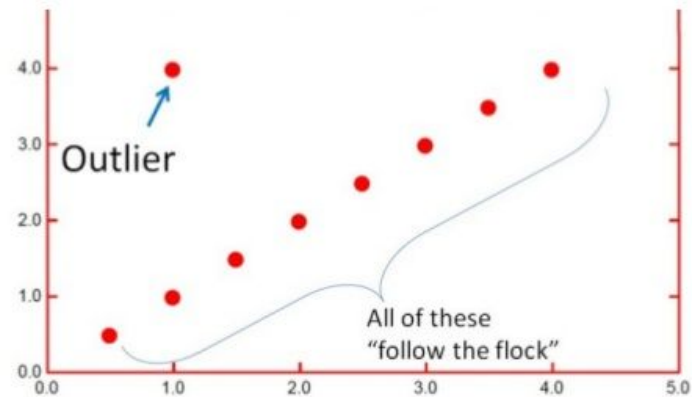


WHAT IS AN OUTLIER?




Never mind what the axes mean...

OUTLIERS




¿QUÉ SON LOS OUTLIERS?

- **Son valores extremos**, es decir valores que se alejan significativamente de los valores que se pueden calificar como normales.
 - Pueden ser outliers **altos o bajos**, en el caso de que se dé un valor anormalmente bajo.
- 




POSIBLES CAUSAS

- **Los outliers se pueden producir por varias razones:**
 - Puede tratarse de un dato erróneo y por ello puede ser un valor diferente al resto.
 - Pueden deberse a acontecimientos reales pero aislados y difíciles de repetir.
 - Pueden ser valores anormales pero que se dan con cierta frecuencia.
- 



EFFECTOS QUE GENERAN

- Los valores extremos pueden **distorsionar las predicciones** del modelo.
 - Puede **distorsionar** las verdaderas **relaciones entre las variables**.
 - A su vez estas observaciones extremas en muchas ocasiones revelan mucha información.
- 



¿QUÉ HACER CON LOS OUTLIERS?

- No hay ninguna solución óptima para todos los casos.
- Depende de la razón del outlier.
- Se ve afectado por la importancia de la variable.
- Está condicionado por el objetivo final del proyecto.



¿QUÉ HACER CON LOS OUTLIERS?

- **Opciones:**
 - Eliminarlos o modificarlos:
 - Valores erróneos.
 - Observaciones reales pero con baja probabilidad de repetirse.
 - Tratarlos:
 - Valores que se pueden repetir.

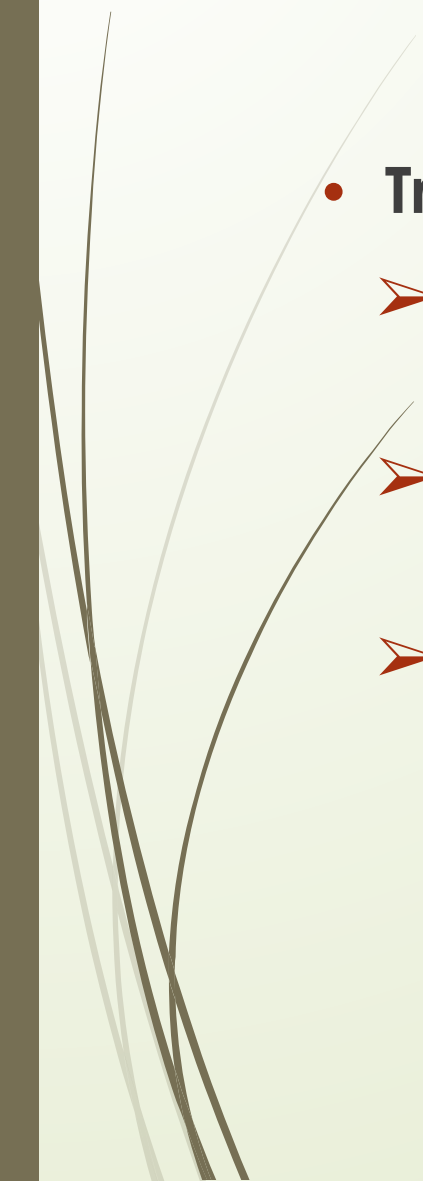


ELIMINARLOS O TRATARLOS

- **Eliminar la observación.**
 - Pérdida de información.
 - Mayor seguridad de no crear falsas relaciones.
- **Modificar la observación:**
 - No se pierde información.
 - Se pueden crear relaciones ficticias.



TRATAMIENTO DE OUTLIERS

- **Tratamiento generalizado:**
 - Separar los outliers de resto de la muestra.
 - No discrimina la variable que genera el outlier.
 - No permite modificar el modelo de manera correcta en función de los outliers.
- 

TRATAMIENTO DE OUTLIERS

- **Tratamiento individualizado:**
 - Crear dos variables (alto y bajo) para marcar los outliers de cada una de las variables.
 - Muestra para cada observación en qué variables presenta el outlier y si este es bajo o alto.
 - Modifica el modelo de manera adecuada.
 - Cada variable recoge una relación.
 - Aumenta el riesgo de overfitting.



COMBINACIÓN

- Se puede realizar una combinación de **varios métodos**:
 - Eliminar o modificar variables que se consideren erróneas o irrepetibles.
 - Tratamiento individualizado de las variables que tras una primera modelización se muestren como más importantes.
 - Tratamiento genérico del resto de variables.