



# DATOS, ETL Y ELT

ING. JAIRO SALAZAR

# Exploracion de Datos

---

Para comenzar a aprovechar mejor los datos, considere un aspecto importante, la **exploración**. La clave para una exploración de datos exitosa es formular buenas preguntas.

---

Por ejemplo, quizá le gusten los perros y se preocupe por su salud. Una pregunta que podría hacerse es: “*¿Tener un perro es bueno para la salud?*”

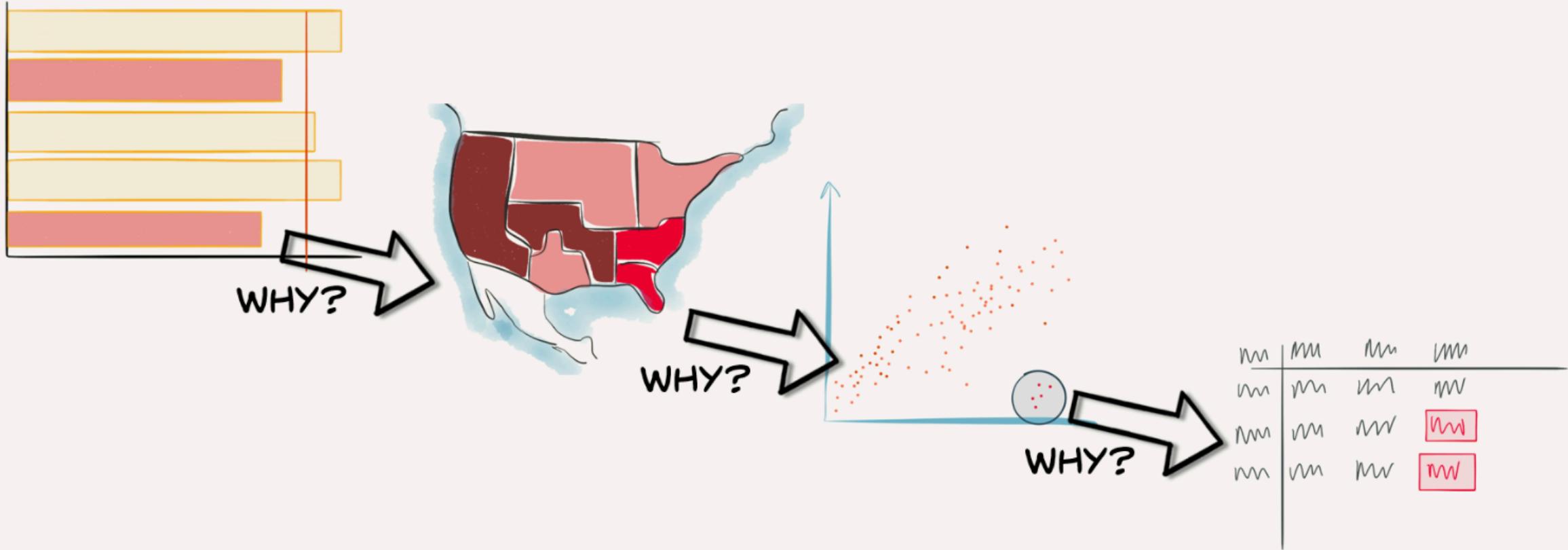
---

Ahora, compare esa pregunta con esta otra: “*¿Cómo son los resultados de salud de las personas que tienen perros en comparación con las que no tienen entre la población con enfermedades crónicas en los Estados Unidos?*”.

# Exploracion de Datos

La primera pregunta es amplia y no establece criterios claros para lo que se considera “bueno”. La segunda pregunta es mucho más específica.

Utiliza términos claramente definidos y limita el enfoque a una población específica. Será mucho más fácil explorar los datos para responder la segunda pregunta que la primera.



## LA TECNICA DE LOS “5 PORQUES” - Sakichi Toyoda

En pocas palabras, se debe preguntar el “porqué” de un problema que haya identificado y continuar haciéndolo por cada respuesta o explicación. El objetivo principal de la técnica es determinar la raíz de un defecto para poder solucionarlo.

## SALES METRICS ANALYSIS AND KPIs

Region  
All

Profit  
Sales

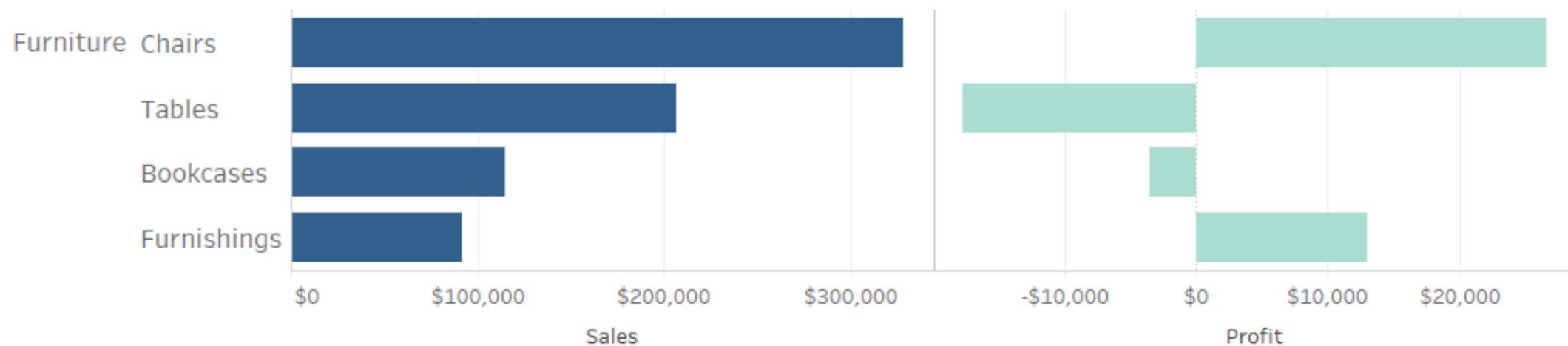
How are sales and profit in our main categories? *(most recent period)*



Aplicaremos un nuevo formato a los datos para dar respuesta a nuestras preguntas. Por ejemplo, podríamos analizar las ventas y los beneficios de los diversos tipos de mobiliario que vendemos:

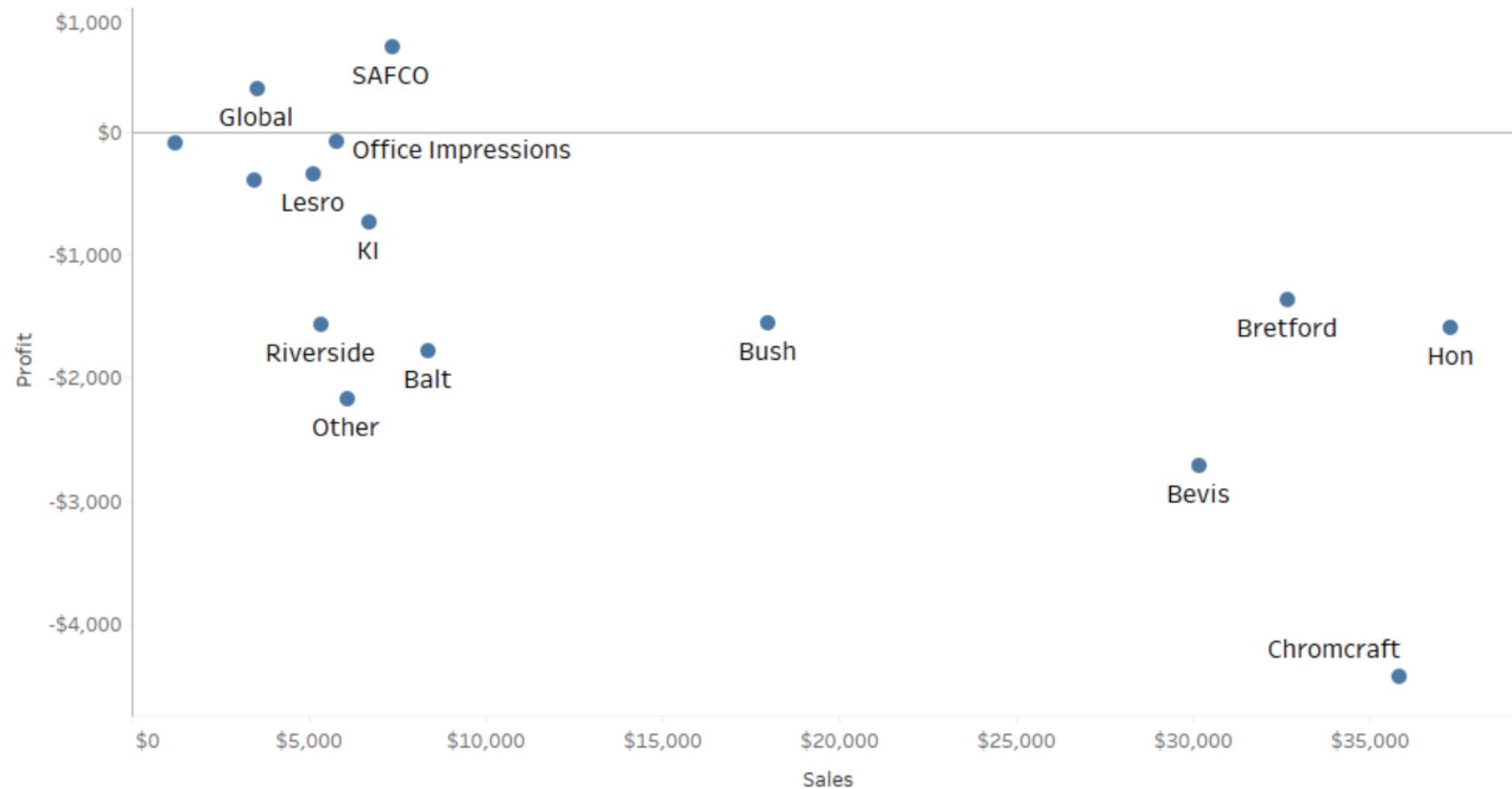
### *Asking why, step 1*

#### *Sales and profit of Furniture categories*



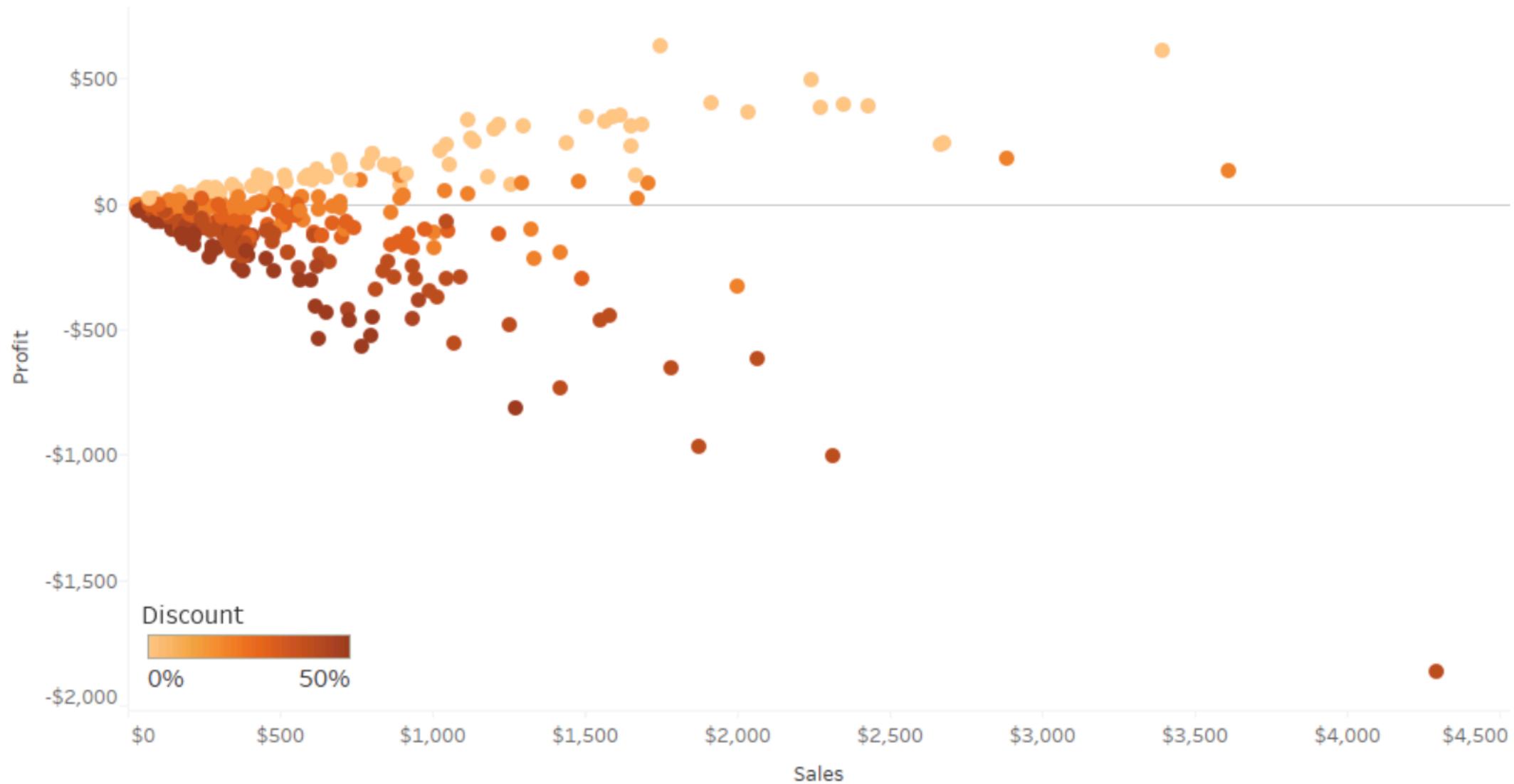
Quizás debamos investigar las ventas y los beneficios según los fabricantes de mesas:

*Asking why, step 2:  
Sales and Profit for all Tables*



*Asking why, step 4:*

*Sales and Profit for all Tables, and Orders. Colour shows Discount.*



---

**“Creemos en el triunfo de los hechos. Un conjunto de datos es un conjunto de hechos, y los hechos son los cimientos del progreso de la humanidad”.**

-- Sitio web de Tableau Software

# Que son exactamente los datos?

Una recopilación de datos es un conjunto de hechos. Explore esta definición aún más específica. Jeffrey Leek, un científico de datos que trabaja como profesor en la Johns Hopkins Bloomberg School of Public Health, adaptó esta definición ampliada de Wikipedia:

*“Los datos se componen de valores de variables cualitativas o cuantitativas que pertenecen a un conjunto de elementos”.*

ningun orden nominales

ordinales que llevan un orden

# Definiciones Varias

---

**Conjunto de elementos:** Hace referencia al grupo de objetos en el que centrará su atención. A veces recibe el nombre de “población”.

**Variable:** Hace referencia a una medida, propiedad o característica de un elemento que puede variar o cambiar. Nota: Esto se opone a una medida constante, como el número Pi, que no varía.

**Variable cualitativa:** Una variable cualitativa describe cualidades o características, como el país de origen, el sexo, el nombre o el color de cabello.

**Variable cuantitativa:** Una variable cuantitativa hace referencia a características medibles, como la altura, el peso o la temperatura.

# ¿Cómo se recopilan los datos?

Los datos se pueden recopilar de diversas formas, por ejemplo por medio de cuestionarios, entrevistas, observaciones, análisis de documentos, extracción web y mediciones automatizadas.

Los datos recibidos o recopilados se conocen como **datos sin procesar**. Los datos sin procesar, que también se conocen como datos de origen o datos primarios, son aquellos que no se han procesado de ninguna manera.

Esto significa que no se ejecutaron con ningún software, que no se modificaron variables, que no se eliminaron datos del conjunto y que no se resumieron de ninguna manera. Los datos sin procesar representan el conjunto más completo posible de datos para el análisis, ya que no se eliminaron ni resumieron.

# Características de los datos relevantes

## Volumen

Contar con una gran cantidad de datos relevantes disponibles significa que existen más posibilidades de que cuente con lo que necesita para responder sus preguntas.

**Nota:** No es necesario acumular datos porque sí. La relevancia es importante.

## Historial

Los datos que se remontan al pasado permiten ver cómo surgió una situación actual como consecuencia de patrones a lo largo del tiempo. Por ejemplo, al analizar las tendencias de ventas en los últimos 10 años para identificar aumentos o disminuciones.



# Características de los datos relevantes

---

## Coherencia

A medida que las cosas cambian, los datos deben adaptarse para mantener la coherencia. Los datos sobre salarios y precios ajustados en función de la inflación son un buen ejemplo de ello.

## Múltiples variables

Los datos deben incluir variables cuantitativas (medibles numéricamente) y cualitativas (características no medibles numéricamente). Cuantas más variables haya en los datos, más información podrá obtener de ellos.

# Características de los datos relevantes

## **Nivel de detalle**

Cuanto más detallados son los datos, más fácil es examinarlos en varios niveles de detalle. Por ejemplo, si quisiera conocer las tendencias en el uso de bicicletas en su estado, resultaría útil ver estas tendencias en función del condado, la ciudad y el barrio.

## **Pulcritud**

Para que los datos sean significativos, no deben ser inexactos ni estar incompletos, ni tampoco deben contener errores.

## **Claridad**

Los datos deben escribirse en términos que se puedan entender fácilmente, no en código.

# Características de los datos relevantes

## Estructura dimensional

Una buena forma de estructurar los datos es organizarlos en dos tipos: **dimensiones** (valores cualitativos) y **medidas** (valores cuantitativos).

## Segmentos

Los grupos, basados en características similares, deben integrarse en los datos para facilitar el análisis. Por ejemplo, los datos sobre películas pueden agruparse por género (acción, ciencia ficción, románticas, comedia, etc.).

- **Origen transparente**

Para poder confiar en los datos, es necesario saber que provienen de una fuente confiable y que se han administrado de manera segura.



# Tipos de variables cualitativas

Las variables cualitativas hacen referencia a características o cualidades. Estas se pueden clasificar además en dos tipos: nominales y ordinales.

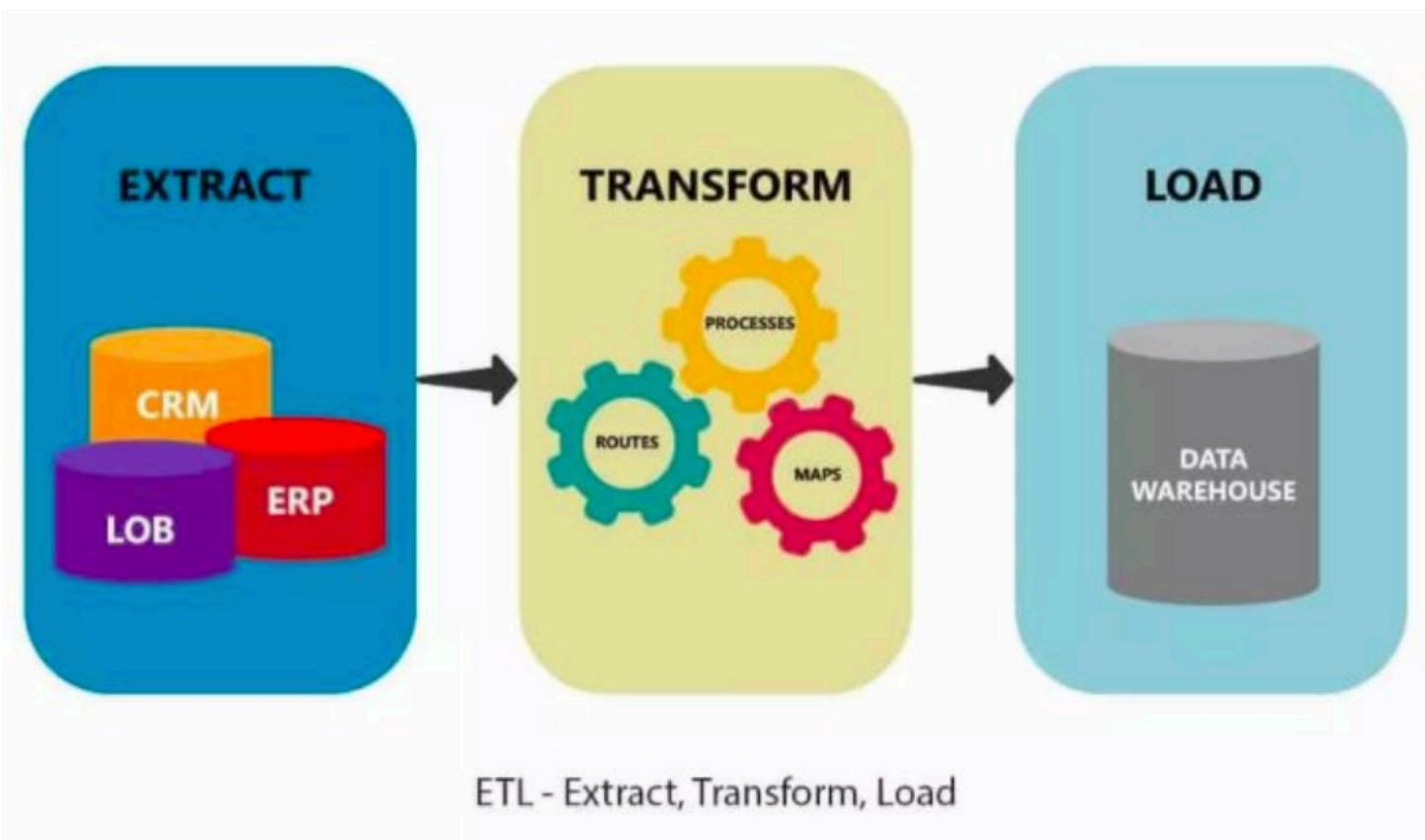
Las variables cualitativas nominales son categorías que no se pueden clasificar. Veamos, por ejemplo, algunos tipos de frutas: naranjas, uvas, peras y manzanas. Son variables nominales porque no existe un orden de clasificación implícito entre ellas. Una naranja, por ejemplo, no ocupa un lugar más alto en la clasificación que una pera.

Las variables cualitativas ordinales se pueden clasificar. Son cualitativas porque no se pueden medir numéricamente, pero hay un orden lógico entre ellas. Por ejemplo, piense en las encuestas que haya respondido. Algunos ejemplos de valores cualitativos ordinales en encuestas son:

Nunca, A veces, Con frecuencia, Siempre

Extremadamente insatisfecho, Insatisfecho, Ni satisfecho ni insatisfecho, Satisfecho, Extremadamente satisfecho

# ETL - EXTRACT TRANSFORM LOAD -



ETL (Extract/Transform/Load) ETL (Extract/Transform/Load) es una técnica de integración que obtiene la información de fuentes de datos remotas, la transforma en los formatos y estilos definidos y luego lo carga en las bases de datos, fuentes de datos o data warehouses.

El proceso de extracción, implica obtener datos de las fuentes de datos. Durante esta fase, los datos son leídos y reunidos, a veces de numerosas y diversas fuentes de datos.

Durante la transformación, los datos son extraídos y luego convertidos en un formato que sea aceptable para otra base de datos. En esta fase, los datos son transformados utilizando expresiones, reglas, buscándolas en tablas o uniéndose con otros sets de datos.

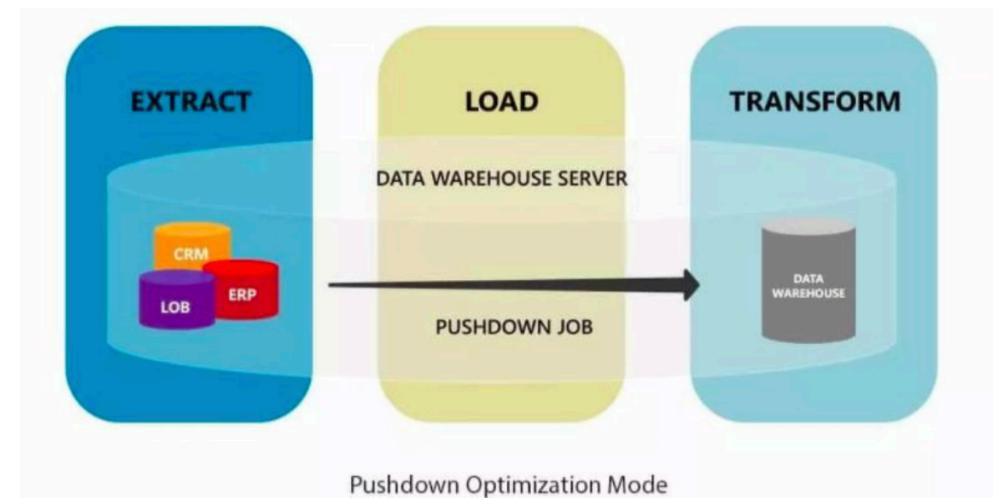
El último paso es la carga, que es el procedimiento de escritura de los datos en el datawarehouse elegido.

# ELT – EXTRACT LOAD TRANSFORM

Similar al ETL, Extrae los datos de una o múltiples fuentes de datos, pero lo carga en la base de datos de datawarehouse sin ningún formato.

El proceso de transformación, ocurre ya en la base de datos objetivo. A diferencia del ETL, donde la transformación de los datos ocurre en un área denominada “staging”, antes de ser cargada en el sistema objetivo.

Bajo este esquema, se reduce considerablemente la carga de los datos. Es un método más eficiente en relación a la utilización de los recursos, pues aprovecha mejor los recursos, disminuyendo el tiempo invertido en la transferencia de datos.



# Tableau

