*Exploratory Data Analysis (EDA)*

**¿QUÉ ES UN EDA?**

El Análisis Exploratorio de Datos o también conocido como EDA de sus siglas en inglés: Exploratory Data Analysis, tiene como finalidad examinar los datos previamente a la aplicación de cualquier técnica estadística. De esta forma el Científico de Datos, consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas.

El EDA, proporciona métodos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de datos, tratamiento y evaluación de datos ausentes, identificación de casos atípicos y mucho más.

Resulta importante destacar, que el examen previo de los datos es un paso necesario, que lleva tiempo, y que habitualmente se descuida por parte de los analistas de datos. Las tareas implícitas en dicho examen pueden parecer insignificantes y sin consecuencias a primera vista, pero son una parte esencial de cualquier análisis estadístico.



**UTILIDADES del EDA**

Algunas de las preguntas que podemos responder gracias a realizar un EDA, son las siguientes:

* ¿Existe algún sesgo en los datos recogidos?
* ¿Hay errores en la codificación de los datos?
* ¿Cómo se sintetiza y presenta la información contenida en un conjunto de datos?
* ¿Existen datos atípicos (outliers)? ¿Cuáles son? ¿Cómo tratarlos?
* ¿Hay datos ausentes (missing)? ¿Tienen algún patrón sistemático? ¿Cómo tratarlos?

***Etapas del EDA – Parte 1***

Para realizar un EDA conviene seguir las siguientes etapas o fases:

1) Preparar los datos para hacerlos accesibles a cualquier técnica estadística.

2) Realizar un examen gráfico de la naturaleza de las variables individuales a analizar y un

análisis estadístico que permita cuantificar algunos aspectos gráficos de los datos.

3) Analizar correlaciones entre las variables y dependencias.

4) Evaluar, si fuera necesario, algunos supuestos sobre la distribución de las variables, asimetrías, formas, etc.

5) Identificar los posibles casos atípicos (outliers) y determinar el impacto potencial que

puedan ejercer en análisis estadísticos posteriores.

6) Establecer si fuera necesario, el impacto potencial que pueden tener los datos ausentes (missing) sobre la representatividad de los datos analizados.

Entendamos juntos un poco más, en qué consiste cada fase 🤓

**Etapas del EDA – 1. Prep. de los Datos**

Como bien comentamos, el primer paso de un EDA es hacer accesible los datos a cualquier técnica estadística. Para ello, tendremos que realizar un input de los datos, los cuales recordemos pueden provenir de diferentes orígenes como ser por ejemplo: Excel, csv, Bases de Datos, etc. Luego tendremos que elegir el software de analítica de datos que utilizaremos para la manipulación y el procesamiento del dataset. En nuestro caso utilizaremos Python.



La gran mayoría de los softwares orientados al análisis de datos, permiten realizar manipulaciones de los datos previas a un análisis de los mismos. Algunas operaciones útiles para realizar son las siguientes:

- Combinar conjuntos de datos de dos o más archivos distintos.

- Seleccionar subconjuntos de los datos.

- Dividir el archivo de los datos en varias partes.

- Transformar variables.

- Filtrar y ordenar el dataset.

- Agregar nuevos datos y/o variables.

- Eliminar datos y/o variables.

- Guardar datos y/o resultados.

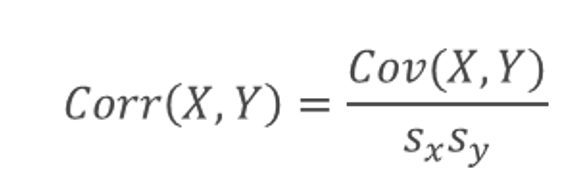
**Etapas del EDA – 2. Análisis Estadístico**

Una vez organizados los datos, el segundo paso dentro de un EDA consiste en realizar un análisis estadístico gráfico y numérico de las variables del dataset, con el fin de tener una idea inicial de la información que se encuentra contenida en el conjunto de datos, así como detectar también en el caso de que existan posibles errores de codificación.

Dentro de este contexto, es importante entender que el tipo de análisis que deberemos realizar va a depende de la escala de medida de la variable analizada.

***Etapas del EDA –* 3. Correlaciones y Dependencias**

La correlación es la covarianza pero dividida por los desvíos estándares de las dos variables. Presenta la siguiente fórmula matemática:



La correlación siempre va a darnos un número entre **-1 y 1**

* Mientras más cerca nos de del valor 1, más fuerte es la **relación lineal directa** entre las variables.
* Mientras más cerca nos de del valor -1, más fuerte es la **relación lineal inversa** entre las variables.
* Si nos da 0 entonces **no hay relación lineal** entre las variables.

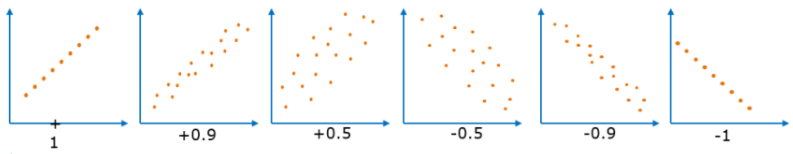
En lo que respecta a la fuerza de la correlación, hablamos siempre de una correlación:

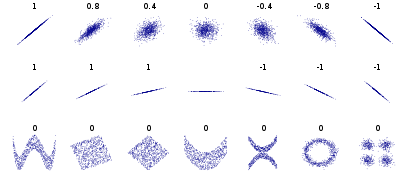
* Nula.
* Débil.
* Fuerte.

También es importante tener en cuenta 2 aspectos relevantes de destacar:

1. La ausencia de correlación significa que no hay una relación lineal, pero no que no hay relación.
2. Correlación no es, ni implica, causalidad.

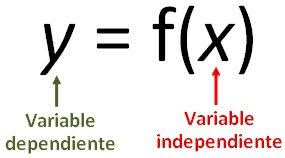
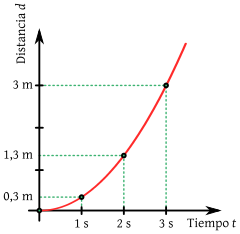
**Correlaciones:**

****

****

**Dependencias:**

La variable dependiente es aquella cuyo valor depende del valor numérico que adopta la variable independiente dentro en la función matemática.



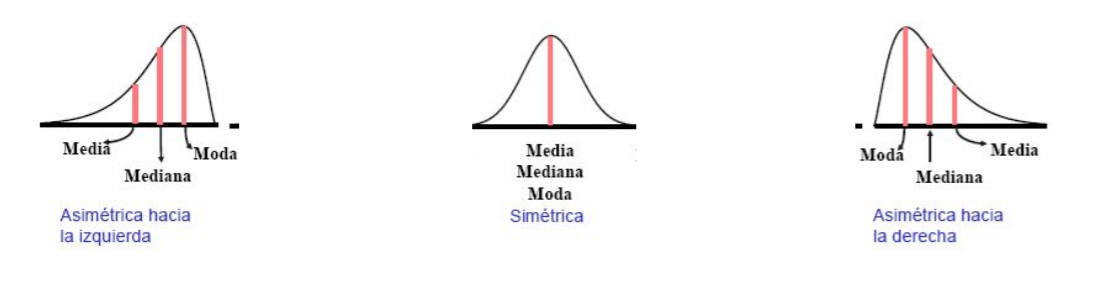
**Etapas del EDA – 4. Distribución de las variables**

En este paso resulta importante estudiar por ejemplo, las “Medidas de Forma” dentro del ámbito de la Estadística.

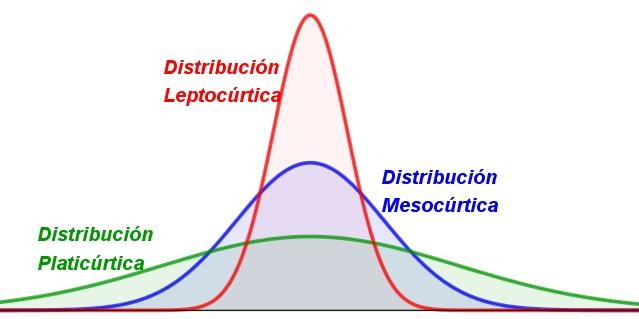
Pero ¿Qué son las medidas de forma? Son aquellas que estudian las características de la distribución de probabilidades observada. Podemos destacar:

* Simetría.
* Curtosis.

**Simetría:** Una variable es simétrica, si los valores que equidistan de la media son iguales. Para una mayor comprensión observemos la siguiente imagen:

****

**Curtosis:** La curtosis mide el grado de apuntamiento o achatamiento de la distribución de frecuenta. Es decir, nos ayuda a entender “cuán empinada está la curva”. Adicionalmente, existen diferentes tipos de curtosis:

****

***Etapas del EDA –* 5. Identificación de Outliers**

Como ya hemos estudiado en otras unidades del curso, tenemos que prestar especial atención a los outliers, dado que pueden tener un potencial negativo dentro de nuestro EDA.

También, es muy importante aclarar que no debemos eliminar los outliers por el sólo hecho de ser outliers. A menos que estemos 100% seguros que ese valor extremo se debe a un error de registro, una falla en el instrumento de medición o algún problema externo que sea verificable, los outliers son observaciones tan válidas como cualquier otra y forman parte de la realidad de nuestros datos. 😀

***Etapas del EDA –* 6. Valores Missings**

Una situación a la que se enfrenta frecuentemente cualquier científico de datos es el tratamiento de los valores perdidos. Los valores faltantes son aquellos que para una variable determinada no constan en algunas filas o patrones.

Los 3 motivos principales por los que se suelen tratar los valores perdidos son:

* Pueden introducir un sesgo considerable (una diferencia notable entre los datos observados y los no observados).
* Hacen el análisis y el manejo de los datos más complicado.
* Generalmente ocasionan pérdidas de información.

Existen multitud de procedimientos para aplicar cuando tenemos valores perdidos. Aunque básicamente existen dos aproximaciones posibles:

* Eliminar muestras o variables que tienen datos faltantes.
* Imputar los valores perdidos, es decir, sustituirlos por estimaciones.

