Como identificar datos atipicos

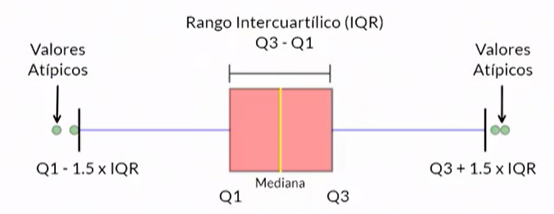
1) Z-score: distancia de media

2) DBSCAN: tecnicas de clustering

3) Cuartiles

Si q<Q1 - 1.5 \* IQR

Si q>Q3 - 1.5 \* IQR



Esta mas abajo o mas arriba del inicio o final de caja

Dispersion

Boxplot (diagrama de caja)

* **BSCAN:** **Density-Based Spatial Clustering of Applications with Noise.**

Consiste en considerar a zonas muy densas como clusters, mientras que los puntos que carecen de ‘vecinos’ no pertenecen a ningún conjunto y por lo tanto se clasifican como ruido (o outliers).

REGRESIONES ROBUSTAS PARA MANEJAR VALORES ATIPICOS

**Resumen**

* **Ransac:** selecciona una muestra aleatoria de los datos asumiendo que esa muestra se encuentra dentro de los valores inliners, con estos datos se entrena el modelo y se compara su comportamiento con respecto a los otros datos. Este procedimiento se repite tantas veces como se indique y al finalizar el algoritmo escoge la combinación de datos que tenga la mejor cantidad de inliners, donde los valores atípicos puedan ser discriminados de forma efectiva.

Ejemplo:

Chart, scatter chart

Description automatically generated

* **Huber Reggresor**: no elimina los valores atípicos sino que los penaliza. Realiza el entrenamiento y si el error absoluto de la perdida alcanza cierto umbral (epsilon) los datos son tratados como atípicos. El valor por defecto de epsilon es 1.35 ya que se ha demostrado que logra un 95% de eficiencia estadística.