

Short Course. Exploratory Data Analysis

Isaac Martín de Diego

Data Science Lab. URJC

Nov. 2022

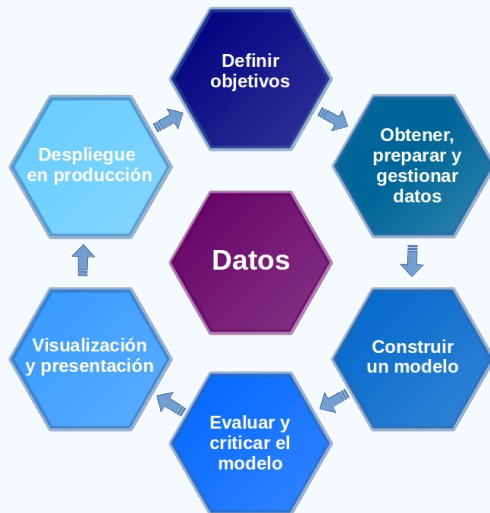


- 1 Entender el problema. Objetivos de negocio
- 2 Técnicas
- 3 Paso a paso funcional
- 4 Para recordar

Entender el problema. Objetivos de negocio

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.

- Este paso es muy importante, especialmente para el siguiente paso, que será modelar los datos para aplicar técnicas de Aprendizaje Máquina. ["Garbage in, garbage out"]
- Los gráficos en EDA consisten en histogramas, gráficos de caja, gráficos de dispersión y muchos más.
- A menudo se necesita mucho tiempo para explorar los datos. Más del 80% del tiempo del proyecto se gasta en EDA.
- A través del proceso de EDA, podemos pedir que se redefina el enunciado del problema o la definición de nuestro conjunto de datos, lo cual es muy importante.



Lo ideal es contar con un objetivo que se haya definido junto con los datos, indicando qué se quiere conseguir a partir de ellos. Por ejemplo, "predecir las ventas en los próximos 30 días", "estimar el riesgo que tiene un paciente de no superar una determinada operación quirúrgica", "clasificar como fraudulenta, o no, una página web", etc.

- Entender los datos.
- Entender las variables.
- Adquirir conocimiento sobre el problema.
- Detectar errores en la adquisición de los datos.
- **Enriquecer los datos.**

"The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data."

John Tukey (chemist, topologist, educator, consultant, information scientist, researcher, statistician, data analyst, executive. [1])

"Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there."
John Tukey

"Make big data as small as possible as quickly as possible."
Robert Gentleman (one of the originators of the R programming language)

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.

Métodos numéricos

- Medidas resumen: media, mediana, moda.
- Medidas de dispersión: desviación típica.
- Detección de atípicos.
- Tratamiento de atípicos.

Métodos gráficos

- Histograma.
- Boxplot.
- Scatter plot.
- Diagrama de barras.
- Diagramas de densidad.

Univariante, bivalente, multivariante.

Enlace al material: [aquí](#).

Paso a paso funcional

Antes de llevar a cabo un análisis de datos es necesario definir el tipo de pregunta que se quiere responder [2].

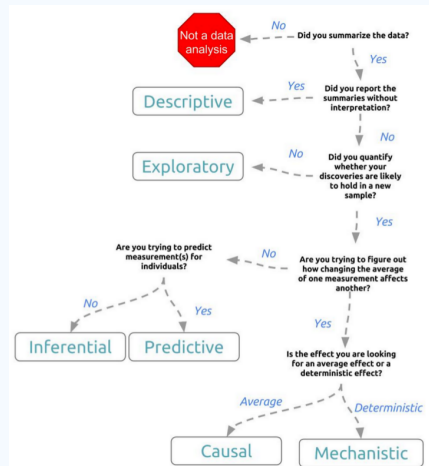


Figure 2.1 The data analysis question type flow chart

- Descriptivo: un análisis de datos descriptivo busca resumir las mediciones en un único conjunto de datos sin mayor interpretación.
- Exploratorio: un análisis exploratorio de datos se basa en un análisis descriptivo buscando descubrimientos, tendencias, correlaciones o relaciones entre las mediciones de múltiples variables para generar ideas o hipótesis.
- Inferencial: un análisis inferencial de datos va más allá de un análisis exploratorio al cuantificar si un patrón observado se mantendrá probablemente más allá del conjunto de datos en cuestión.
- Predictivo: un análisis de datos predictivo utiliza un subconjunto de medidas (las características) para predecir otra medida (el resultado) en una sola persona o unidad.
- Casual: un análisis de datos causal busca averiguar qué ocurre con una medida si se hace cambiar otra medida.
- Mecanicista: un análisis de datos mecanicista busca demostrar que el cambio de una medida conduce siempre y exclusivamente a un comportamiento específico y determinista en otra.



- La respuesta es que depende de las características particulares del conjunto de datos con el que se trabaje.
- No hay un método o métodos comunes para realizar EDA. En este curso presentamos algunos métodos comunes que se utilizarían en el proceso de EDA.
- A continuación planteamos algunas de las primeras preguntas que deberíamos de plantearnos.

- ¿Cuál es el tamaño de la base de datos?. Es decir:
 - ¿Cuántas observaciones hay?
 - ¿Cuántas variables/características están medidas?
- ¿Disponemos de capacidad de cómputo en nuestra máquina para procesar la base de datos o necesitamos más recursos?
- ¿Existen valores faltantes?
- ¿Qué tipo variables aparecen en la base de datos?
 - ¿Qué variables son discretas?
 - ¿Cuáles son continuas?
 - ¿Qué categorías tienen las variables?
 - ¿Hay variables tipo texto?

- Variable objetivo: ¿Existe una variable de "respuesta"?
 - ¿Binaria o multiclase?
- ¿Es posible identificar variables irrelevantes?.
 - Estudiar variables relevantes requiere, habitualmente, métodos estadísticos.
- ¿Es posible identificar la distribución que siguen las variables?
- Calcular estadísticos resumen (media, desviación típica, frecuencia,...) de todas las variables de interés.
- Detección y tratamiento de valores atípicos.
 - ¿Son errores de media?
 - ¿Podemos eliminarlos?
- ¿Existe correlación entre variables?

Para recordar

EDA Cheat Sheet: [aquí](#).

-  Brillinger, D.R.: John w. tukey: His life and professional contributions. The Annals of Statistics **30**(6), 1535–1575 (2002).
URL <http://www.jstor.org/stable/1558729>
-  Leek, J.: The Elements of Data Analytic Style (2015)