEM 17		

Proyecto EDA: Hoja de ruta

Entrega 1

SEM 1 - 4

- Python
- Numpy
- Pandas
- Visualización de datos: Matplotlib y Seaborn
- Estadística

SEM 5

- Ejercicio/s Integradores
- Programación: Funciones
- Transformación de Datos
- Transformación de Datos con Pandas

Entrega 2

SEM 6

- Programación: Clases
- Cómo dar Presentaciones Orales
- Transformación de Datos con Scikit-Learn

Usted
está aquí

SEM 7

- Transformación de Datos con Scikit-Learn
- Outliers
- Puesta en Común Proyecto 1

SEM 8

- ¡Arrancamos con Machine Learning!

Repaso: Glosario Estadístico



Repaso: Glosario

Con dos compañeros/as dar una (breve) definición y un ejemplo de los siguientes conceptos estadísticos:

Distribución

Mediana

Percentil

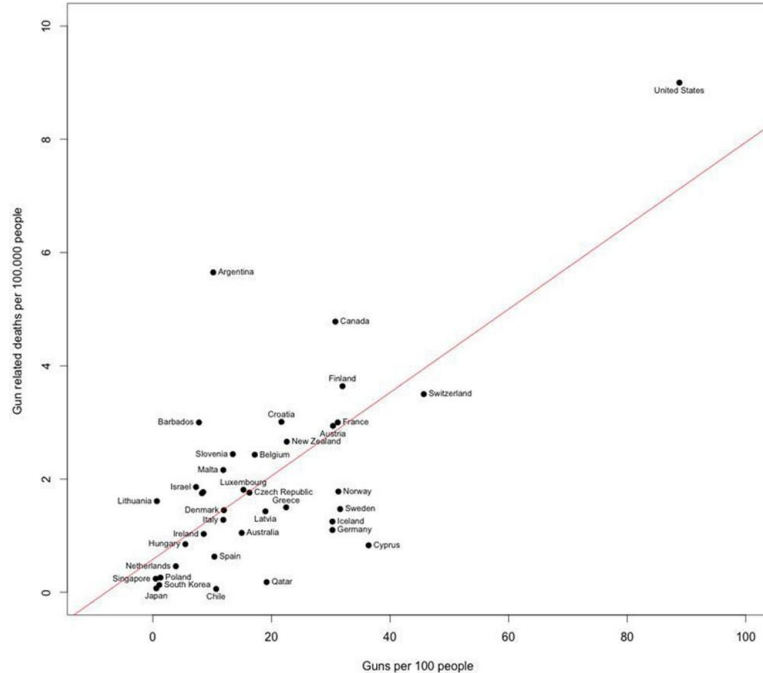
Cuartil

Asimetría estadística (skewness)

Detección de Outliers

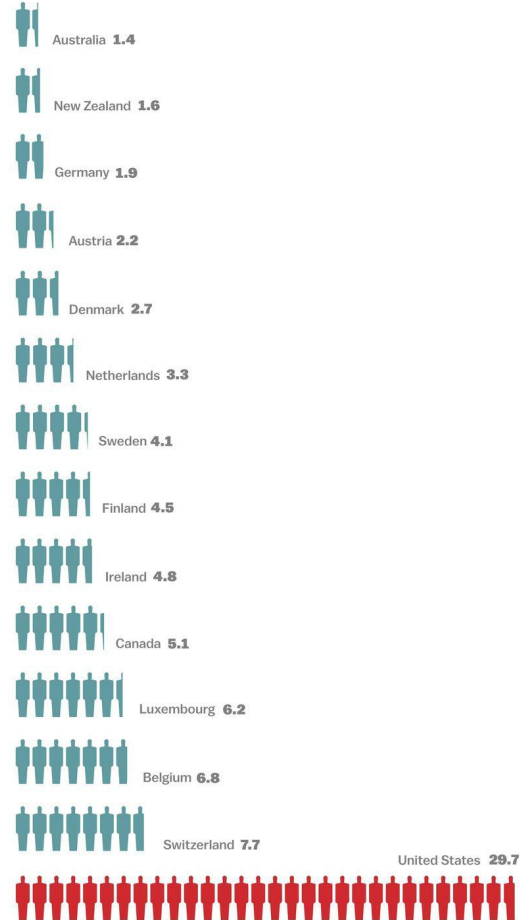


America's unique gun violence problem, explained in 16 maps and charts



Homicides by firearm per 1 million people

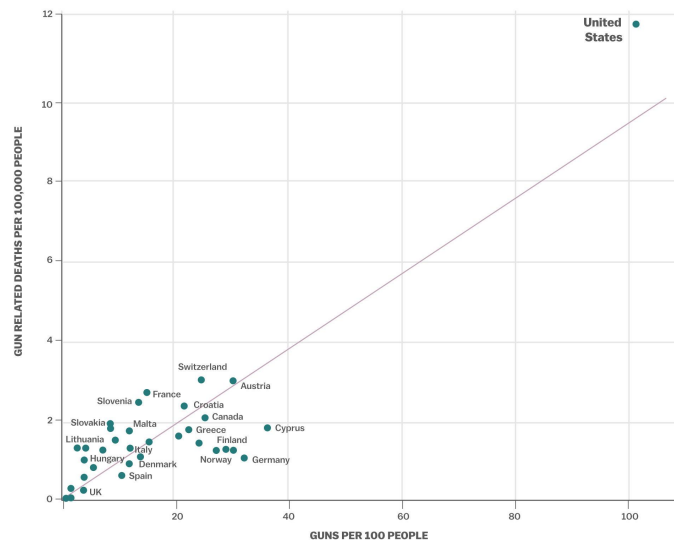
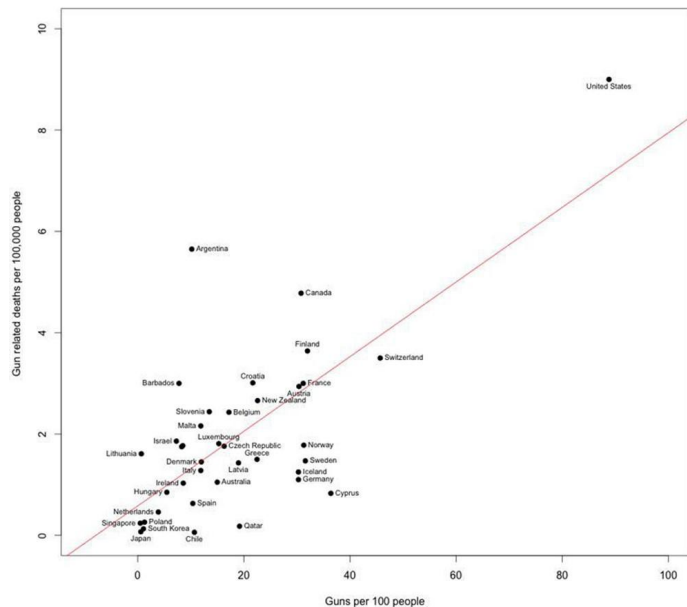
In advanced countries according to the Human Development Index. Numbers are for 2012.



SOURCE: UNODC, Small Arms Survey, via The Guardian.

Nota de Color

La figura original (izquierda) fue reemplazada por la de la derecha. Notar que no aparece Argentina.



SOURCE: Gunpolicy.org , United Nations Development Programme

Vox

Si les intriga saber por qué, acá, acá y acá tal vez encuentren algunas pistas y unas discusiones interesantes.

¿Qué es un
Outlier?
¿Por qué
ocurren?

OUTLIER = valor atípico que difiere
significativamente del resto de las
observaciones.

¿Por qué difiere?

- Error de medición del instrumento.
- Error al introducir un dato.
- Estamos trabajando con muestras/poblaciones que no son tan homogéneas como creíamos.



¿Qué es un
Outlier?
¿Por qué
ocurren?

¡Muchas veces los **OUTLIERS** son una manifestación del proceso que estamos estudiando!

Ejemplos:

- Transacción fraudulenta con una tarjeta de crédito.
- Persona enferma en un conjunto de personas sanas.
- Mayor incidencia de una enfermedad en una ciudad.
¿Esperable o outlier?





SUGERENCIA

Siempre es importante pensar por qué hay un outlier en nuestro dataset

Tipos de valores atípicos

univariante

Se desvía de los valores típicos de un feature (columna)

multivariante

Se desvía de los valores típicos que hay en la relación entre dos o más columnas

Los valores atípicos suelen *confundir* la estadística que hacemos sobre los datos, ya que nos indican que no estamos trabajando con poblaciones homogéneas.

A veces, **detectar outliers** es el objetivo de nuestro estudio.

¿Se les ocurre algún ejemplo?

¿Cómo detectar outliers?

Muchas veces no existe una manera *obvia* de detectar outliers, y, en general, ¡depende del problema!

Algunas técnicas son



- Visualización: Boxplots
- Por rango intercuartílico (Interquartile Range)
- Regla de las tres sigmas
- ¡Y más!

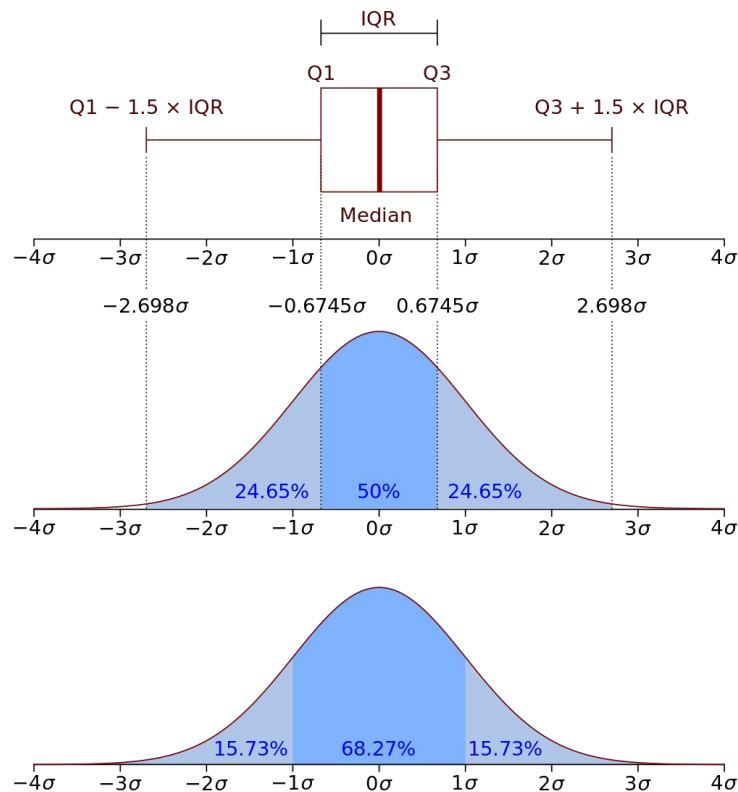
Boxplot

Rango intercuartílico

Regla de las tres sigmas

El **diagrama de cajas** es una forma de visualizar un conjunto de valores.

Muchas veces resulta más **informativo** que simplemente dibujar un punto por cada valor, ya que nos permite tener una idea de como es la distribución subyacente.

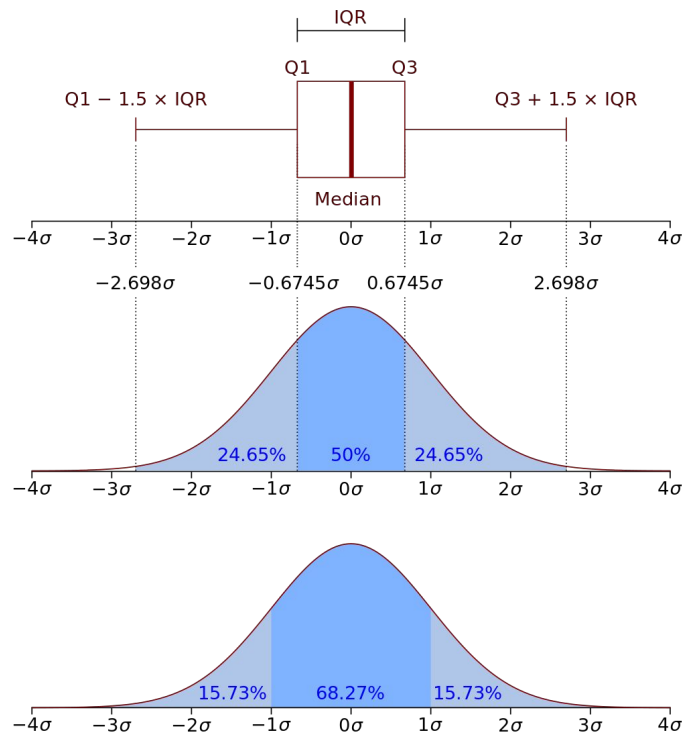


Elegimos un valor mínimo y un valor máximo para los valores “permitidos”.

Marcamos como outliers aquellos valores que estén por debajo del mínimo o por arriba del máximo.

¿Cómo elegimos el mínimo y el máximo?

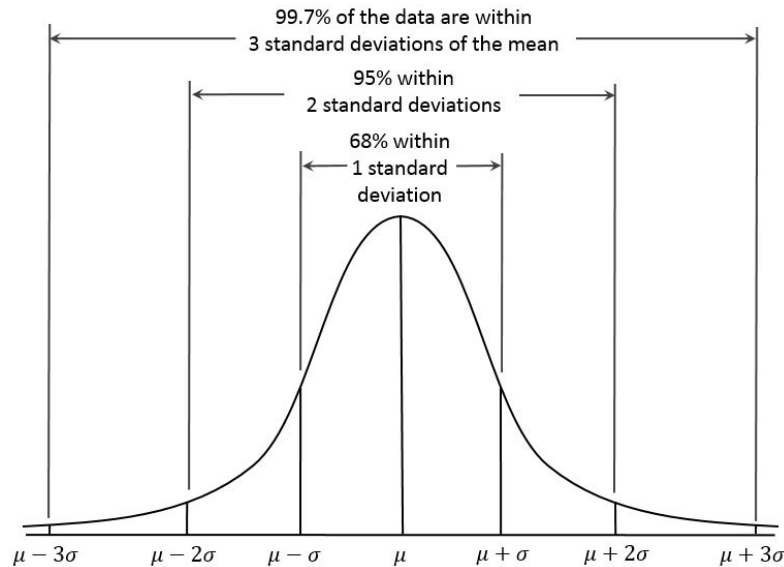
- A veces, es la variable la que nos lo indica. Por ejemplo, la asistencia a un curso no puede ser menor que cero o mayor al número de alumnos que tiene el curso.
- Un criterio estandarizado es usar
mínimo = $Q1 - 1.5 \times IQR$
máximo = $Q3 + 1.5 \times IQR$



¿Y si en lugar de usar los cuartiles
usamos las desviaciones estándar?

mínimo = valor medio - 3 x SD

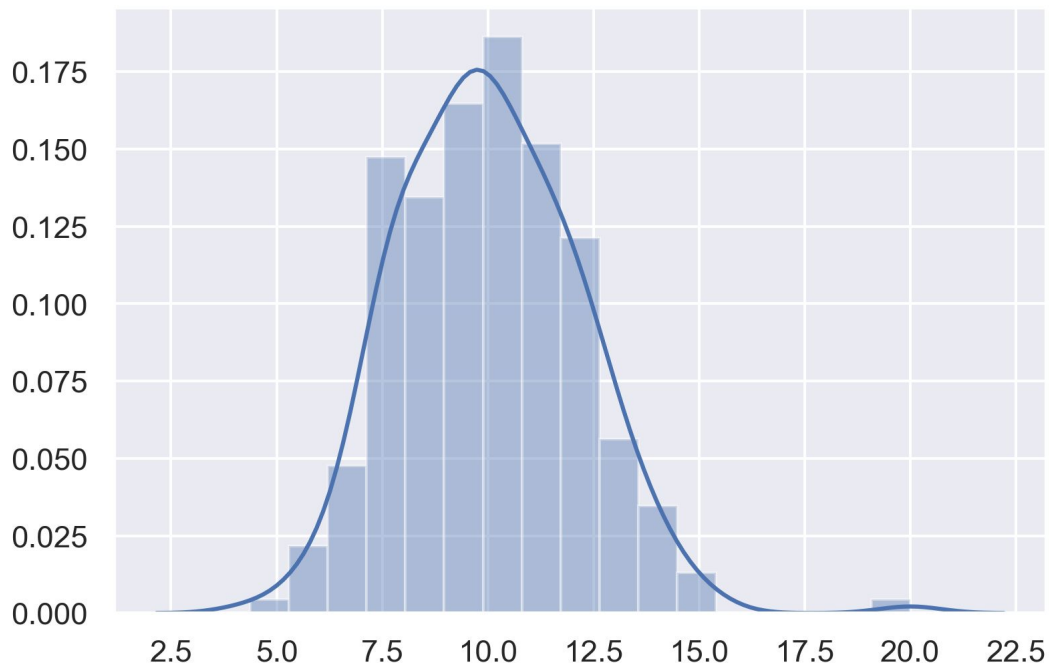
máximo = valor medio + 3 x SD



Boxplot

Rango intercuartílico

Regla de las tres sigmas

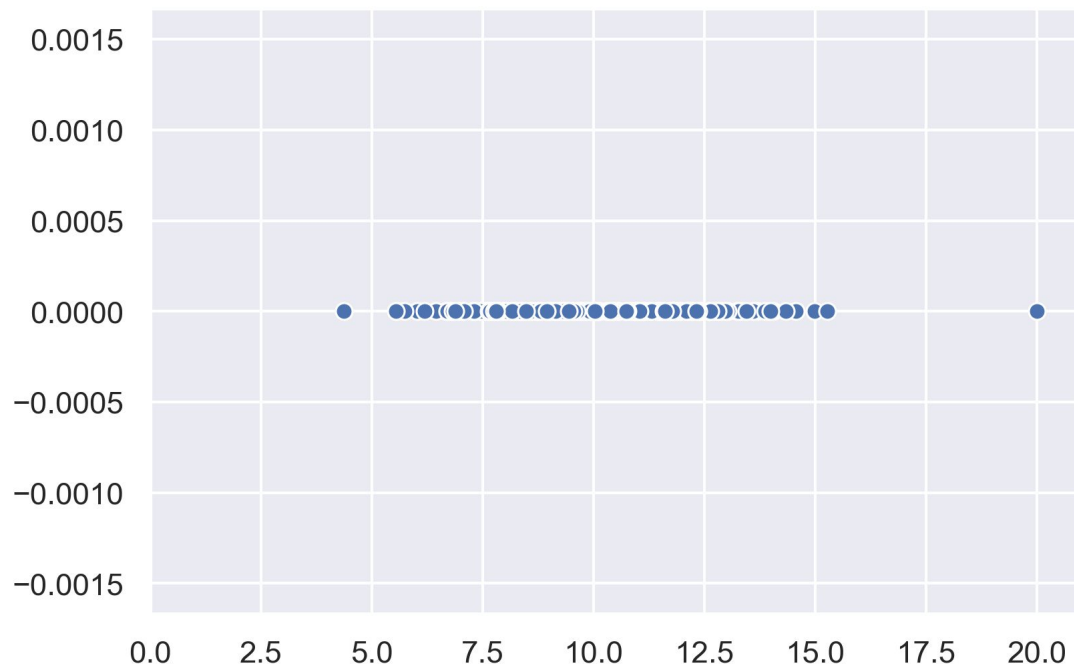


Todos los valores que estén a distancia mayor a tres desviaciones estándar de la media son seleccionados como outliers.

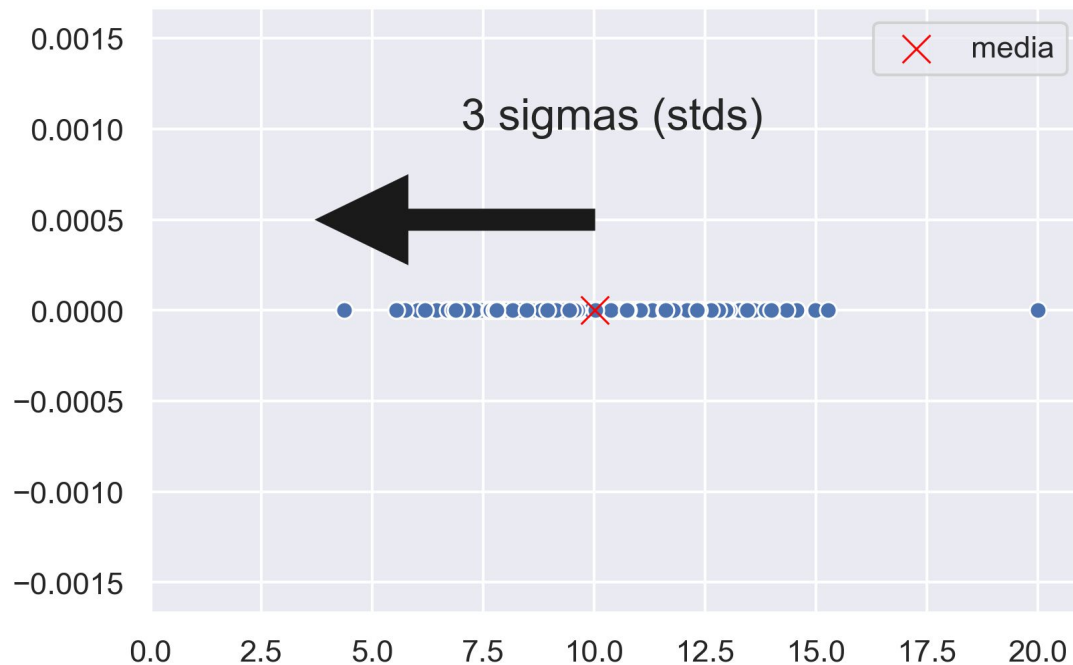
Boxplot

Rango intercuartílico

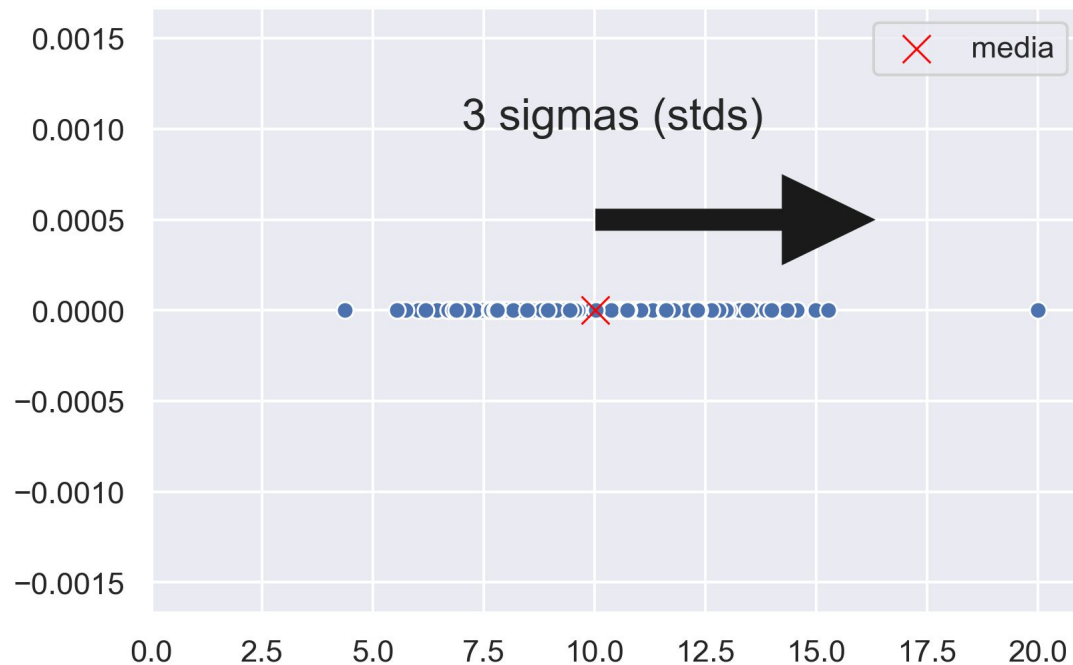
Regla de las tres sigmas



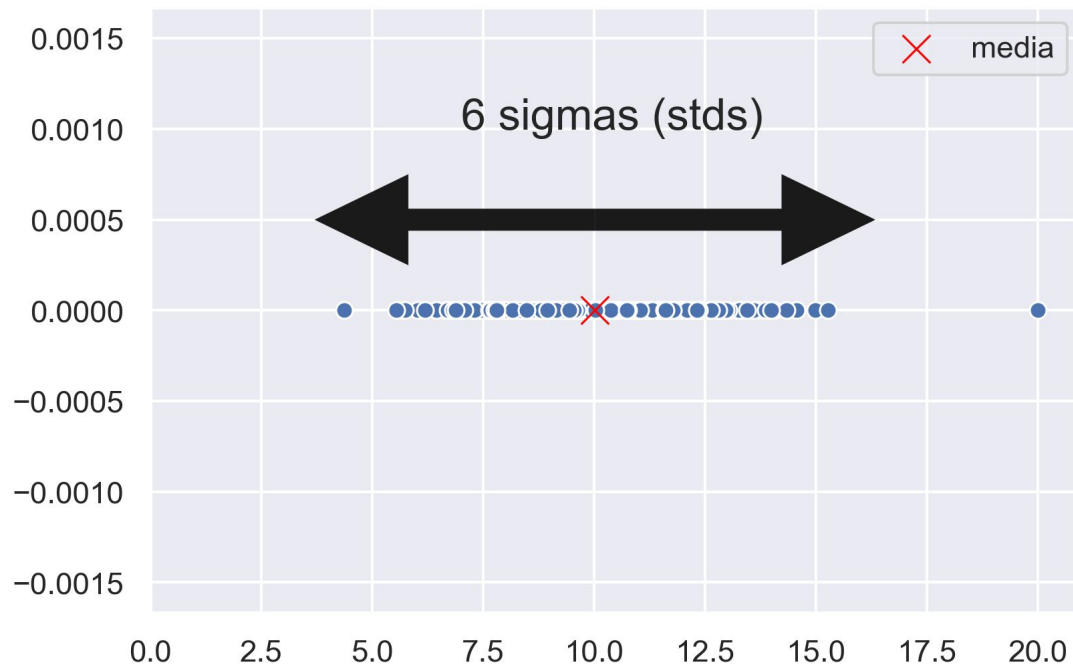
Todos los valores que estén a distancia mayor a tres desviaciones estándar de la media son seleccionados como outliers.



Todos los valores que estén a distancia mayor a tres desviaciones estándar de la media son seleccionados como outliers.



Todos los valores que estén a distancia mayor a tres desviaciones estándar de la media son seleccionados como outliers.



Todos los valores que estén a distancia mayor a tres desviaciones estándar de la media son seleccionados como outliers.

**A veces, este método se
aplica a través del Z-Score**

Boxplot

Rango intercuartílico

Regla de las tres sigmas



método Z-Score

Tenemos un conjunto de números $x_1, x_2, x_3, \dots, x_n$. Su media es μ , y su desviación estándar σ .

$$Z = (x_i - \mu) / \sigma$$

Es una medida de cuánto se desvía un valor del promedio, medido en desviaciones estándar.

Ejemplo: $x_1 = 1, x_2 = 2, x_3 = 1.5$

- Media, $\mu = 1.5$
- Desviación estándar, $\sigma = 0.5$

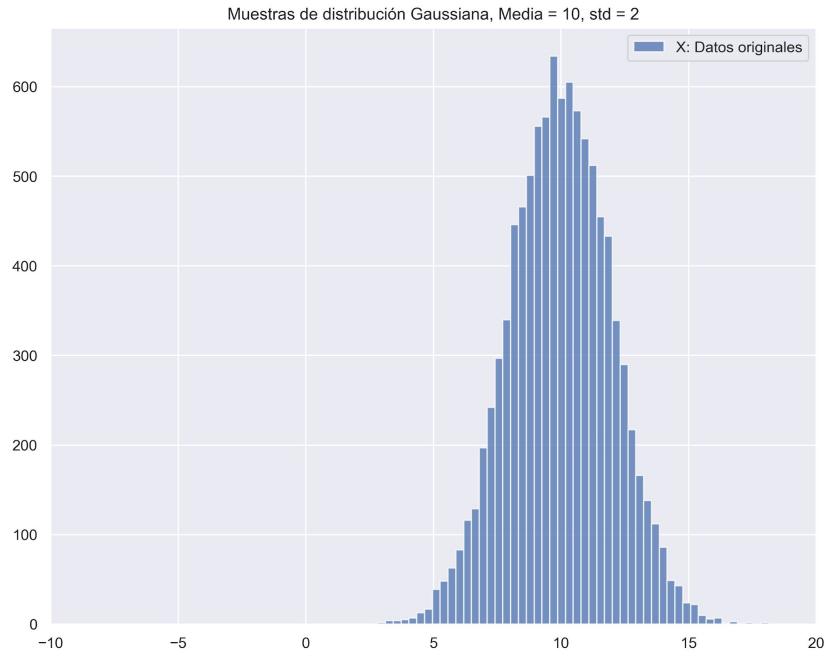
$x_1 = 1$	→	$z_1 = (1 - 1.5) / 0.5 = -1$
$x_2 = 2$	→	$z_2 = (2 - 1.5) / 0.5 = 1$
$x_3 = 1.5$	→	$z_3 = (1.5 - 1.5) / 0.5 = 0$

Boxplot

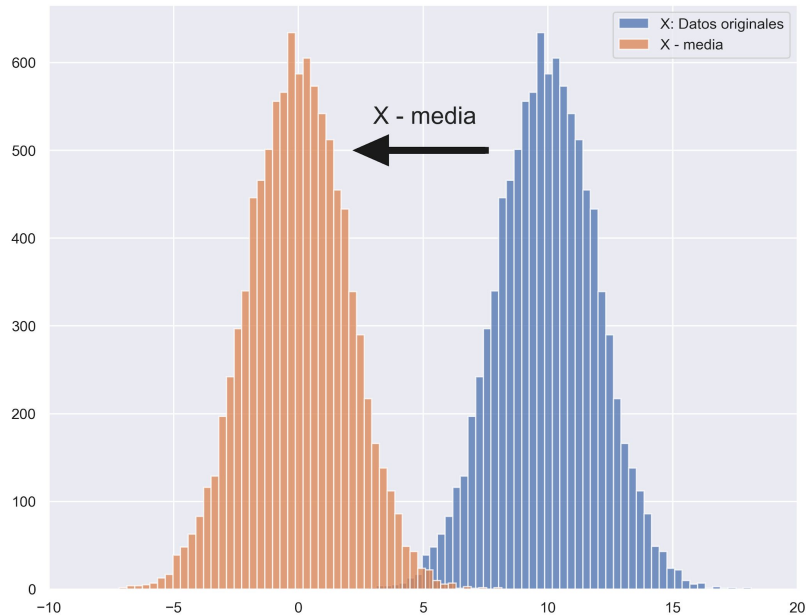
Rango intercuartílico

Regla de las tres sigmas

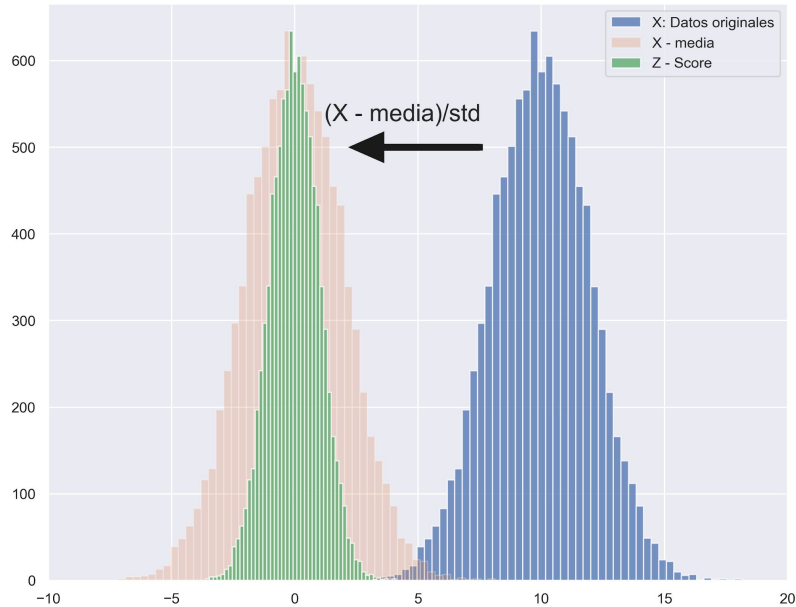
método Z-Score



método Z-Score



método Z-Score



método **Z-Score**

- El Z-Score es una medida de cuán lejos está un dato del promedio de la muestra a cual pertenece, medido en desviaciones estándar.
- También nos va a servir, más adelante, para **Reescalar Datos**. A veces lo podrán encontrar por el nombre de Estandarización o Normalización.
- En Scikit-Learn, existe una clase [StandardScaler](#) del módulo *preprocessing* que lo implementa.

Hands-on training





DS_Clase_13_Outliers.ipynb

A close-up photograph of a white ceramic cup filled with a latte. The surface of the milk is decorated with intricate latte art, featuring a central heart shape surrounded by concentric, wavy lines. The cup is placed on a matching white saucer. In the background, a white napkin and a silver spoon are visible, though they are out of focus. The overall lighting is soft and even, highlighting the textures of the coffee and the smooth surface of the cup.

¡BREAK!



Sabías que...



Valores
Atípicos
vs. Valores
Extremos



VALOR EXTREMO = valor distante del resto de las observaciones pero comprendido dentro de los valores esperados en mi distribución.

En general, son más comunes en distribuciones con alta curtosis.

Valores Atípicos vs. Valores Extremos



VALOR EXTREMO

¿Debemos sacarlos? ¿Los podemos considerar outliers?

Eso dependerá del problema que estemos estudiando.

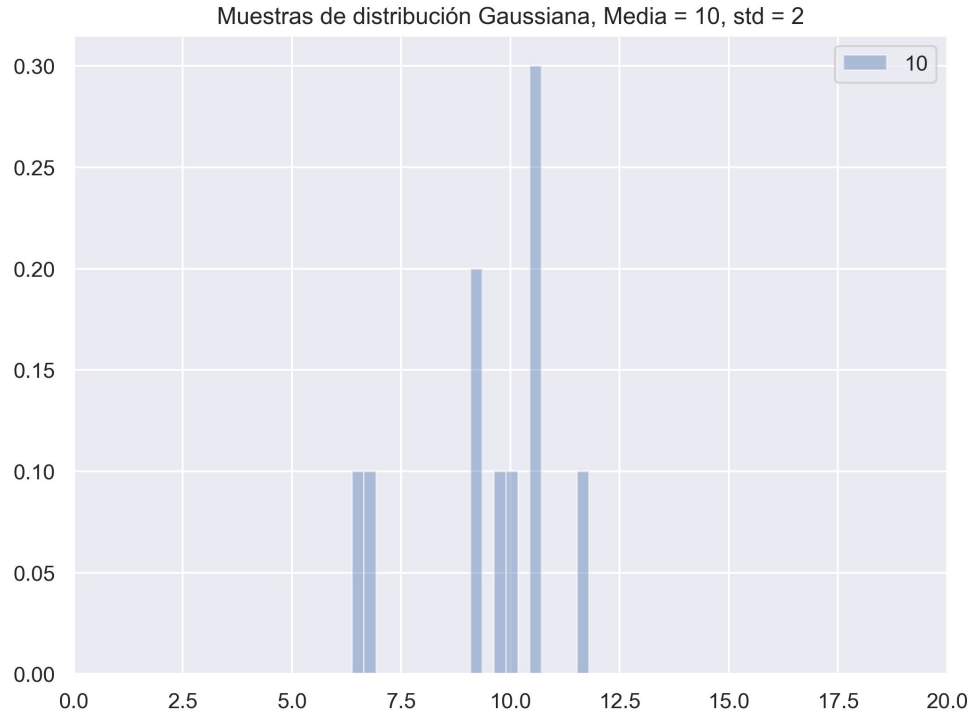
Pero es importante recalcar la diferencia entre:

- **un valor atípico** porque estamos mezclando poblaciones (transacciones no-fraudulentas/fraudulentas, personas sanas/no-sanos, etc.)
- **valores extremos** de una población homogénea (Ejemplo: usamos siempre la tarjeta de crédito para hacer compras pequeñas y un día compramos un pasaje en avión).

Valores Atípicos vs. Valores Extremos



VALOR EXTREMO: Una demostración

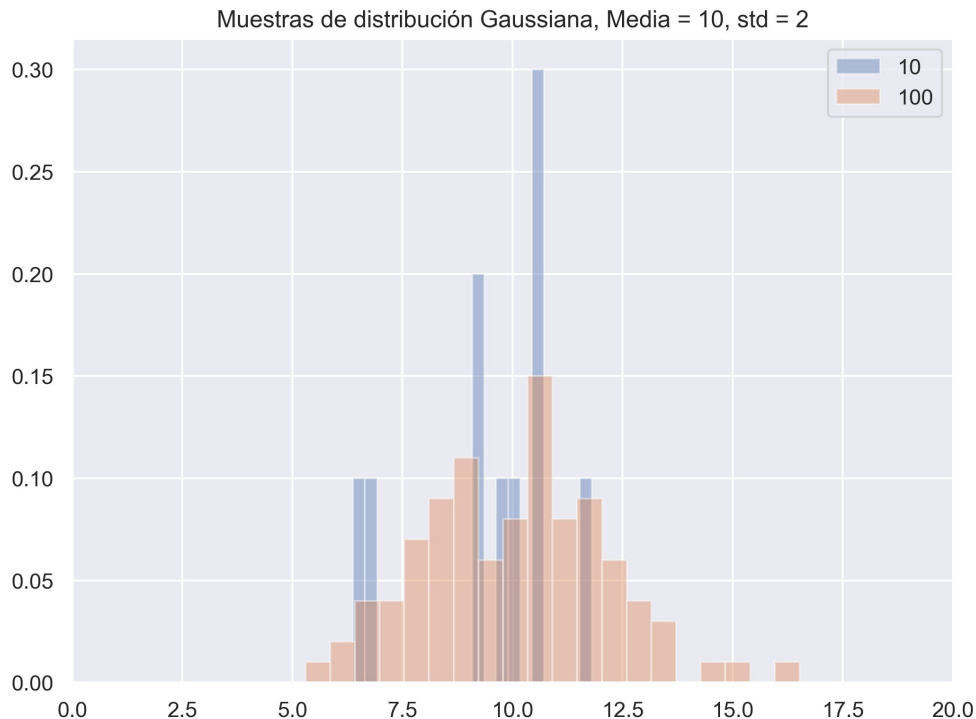


¿Qué pasa si aplicamos la **regla de tres sigmas** en estos casos?

Valores Atípicos vs. Valores Extremos



VALOR EXTREMO: Una demostración

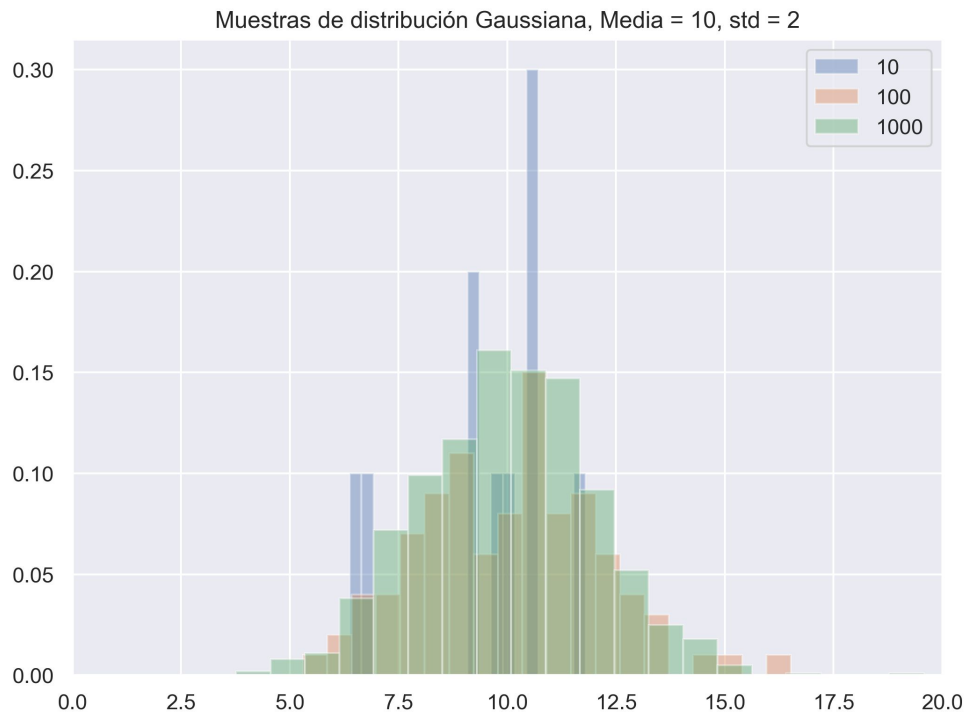


¿Qué pasa si aplicamos la **regla de tres sigmas** en estos casos?

Valores Atípicos vs. Valores Extremos



VALOR EXTREMO: Una demostración

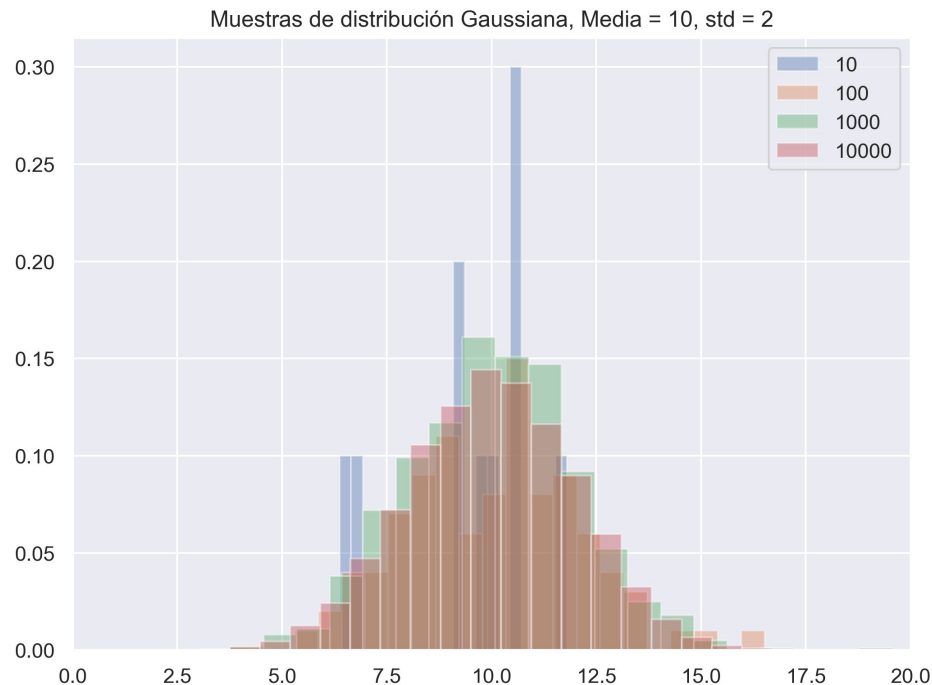


¿Qué pasa si aplicamos la **regla de tres sigmas** en estos casos?

Valores Atípicos vs. Valores Extremos



VALOR EXTREMO: Una demostración

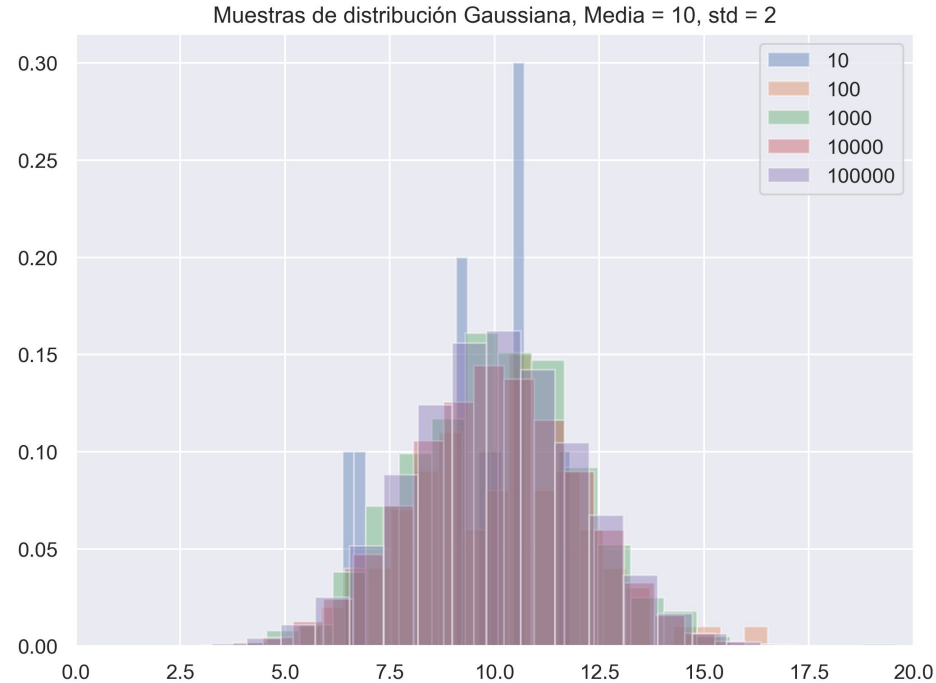


¿Qué pasa si aplicamos **la regla de tres sigmas** en estos casos?

Valores Atípicos vs. Valores Extremos



VALOR EXTREMO: Una demostración

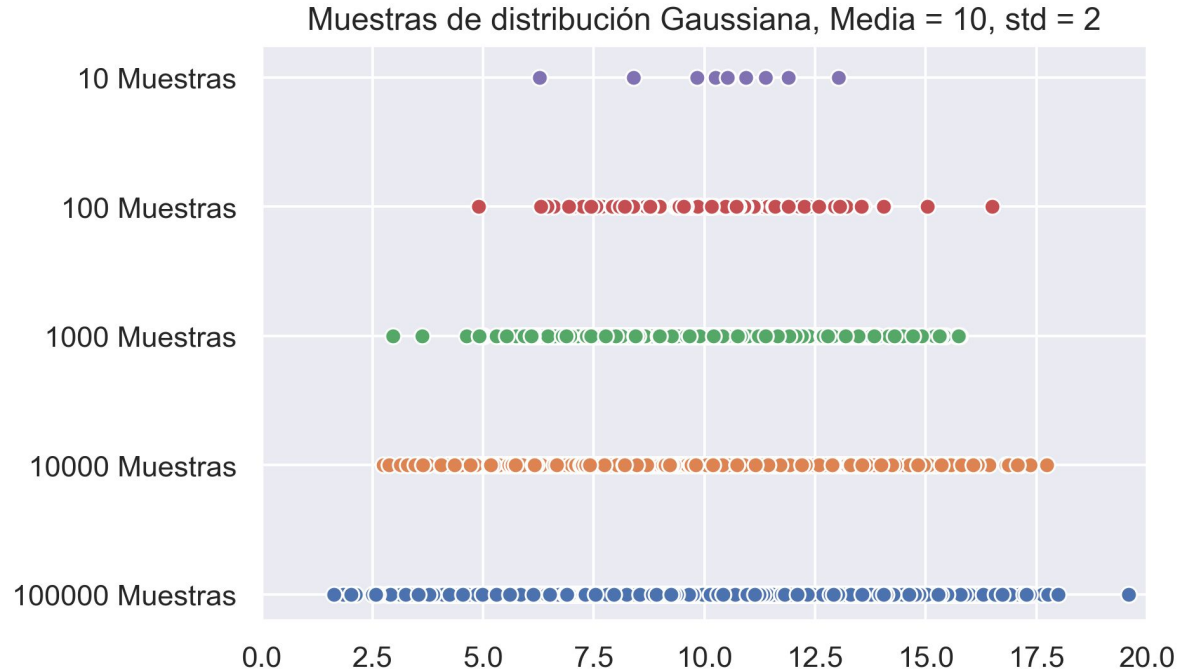


¿Qué pasa si aplicamos la regla de tres sigmas en estos casos?

Valores Atípicos vs. Valores Extremos

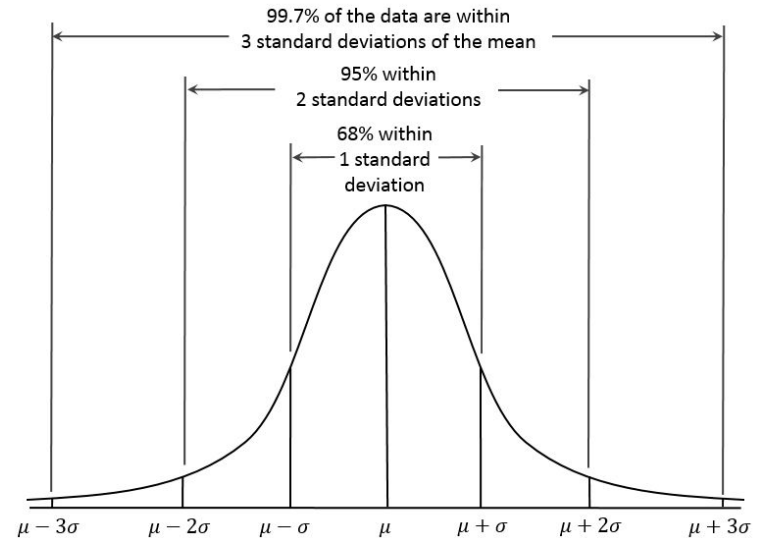


VALOR EXTREMO: Una demostración



Regla de las tres sigmas

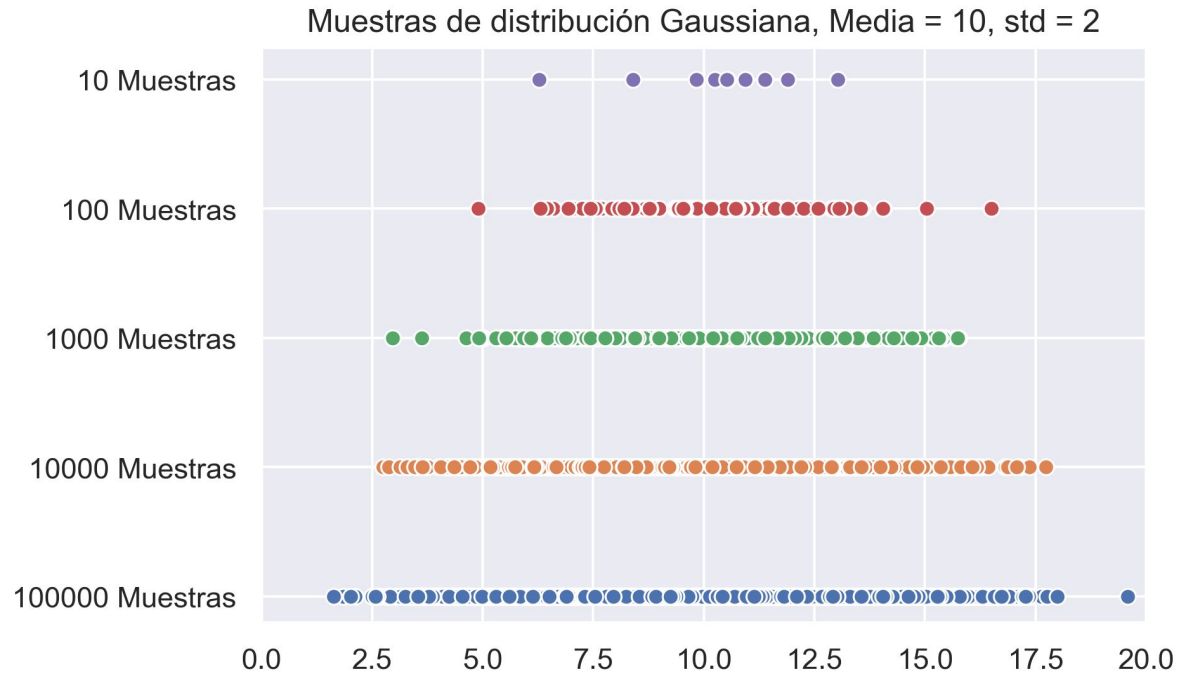
n	p	1 - p	Aprox. 1 en...
1	0.682	0.317	3
2	0.954	0.046	22
3	0.997	0.003	370
4	0.999936658	0.000063342	15787
5	0.999999427	0.000000573	1744277
6	0.999999998	0.000000002	506797346



Valores Atípicos vs. Valores Extremos



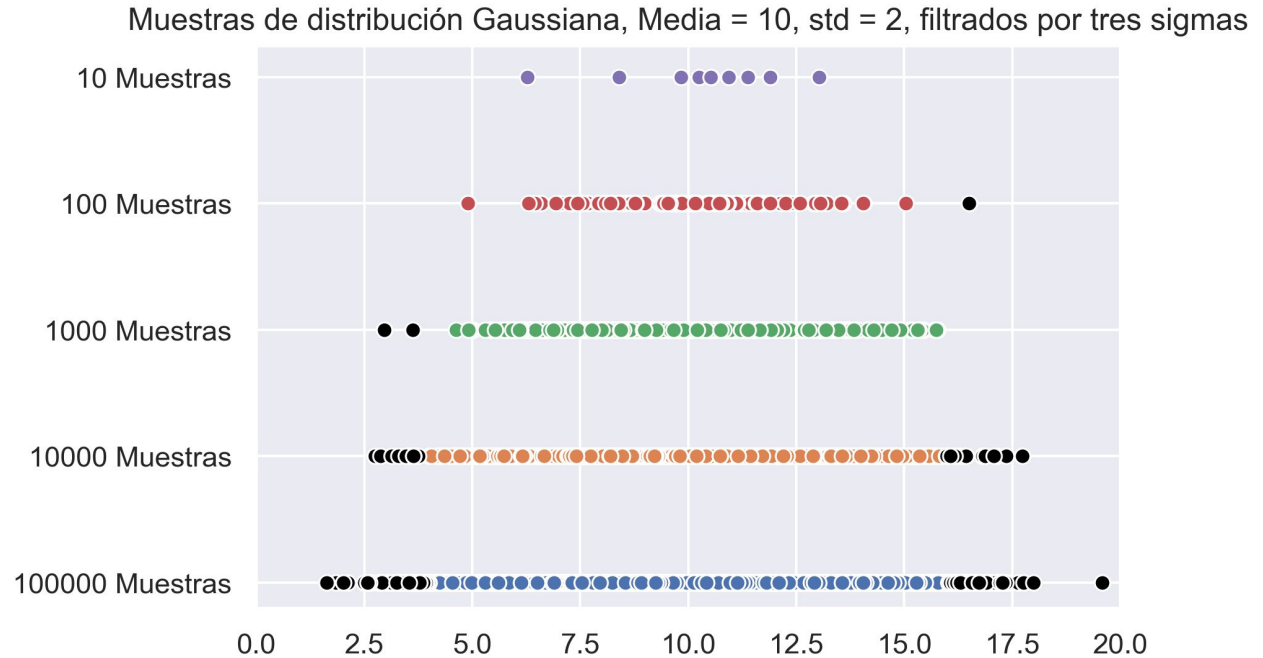
VALOR EXTREMO: Una demostración



Valores Atípicos vs. Valores Extremos



VALOR EXTREMO: Una demostración



Resumen

- Un *Outlier* generalmente indica que la distribución subyacente no es homogénea. Sin embargo, hay un grado de subjetividad en qué es un valor atípico.
- Cómo detectarlos va a depender de nuestro problema. Sin embargo, hay algunas técnicas estandarizadas
- ¡Mucho cuidado! No siempre hay que tirarlos. A veces es lo que buscamos. Ej: detección de fraudes.

Actividad: Explorando mis datos



¡Seguimos explorando!

¿Qué preguntas me *gustaría* responder?

¿Qué preguntas *podrán* responder con ese dataset?



**Lamentablemente no
siempre ambas coinciden**

Para la próxima

1. Si no lo hicieron, terminar de ver los videos sobre “Transformación de Datos”.
2. Terminar la Entrega 02.
3. Seguir explorando el dataset que eligieron.
4. Completar el notebook de hoy si no lo terminaron.

ACÀMICA