PLAYING WITH DATA

short line

# Dataset

Dataset que contiene datos de animales

# Preguntas

1. Cuántos datos tenemos: filas y columnas? (26729, 10)
2. Hay datos nulos? En caso afirmativo, en qué porcentaje?

AnimalID 0.000000

Name 28.773991

DateTime 0.000000

OutcomeType 0.000000

OutcomeSubtype 50.925961

AnimalType 0.000000

SexuponOutcome 0.003741

AgeuponOutcome 0.067343

Breed 0.000000

Color 0.000000

1. El dataset se compone de algunas variables: ¿cuáles son numéricas, categóricas, fechas, ...? En el caso de que sean categóricas, de qué tipo?

AnimalID String

Name String

DateTime DateTime

OutcomeType Categorico

OutcomeSubtype Categorico

AnimalType Categorico

SexuponOutcome Categorico

AgeuponOutcome Numérico

Breed Categorico

Color Categorico

1. ¿Existen duplicados? **No**
2. Las variables tiene diferentes valores. Calcula cuántos tienen cada una y cuales son.
3. La variable objetivo es 'OutcomeType': qué valores tiene y en qué probabilidad?
4. Ahora que sabes cuál es el posible futuro de estos animales ... ¿ocurre lo mismo con gatos y perros? Podrías decir que estos datos se deben a la proporción entre el número e perros y gatos? **Que no se deben únicamente a la proporción entre perros y gatos.**
5. Además de la variable 'OutcomeType' existe otra 'OutcomeSubtype'. Estudia su distribución. Haz lo mismo con la variable 'AgeuponOutcome'.
6. ¿Tiene algo que ver el género del animal con la variable objetivo?
7. ¿Hay algún momento del año en el que se produzcan más adopciones?
8. Toma una muestra de 10000 datos de manera aleatoria. Haz un test de hipótesis sobre la variable 'AgeuponOutcome' para ver si la media corresponde a la media total.
9. ¿Qué otros datos se pueden extraer? Describe al menos 2.

short dash