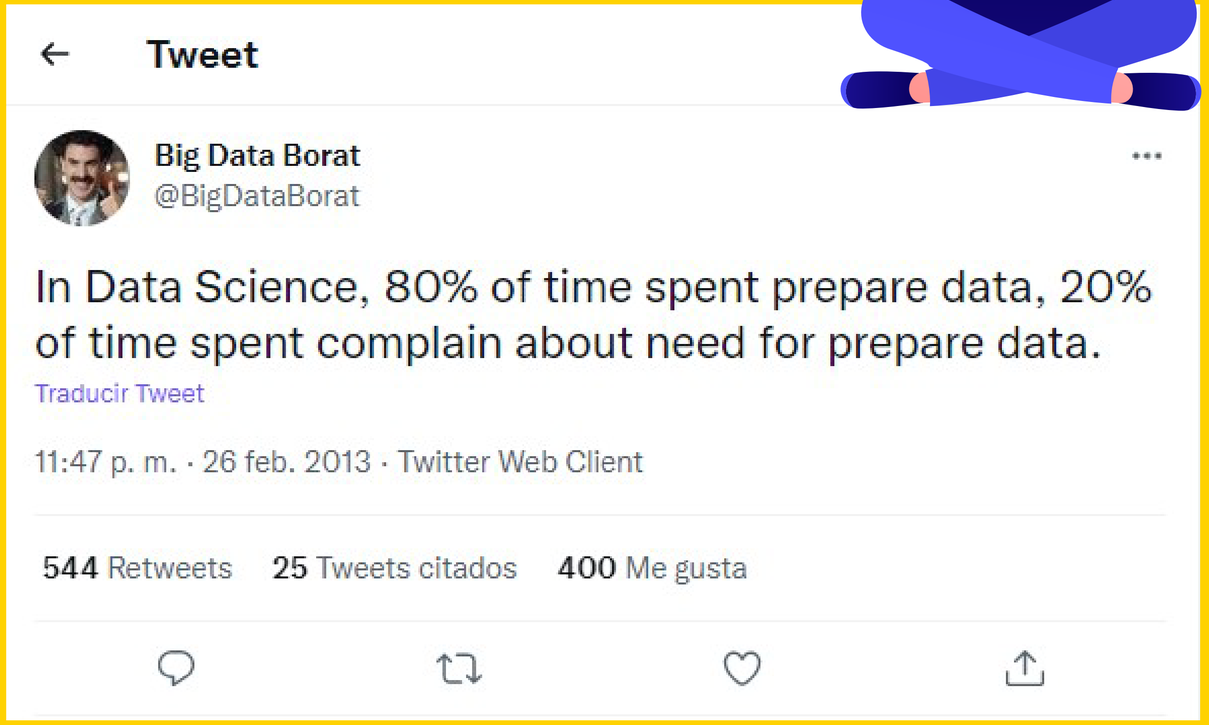
Introducción a datos abiertos

Clase 04

Tabla de contenidos

Introducción

En 2009, Mike Driscoll (data scientist y CEO de Metamarkets) popularizó el término ‘data munging’ para referirse al arduo proeso de limpiar, preparar y validar los datos.

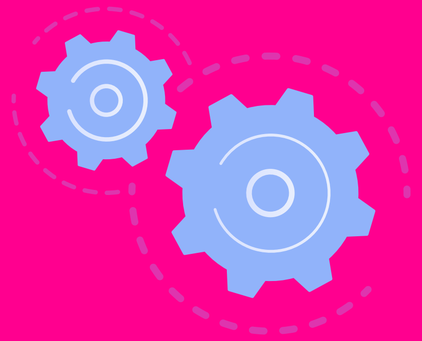


Traducción: En Data Science, se invierte un 80% del tiempo en preparar los datos y el 20% restante en quejarse de la necesidad de preparar los datos.

1. Resolver problemas de formato y asignar los tipos correctos de datos.

El formato en que se encuentran los datos va a afectar nuestro análisis por varias razones, como pueden ser las operaciones que se pueden realizar dependen del tipo de datos. Además algunos tipos ocupan menos espacio en memoria que otros.

Por ejemplo cuando al pasar de CSV a Pandas una fecha no se importa correctamente. Ej: 20090609231247 en lugar de 2009-06-09 23:12:47.



2. Estandarizar categorías

Cuando los datos se recolectaron con un sistema que no tiene los valores tipificados, valores que representan las mismas categorías pueden estar expresados de forma distinta, por ejemplo Arg, AR y

3. Corregir valores erróneos.

Por ejemplo, un valor numérico o inválido para describir el género. O una edad representada por un número negativo o mucho mayor que 100.

4. Completar datos faltantes

Los datasets del mundo real suelen venir con datos faltantes que responden a información que se perdió o nunca se recolectó. Existen varias técnicas para completar datos faltantes. Al proceso de completar datos faltantes se lo llama ‘imputación’.

Los datos completos

La imputación simple

La imputación por media

5. Organizar correctamente el dataset.

Es importante estructurar las filas y columnas de la forma más conveniente. Para hacerlo se pueden aplicar las reglas del "tidy data".

Decimos que un dataset está ordenado cuando:

- Cada variable es una columna
- Cada observación es una fila
- Cada tipo de unidad observacional forma una tabla

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Ejercicio de clase:

A partir de los datasets ‘llamados al 107 por Covid’:
1- Corregir los valores erróneos (fecha).
2- Estandarizar las categorías.
3- Resolver problemas en el formato (fecha).
4- Completar los datos faltantes.

Ejercicio práctico:

1- Cargar el recurso Molinetes - Agosto 2021 (XLSX) del dataset homónimo en Google Colaboratory
2- Seleccionar la columna “FECHA” y unirla con la columna “DESDE” para convertir en un objeto datetime.
3- Crear una nueva columna con el contenido de la anterior con el nombre: “fecha_desde”
4- Cambiar el nombre de las demás columnas y expresar en letras minúsculas*.

*(¿Cómo hacerlo?) Ver atributo .columns
nombre_dataset.columns
pista:
nombre_dataset.columns =[‘nombre_col1’,’nombre_col2’,...]