Proceso ETL (Extract Transform Load)

Proceso EDA (Exploratori Data Analysis)

* Descripción de las variables (fuera del perímetro del proyecto):
  + Tipo de dato
  + Rango (para datos numéricos) / Valores (para datos discretos)
  + Descripción
* Análisis estadístico del dataset:
  + Número de registros y variables
  + dataset.describe():
    - medidas de tendencia central: media, mediana y moda
    - medidas de dispersión: rango, varianza y desviación estándar
    - medidas de asimetria: skewness, kurtosis
  + ...
* Data Cleaning:
  + Identificar valores perdidos
  + Imputación de valores perdidos (si procede), en base a la media, la mediana, la moda o haciendo usa de una técnica avanzada de imputación
  + Eliminar registros duplicados
  + Detección de outliers
  + Eliminar variables que no se van a utilizar
  + Discretizar variables continuas
* Análisis inferencial (fuera del perímetro del proyecto): (DUDA: ¿DEBE HACERSE ANTES O DESPUÉS DE LIMPIAR? > LO ANTES Y DESPUÉS PARA COMPARAR)
  + Test de hipótesis: t-test, chi cuadrado y análisis de la varianza (ANOVA)
  + Intervalos de confianza
  + ANOVA
  + Distribuciones probables
* Visualización

Fuentes:

<https://medium.com/@snehalkore016/exploratory-data-analysis-eda-d30967cb5589>

<https://medium.com/@navamisunil174/exploratory-data-analysis-of-breast-cancer-survival-prediction-dataset-c423e4137e38>

<https://medium.datadriveninvestor.com/250-python-and-data-science-tips-covering-pandas-numpy-ml-basics-sklearn-jupyter-and-more-e33074b92d58>