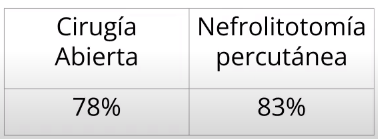
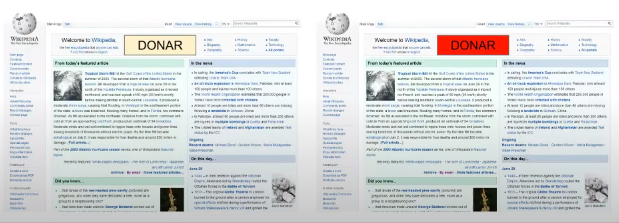
# Falacias en los datos

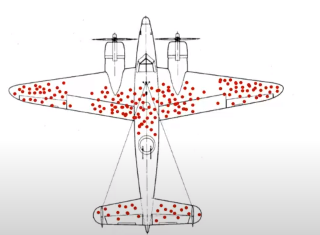
* Paradoja de Simpson: paradoja en la cual una tendencia que aparece en varios grupos de datos desaparece cuando estos grupos se combinan y en su lugar aparece la tendencia contraria para los datos agregados. La paradoja desaparece cuando se analizan las relaciones causales presentes.



* A/B Testing: Enviar aleatoriamente la mitad de las request hacia un sitio y la otra mitad al otro.



* Sesgo de supervivencia: Consiste en concentrarse en las personas o cosas que superaron un proceso de selección pasando por alto a aquellas que no lo hicieron, típicamente por su falta de visibilidad. Esto puede conllevar a conclusiones falsas de muy diferentes formas. En estos casos hay que pensar cuál es el origen de los datos.



# Introducción a la ciencia de datos

# 

# Variables

Tipos

Vamos a contar con dos tipos de variables:

* Independientes: entradas
  + Cualitativas (problemas de clasificación)
    - texto
      * nominales
      * ordinales
    - numéricas
      * nominales
      * ordinales
  + Cuantitativas
    - discretas
    - continuas
* Dependientes: salidas

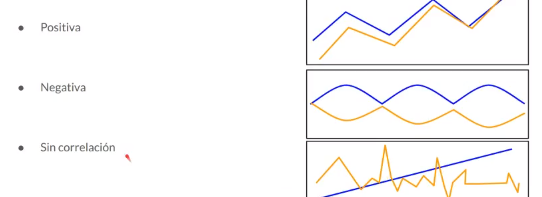
Problemas

* De clasificación: variables dependientes cualitativas
* De regresión: variables dependientes cuantitativas
* De agrupamiento: no hay variables

**Outliers**: valores atípicos

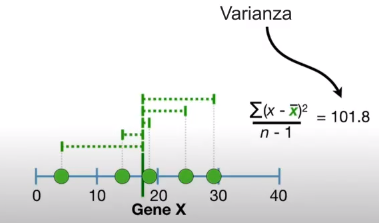
Correlación de variables

No implica causalidad



Varianza

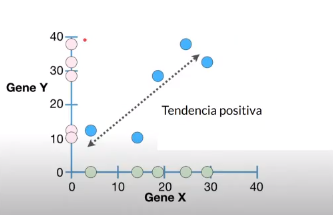
Promedio de la diferencia entre todas las observaciones respecto a su media. Qué tan dispersos están los datos.



Covarianza

Valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias. Puede ser positiva, negativa o neutra.





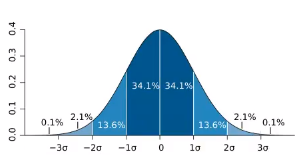
Correlacion de Pearson

Mide la correlación lineal. ¿Qué tan relacionadas están las variables?

* 0: no hay correlación
* 1: correlación perfecta
* -1: correlación negativa perfecta

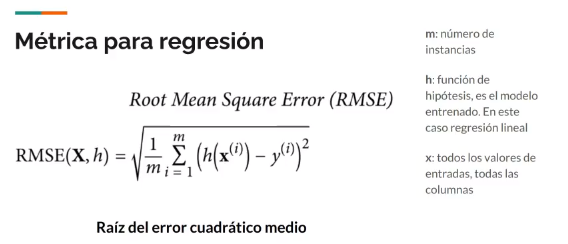
Desvío estándar

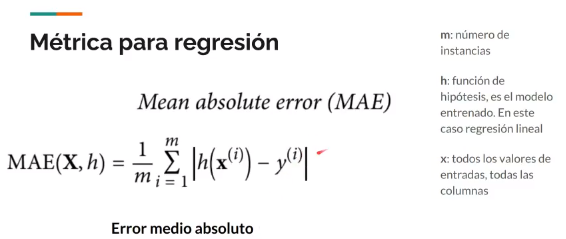
Se usa para cuantificar la variación de un conjunto de datos numéricos.

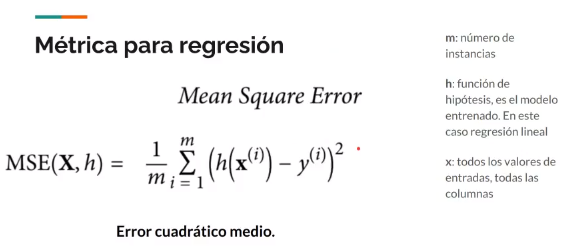


Métodos de regresión

Métricas para el error de regresión (mínimos cuadrados)







(Esta última es la que mas vamos a usar)