guayerd

Fundamentos IA

Análisis con Python

Clase 7

En colaboración con

IBM **SkillsBuild**

¡Bienvenidos!

¿Nos presentamos?

- ¿Qué recuerdan de la clase anterior?
- ¿Qué esperan aprender?
- ¿Tienen alguna pregunta?

guayerd

En colaboración con

IBM **SkillsBuild**

Contenidos

Por temas

05
- Copilot Chat y prompts
- Demo asincrónica

06
- Limpieza y transformación

07
- Estadística aplicada

08
- Visualización

guayerd

En colaboración con

IBM **SkillsBuild**

Objetivos de la clase

- Estadística descriptiva básica
- Distribuciones de datos
- Correlaciones

guayerd

En colaboración con

IBM **SkillsBuild**

Análisis con Python

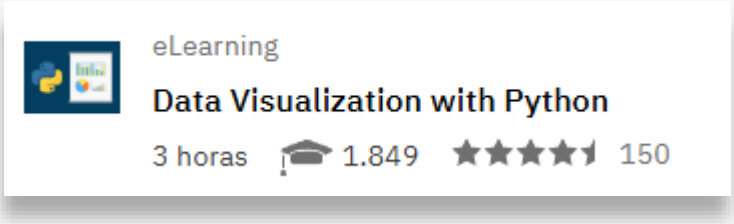
Estadística aplicada

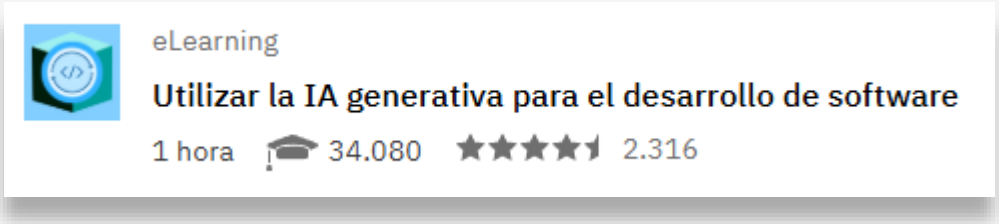
guayerd

En colaboración con

IBM **SkillsBuild**

Plataforma Skill Build: Python





guayerd

En colaboración con

IBM **SkillsBuild**

Estadística aplicada

La **estadística** es el arte y la ciencia de reunir datos, analizarlos, presentarlos e interpretarlos.

Esto ayuda a las personas que deben tomar decisiones una mejor comprensión del entorno, permitiéndoles así tomar mejores decisiones con base en mejor información.

Estadística aplicada

Conjunto de **técnicas para entender y resumir datos**.

- Describe características principales
- Detecta patrones y tendencias
- Mide relaciones entre variables
- Soporta la toma de decisiones

Estadística descriptiva

La mayor parte de la información estadística en periódicos, revistas, informes de empresas y otras publicaciones consta de datos que se resumen y presentan en una forma fácil de leer y de entender. A estos resúmenes de datos, que pueden ser tabulares, gráficos o numéricos se les conoce como **estadística descriptiva**.

Inferencia Estadística

Una de las principales contribuciones de la estadística es emplear datos de una muestra para hacer estimaciones y probar hipótesis acerca de las características de una población mediante un proceso al que se le conoce como **inferencia estadística.**

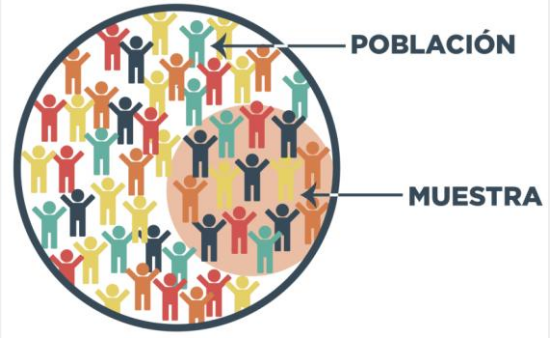
Población y muestra

Población
Cuando se examina un grupo entero o universo completo de observaciones.

Muestra
Cuando se examina una pequeña parte del grupo.

Población y muestra

El concepto de **población** en Estadística va más allá de la clásica definición que se da en la Demografía, esto es, la población de seres humanos exclusivamente. En la actividad estadística una población puede estar constituida por **elementos de cualquier tipo**, no solamente por seres humanos.



Distribución de frecuencias

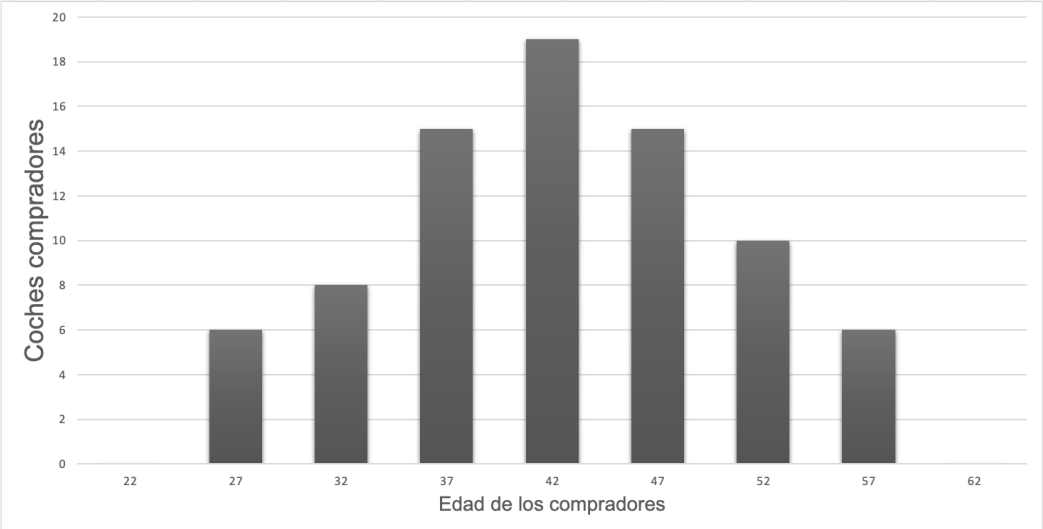
- Forma de presentación de los datos que facilita su tratamiento conjunto y permite una comprensión diferente de ellos.
- Es una tabla de datos con base en observaciones (frecuencias).
- La frecuencia es el número de casos que pertenecen a un valor determinado.

Edades de los compradores de automóviles

Edades	Nº de autos f_i	Verdadero Límite VL	Punto Medio x_i	Frecuencia Acumulada Menor que $F_i^{(-)}$	Frecuencia Acumulada Mayor que $F_i^{(+)}$	Frecuencia Relativa h_i	Frecuencia Relativa Acumulada Menor que $H_i^{(-)}$	Frecuencia Relativa Acumulada Mayor que $H_i^{(+)}$
25 - 29	6	24,5	27	*6*	*80*	7,50%	7,50%	100,00%
30 – 34	9	29,5	32	15	74	11,25%	18,75%	92,50%
35 – 39	15	34,5	37	30	65	18,75%	37,50%	81,25%
40 – 44	18	39,5	42	48	50	22,50%	60,00%	62,50%
45 – 49	15	44,5	47	63	32	18,75%	78,75%	40,00%
50 – 54	10	49,5	52	73	17	12,50%	91,25%	21,25%
55 - 59	7	54,5	57	80	7	8,75%	100,00%	8,75%
	80					**100,00%**		

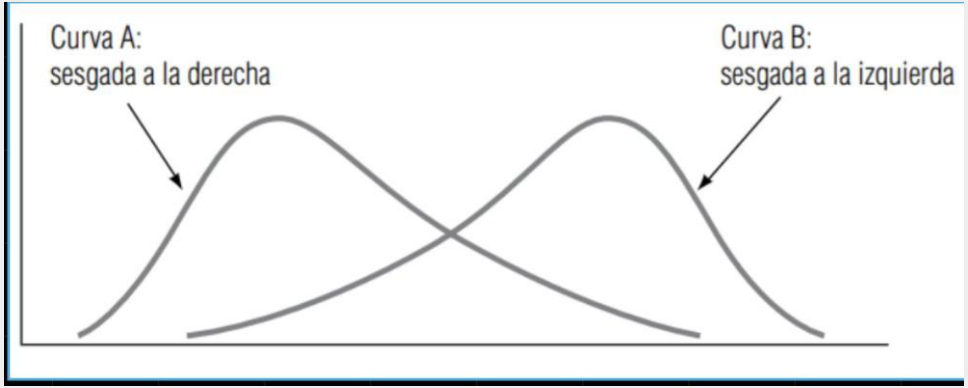
Histograma

Gráfico de la distribución de frecuencias, que se construye con rectángulos de superficie proporcional al producto de la amplitud por la frecuencia absoluta (o relativa) de cada uno de los intervalos de clase.



Tendencia Central

Se refiere al **punto medio**de una distribución.
El **sesgo** se produce cuando al trazar una línea vertical que pase por el punto más alto de la curva dividirá su área en dos partes que no son iguales.



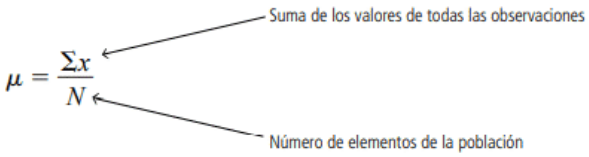
- Cuando se da el caso de que cada parte es una imagen de espejo de la otra, esta curva se denomina simétrica.
- Si la curva está sesgada hacia la derecha, se considera positivamente sesgada y si el sesgo se pronuncia hacia la izquierda, se denomina negativamente sesgada.

Media

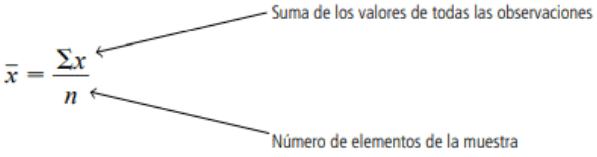
Media aritmética (Promedio):
Es la suma de los valores de todas las observaciones, dividido la cantidad de elementos de la muestra.

Casi siempre, cuando nos referimos al "**promedio**" de algo, estamos hablando de la **media aritmética**. En una muestra de una población que consiste en n observaciones (con 'n' minúscula), la media se denomina con 'x' (x barra).

Media aritmética de la población

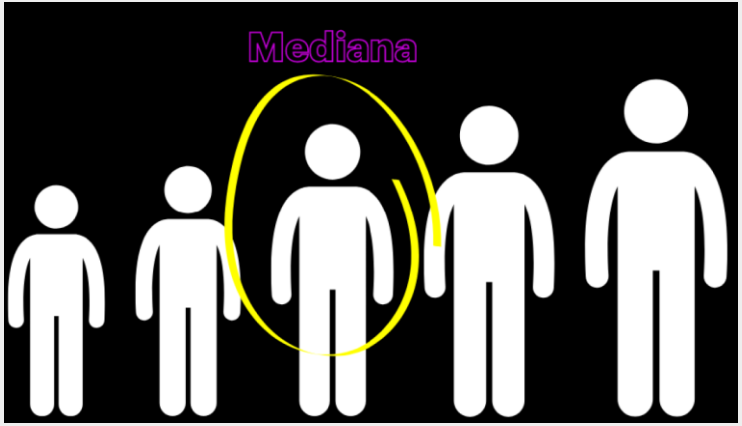


Media aritmética de la muestra



Mediana

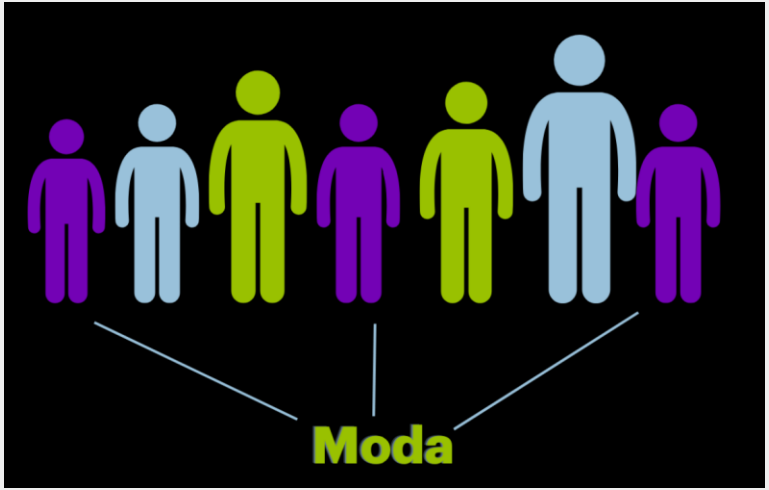
- Mide la observación central del conjunto.
- Para hallar la mediana de un conjunto de datos, primero se organizan en orden descendente o ascendente.
- El elemento que está más al centro del conjunto de números, la mitad de los elementos están por arriba de este punto y la otra mitad está por debajo.



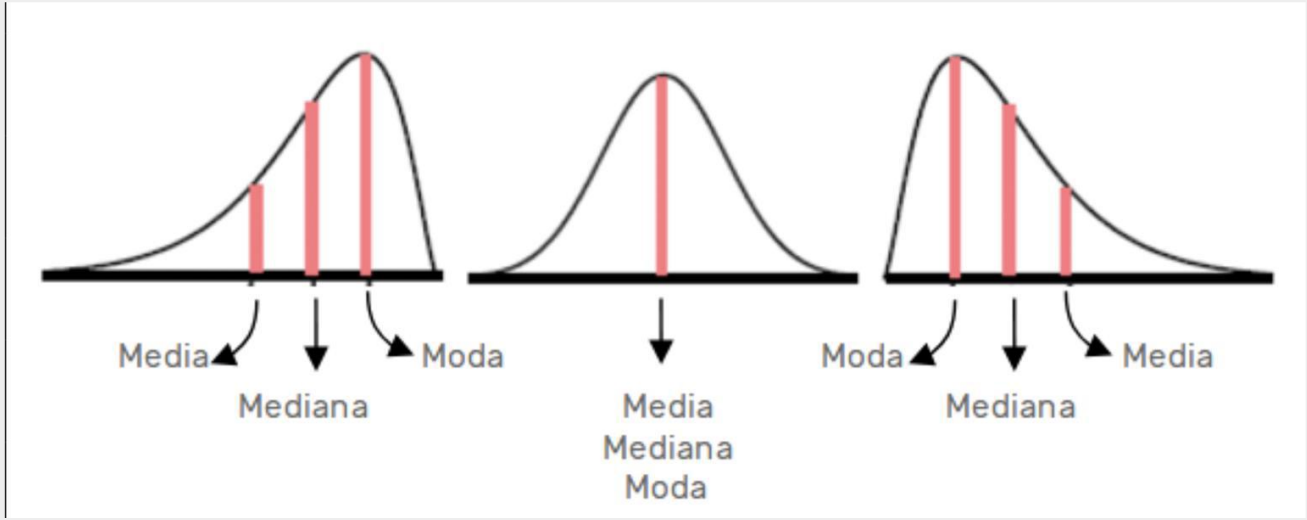
Si el conjunto de datos contiene un número impar de elementos, el de en medio en el arreglo es la mediana; si hay un número par de observaciones, la mediana es el promedio de los dos elementos de en medio.

Moda

La moda es el valor que más se repite en el conjunto de datos.



Media, Mediana y Moda



Medidas de Dispersión

Concepto general:

- Las **medidas de dispersión** indican **cuánto varían o se alejan los datos** entre sí respecto de una medida de tendencia central (por lo general, la media).
- Complementan la información de la media, permitiendo evaluar **la homogeneidad o heterogeneidad** del conjunto de datos.

Principales medidas:

- **Rango:** diferencia entre el valor máximo y el mínimo.
- **Varianza (σ^2 o s^2):** promedio de los cuadrados de las desviaciones respecto de la media.
- **Desviación estándar (σ o s):** raíz cuadrada de la varianza.

💡 **Importancia:** permiten comparar la variabilidad entre distintos grupos o distribuciones.

Varianza

Concepto:

- Mide **cuánto se dispersan los datos respecto de la media**.
- Cuanto mayor sea la varianza, **más alejados están los valores** del promedio.
- Se expresa en **unidades al cuadrado**, por lo que no es directamente interpretable.

Fórmulas:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} \quad \text{(Población)}$$

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} \quad \text{(Muestra)}$$

Donde:

- x_i: cada valor observado
- μ: media poblacional
- $\bar{x}$: media muestral
- N: tamaño de la población
- n: tamaño de la muestra
- $\sum$: suma de todos los valores

Desviación Estándar (σ o s)

Concepto:

- Es la **raíz cuadrada de la varianza**.
- Indica en **promedio cuánto se desvía cada dato de la media**.
- Se mide en **las mismas unidades que los datos**, por lo que facilita su interpretación.

Fórmulas:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \quad \text{(Población)}$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} \quad \text{(Muestra)}$$

Donde:

- x_i: cada valor observado
- μ: media poblacional
- $\bar{x}$: media muestral
- N: tamaño de la población
- n: tamaño de la muestra
- $\sum$: suma de todos los valores

Interpretación:

- **s o σ pequeña:** poca variabilidad, datos concentrados.
- **s o σ grande:** gran dispersión.
- En distribuciones normales:
  - ≈ **68%** de los valores están dentro de ±1σ
  - ≈ **95%** dentro de ±2σ

Ejemplo:

Si $\bar{x} = 50$ y $s = 5$, la mayoría de los valores se encuentran entre **45 y 55**.

Exploración con Pandas

Herramientas para exploración y análisis estadístico

Función	Propósito	Resultado
.describe()	Resumen completo	Todas las estadísticas principales
.info()	Información general	Tipos y valores nulos
.value_counts()	Frecuencias	Conteo por categoría
.groupby().agg()	Estadísticas agrupadas	Métricas por segmento

Medidas descriptivas

Resumen numérico de características principales

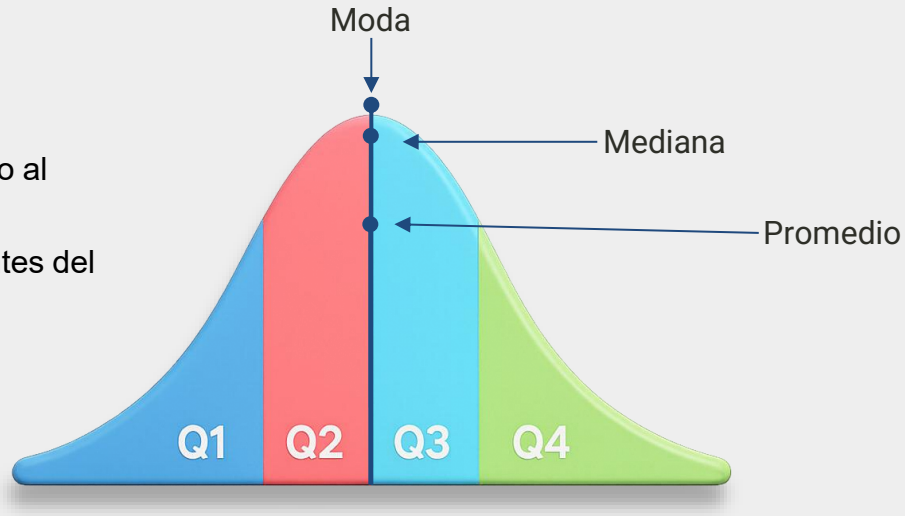
Medida	Comando	Descripción
Media	df['columna'].mean()	Promedio aritmético
Mediana	df['columna'].median()	Valor central ordenado
Moda	df['columna'].mode()	Valor más frecuente
Desviación estándar	df['columna'].std()	Dispersión promedio

Estadística descriptiva

Ejemplo

Distribución “Normal”

- Moda = Mediana = Promedio
- La curva es simétrica respecto al centro.
- Los cuartiles están equidistantes del centro.



Cuartiles

Medidas de posición

Ubicación de valores en la distribución

Medida	Comando	Interpretación
Mínimo	df['columna'].min()	Valor más bajo
Máximo	df['columna'].max()	Valor más alto
Cuartiles	df['columna'].quantile([0.25, 0.5, 0.75])	Divide datos en 4 partes
Rango	df['columna'].max() - df['columna'].min()	Amplitud total

¿Qué significan estos datos?

Dataset de salarios (en miles USD)

Media: 45, **mediana**: 38, **moda**: 35

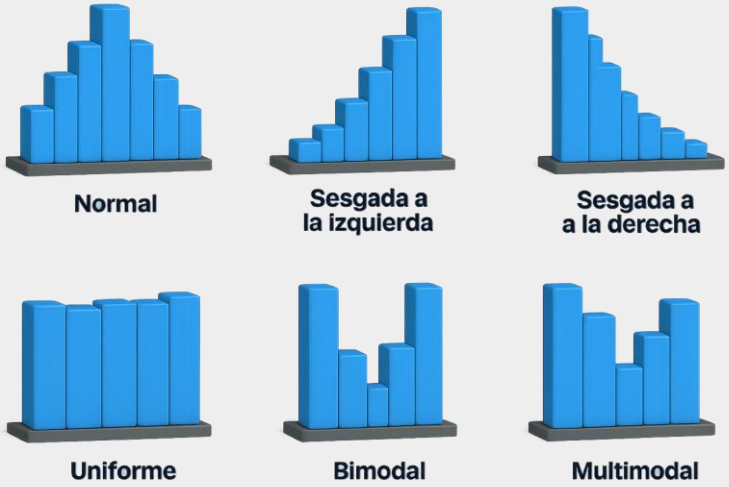
Desviación estándar: 15

Rango: 80 (min: 20, máx: 100)

Distribuciones de datos

Muestran la forma en que se organizan los valores dentro de un conjunto de datos.

- **Normal:** campana simétrica
- **Sesgada:** cola hacia un lado
- **Bimodal:** dos picos de frecuencia
- **Multimodal:** múltiples picos de frecuencia
- **Uniforme:** frecuencias similares



Identificación de distribución

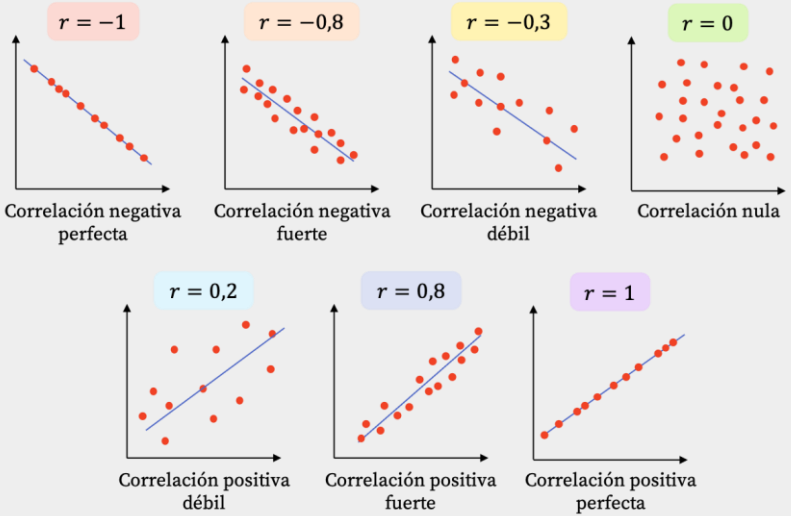
El tipo de distribución se deduce comparando media y mediana.

Tipo de Distribución	Relación Media–Mediana	Ejemplo Real
Normal	Media ≈ Mediana	Alturas, pesos
Sesgada	Media muy diferente de mediana	Ingresos, precios
Bimodal	Depende de los picos	Horarios de tráfico
Multimodal	Variable según picos	Preferencias múltiples
Uniforme	Media ≈ Mediana	Números aleatorios

Correlaciones

Medida de **qué tanto dos variables cambian** juntas.

- Valores entre -1 (inverso) y +1 (directo)
- 0 indica que no hay relación lineal
- **Comando:** df[['var1', 'var2']].corr()



Errores comunes de interpretación

Media vs. Mediana

- Diferencias grandes indican presencia de valores extremos
- Siempre reportar ambas métricas

Correlación ≠ Causación

- Una correlación alta no implica causalidad
- Considerar variables ocultas que puedan explicar la relación

Evaluación de confiabilidad

Desviación estándar

- **Baja:** datos consistentes, resultados predecibles
- **Alta:** datos dispersos, mayor incertidumbre

Outliers (valores extremos)

- Analizar antes de eliminarlos
- Pueden ser errores de medición o información relevante

Rendimiento E-commerce

- Identificar mes con mayor eficiencia (ventas/gasto publicidad)
- Determinar mes con mejor tasa de conversión y analizar causa
- Calcular ticket promedio (ventas/productos) por mes
- Evaluar relación entre visitantes y ventas

Mes	Ventas ($)	Visitantes	Conversión (%)	Gasto Publicidad ($)	Productos Vendidos
Ene	45,000	15,000	3.2	8,500	450
Feb	52,000	18,200	2.9	9,800	520
Mar	38,000	12,500	3.8	7,200	380
Abr	61,000	20,500	3.1	11,200	610
May	48,000	16,800	2.7	9,500	480

Proyecto

Tienda Aurelion

- **Documentación:** notebook Markdown
- **Desarrollo técnico:** programa Python
- **Visualización de datos:** dashboard en Power BI
- **Presentación oral:** problema, solución y hallazgos

guayerd

En colaboración con

IBM **SkillsBuild**

Análisis estadístico descriptivo

Trabajo en equipo

1. Calcular **estadísticas básicas**
2. Identificar **tipo de distribución**
3. Calcular **correlaciones** entre variables principales
4. Analizar **outliers**
5. **Interpretar resultados** para el problema de negocio
6. **Documentar con Copilot** cada paso y resultado

Retro

¿Cómo nos vamos?

- ¿Qué fue lo más útil de la clase?
- ¿Qué parte te costó más?
- ¿Qué te gustaría repasar o reforzar?

guayerd

En colaboración con

IBM **SkillsBuild**