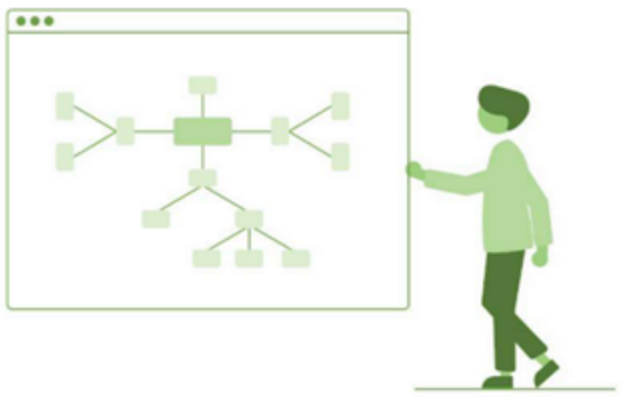
¿Cómo hacer un análisis exploratorio de datos?

Antes de preguntarnos ¿cómo hacer un análisis exploratorio de datos? es bueno preguntarse **¿por qué deberías realizar un análisis exploratorio de datos?**, entonces

¿qué razones tenemos para hacerlo?

1. **Organizar y entender las variables:** Al realizar un análisis exploratorio de datos, lograremos identificar los tipos de variables, las categorías a las que pertenecen y por consiguiente el tipo de análisis que podemos realizar con ellas.

2. **Establecer relaciones entre las variables:** Además al realizar un análisis exploratorio de datos lograremos establecer las relaciones entre las variables; esta variable se modifica si la gráfico junto con esta otra, estás 2 variables no tienen ningún tipo de relación entre ellas. Así lograremos identificar algunas relaciones.



3. **Encontrar patrones ocultos en los datos:** Nos percataremos de patrones; cosas que si no hubiéramos analizado, explorado nadie mas se hubiera dado cuenta. En el momento de realizar un análisis exploratorio de datos estamos en un punto de descubrir nuevas cosas.

4. **Ayudarte a escoger el modelo correcto para la necesidad correcta:** Muchas veces nosotros queremos empezar con un conjunto de datos y crear un modelo que nos prediga x cosa, sin embargo si no hacemos un análisis exploratorio de datos para identificar que variables sirven y que variables no sirven o como están relacionadas, cuales son las variables que mejor se ajustan ya sea a un tipo de modelo u otro, muy probablemente el modelo que vayamos a construir probablemente sea deficiente. Al hacer el análisis exploratorio de datos podremos encontrar el modelo correcto para la necesidad correcta.

5. **Ayudarte a tomar decisiones informadas:** Así podrás tomar decisiones basadas en datos, en la exploración y las relaciones que encontremos en nuestras variables; es decir esos patrones ocultos y esos modelos que generemos a través de la exploración.

¿Cuales son los pasos para realizar un análisis exploratorio de datos?

Los podemos dividir principalmente en 5 pasos:

1. Hacer preguntas.
2. Determinar el tamaño de los datos.
3. Categorizar las variables.
4. Limpieza y validación de datos.
5. Establecer relaciones.

Hacer preguntas

Necesitamos tener preguntas para nuestros datos. Una vez visualicemos el conjunto de datos, empecemos a hacernos preguntas, es decir toda clase de preguntas:

- ¿qué te gustaría encontrar en tus datos?
- ¿qué quieras saber sobre los datos?
- ¿cuál es la razón para realizar el análisis?

A traves de generar estas preguntas seremos capaces de dar una guía a todo el proceso de exploración de datos.

Determinar el tamaño de los datos

Con este paso nos vamos a hacer preguntas como:

- ¿cuantas observaciones existen?
- ¿cuantas variables hay?
- ¿necesito todas las observaciones?
- ¿necesito todas las variables? Digamos ¿podré tener variables que no me aportan información útil?, este tipo de preguntas serán muy recurrentes.

Categorizar nuestras variables

Aquí nos vamos a preguntar:

- ¿cuántas variables categóricas existen?
- ¿cuántas variables continuas existen?
- ¿cómo puedo explorar cada variable dependiendo de su categoría? Es decir preguntarnos y cuestionarnos que análisis puedo hacer.

Limpieza y validación de datos

Ahora que ya sabemos cual es el tamaño de los datos, cuantas variables hay. Nos tenemos que preguntar:

- ¿tengo valores faltantes?
- ¿cuál es la proporción de datos faltantes?
- ¿cómo puedo tratar los datos faltantes?
- ¿cuál es la distribución de los datos?
- ¿tengo valores atípicos? Es decir ver si hay una razón por la que mis datos están ausentes, y en caso de que falten ¿será que puedo rellenarlos con otros valores de mi conjunto?, así podemos cuestionarnos para saber como lidiar con ellos. También es bueno para saber que hacer con valores que están en los extremos, ya sea muy altos o muy pequeños.

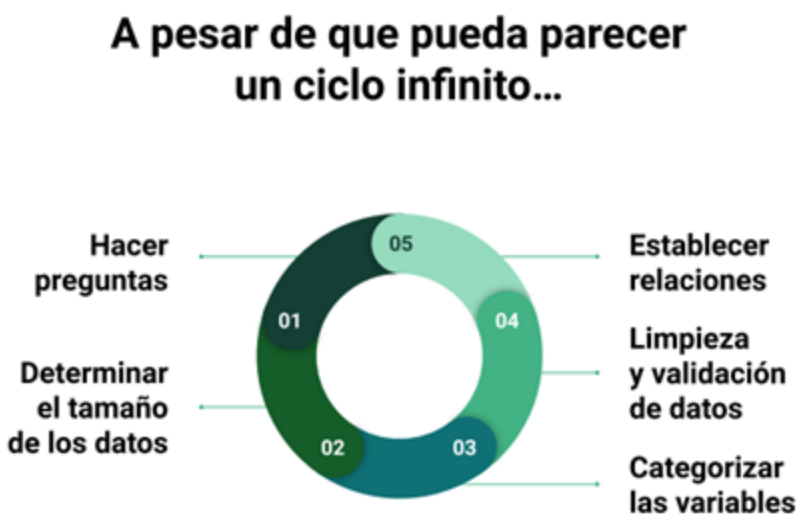
Este paso nos va a ayudar a determinar la respuesta a todas estas preguntas.

Establecer relaciones

Finalmente tenemos un paso en el que vamos a establecer las relaciones entre nuestros conjuntos de datos y nos vamos a hacer preguntas como:

- ¿existe un relación entre mi variable X e Y?
- ¿qué pasa si ahora considero a la variable Z en mi análisis?
- ¿qué significa que las observaciones se agrupen?
- ¿qué significa el patrón que se observa? Cosas extrañas podrían pasar en este proceso, también podremos preguntarnos que pasa con las proporciones; es decir para un conjunto de datos tengo miles de observaciones y para otra solo tengo unas cuantas. Es decir buscar si hay o no explicaciones con mis datos a los patrones que presenten, es decir para encontrar o extraer más información perdida dentro de mis datos y así dar una respuesta a estas preguntas.

Estos son los pasos que debemos seguir, pero hay que contemplar que no es un proceso lineal, mas bien puede parecer un proceso cíclico. Y podremos repetir ciertos pasos nuevamente, es decir:



Una vez que terminemos lo vamos a volver a realizar; vamos a filtrar datos, realizar nuevas preguntas, volver a calcular el tamaño de los datos y así nuevamente.

> **A pesar de que pueda parecer un ciclo infinito y que no se acaba** seguir en el no tiene sentido, y tiene que existir un momento en el que debemos romper el ciclo y continuar.

Un análisis exploratorio de datos que nunca termina en realidad tiene un valor nulo; porque nunca le pudimos entregar la información a alguien o nunca pudimos realizar algo que genere impacto con estos datos en la empresa, en el trabajo o en alguna determinada razón que tengamos para realizarlo.