




# EXPLORACIÓN DE LOS DATOS Y ANÁLISIS DESCRIPTIVO UNIDAD TEMA 6

*El propósito de esta tarea es que los estudiantes apliquen herramientas y técnicas de análisis descriptivo de datos, comprendiendo las características principales de los datos, realizando un análisis exploratorio y utilizando sistemas para crear visualizaciones y cuadros de mando.*



Andrei Alexandru Miu



## Índice

Caracterización de los datos .....	3
• Ejercicio 1: .....	3
Análisis exploratorio de datos .....	5
• Ejercicio 2: .....	5
• Ejercicio 3: .....	6
Análisis multidimensional .....	7
• Ejercicio 4: .....	7
Sistemas para análisis descriptivo .....	8
• Ejercicio 5: .....	8

## Caracterización de los datos

- Ejercicio 1:

1. Determinar:

- El tamaño del dataset (número de filas y columnas).

```
Filas: 9994, Columnas: 13
```

- Los tipos de datos de cada variable.

```
Data columns (total 13 columns):
#      Column      Non-Null Count  Dtype
---  -
0      Ship Mode    9994 non-null    object
1      Segment        9994 non-null    object
2      Country         9994 non-null    object
3      City            9994 non-null    object
4      State           9994 non-null    object
5      Postal Code     9994 non-null    int64
6      Region          9994 non-null    object
7      Category        9994 non-null    object
8      Sub-Category    9994 non-null    object
9      Sales           9994 non-null    float64
10     Quantity        9994 non-null    int64
11     Discount         9994 non-null    float64
12     Profit           9994 non-null    float64
dtypes: float64(3), int64(2), object(8)
```

- Los valores faltantes, si existen, y cómo se podrían manejar.

```
Valores nulos por columna:
Ship Mode      0
Segment        0
Country        0
City           0
State          0
Postal Code    0
Region         0
Category       0
Sub-Category   0
Sales          0
Quantity       0
Discount       0
Profit         0
```

No hay.

- La descripción estadística de las variables numéricas (mínimo, máximo, media, mediana, desviación estándar).

	Postal Code	Sales	Quantity	Discount	Profit
count	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000
mean	55190.379428	229.857901	3.789574	0.156203	28.656599
std	32063.693350	623.245124	2.225110	0.206452	234.260115
min	1040.000000	0.444000	1.000000	0.000000	-6599.978000
25%	23223.000000	17.280000	2.000000	0.000000	1.728750
50%	56430.500000	54.490000	3.000000	0.200000	8.666500
75%	90008.000000	209.940000	5.000000	0.200000	29.364000
max	99301.000000	22638.480000	14.000000	0.800000	8399.976000

## 2. Responder las preguntas:

- ¿Cuáles son las variables categóricas y cuáles las numéricas?
- **Variables categóricas (tipo object):** Ship Mode, Segment, Country, City, State, Region, Category, Sub-Category.
- **Variables numéricas (tipo float64 e int64):** Postal Code, Sales, Quantity, Discount, Profit.
  - ¿Existen valores atípicos en las variables? ¿Cómo los identificaste?
  - Valores atípicos identificados:
    - Sales: 1167 valores atípicos
    - Profit: 1881 valores atípicos
    - Quantity: 170 valores atípicos
    - Discount: 856 valores atípicos

Lo calculé sacando el IQR

---

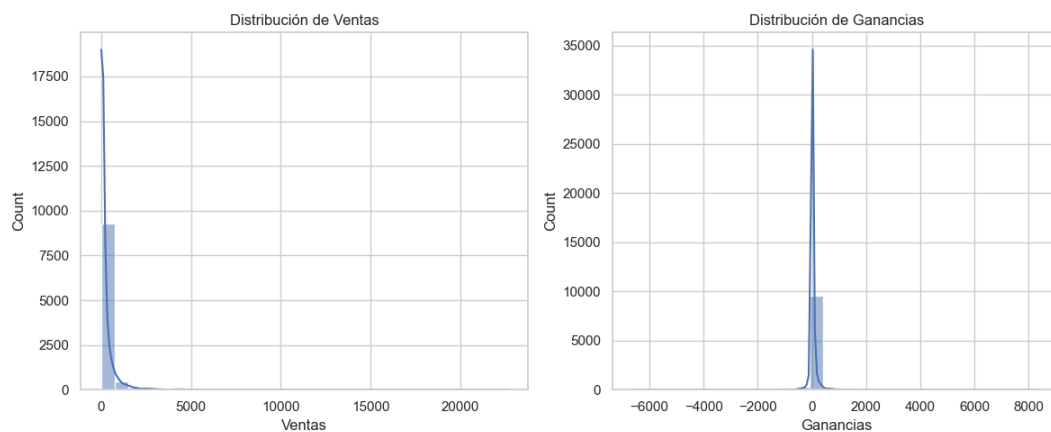
## Análisis exploratorio de datos

### A) Análisis univariante

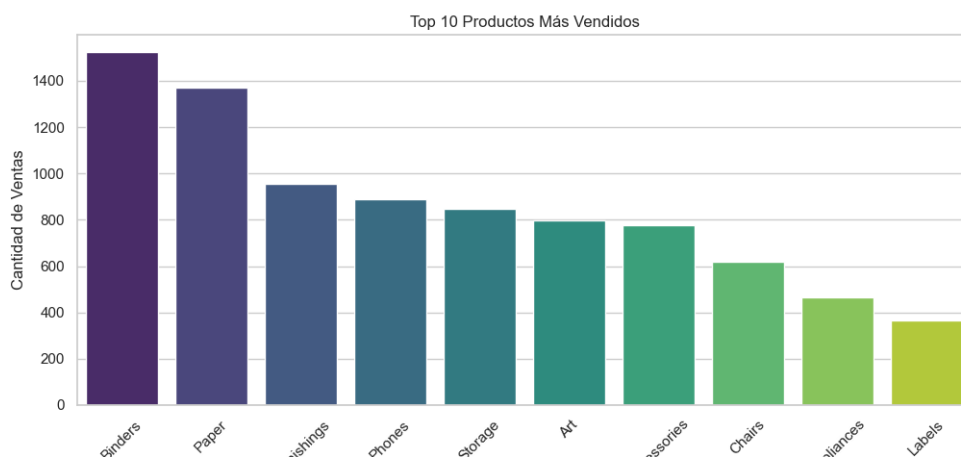
- Ejercicio 2:

1. Crear gráficos para las variables principales:

- **Histogramas para las variables numéricas (Ventas, Ganancia)**



- **Diagramas de barras para variables categóricas (Región, Producto).**



2. Interpretar los gráficos:

- **¿Qué distribución tienen las ventas?**

Basándonos en él, es probable que la distribución sea **sesgada a la derecha**, lo que indica que la mayoría de las ventas son bajas, pero hay algunos valores extremadamente altos que actúan como valores atípicos.

- **¿Qué producto es el más vendido?**

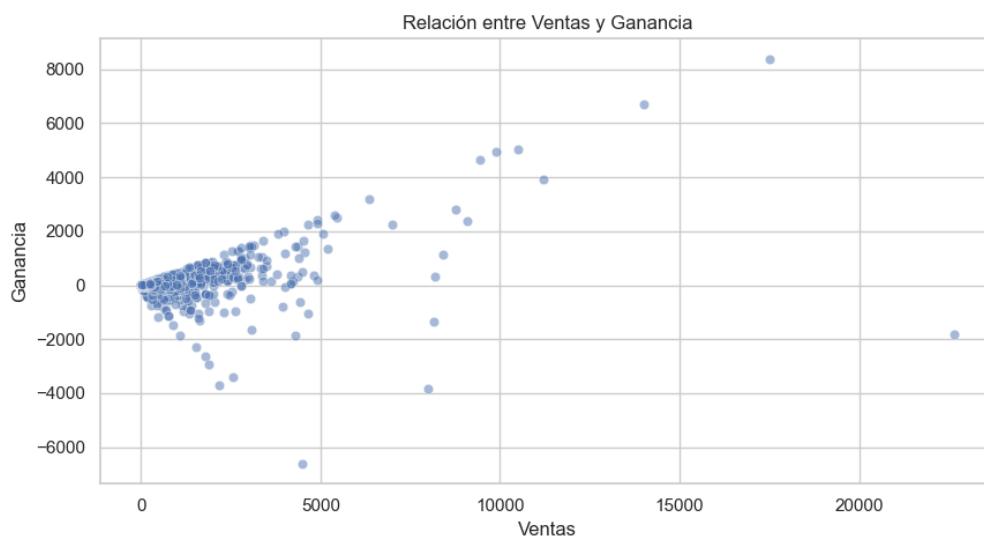
Producto más vendido: Binders

## B) Análisis multivariante

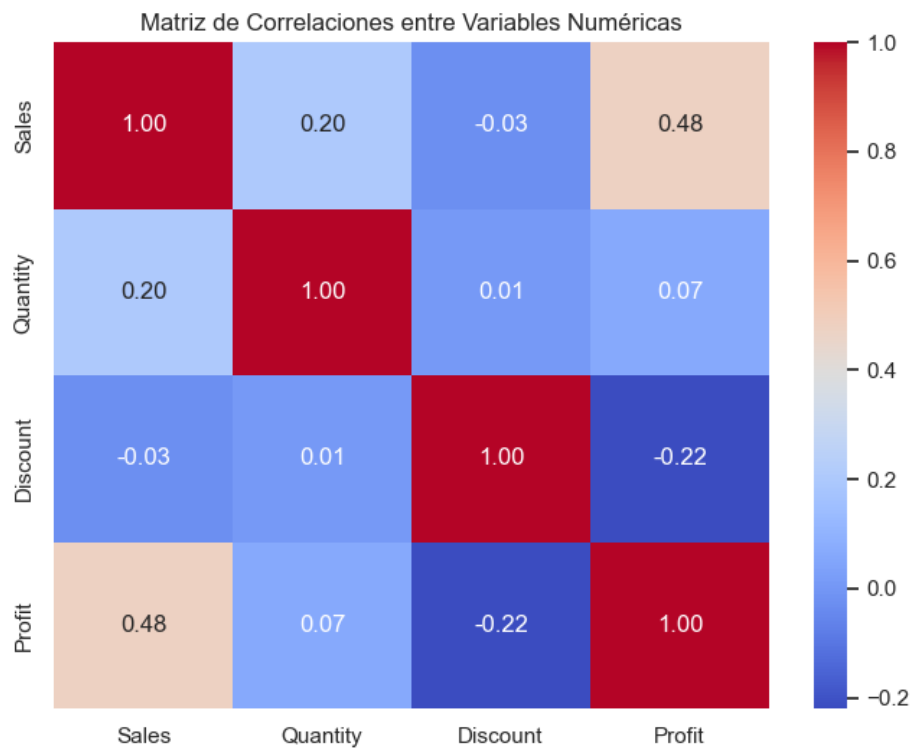
- Ejercicio 3:

1. Generar gráficos como:

- Diagramas de dispersión para analizar la relación entre Ventas y Ganancia.



2. Calcular las correlaciones entre las variables numéricas.



3. Interpretar:

- ¿Hay una relación entre las ganancias y las ventas?

Si, a más ventas, más ganancias

- ¿En qué región se venden más productos?

En la región West

---

**Análisis multidimensional**

- **Ejercicio 4:**

1. Seleccionar indicadores como:

- **Ventas totales por región.**

Ventas totales por región:

- Central 501239.8908
- East 678781.2400
- South 391720.9050
- West 725457.8245

- **Producto más vendido.**

Producto más vendido: Binders

2. Responder:

- ¿Qué regiones tienen mejor desempeño?

Promedio de ganancia por región:

- Central 17.091848
- East 32.135808
- South 28.857673
- West 33.848729

- ¿Qué áreas necesitan más atención según los indicadores?

La de profit

---

## Sistemas para análisis descriptivo

- **Ejercicio 5:**

1. Realizar:

- **Una reflexión sobre qué herramienta prefieres y por qué.**

Personalmente, prefiero utilizar **Python** junto con librerías como **Pandas** y **Matplotlib/Seaborn**, primero que nada, porque es la única herramienta que hemos visto y con la que estamos trabajando.

Creo que facilita la visualización, interpretación y extracción de los datos, o al menos lo considero que lo hace mejor que las otras herramientas.

---