# ANÁLISIS DESCRIPTIVO: EXPLORACIÓN DE DATOS

Partimos de la base de que disponemos ya de un conjunto de datos susceptible de ser explotado. El grado de acondicionamiento y modelado de estos dependerá de la arquitectura elegida para el entorno de información. Esto es, podemos disponer de datos en bruto,donde el grado de transformación ha sido nulo o minimo y el esquema se establece en el momento de lectura,como en un data lake, o de datos procesados, con un mayor o menor grado de estructura de cara a la optimización del acceso, como en un
data warehouse o lakehouse. Simultáneamente,estos datos podrán ser estáticos,hasta la próxima carga ETL,o formarán parte de un flujo continuo de eventos, que hay que consumir al vuelo. En principio,cualquier combinación de estos escenarios es posible, estando nuestro objetivo en la explotación efectiva de toda esa información,empleando para ello tanto recursos locales como servicios en la nube.

## 1. MOTIVACIÓN Y OBJETIVOS

El objetivo del análisis descriptivo es estudiar de forma retrospectiva los datos con el fin de entender qué ha sucedido y las causas que hay detrásis1. Para ello se vale de la estadistica descriptiva,que le proporciona técnicas y medidas para resumir y explicar cómo se distribuyen y relacionan los datos entre si, y de las técnicas de representación gráfica,que permiten una visualización efectiva de la información con el fin de entenderla y, todavía más importante,transmitirla de forma adecuada.

Aunque uno de los objetivos del análisis descriptivo es la identificación de tendencias, ien cualitativo, de ahí la importancia de las técnicas de visualización. La cuantificación concretamente de la mineria de datosiss, aunque las fronteras siempre son difusas. No hay conceptual,En la práctica, in analista de datos puede utilizar una combinación de ellas a lo largo de las diferentes etapas de su estudio.

En definitiva,el análisis descriptivo trata de obtener un resumen y una representación adecuada de los datos que permita establecer un conocimiento claro y contextualizado del estado del negocio y de su evolución temporal. Para esta última podemos tener en cuenta tanto una alta profundidad histórica, como eventos relevantes que se están produciendo en tiempo real. Vamos a empezar por entender cuál es el foco de nuestro análisis y las formas en las que podemos describir los datos, tanto numérica como gráficamente.

## 2. CARACTERIZACIÓN DE LOS DATOS

Desde un punto de vista estadístico, muchos de los análisis que hacen las organizaciones empresariales son **estudios observacionales**. Esto quiere decir que los datos analizados no son el resultado de un experimento que se puede diseñar y repetir de forma controlada estando hasta cierto punto indefinido el conjunto de sus posibles resultados. Por el contrario, son fenómenos aleatorios que son autónomos, no se pueden reproducir a voluntad, ni tampoco es posible controlar unos factores para calibrar la influencia que tiene sobre otro otros. El análisis del comportamiento de un cliente en cuanto a sus preferencias, hábitos y percepciones, o la evolución de las ventas en función de la geografía y de la época del año, son ejemplos de estudios observacionales. Por el contrario, encuestas de opinión para conocer la percepción que tienen los clientes sobre un nuevo producto, o purebas A/B para determinar que interfaz de aplicación web

favorece más la compre en linea, son experimentos donde existe un control sobre la reproducibilidad de los resultados.

Con independencia de si el fenómeno a estudiar es de carácter observacional experimental, el objetivo del análisis exploratorio de los datos (EDA, Exploraton Data Analysis) es investigar los datos recopilados con el fin de resumir y presentar sus principales características, identificando posibles relaciones y tendencias dentro de ellos. Es importante partir de la base que el EDA no tiene una vocación formal. Como comentábamos al principio del capitulo,la cuantificación y confirmación de posibles patrones y asociaciones dentro de los datos requiere el desarrollo y validación de un modelo de base estadística, algo que cubriremos en el siguiente capitulo.

| Familia | Tipo | Descripción y operación | Ejemplos |
|---|---|---|---|
| Cualitativos | Nominal | Los valores son símbolos diferentes, proporcionando la información necesaria para distinguir entre observaciones<br>Distinción(=, !=) | ·Código postal<br>·Identificador de empleado<br>·Color de los ojos |
| | Ordinal | Los valores proporcionan suficiente información para ordenar las observaciones<br>Ordenación(<, <=, >=, >) | ·Dureza en un mineral ((baja,media,alta})<br>·Calificaciones<br>·Número de vivienda |
| Cuantitativos | Intervalo | Las diferencias entre valores tienen sentido,existiendo una unidad de medida. El valor cero es arbitrario.<br>Adición(+,-) | ·Fechas y horas<br>·Temperatura en grados Celsius<br>·pH |
| | Ratio | Tanto las diferencias como los ratios tienen sentido. El valor cero es absoluto;no existen valores negativos<br>Multiplicación(x,÷) | ·Número de aviones<br>·Importe monetario<br>·Temperatura en grados Kelvin |

Tabla 6-1. Clasificación de atributos según el tipo de operaciones que soportan.

## 2.1. Observaciones y atributos

El resultado de un fenómeno objeto de estudio es, por lo tanto, uno o varios **conjuntos de datos**. Cada uno de ellos está organizado como una colección de **observaciones**, descritas a su vez por una serie de **atributos**, cada uno de los cuales captura una característica o una medida de estas. Los atributos acostumbran a ser consistentes a lo largo de las distintas observaciones, tanto en significado, tipología y número. Conceptualmente, un conjuntos de datos lo representamos como una arreglo tabular, donde cada columna contiene un atributo particular, cada fila se corresponde con una observación, y la intersección entre ambas es el valor que toma el atributo para esa observación.

| Rol | Descripción |
|---|---|
| **Identificativo** | Atributo que participa en la identificación de la observación, de forma individual o formando parte de una clave única. Puede aportar algún tipo de significado o ser artificial. |
| **Objetivo** | Atributo dependiente (típicamente solo uno) sobre el cual se deseen probar los efectos de una cierta interacción en otros atributos. Los estudios puramente descriptivos no contienen atributos objetivo. |
| **Explicativo** | Atributo independiente susceptible de afectar a los atributos objetivo en el análisis, con independencia de si se manipula de forma explícita o no, y que es objeto de interés. |
| **Control** | Atributo que permanece constante durante la realización del análisis con el fin de minimizar su efecto sobre si misma, no siendo objeto de interés. |
| **Extrínseco** | Todo atributo que no forma parte del análisis y no está bajo control, pudiendo influir en la variable objetivo y alterar la interpretación de los resultados obtenidos. |

Tabla 6-2. Tipos de atributos según su rol en el análisis.

Las observaciones vendrán determinadas por la naturaleza de la población objeto de estudio y su contexto, pudiéndose corresponder con clientes, cestas de la compra, llamadas telefónicas, transacciones de compra, fotografías en una cadena de montaje, etc. Respecto a los atributos, estos se pueden clasificar según las operaciones que soportan (Tabla6-1), y también según el número de distintos valores que pueden tomar. En este último caso hablamos de **atributos discretos**, cuando el número es finito o contable, **binarios**, cuando el número de posibles valores se restringe a dos, o **continuos**, cuando los valores son números reales (*). La caracterización de un atributo según estas clasificaciones es importante,ya que no solo indica el tipo de información que recoge, sino que define también el tipo de operaciones y transformaciones que puede soportar, así como la forma más adecuada a la hora de describirlo, resumirlo y representarlo gráficamente.

Otra clasificación importante de los atributos es según el **rol** que desempeñan en el análisis,de forma que este puede variar de uno a otro. La Tabla 6-2 contiene los 5 roles principales. Hay que tener en cuenta que en estudios observacionales es difícil, por no decir imposible,tener atributos de control, de forma que los atributos extrinsecos pueden llegar a dominar el estudio, distorsionando los resultados y llevándonos a conclusiones erróneas.

```
(*)Si bien no existe una relación univoca entre la clasificación de los atributos
según la tipología y según el distinto número de valores que puedan tomar, los
atributos continuos se corresponden con tipos cuantitativos y los binarios con
cualitativos-nominales. Respecto a los valores discretos estos acostumbran a ser
cualitativos, aunque un conteo es un atributo discreto de tipo cuantitativo-
ratio.
```

Tomemos como ejemplo una investigación sobre el crecimiento de un conjunto de plantas. Se desea analizar si existen especies de plantas que son más tolerantes a las sales que otras, teniendo en cuenta la cantidad de estas que se añade al agua de riego, la especie de la planta, así como otros factores relacionados con la salud del ejemplar, como el crecimiento y el marchitado.

| # | Especie | Transgénica (1:si,0:no) | Sal añadida (mg/l) | Altura inicial (cm) | Crecimiento (cm) | Marchitado (rango 0-5) |
|---|---|---|---|---|---|---|
| 1 | A | 0 | 0 | 12 | 5 | 0 |
| 2 | A | 1 | 100 | 13 | 10 | 0 |
| 3 | A | 1 | 250 | 11 | 2 | 3 |
| 4 | B | 0 | 0 | 25 | 8 | 0 |
| 5 | B | 0 | 100 | 26 | 12 | 0 |
| 6 | B | 0 | 250 | 25 | 3 | 3 |
| 7 | B | 1 | 350 | 30 | 0 | 4 |

Tabla 6-3. Crecimiento de distintos ejemplares de plantas registrados al cabo de 1 mes.

Para ello se recogen distintas observaciones después de 1 mes de riego cada 5 dias (Tabla 6-3). En este análisis, el numero de planta seria un elemento identificativo: unicamente sirve para identificar la observacion y no es de interes de cara el analisis. EI crecimiento y el nivel de marchitado serian los atributos objetivo, ya que representan el resultado del experimento. La especie y la sal añadida son los factores que se manipulan para alterar el resultado y son, por lo tanto, atributos explicativos. La condicion de planta transgenica seria aqui tambien un atributo explicativo: aunque no se altera de forma xplicita (las distintas observaciones que participan en el experimento no han sido elegidas teniendo en cuenta esta condicion), se desea estudiar resultados. La altura inicial no tiene un rol definido si afecta de alguna manera en el analisis, pero es necesaria pa I calcular el crecimiento al final de este. Por que se mantienen constantes durante el experimento: la temperatura, la luz de la sal donde estan las planas y el volumen de agua en cada riego.

Si miramos la tipología de los atributos, el identificador del ejemplar y la especie son atributos nominales y discretos. El indicador de planta transgénica también es nominal, pero en este caso es binario. La sal añadida, la altura inicial y el crecimiento son ratios continuos, mientras que el marchitado es un ordinal discreto. Respecto a los atributos de control, el nivel de luz y el volumen de agua en el riego serian ratios continuos.  continuos. La temperatura de la sala también es un intervalo es un atributo continuo, pero en este caso es un intervalo.

## 2.2. Relaciones entre atributos

Parte del objetivo del EDA consiste en la caracterización de las posibles relaciones entre pares de atributos. Dos atributos son **independientes** cuando el valor de uno no proporciona ninguna información sobre el del otro: las opciones de medir un posible valor en el primero no están afectadas por el valor del segundo, y viceversa. Cuando si existe esta afectación, se dice que los atributos son dependientes, existiendo una **asociación** entre ambas.

Cuando dos atributos asociados exhiben además una tendencia creciente o decreciente se dice que existe CORRELACIÓN entre ambos. Hay que resaltar que no toda asociación no es una correlación. En cualquier caso una relación de dependencia entre dos atributos no implica causalidad: la modificación de uno no es necesariamente la causa que provoca un efecto en el otro. Como hemos visto la existencia de un atributo extrínseco puede ser la explicación a la dependencia detectada.

La correlación es un tipo de asociación, pero ninguna de las dos implica la existencia de una relación causa-efecto. En sentido inverso, la causalidad siempre es el resultado de una asociación, pero no necesariamente de una correlación.

## 3. ANÁLISIS EXPLORATORIO

El análisis exploratorio de datos (AED) es el proceso inicial en el análisis de cualquier conjunto de datos que tiene como objetivo descubrir patrones, detectar anomalías, probar hipótesis y verificar supuestos a través de métodos estadísticos y gráficos. En esencia, se trata de "explorar" la información para entenderla a fondo antes de aplicar modelos más complejos o sacar conclusiones definitivas.

A continuación, te doy algunos ejemplos prácticos de lo que se puede hacer en un AED:

1. **Resumen de estadísticas descriptivas:**
    - **Ejemplo:** Supongamos que tienes un conjunto de datos con la edad y el salario de empleados. Puedes calcular la media, mediana, moda, desviación estándar y rango de estas variables para conocer la tendencia central y la dispersión de los datos.
    - *Interpretación:* Si la media de la edad es de 35 años pero la mediana es de 30, podría indicar que existen algunos valores muy altos (empleados mayores) que están sesgando la media.
2. **Visualización de la distribución de datos:**
    - **Ejemplo:** Para la variable "salario", puedes construir un histograma que muestre cómo se distribuyen los salarios entre los empleados.
    - *Interpretación:* Si el histograma muestra una cola larga hacia la derecha, indica que hay algunos empleados con salarios muy altos, lo cual puede ser relevante para análisis posteriores.
3. **Detección de valores atípicos (outliers):**
    - **Ejemplo:** Utiliza diagramas de caja (box plots) para la variable "edad" o "salario".
    - *Interpretación:* Si en el box plot observas puntos aislados fuera del "bigote" principal, estos representan valores atípicos que podrían necesitar una revisión adicional para decidir si se trata de errores de registro o casos genuinamente extremos.
4. **Análisis de relaciones entre variables:**
    - **Ejemplo:** Si además de "edad" y "salario" cuentas con variables como "años de experiencia", puedes crear un diagrama de dispersión (scatter plot) entre "años de experiencia" y "salario" para identificar si existe alguna relación lineal o patrón.
    - *Interpretación:* Una tendencia ascendente en el scatter plot podría indicar que, en general, a mayor experiencia, mayor es el salario.
5. **Estudio de la forma de la distribución:**
    - **Ejemplo:** Analiza la simetría (sesgo) y la curtosis de la variable "edad".
    - *Interpretación:* Si la distribución es asimétrica o tiene una curtosis elevada, esto puede influir en la elección de técnicas estadísticas o de modelado en pasos posteriores.
6. **Verificación de la calidad de los datos:**
    - **Ejemplo:** Revisa si hay datos faltantes o inconsistentes en el conjunto de datos.
    - *Interpretación:* Detectar celdas vacías o valores fuera de rango (como una edad negativa) es crucial para limpiar los datos antes de realizar análisis más profundos.

En resumen, el análisis exploratorio es fundamental para obtener una visión global de los datos, identificar problemas potenciales y generar ideas o hipótesis sobre lo que realmente está ocurriendo en la información. Este proceso te prepara para aplicar técnicas de análisis más avanzadas con una base sólida y bien comprendida.

## 3.1. Análisis Univariante

El análisis univariante es el proceso estadístico que se enfoca en describir y resumir las características de una única variable. Es el primer paso en cualquier análisis de datos, ya que permite entender la distribución, tendencias centrales y la dispersión de los datos. Aquí algunos puntos clave:

1. **Medidas de tendencia central**: Se calculan indicadores como la media, mediana y moda para conocer el "centro" de la distribución.
2. **Medidas de dispersión**: Se evalúa la variabilidad a través del rango, la varianza, la desviación estándar, entre otros, lo que ayuda a entender la extensión y la dispersión de los datos.
3. **Medidas de forma**: Se analiza la simetría (sesgo) y la curtosis para saber si la distribución es simétrica o si tiene colas pesadas.
4. **Visualización de datos**: Se utilizan gráficos como histogramas, diagramas de caja (box plots) y gráficos de barras para representar de manera visual la distribución de la variable.

En resumen, el análisis univariante es fundamental para obtener una primera impresión de los datos y para identificar posibles errores, valores atípicos o patrones que pueden influir en análisis posteriores.

En el contexto del análisis de datos, un **estadístico** es un valor numérico calculado a partir de una muestra de datos que resume alguna de sus características. Dicho de otro modo, es una medida que se extrae de los datos para describir, resumir o inferir propiedades sobre la población de la que provienen. El término "**estadístico univariante**" se refiere a cualquier medida o resumen numérico que se utiliza para describir y analizar la distribución de una única variable.

En función de las características que describen, podemos hablar de tres familias de estadísticos univariantes. Es importante resaltar que, a excepción de la moda, todas estas medidas solo se pueden calcular para atributos cuantitativos.

1. **Medidas de tendencia central:**
    - **Media:** Es el promedio de todos los valores de la variable. Por ejemplo, si tienes las edades de un grupo de personas, la media te indica la edad promedio.
    - **Mediana:** Es el valor que se encuentra en el centro al ordenar los datos. Es útil cuando la distribución tiene valores extremos, ya que no se ve tan afectada por ellos.
    - **Moda:** Es el valor que más se repite en el conjunto de datos.
2. **Medidas de dispersión:**
    - **Rango:** Diferencia entre el valor máximo y el mínimo.
    - **Desviación estándar y varianza:** Estas medidas indican cuánto varían o se dispersan los datos respecto a la media.
3. **Medidas de forma:**
    - **Sesgo (asimetría):** Indica si la distribución está sesgada hacia un lado.
    - **Curtosis:** Mide el grado de apuntamiento de la distribución comparada con una distribución normal.

Estos estadísticos son fundamentales en cualquier análisis porque permiten obtener una primera impresión sobre la distribución de los datos y ayudan a identificar posibles errores, valores atípicos o patrones importantes. En resumen, un estadístico es una herramienta clave para transformar grandes cantidades de datos en información comprensible y útil para la toma de decisiones.

Vamos a ver lo que representa uno de estos estadísticos, por ejemplo vamos a ver el rango intercuartílico.

Primero tenemos que saber qué es un **rango intercuartílico**:

El **rango intercuartílico (IQR)** es una medida de dispersión que indica la amplitud del "centro" de la distribución de datos, es decir, el rango en el que se encuentra el 50% central de los valores. Se calcula como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1):

$$IQR = Q3 - Q1$$

**¿Qué son los cuartiles?**

- **Q1 (Primer cuartil):** Es el valor que separa el 25% inferior de los datos del resto.
- **Q3 (Tercer cuartil):** Es el valor que separa el 25% superior de los datos del resto.
- **Mediana (Q2):** Divide la distribución en dos partes iguales.

**Ejemplo práctico**

Imagina que tienes la siguiente serie de datos ordenados:

```
10, 12, 14, 16, 18, 20, 22, 24, 26
```

1. **Mediana (Q2):**
   El valor central es 18.
2. **Q1 (primer cuartil):**
   Es la mediana de la mitad inferior (10, 12, 14, 16), que es el promedio entre 12 y 14, es decir, 13.
3. **Q3 (tercer cuartil):**
   Es la mediana de la mitad superior (18, 20, 22, 24, 26) o, en este caso, para igual número de datos se toma la mediana de (20, 22, 24, 26) calculada como el promedio entre 22 y 24, es decir, 23.
4. **Cálculo del IQR:**

$$IQR = Q3 - Q1 = 23 - 13 = 10$$

**¿Por qué es útil el IQR?**

- **Robustez:** Es una medida robusta, ya que no se ve afectada por valores atípicos o extremos.
- **Representación del núcleo de la distribución:** Permite conocer la dispersión del 50% central de los datos, brindando una idea de la variabilidad sin la influencia de datos extremos.
- **Identificación de outliers:** Se utiliza para detectar valores atípicos. Por ejemplo, se considera un valor fuera de rango si está por debajo de  $Q1 - 1.5 \times IQR$ o por encima de $Q3 + 1.5 \times IQR$.

En resumen, el rango intercuartílico es una herramienta muy útil en el análisis exploratorio de datos para comprender cómo se distribuyen los valores en el centro de una muestra, ayudando a evaluar la consistencia y detectar posibles anomalías.

Ejemplo:

Imaginemos que tenemos dos tiendas, la Tienda A y la Tienda B, que venden el mismo producto (por ejemplo, laptops). Queremos comparar la variabilidad de las ventas mensuales a lo largo de un año utilizando como estadístico el **rango intercuartílico (IQR)**.

**Paso 1: Recolección de datos**

Supongamos que para cada tienda se registraron las ventas mensuales (en unidades) durante 12 meses. Los datos (inventados) podrían ser:

- **Tienda A:** 45, 50, 48, 52, 47, 51, 49, 53, 46, 50, 48, 52
- **Tienda B:** 30, 65, 40, 55, 35, 60, 45, 70, 38, 68, 42, 66

**Paso 2: Ordenar los datos**

Para calcular el IQR, primero debemos ordenar los datos de cada tienda de menor a mayor.

- **Tienda A (ordenada):** 45, 46, 47, 48, 48, 49, 50, 50, 51, 52, 52, 53
- **Tienda B (ordenada):** 30, 35, 38, 40, 42, 45, 55, 60, 65, 66, 68, 70

**Paso 3: Calcular el primer (Q1) y tercer cuartil (Q3)**

Dado que tenemos 12 datos (n = 12), podemos dividir el conjunto en dos mitades.
Para conjuntos con un número par de datos, a menudo se calcula Q1 y Q3 como la mediana de la primera mitad y la segunda mitad, respectivamente.

**Tienda A:**

- Primera mitad:

  45, 46, 47, 48, 48, 49

  - La mediana de estos 6 datos se calcula como el promedio del 3er y 4to valor: (47 + 48)/2 = 47.5 → **Q1 = 47.5**
- Segunda mitad:

  50, 50, 51, 52, 52, 53

  - La mediana es (51 + 52)/2 = 51.5 → **Q3 = 51.5**

**Tienda B:**

- Primera mitad:

  30, 35, 38, 40, 42, 45

  - La mediana es (38 + 40)/2 = 39 → **Q1 = 39**
- Segunda mitad:

  55, 60, 65, 66, 68, 70

  - La mediana es (65 + 66)/2 = 65.5 → **Q3 = 65.5**

**Paso 4: Calcular el rango intercuartílico (IQR)**

El IQR se calcula como la diferencia entre Q3 y Q1.

- **Tienda A:** IQR = 51.5 − 47.5 = **4**
- **Tienda B:** IQR = 65.5 − 39 = **26.5**

**Paso 5: Interpretación**

- **Tienda A:** El IQR de 4 unidades indica que el 50% central de las ventas mensuales está concentrado en un rango bastante estrecho, lo cual sugiere que sus ventas mensuales son consistentes y con poca variabilidad.
- **Tienda B:** Un IQR de 26.5 unidades indica una mayor dispersión en las ventas mensuales. Esto sugiere que, en el 50% central, las ventas varían significativamente, lo que podría deberse a factores como promociones, variaciones en la demanda o diferencias en el flujo de clientes.

**Conclusión**

Utilizando el IQR como estadístico, podemos concluir que la **Tienda A** presenta una mayor consistencia en las ventas mensuales durante el año, mientras que la **Tienda B** muestra una mayor variabilidad en sus ventas. Este análisis ayuda a identificar que la Tienda B podría necesitar investigar las razones detrás de esta variabilidad para, por ejemplo, estabilizar sus ingresos o aprovechar picos de ventas, mientras que la Tienda A podría estar operando de manera más estable.

Me faltaría poner el esquema de cajas y bigotes, lo que es un histograma, gráfico de columnas, gráficos de barras, gráfico circular. mapa de árbol.

## 3.2. ANÁLISIS MULTIVARIANTE

Una **tabla de contingencia** (o tabulación cruzada) es una herramienta descriptiva que permite representar la distribución conjunta de dos o más variables categóricas. En otras palabras, muestra cómo se combinan las categorías de diferentes variables y cuántas veces ocurre cada combinación en el conjunto de datos.

**¿Para qué sirve?**

- **Visualizar relaciones:** Permite identificar si existe alguna asociación o relación entre las variables.
- **Resumir información:** Consolida la información de forma sencilla, mostrando conteos o frecuencias relativas para cada combinación de categorías.
- **Facilitar comparaciones:** Ayuda a comparar la distribución de una variable a través de los niveles de otra.

**Cómo se construye**

1. **Estructura básica:**
   Imagina dos variables:
   - **Variable A:** Género (Masculino, Femenino)
   - **Variable B:** Preferencia de producto (Producto X, Producto Y)

   La tabla de contingencia tendría filas y columnas. Por ejemplo, podríamos poner el género en las filas y la preferencia de producto en las columnas.

2. **Llenado de la tabla:**
   Cada celda muestra el número de casos que corresponden a la combinación de la categoría de la fila y la categoría de la columna. Por ejemplo:

| | **Producto X** | **Producto Y** | **Total** |
|---|---|---|---|
| Masculino | 30 | 20 | 50 |
| Femenino | 25 | 35 | 60 |
| **Total** | 55 | 55 | 110 |

   - Aquí, la celda (Masculino, Producto X) indica que 30 hombres prefirieron el Producto X.
   - Los totales de cada fila y columna permiten obtener una visión global de la distribución.
3. **Frecuencias relativas:**
   Además de los conteos absolutos, se pueden calcular frecuencias relativas (por ejemplo, el porcentaje que representa cada celda respecto al total) o incluso porcentajes marginales y condicionales, lo que enriquece el análisis.

**Ejemplo práctico**

Supongamos que en una encuesta a 110 personas se preguntó sobre su género y su preferencia de producto. Con los datos anteriores, la tabla nos permite observar:

- La distribución de preferencias entre hombres y mujeres.
- Que, por ejemplo, el 60% de las mujeres (35 de 60) prefirieron el Producto Y, mientras que en hombres, el Producto X es más popular (30 de 50).

**Importancia en el análisis de datos**

- **Exploración de relaciones:** Antes de realizar análisis más complejos, la tabla de contingencia ayuda a visualizar patrones, identificar posibles relaciones y detectar si la asociación entre variables es fuerte o débil.
- **Base para análisis inferencial:** Muchas pruebas estadísticas (como la prueba de chi-cuadrado) se basan en tablas de contingencia para determinar si las diferencias observadas entre categorías son estadísticamente significativas.

En resumen, la tabulación cruzada es una forma efectiva de representar y resumir la distribución conjunta de variables categóricas, permitiendo a los analistas explorar y comprender la relación entre estas variables de manera clara y visual.