

Data Cleaning *

Descripción de los pasos usuales que se deben seguir en un dataset para realizar una limpieza completa.

Identificar el tipo de información

Etiquetar las variables de acuerdo a su tipo, esto nos ayudará para definir los siguientes pasos para cada tipo de variable

1. Identificar la unidad muestral
2. Eliminar duplicados globales
3. Eliminar duplicados de acuerdo a la unidad muestral

Duplicidad

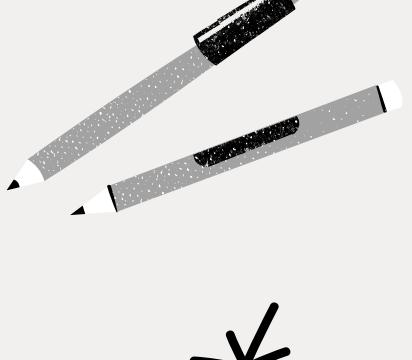
Compleitud

Conocer el número de registros en los cuales no poseemos información y eliminar aquellas variables que no se pueden utilizar por falta de información



1. Identificar variables UNITARIAS (una categoría posee la mayoría de los registros)
2. Investigar si aporta info valiosa, si no es así eliminarla
3. Identificar variables IRRELEVANTES (variables que no apuntan al objetivo)
4. Eliminar variables irrelevantes

Variables REPETITIVAS O IRRELEVANTES



[Desliza para ver el resto]

[Continúa aquí]



- Continuas - Valores enteros o flotantes en un formato de python
- Categóricas -
 - Limpieza de texto
 - Agrupación de categorías(NORMALIZACION)



- Texto - Eliminar hapaxes, stopwords, caracteres especiales, etc.
- Ubicaciones -
 - Definir formato para direcciones
 - validar códigos postales,
 - validar lat y lng dentro del rango posible
 - etc.
- Identificadores Únicos
 - Validar su unicidad

OPCIONES:

- *Eliminar inconsistencias de tipo de dato
- *Reemplazar inconsistencias por nulos



3.

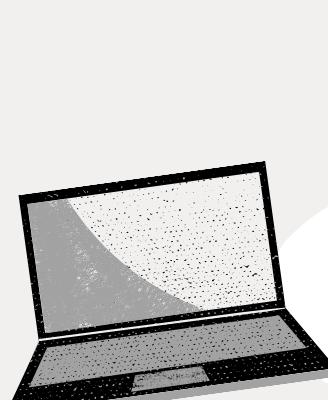
Consistencia
-Precisión

- Identificar irregularidades en los datos:
 - Coincidencia de fechas
 - Coincidencia de ubicaciones
 - Revisión de rangos
 - Etc

OPCIONES:

- *Eliminar inconsistencias de tipo de dato
- *Reemplazar inconsistencias por nulos

LISTO



Cada conjunto de datos tiene particularidades, por lo tanto es complicado englobar todos los casos posibles

