Exploratory Data Analysis

{ Isaac Martín } Data Science Lab. URJC

CHEAT SHEET

Introducción

Definición

El **Análisis Exploratorio de Datos** o (EDA, del inglés "Exploratory Data Analysis") consiste en comprender los conjuntos de datos resumiendo sus características principales, a menudo representándolas visualmente.

Es necesario involucrar (al menos) a tres actores principales: analista de datos, experto en el dominio, responsable de la obtención de datos.

Objetivos de negocio

- Objetivo de negocio definido junto con los datos.
- Entender los datos.
- Entender las variables.
- Adquirir conocimiento sobre el problema.
- Detectar errores en la adquisición de los datos.
- **Enriquecer los datos**.
- **Obtención de Insights**.

Técnicas

Limpieza de Datos

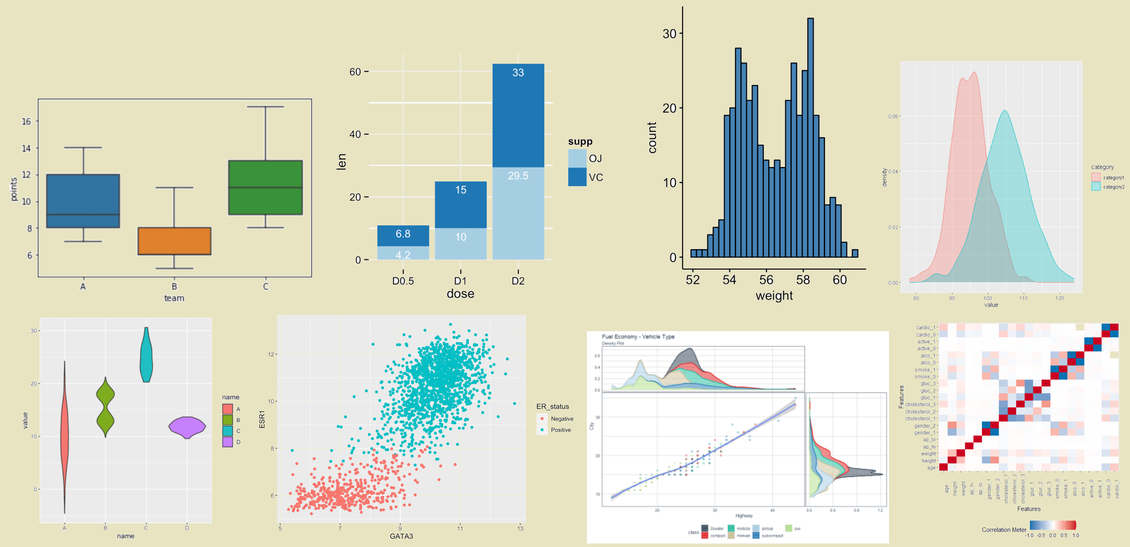
Es el proceso de garantizar que los datos son correctos y utilizables, identificando cualquier error en los datos, o los datos que faltan, corrigiéndolos o eliminándolos.

Preprocesado de Datos

Técnica de extracción de datos que consiste en transformar los datos brutos en un formato comprensible. Incluye la normalización y estandarización, la transformación, la extracción y selección de características, etc. El producto del preprocesamiento de datos es el conjunto de datos de entrenamiento final.

Visualización de Datos

Representación gráfica de información y datos. Utiliza gráficos estadísticos, diagramas, gráficos informativos y otras herramientas para comunicar la información de forma clara y eficaz.



Medidas numéricas

- Media, varianza, moda, rangos, intervalos de confianza, tablas.
- T-test, χ^2 test.
- Imputación de valores faltantes.
- Correlación.

Librerías y Paquetes

pandas, dtale, sweetviz, autoviz

Dataexplorer, GGally, SmartEDA, dplyr, ggplot2

Paso a Paso Funcional

1. Características de los datos
2. Tamaño del problema
   - Evaluar recursos de computación.
   - Evaluar submuestreo.
3. Características de las variables.
4. Análisis univariante.
   - Numérico.
   - Gráfico.
5. Detección de valores faltantes.
   - Imputación de valores faltantes.
6. Transformación de variables.
7. Detección de relaciones entre las variables.
   - Variable respuesta.
   - Análisis bivariante.
     - Contraste de hipótesis.
   - Análisis Multivariante.
     - Componentes principales.
     - t-SNE UMAP.
8. Detección de atípicos.