**¿Qué es un valor atípico?**

Un valor atípico es cualquier medición que se encuentra por fuera del comportamiento general de una muestra de datos.**Datos que no encajan con el comportamiento general de los demás datos.**

Pueden indicar variabilidad, errores de medición o novedades.

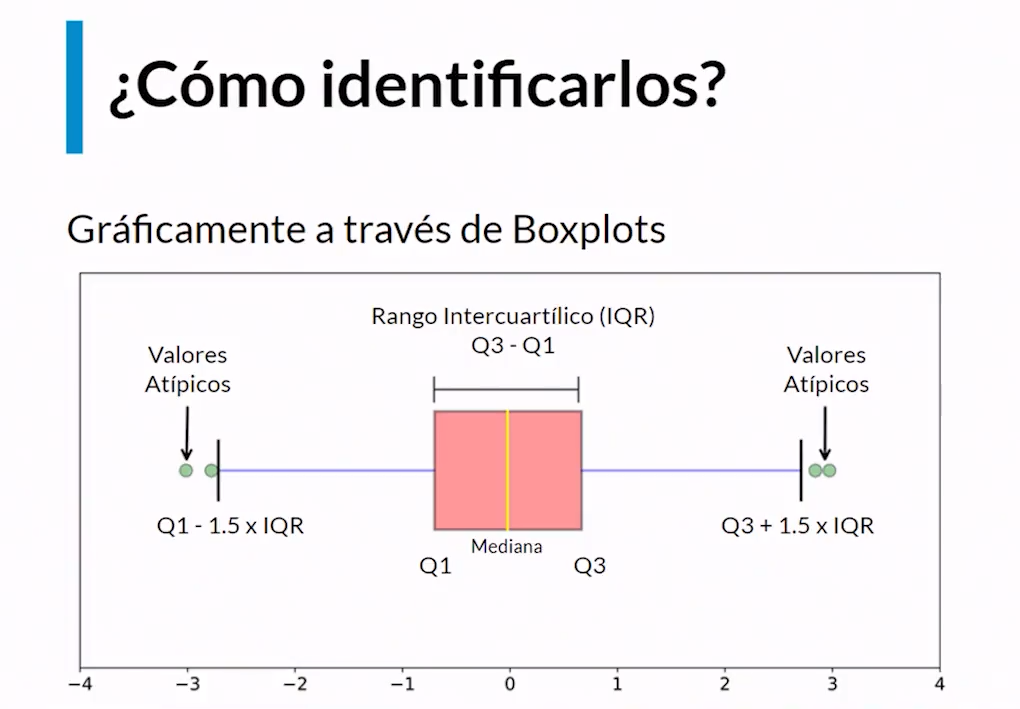
**¿Por qué son problemáticos?**

1. Pueden generar sesgos importantes en los modelos de machine learning.
2. A veces contienen información relevante sobre la naturaleza de los datos.
3. Detección temprana de fallos.

**¿Cómo identificarlos?**

A través de métodos estadísticos:

1. **Z-Score:** Mide la distancia (en desviaciones estándar) de un punto dado la media.
2. Técnicas de clustering como **DBSCAN.**
3. Si q< Q1 -1.5\*IQR ó q> Q3 +1.5\*IQR ( Estos son los puntos fuera de las cajitas estadísticas (BoxPlots)



Las “Q” mayúsculas vistos en el punto 4 son los llamados **Cuartiles**

**¿Qué es un Cuartil?**

Los cuartiles son los 3 valores de la variable que **dividen**  a un conjunto de datos ordenados **en cuatro partes iguales.**

**Q1, Q2 y Q3**  determinan los valores correspondientes al 25%, al 50% y al 75% de los datos.

**Q2** coincide con la **mediana.**

**IRQ:** Lo definimos como la diferencia entre (Q3 -Q1)

**Regresiones Robustas en Scikit-Learn**

Si bien es cierto que hay veces en que los valores atípicos los podemos tratar desde la etapa de preprocesamiento( Eliminarlos o transformarlos) , hay muchas veces en que la única opción que tenemos es lidiar con ellos cuando estamos aplicando nuestro modelo de Machine Learning.

Scikit Learn nos ofrece algunos modelos específicos para abordar el problema de los valores atípicos: **Ransac** ó **Huber Regressor (**Por ahora solo veremos 2)

**RANSAC**

Lo que hacemos es que tomamos una muestra aleatoria de todo nuestro conjunto de datos y asumimos que todos ellos son nuestros “inliners”(es decir, que son valores que no son atípicos o los que se comportan de acuerdo a la distribución estadística del modelo).Luego, procedemos a entrenar a nuestro modelo con los datos de nuestra muestra para después compararlas con muchos muestreos de este tipo. Al final de muchas pruebas podremos encontrar los cuál es la combinación de datos que más se ajusta o que tenga la mejor cantidad de “inliners” para que , en base a ello, podemos eliminar o discriminar nuestros datos atípicos

*“ Usamos una muestra aleatoria sobre el conjunto de datos que tenemos y luego buscamos la muestra que más datos “buenos” logre incluir.*

*El modelo asume que los “malos valores” no tienen patrones específicos. ”*

**Huber Reggresor**

Aquí nuestro modelo trabaja de la manera normal y, en base a el valor de “la pérdida”, va a escoger los valores que están debajos de cierto umbral. Si estan por encima de nuestro umbral *“ épsilon ”* , va a considerarlos como atípicos y no los va a tener en cuenta dentro de nuestro modelo.

*“No ignora los valores atípicos, sino disminuye su influencia en nuestro modelo.*

*Los datos son tratados como atípicos si el error absoluto de nuestra pérdida está por encima de un umbral llamado épsilon.*

*Se ha demostrado que un valor de épsilon = 1.35 logra un 95% de eficiencia estadística”*