Análisis Exploratorio de Datos

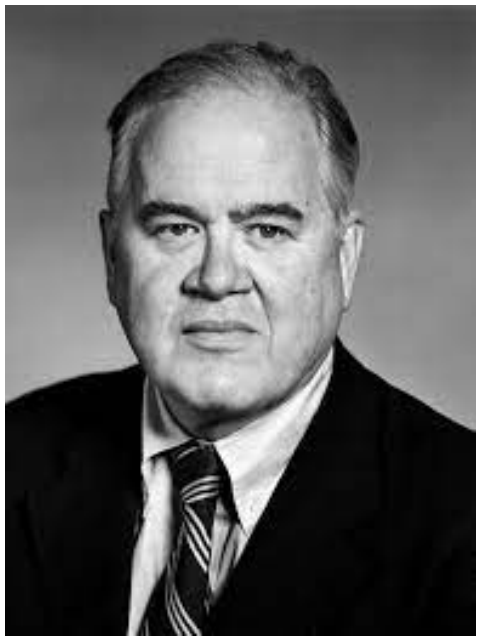
Camilo Esteban Núñez Fernández

INF396 - Introducción a la Ciencia de Datos
Departamento de Informática

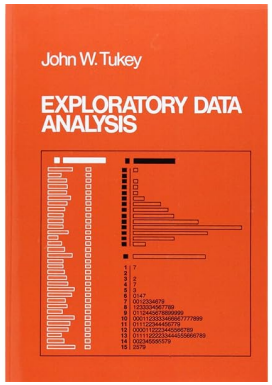
2025-03-14

Fast Intro

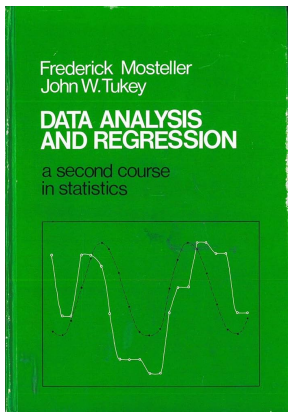
John Wilder Tukey



- Exploratory Data Analysis. 1977. Addison-Wesley



- Data Analysis and Regression: A Second Course in Statistics. 1977 (+ Frederick Mosteller). Addison-Wesley.



- Boxplots, Fast Fourier Transform (FFT), Cooley–Tukey FFT algorithm ...

I.I ▷ EDA: Exploratory Data Analysis

- El Análisis Exploratorio de Datos es un enfoque que emplea diversas técnicas para:
  - Maximizar la comprensión de un conjunto de datos
  - Extraer variables importantes
  - Detectar valores outliers y anomalías
  - Verificar supuestos subyacentes
  - Determinar configuraciones óptimas de factores para modelar los datos
- La mayoría de las técnicas aplicadas en EDA son gráficas, con ciertas técnicas cuantitativas.

I.I ▷ EDA: Exploratory Data Analysis

¿En qué se diferencia de otros análisis estadísticos?

- Tres enfoques principales: (I) Clásico, (II) Bayesiano, (III) Exploratorio (EDA)
- **Clásico**: Asume un modelo a priori. Flujo:
  Problema → Datos → Modelo → Análisis → Conclusiones
- **Bayesiano**: Incorpora conocimiento experto (prior). Flujo:
  Problema → Datos → Modelo → Prior → Análisis → Conclusiones
- **Exploratorio (EDA)**: Análisis sin modelo previo para determinar el modelo adecuado. Flujo:
  Problema → Datos → Análisis → Modelo → Conclusiones

Por donde comenzar?

I.II ▷ Por donde comenzar?

Formula tus preguntas o requisitos !

- Útil para guiar el proceso exploratorio.
- Útil para reducir el espacio de búsqueda.
- Acota tus caminos de exploración.

Preguntas sencillas y concisas pueden darte una rápida reducción en la dimensionalidad de los datos !

I.II ▷ Formulación de preguntas

```
pdf_suicide_rates.sample(n=10)
```

	country	year	sex	age	suicides_no	population	suicides_100k_pop	country_year	hdi_for_year	gdp_for_year	gdp_per_capita	generation
16488	Mauritius	2010	female	35-54 years	8	184083	4.35	Mauritius2010	0.756	10,003,670,690	8587	Generation X
18712	Paraguay	2000	female	15-24 years	23	521651	4.41	Paraguay2000	0.623	8,195,993,231	1782	Generation X
22671	Singapore	2007	female	35-54 years	46	608800	7.56	Singapore2007	NaN	179,981,288,567	53098	Boomers
24669	Sweden	2012	male	55-74 years	256	1074267	23.83	Sweden2012	0.904	543,880,647,757	60776	Boomers
21566	Saint Lucia	2009	female	35-54 years	0	23037	0.00	Saint Lucia2009	NaN	1,262,973,407	7902	Boomers
25106	Thailand	1998	female	55-74 years	142	3595775	3.95	Thailand1998	NaN	113,675,706,127	2005	Silent
22790	Slovakia	1993	female	55-74 years	36	491822	7.32	Slovakia1993	NaN	16,452,201,101	3334	Silent
716	Argentina	1995	female	15-24 years	101	3053300	3.31	Argentina1995	0.731	258,031,750,000	8232	Generation X
23069	Slovenia	1997	female	15-24 years	5	144100	3.47	Slovenia1997	NaN	20,749,140,606	11014	Generation X
10881	Guatemala	2000	male	5-14 years	5	1621778	0.31	Guatemala2000	0.552	19,288,827,159	1977	Millenials

- Iteración 0: ¿Es la tasa de suicidios mayor en América del Sur que en América del Norte?
- Iteración 1: ¿Es la tasa de suicidios por cada 100K hab. mayor en USA que en Chile para la década del 2000?

Estadística Descriptiva

- Busca resumir y describir características importantes en los datos.
- Podemos encontrar dos representaciones clásicas:
  1. **Representaciones Numéricas**: Medidas de tendencias y dispersión.
  2. **Representaciones Gráficas**: Histogramas, Scatter plots, Boxplots, etc.

II.I ▷ Estadística Descriptiva - Rep. Numéricas

Medidas de Tendencias

- *Moda*
- *Media Muestral*
- *Mediana Muestral*

Medidas de Dispersión

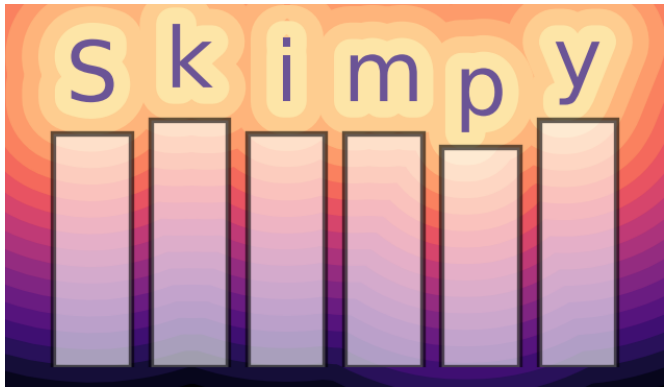
- *Rango*
- *Indice de Variación*
- *Varianza Muestral*
- *Desviación Estándar Muestral*

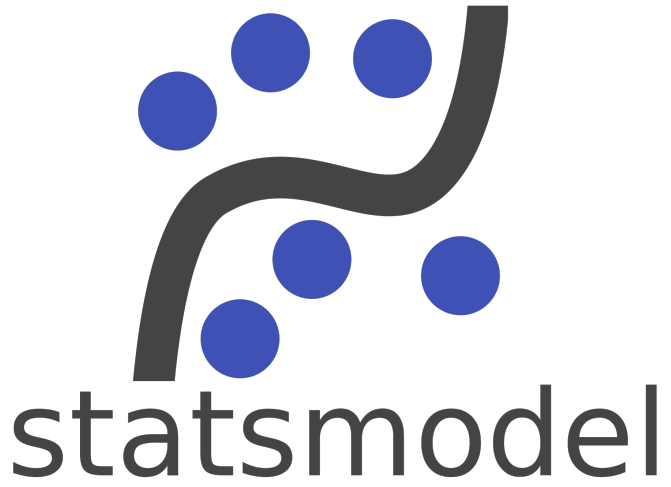
- **Five-Number Summary**: Clásico reporte cuantitativo que incluye, Q_1 (percentil 25), Q_2 (percentil 50, mediana), Q_3 (percentil 75), máximo, mínimo.

Nota sobre la Desviación Estándar (σ)

- σ mide la dispersión alrededor de la **media**(μ), por lo que solo debe usarse cuando la media sea la medida de tendencia central elegida.
- Si $\sigma = 0$ **solo**, entonces no hay dispersión (todas las observaciones tienen el mismo valor).
- Tanto la varianza como la desviación estándar tienen **buena escalabilidad** en bases de datos grandes.

Por estas propiedades, σ es un indicador confiable de la dispersión de un conjunto de datos. (Ej. clásico: Análisis de varianza (ANOVA))


skimpy


statsmodel

II.I ▷ Estadística Descriptiva - Rep. Numéricas

```
1 from skimpy import skim
```

```
1 skim(pdf_suicide_rates)
```

skimpy summary

Data Summary

Dataframe	Values
Number of rows	27820
Number of columns	12

Data Types

Column Type	Count
string	6
int64	4
float64	2

number

column	NA	NA %	mean	sd	p0	p25	p50	p75	p100	hist
year	0	0	2001	8.469	1985	1995	2002	2008	2016	
suicides_no	0	0	242.6	902	0	3	25	131	22340	
population	0	0	1845000	3912000	278	97500	430200	1486000	43810000	
suicides_100k_pop	0	0	12.82	18.96	0	0.92	5.99	16.62	225	
hdi_for_year	19456	69.9352983465133	0.7766	0.09337	0.483	0.713	0.779	0.855	0.944	
gdp_per_capita	0	0	16870	18890	251	3447	9372	24870	126400	

string

column	NA	NA %	shortest	longest	min	max	chars per row	words per row	total words
country	0	0	Cuba	Saint Vincent and Grenadines	Albania	Uzbekistan	8.62	1.2	34574
sex	0	0	male	female	female	male	5	1	27820
age	0	0	75+ years	15-24 years	15-24 years	75+ years	10.5	2	55640
country_year	0	0	Cuba1992	Saint Vincent and Grenadines1985	Albania1987	Uzbekistan2014	12.6	1.2	34574
gdp_for_year	0	0	98,585,185	10,284,779,000,000	1,002,219,052,968	997,007,926	14.2	1	27820
generation	0	0	Silent	G.I. Generation	Boomers	Silent	9.61	1.4	38442

End

II.I ▷ Estadística Descriptiva - Rep. Numéricas

```python
import statsmodels.stats as ss
```

```python
ss.descriptivestats.describe(pdf_suicide_rates,
                             stats=["range", "coef_var", "std", "mode", "mean", "median", "max", "min", "percentiles"]).T
```

	range	coef_var	std	mode	mode_freq	mean	median	max	min	1%	5%	10%	25%	50%	75%	90%	95%	99%
year	3.100000e+01	0.004232	8.469055e+00	2009.000	0.038390	2.001258e+03	2002.000	2.016000e+03	1985.000	1985.00000	1987.000	1989.000	1995.000	2002.000	2008.000	2013.000	2.014000e+03	2.015000e+03
suicides_no	2.233800e+04	3.718644	9.020479e+02	0.000	0.153882	2.425744e+02	25.000	2.233800e+04	0.000	0.00000	0.000	0.000	3.000	25.000	131.000	496.000	1.050050e+03	3.993670e+03
population	4.380494e+07	2.120443	3.911779e+06	24000.000	0.000719	1.844794e+06	430150.000	4.380521e+07	278.000	1763.09000	7195.600	17303.300	97498.500	430150.000	1486143.250	4960713.500	8.850240e+06	1.999273e+07
suicides_100k_pop	2.249700e+02	1.479507	1.896151e+01	0.000	0.153882	1.281610e+01	5.990	2.249700e+02	0.000	0.00000	0.000	0.000	0.920	5.990	16.620	33.291	5.053050e+01	9.157100e+01
hdi_for_year	4.610000e-01	0.120225	9.336671e-02	0.713	0.010043	7.766011e-01	0.779	9.440000e-01	0.483	0.56326	0.619	0.648	0.713	0.779	0.855	0.897	9.120000e-01	9.323700e-01
gdp_per_capita	1.261010e+05	1.119830	1.888758e+04	1299.000	0.001294	1.686646e+04	9372.000	1.263520e+05	251.000	476.00000	935.000	1524.000	3447.000	9372.000	24874.000	43487.000	5.429400e+04	8.963400e+04

```
"nobs", "missing", "mean", "std_err", "ci", "ci", "std", "iqr",
"iqr_normal", "mad", "mad_normal", "coef_var", "range", "max",
"min", "skew", "kurtosis", "jarque_bera", "mode", "freq",
"median", "percentiles", "distinct", "top", "freq"
```

II.I ▷ Estadística Descriptiva - Rep. Numéricas

```
1 from skimpy import skim
```

```
1 skim(pdf_suicide_rates)
```

skimpy summary

Data Summary

Dataframe	Values
Number of rows	27820
Number of columns	12

Data Types

Column Type	Count
string	6
int64	4
float64	2

number

column	NA	NA %	mean	sd	p0	p25	p50	p75	p100	hist
year	0	0	2001	8.469	1985	1995	2002	2008	2016	
suicides_no	0	0	242.6	902	0	3	25	131	22340	
population	0	0	1845000	3912000	278	97500	430200	1486000	43810000	
suicides_100k_pop	0	0	12.82	18.96	0	0.92	5.99	16.62	225	
hdi_for_year	19456	69.9352983465133	0.7766	0.09337	0.483	0.713	0.779	0.855	0.944	
gdp_per_capita	0	0	16870	18890	251	3447	9372	24870	126400	

string

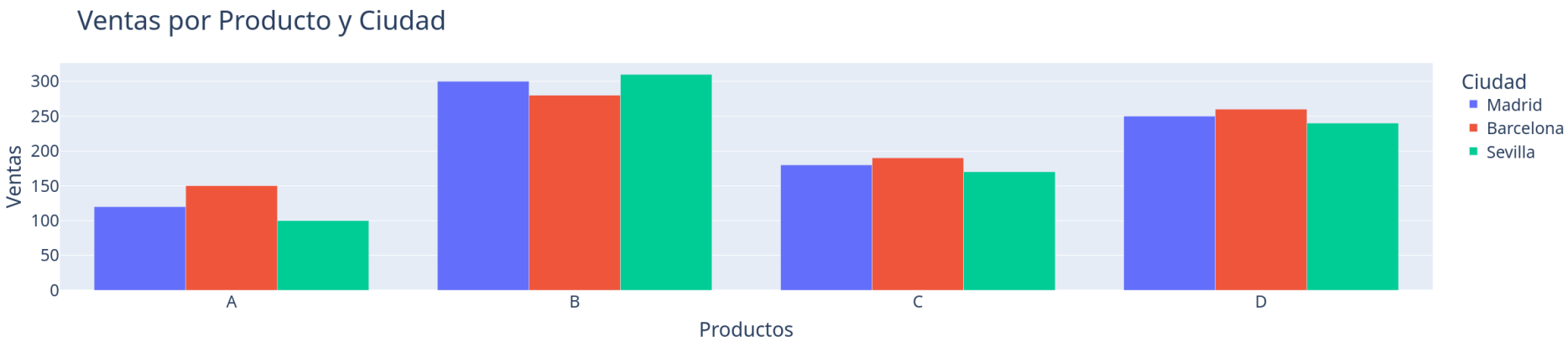
column	NA	NA %	shortest	longest	min	max	chars per row	words per row	total words
country	0	0	Cuba	Saint Vincent and Grenadines	Albania	Uzbekistan	8.62	1.2	34574
sex	0	0	male	female	female	male	5	1	27820
age	0	0	75+ years	15-24 years	15-24 years	75+ years	10.5	2	55640
country_year	0	0	Cuba1992	Saint Vincent and Grenadines1985	Albania1987	Uzbekistan2014	12.6	1.2	34574
gdp_for_year	0	0	98,585,185	10,284,779,000,000	1,002,219,052,968	997,007,926	14.2	1	27820
generation	0	0	Silent	G.I. Generation	Boomers	Silent	9.61	1.4	38442

End

¡Considera lo siguiente!

- Visualizar tus datos mediante gráficos puede facilitar la comprensión de sus propiedades, la detección de patrones y la identificación de estrategias de modelado adecuadas para responder a tus preguntas.
- Los gráficos también pueden servir como una herramienta de depuración (*debugging*) para validar tu análisis descriptivo.
- Es importante diferenciar un gráfico exploratorio de un gráfico final. Los gráficos exploratorios ayudan a inspeccionar los datos en las primeras etapas del análisis, mientras que los gráficos finales están diseñados para comunicar claramente los resultados.
- En un gráfico final, prioriza la claridad y la precisión en la comunicación de tus hallazgos.

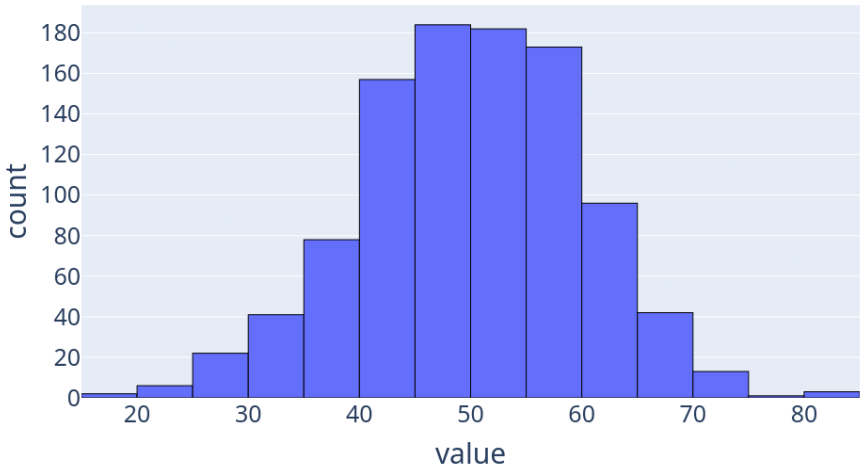
- Representa la frecuencia de las categorías en un conjunto de datos.
- Ideal para visualizar datos categóricos y comparar distribuciones.
- Se puede usar tanto para conteo directo como para representar valores agregados (por ejemplo, ventas promedio por categoría).
- La altura de cada barra refleja la cantidad o proporción de cada categoría.



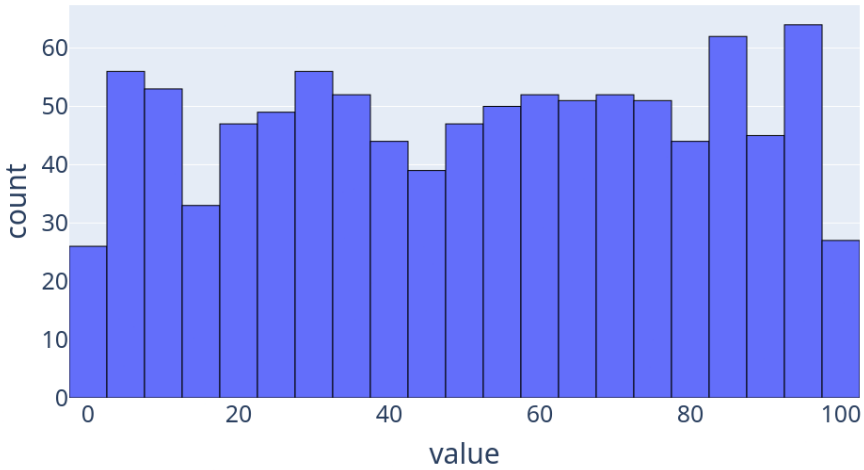
II.II ▷ Estadística Descriptiva - Visualización: Histograma

- Muestra la distribución empírica de todos los datos del conjunto.
- Nos ayuda a identificar:
  - Asimetrías estadísticas (skewness)
  - Simetrías
  - Multi-modalidad

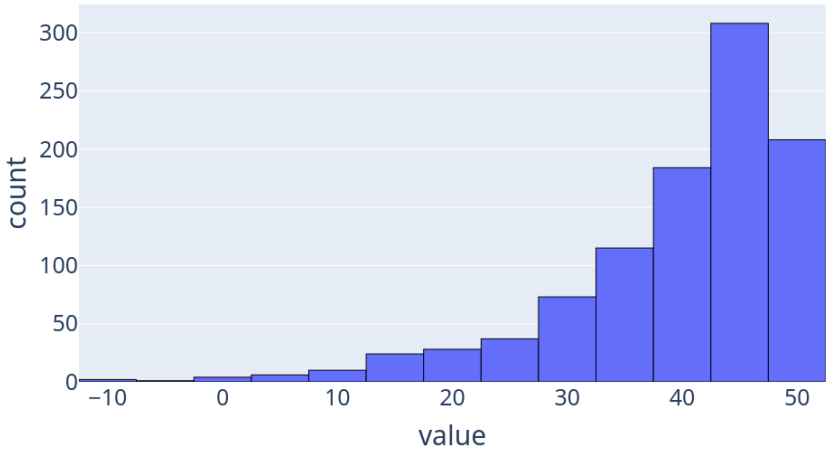
Distribución Normal



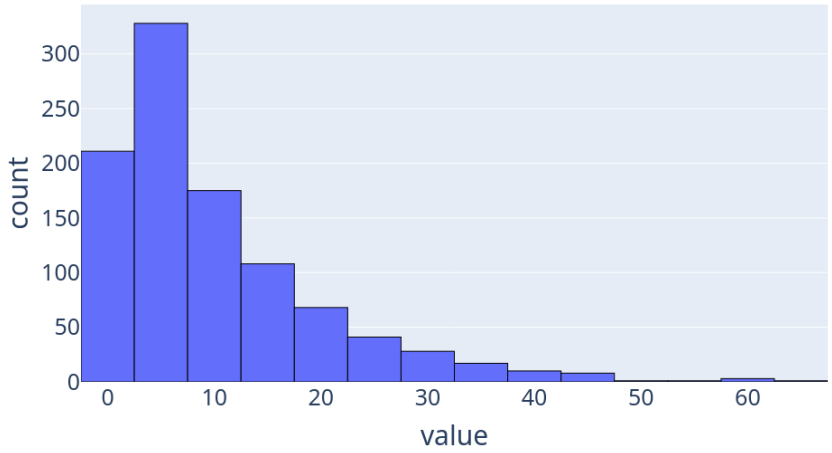
Distribución Uniforme



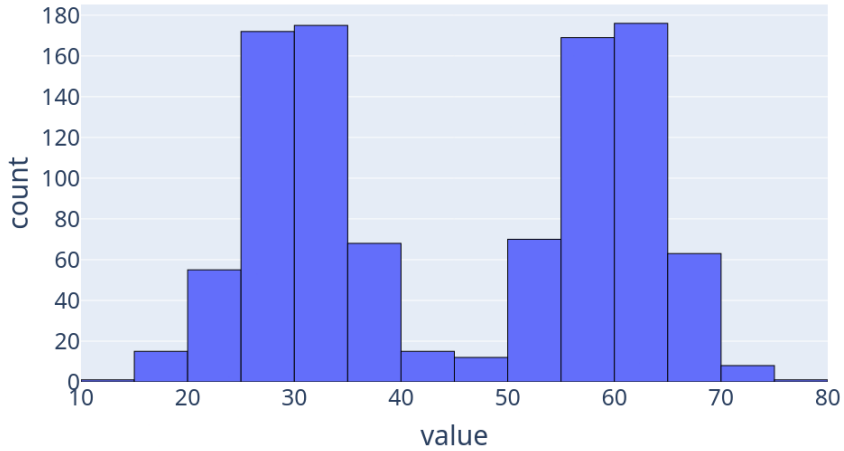
Asimetría a la Izquierda



Asimetría a la Derecha

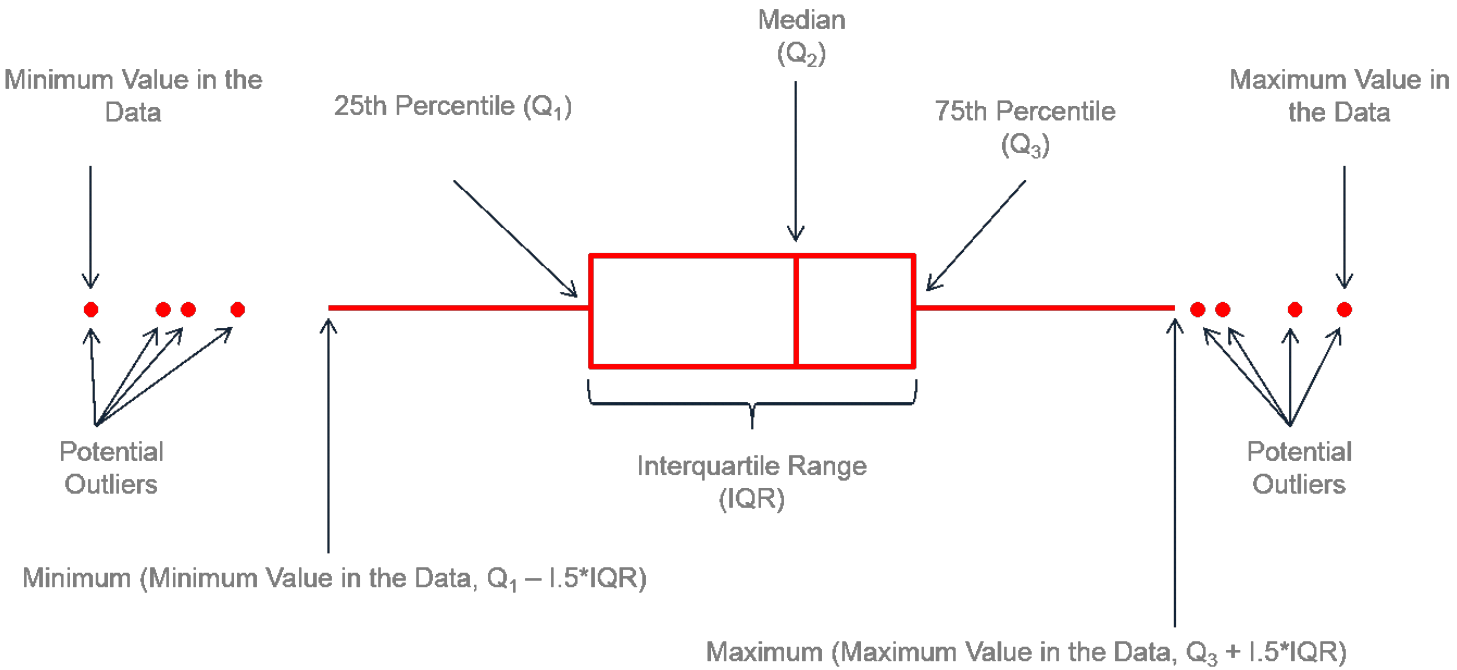


Distribución Multimodal



II.II ▷ Estadística Descriptiva - Visualización: Boxplot

- Representa la tendencia central y la dispersión del conjunto de datos.
- Considera los cuartiles: Q_1 (percentil 25), Q_2 (percentil 50, mediana) y Q_3 (percentil 75), además de los valores máximo y mínimo.
- Permite identificar valores atípicos (outliers) que se encuentran más allá de la *bulk data* (zona central de los datos).



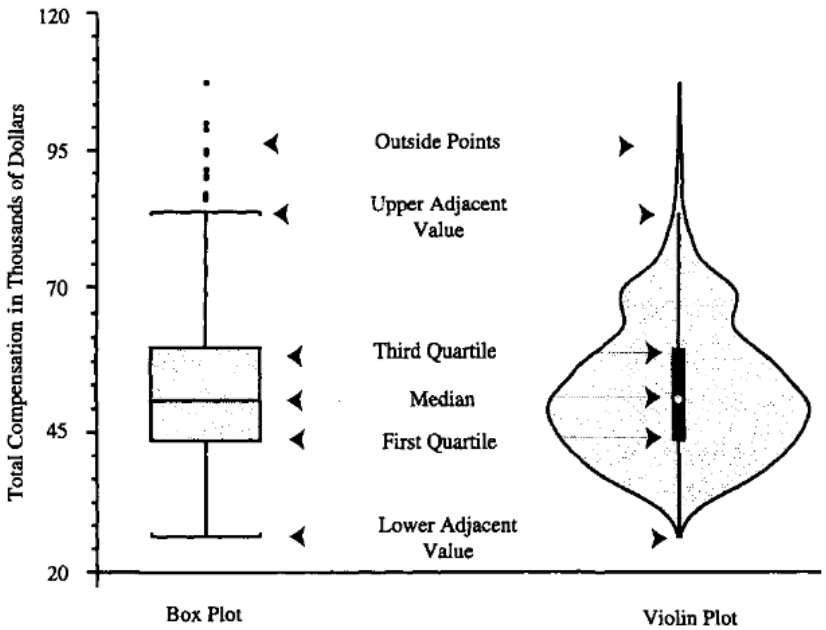
Statistical Computing and Graphic

Violin Plots: A Box Plot-Density Trace Synergism

Jerry L. Hintze & Ray D. Nelson

Pages 181-184 | Received 01 Feb 1997, Published online: 22 Mar 2012

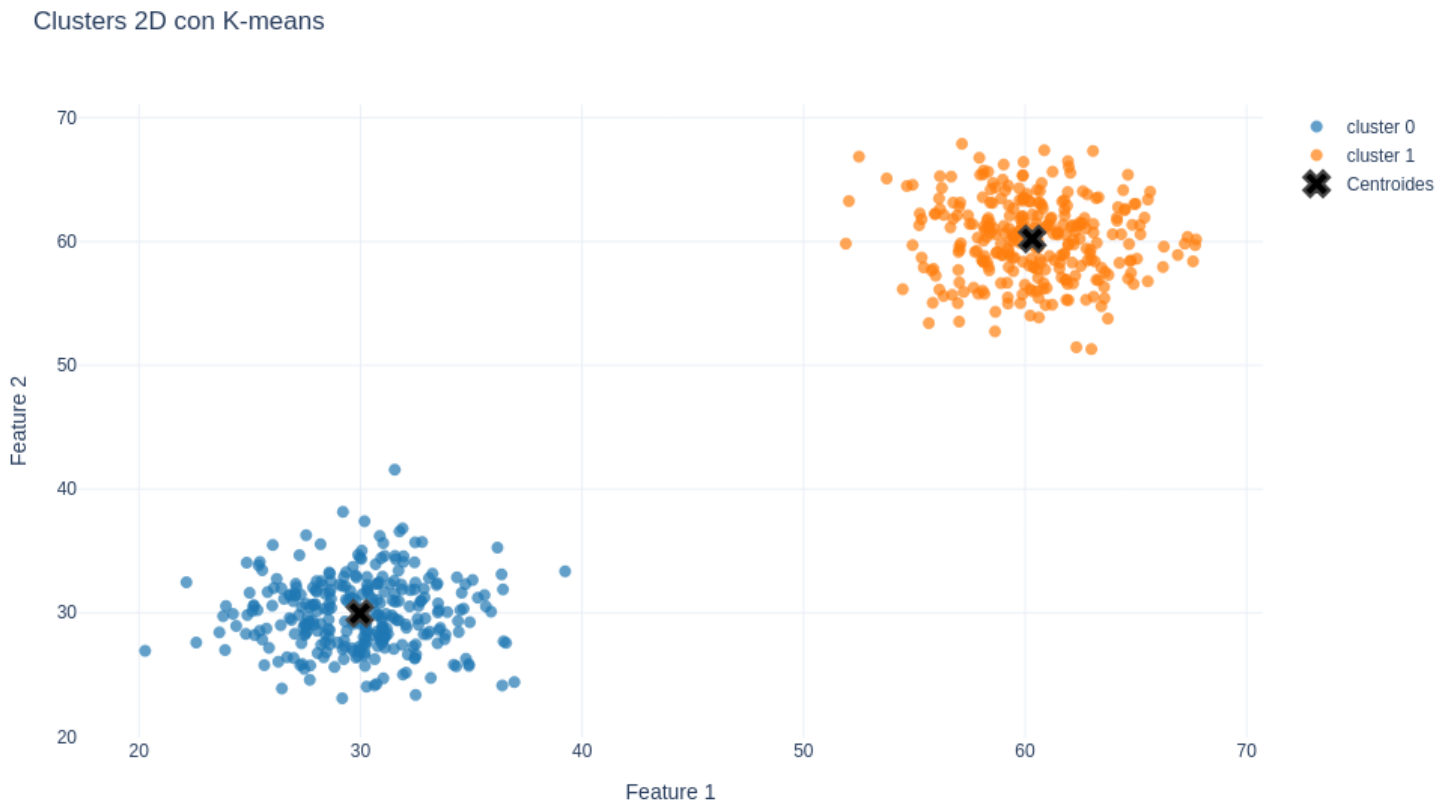
- Combina un Boxplot con una traza de densidad (métodos de *smoothed histogram* como KDE).



- Boxplot:
- Traza de Densidad:

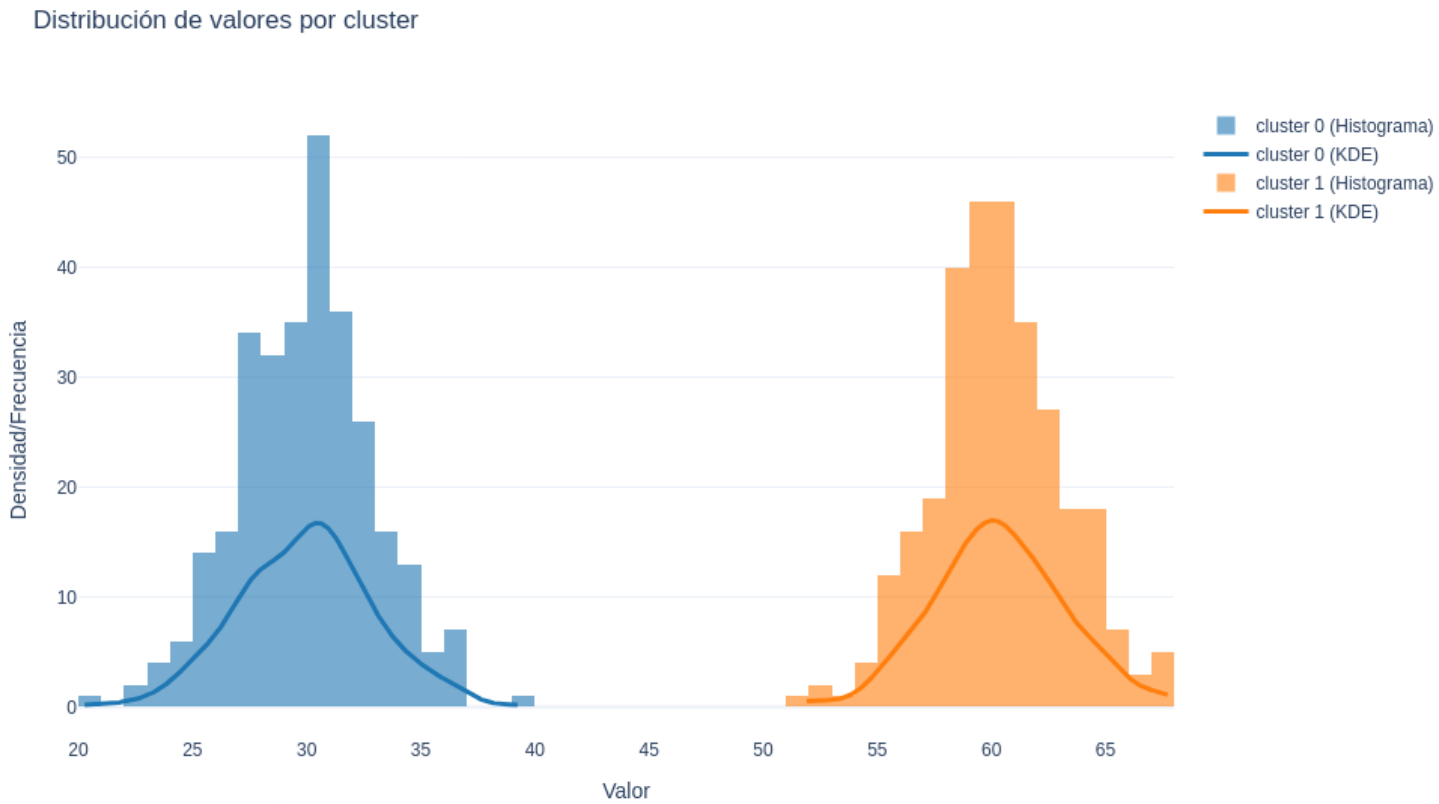
Ejemplo 1

Consideremos dos clusters con las siguiente distribución respectivamente: Cluster 1 $\sim \mathcal{N}(\mu_1 = 30, \Sigma = 3)$ y Cluster 2 $\sim \mathcal{N}(\mu_2 = 60, \Sigma = 3)$.



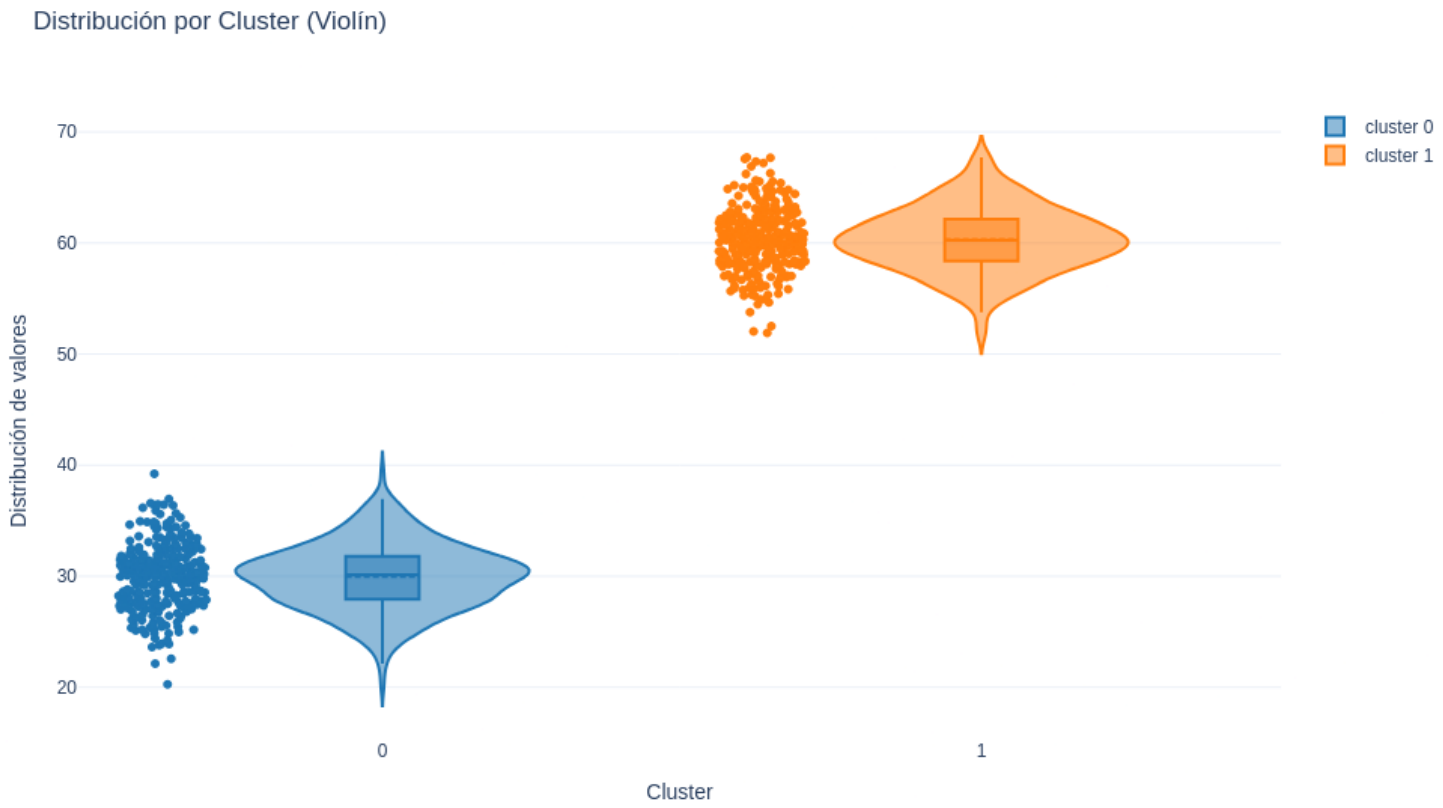
Ejemplo 1

Consideremos dos clusters con las siguiente distribución respectivamente: Cluster 1 $\sim \mathcal{N}(\mu_1 = 30, \Sigma = 3)$ y Cluster 2 $\sim \mathcal{N}(\mu_2 = 60, \Sigma = 3)$.



Ejemplo 1

Consideremos dos clusters con las siguiente distribución respectivamente: Cluster 1 $\sim \mathcal{N}(\mu_1 = 30, \Sigma = 3)$ y Cluster 2 $\sim \mathcal{N}(\mu_2 = 60, \Sigma = 3)$.



¿Qué diferencias hay con un Boxplot?

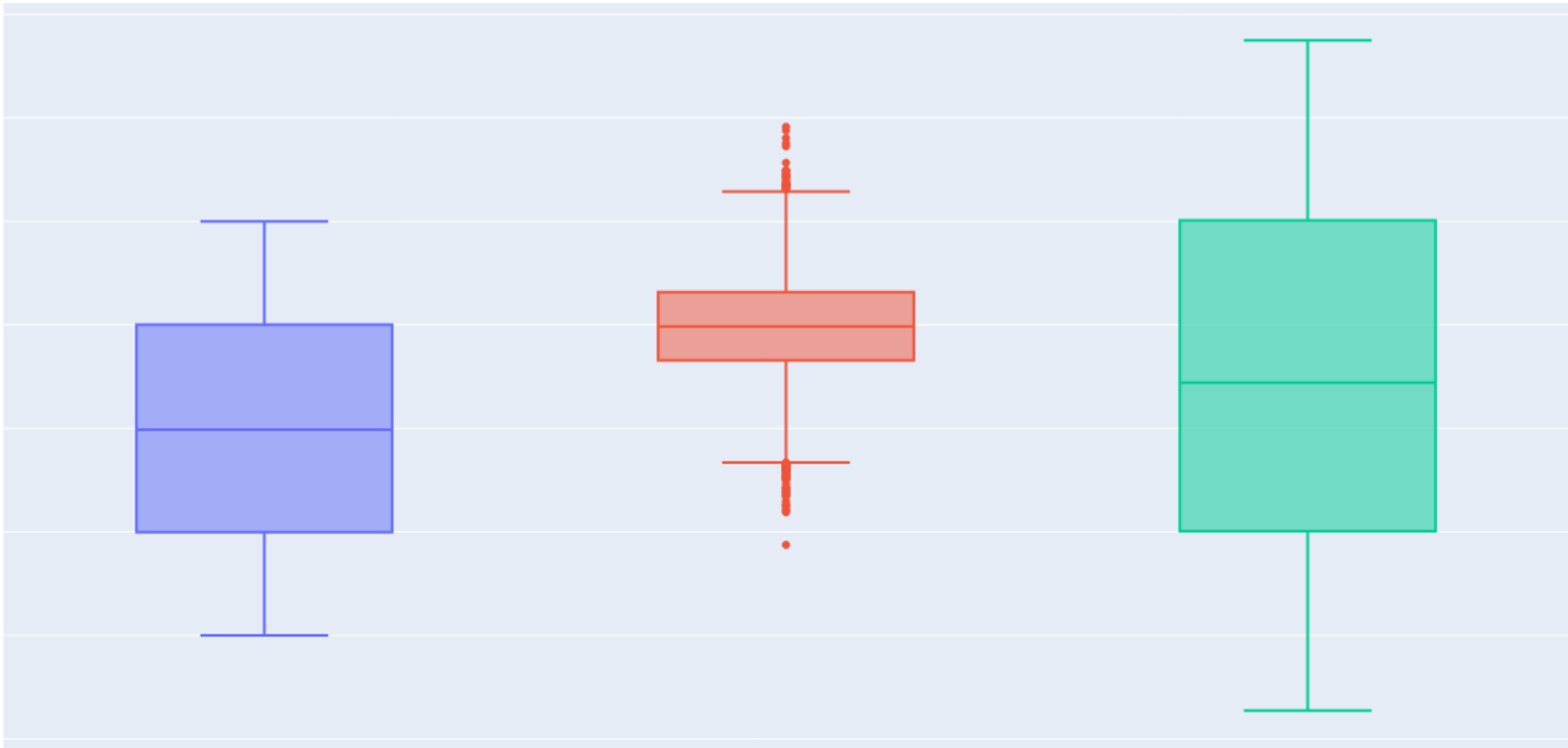
- Entrega un mejor entendimiento respecto a la forma de la distribución de los datos.
- Muestra la existencias de clústeres.
- Resalta los peaks, valles y bumps de la distribución.

Ejemplo 2

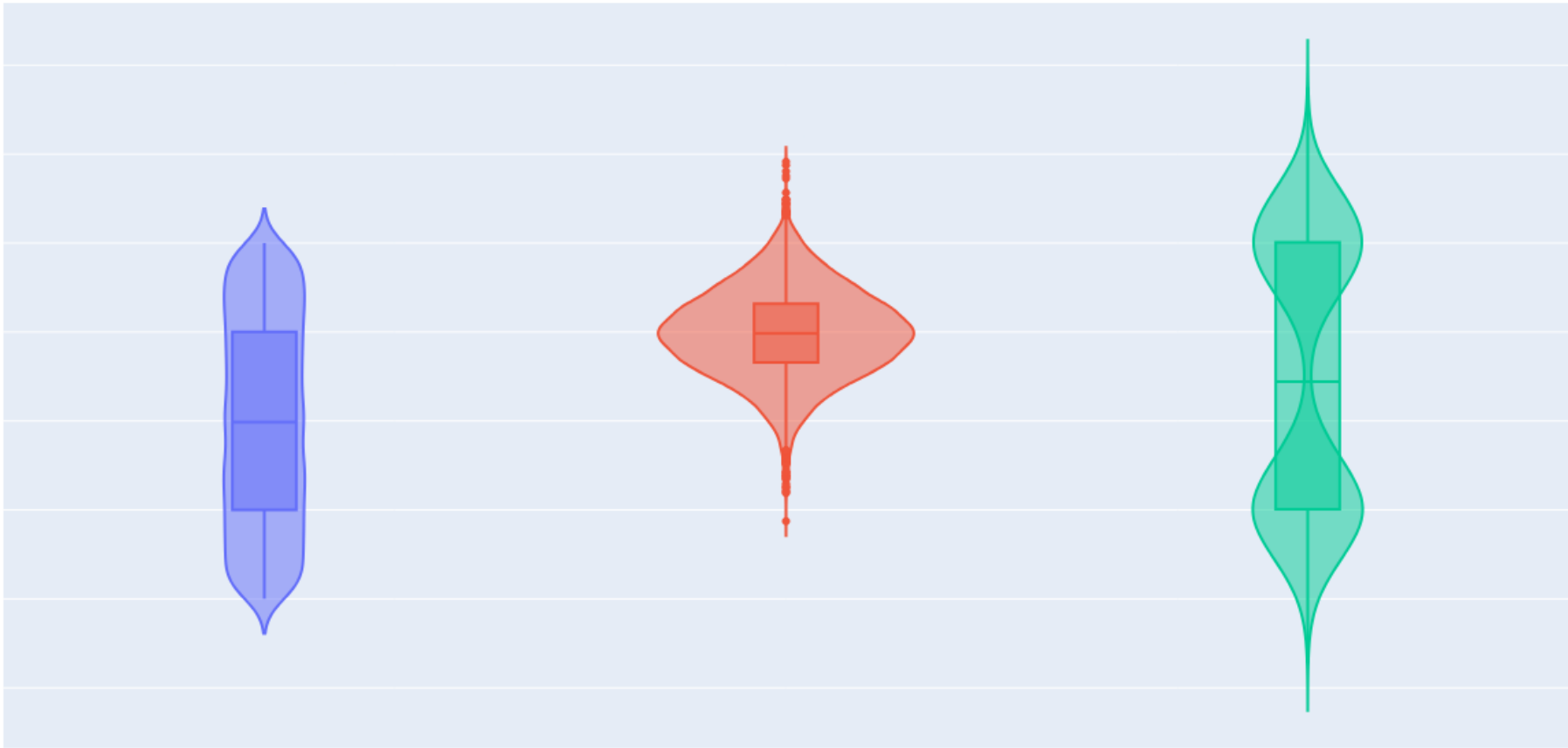
Consideremos los siguientes tres samples de 10.000 elementos cada uno:

- Sample 1 $\sim \mathcal{N}(\boldsymbol{\mu} = 30, \Sigma = 5) + \mathcal{N}(\boldsymbol{\mu} = 60, \Sigma = 5)$
- Sample 2 $\sim \mathcal{N}(\boldsymbol{\mu} = 50, \Sigma = 5)$
- Sample 3 $\sim \mathcal{U}(\text{min} = 20, \text{max} = 60)$

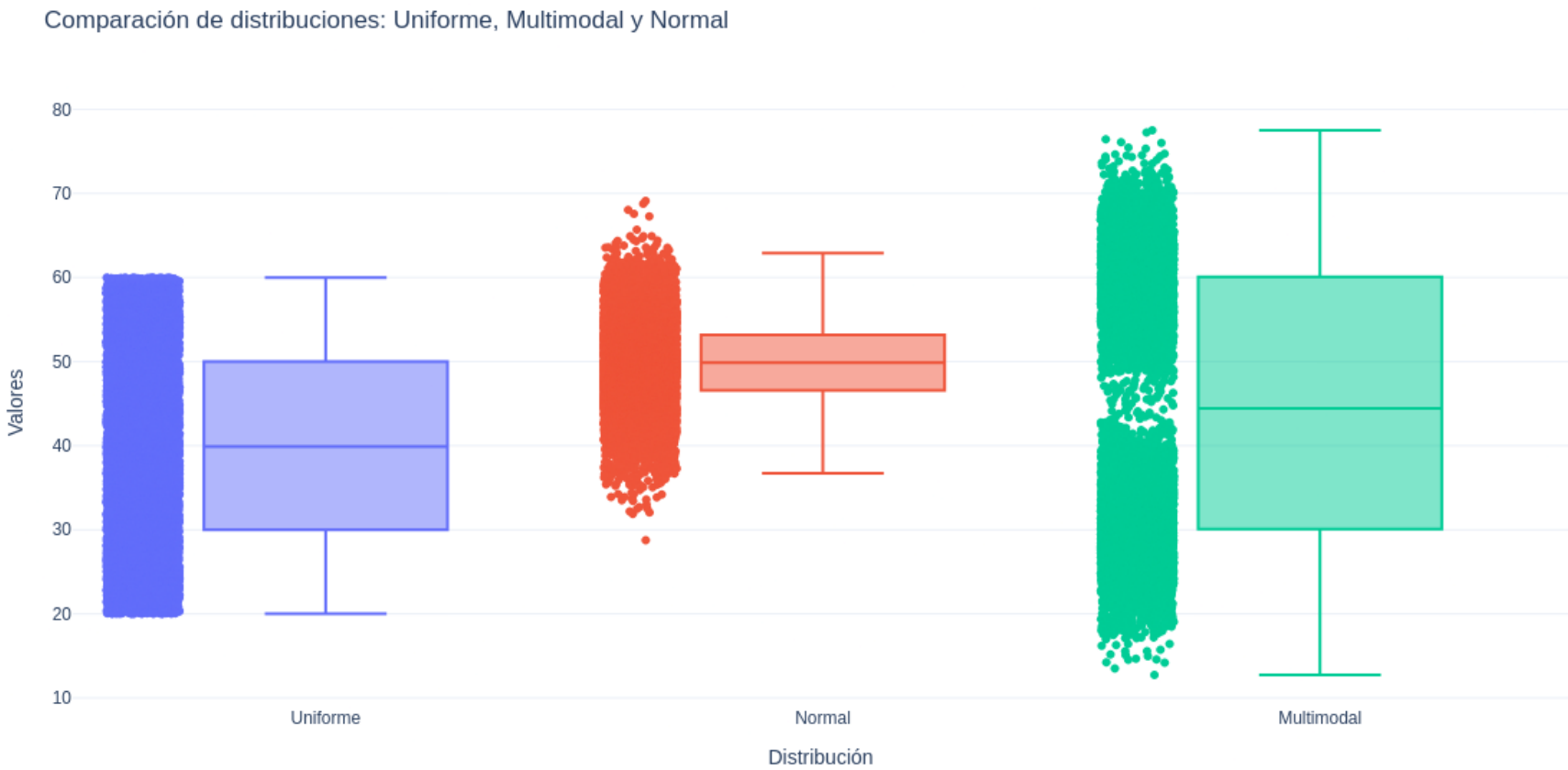
II.II ▷ Estadística Descriptiva - Visualización: Violin Plots



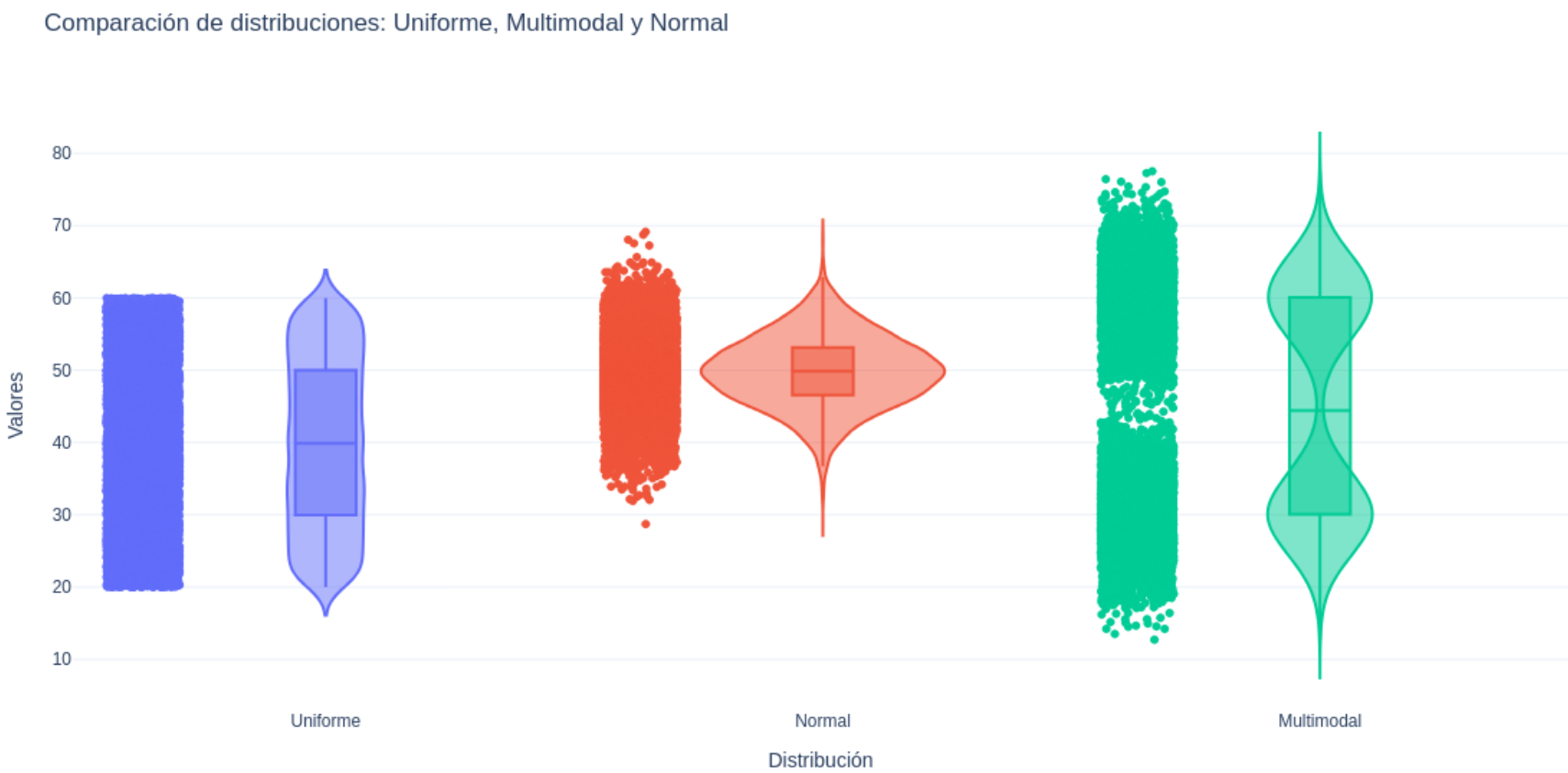
II.II ▷ Estadística Descriptiva - Visualización: Violin Plots



II.II ▷ Estadística Descriptiva - Visualización: Violin Plots



II.II ▷ Estadística Descriptiva - Visualización: Violin Plots



Pre-Procesamiento de Datos

¿Por qué es importante la calidad de los datos?

- Buscamos que nuestros datos tengan las siguientes características:
  - **Precisión** (Accuracy)
  - **Completitud** (Completeness)
  - **Consistencia** (Consistency)

- Los pasos más comunes para el preprocesamiento son:
  - **Data Cleaning**: Manejo de valores faltantes, ruido y outliers
  - **Data Integration**: Combinación de múltiples fuentes

III.I ▷ Data Cleaning

- La limpieza de los datos busca corregir datos faltantes, suavizar el ruido en los datos, o simplemente identificar o remover inconsistencias como los outliers.

- Consideremos los siguientes escenarios clásicos:
  1. Noisy Data
  2. Missing Values

III.I ▷ Data Cleaning - Noisy Data

- El ruido en los datos (*noise*) es un error aleatorio dentro de estos.
- Las Visualizaciones de estadística descriptiva suelen identificarlo, por ejemplo en los boxplot o violin plots.
- Para eliminarlo, debes *'suavizar'* los datos (smoothing process).

- La técnica mas popular de suavizamiento es **Binning**:
  - Dividir los **datos ordenados** en 'bins' (o 'buckets')
  - Reemplazar los valores originales dentro de cada bin por un valor representativo
  - Los reemplazos pueden ser utilizando: media (smoothing by bin means), mediana (smoothing by bin medians), min-max ((smoothing by bin boundaries).

Ejemplo 3 - Binning

Consideremos la siguiente distribución de datos:

```
x = np.linspace(0, 10, 100)
y = np.sin(x) + np.random.normal(0, 0.3, 100)
```


Comparación de Técnicas de Binning para Datos Ruidosos
Datos Originales
Binning por Media
Binning por Mediana
Binning por Fronteras
1.5
1
0.5
0
−0.5
−1
−1.5
0
2
4
6
8
10

Data Cleaning - Noisy Data

Ejemplo 4

```
data = {
    'Nombre': ['Ana', 'Carlos', 'Beatriz', 'David', 'Elena', None],
    'Edad': [25, np.nan, 32, 28, None, 40],
    'Puntuacion': [85, 92, None, 78, 88, np.nan],
    'Departamento': ['Ventas', None, 'IT', 'IT', 'Ventas', 'HR']
}
```

- Los *missing values* suelen verse reflejada en aquellos valores `None` o `np.nan` que vemos en los datsets.
- **Importante !** No siempre los *missing values* implican un error
  $\rightarrow$ e.g. Preguntas opcionales en encuestas.

¿Que alternativas tenemos?

- Ignorar los registros con *missing values*. No es efectivo, a menos que el feature tengo demasiados missing values.
- Rellenar los valores faltantes de manera manual. Costoso en tiempo y recursos.
  - Usar una constante para identificar los *missing values*, como `'Missing'`, `-1`, `0`, `'Unknown'` $\rightarrow$ No siempre es 'foolproof'.

- Imputación de valores:
  - Usar una medida de tendencia central, como *Mean Imputation* o *Median Imputation*. Para datos distribuidos normalmente (simétricos), usar la media. Para datos asimétricos (skewed data), usar la mediana.
  - Usar los *'valores más probables'*. Por medio de una regresión, o usando inferencia bayesiana.

Imputación de valores

- Problema: Inyección de sesgo estadístico en los datos (Data Bias).
- Alternativas actuales → **k-Nearest Neighbors Imputation**: los missing values son imputados por valores más cercanos de acuerdo a una métrica de *similaridad* respecto a patrones en el dataset.

Ejemplo 5 - `sklearn.impute.SimpleImputer` - Mean - Axis 0

```
Datos originales:
[[ 1.  2. nan]
 [ 4. nan  6.]
 [ 7.  8.  9.]
 [nan 11. 12.]]

Datos imputados con la media:
[[ 1.  2.  9.]
 [ 4.  7.  6.]
 [ 7.  8.  9.]
 [ 4. 11. 12.]]

Valores usados para la imputacion (medias):
[4. 7. 9.]
```

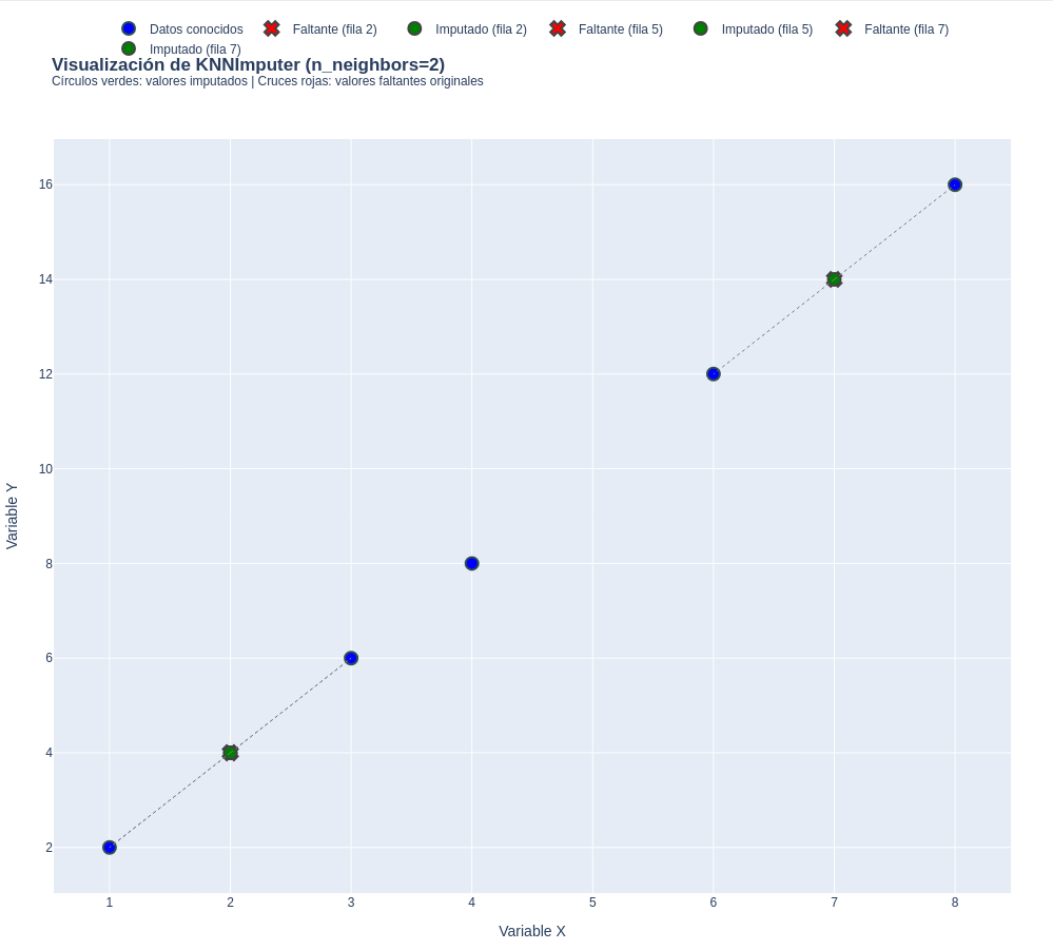
Ejemplo 6 - `sklearn.impute.SimpleImputer` - Median - Axis 0

```
Datos originales:
[[ 10.  20.  nan]
 [ 40.  nan  60.]
 [ 70.  80.  90.]
 [ nan 110. 120.]]

Datos imputados con la mediana:
[[ 10.  20.  90.]
 [ 40.  80.  60.]
 [ 70.  80.  90.]
 [ 40. 110. 120.]]

Valores usados para la imputacion (medianas):
[40. 80. 90.]
```

Ejemplo 7 - `sklearn.impute.KNNImputer` - k=2 - Axis 0




```
Datos originales con
      valores faltantes:
[[ 1.   2.]
 [ 2.  nan]
 [ 3.   6.]
 [ 4.   8.]
 [nan   4.]
 [ 6.  12.]
 [ 7.  nan]
 [ 8.  16.]]

Datos despues de la
      imputacion KNN:
[[ 1.   2.]
 [ 2.   4.]
 [ 3.   6.]
 [ 4.   8.]
 [ 2.   4.]
 [ 6.  12.]
 [ 7.  14.]
 [ 8.  16.]]
```

- Caso típico: fusión de múltiples fuentes de datos en un único conjunto.

- Problemas frecuentes en la integración:
  - **Redundancia**: Ocurre cuando un dato puede derivarse de otros atributos.
  - **Inconsistencia**: Puede surgir como consecuencia de redundancias en los datos.

- Detección de redundancias mediante **análisis de correlación**:
  - **Test de correlación** χ^2: Para variables nominales/categóricas.
  - **Coeficiente de correlación** y **covarianza**: Para variables numéricas.

III.II ▷ Data Integration - Test de correlación χ^2

- Supongamos que buscamos la correlación entre dos atributos A y B, nominales, de un dataset.
- A tiene c valores distintos $a_1, a_2, ..., a_c$, mientras que B tiene r valores distintos $b_1, b_2, ..., b_r$.
- Definimos la **tabla de contingencia** como la matriz de c columnas y r filas, tal que (A_i, B_j) denote **frecuencia observada combinada** de que el atributo a_i de A tome el valor de ocurrencia del atributo b_j de B.

Definición

Dada una tabla de contingencia con r filas y c columnas, el valor del test de correlación χ^2 es:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

donde:

- o_{ij} = Valor observado en la celda (i, j), **frecuencia observada** del evento (A_i, B_j).
- $e_{ij} = \frac{\text{Total fila } i \times \text{Total columna } j}{\text{Gran total}}$ y es la *frecuencia esperada* del evento (A_i, B_j).
- Grados de libertad: $(r-1)(c-1)$

III.II ▷ Data Integration - Test de correlación χ^2

El Test de correlación χ^2 tiene por hipótesis que A y B son **independientes**, osea que no están correlacionados entre ellos.

Ejemplo 8 - `scipy.stats. chi2_contingency`

```
Tabla de Contingencia Observada (o_{ij}):

             Pop  Rock  Clasica  Jazz
Secundaria    45    30       25    10
Pregrado      20    40       30    20
Maestria      10    15       35    40
Doctorado      5    10       20    35
```

```
Resultados del Test:
Estadistico chi-square: 86.6467408190
Valor-p: 0.0000000000
Grados de libertad: 9
```

```
Tabla de Valores Esperados (e_{ij}):
              Pop   Rock  Clasica   Jazz
Secundaria  22.56  26.79    31.03  29.62
Pregrado    22.56  26.79    31.03  29.62
Maestria    20.51  24.36    28.21  26.92
Doctorado   14.36  17.05    19.74  18.85

Interpretacion:
Rechazamos H0 (p=0.0000 < 0.05). Existe relacion significativa entre nivel educativo y
     preferencia musical.
```

- Buscamos evaluar cómo varían conjuntamente dos atributos numéricos A y B respecto a sus **medias**.

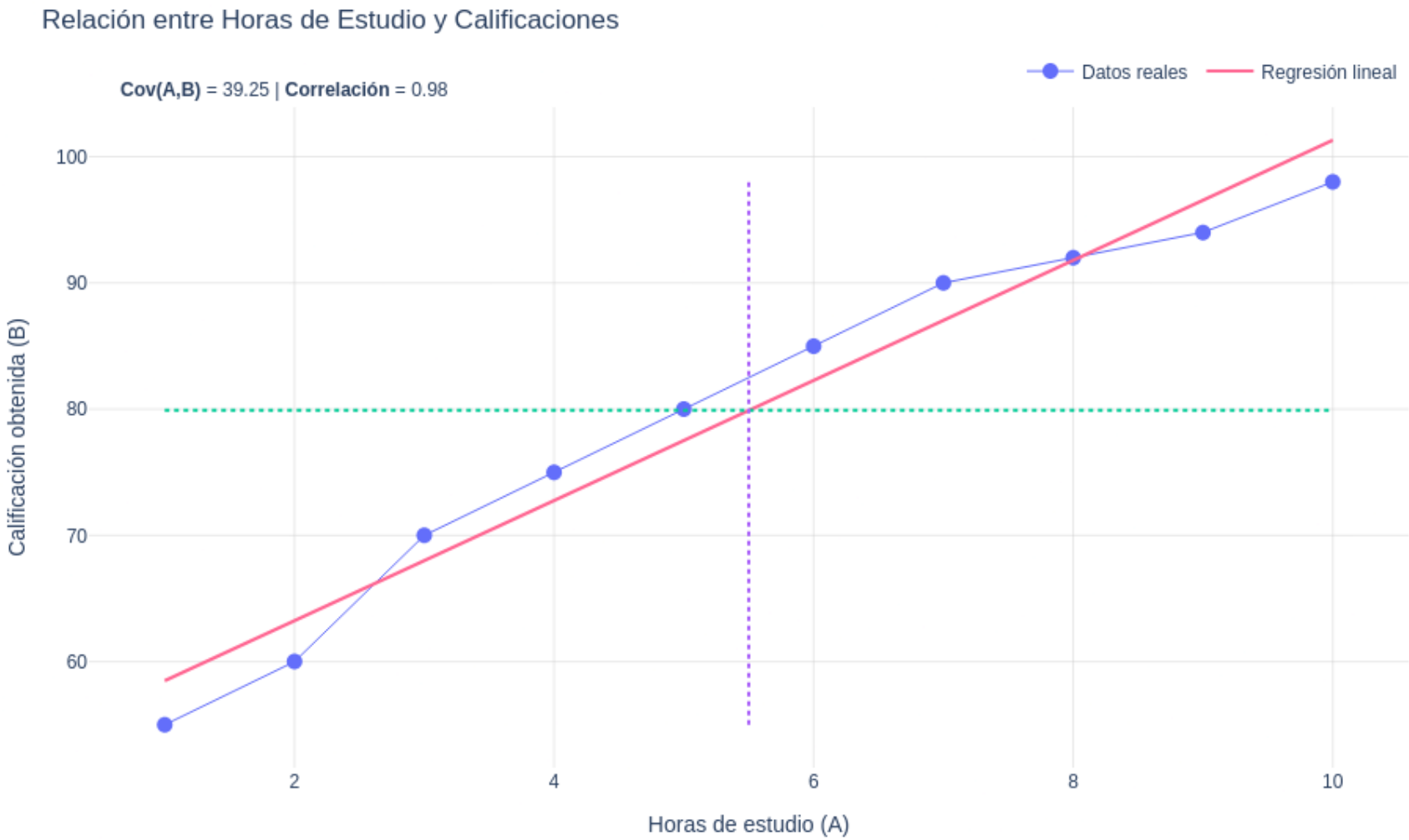
Definición

Dadas dos variables numéricas A y B, con n observaciones cada una, la **covarianza** se define como:

$$\text{Cov}(A, B) = \frac{1}{n} \sum_{i=1}^{n} (A_i - \bar{A})(B_i - \bar{B})$$

- A_i y B_i: Valores individuales de las variables A y B.
- $\bar{A}$ y $\bar{B}$: Medias muestrales de A y B, respectivamente.
- n: Número de observaciones.

- Covarianza > 0: Relación directa (ambas variables tienden a aumentar o disminuir juntas).
- Covarianza < 0: Relación inversa (una aumenta mientras la otra disminuye).
- Covarianza $= 0$: No hay relación lineal (puede haber independencia o una relación no lineal).


Relación entre Horas de Estudio y Calificaciones
Cov(A,B) = 39.25 | Correlación = 0.98
Datos reales
Regresión lineal
Calificación obtenida (B)
100
90
80
70
60
2
4
6
8
10
Horas de estudio (A)

- Conocido como *Pearson correlation coefficient* o *Pearson's product moment coefficient*.
- Buscamos evaluar la correlación entres dos atributos numéricos A y B de un dataset.

Definición

El coeficiente de correlación r_{AB} se define como:

$$r_{AB} = \frac{\text{Cov}(A,B)}{\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{A})^2}\sqrt{\sum_{i=1}^{n}(B_i - \bar{B})^2}}$$

donde:

- $\bar{A}$ y $\bar{B}$ son las medias muestrales
- $\text{Cov}(A,B)$ es la covarianza entre A y B
- σ_A, σ_B son las desviaciones estándar

Propiedades

- $-1 \leq r_{AB} \leq 1$
- $r_{AB} = 1$: Correlación lineal positiva perfecta
- $r_{AB} = -1$: Correlación lineal negativa perfecta
- $r_{AB} = 0$: No hay correlación lineal

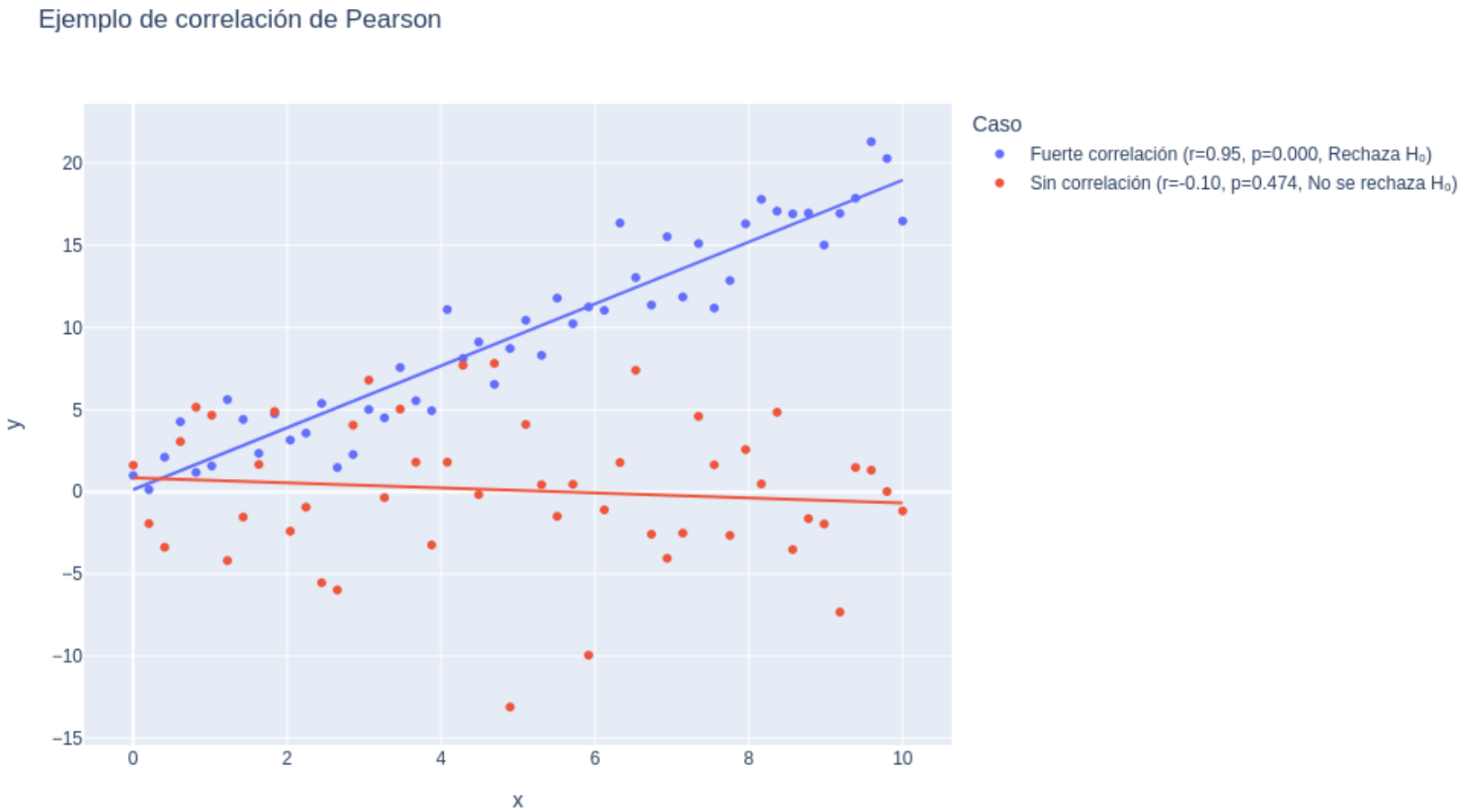
Ejemplo 9

```
x1 = np.linspace(0, 10, 50)
y1 = 2 * x1 + np.random.normal(0, 2, size=len(x1))

x2 = np.linspace(0, 10, 50)
y2 = np.random.normal(0, 5, size=len(x2))
```

III.II ▷ Data Integration - Coeficiente de Correlación

El Coeficiente de Correlación tiene por hipótesis que A y B son **independientes**, osea que **no** están correlacionados entre ellos.[1]



[1] Rechazamos con $p - value < 0.05$.