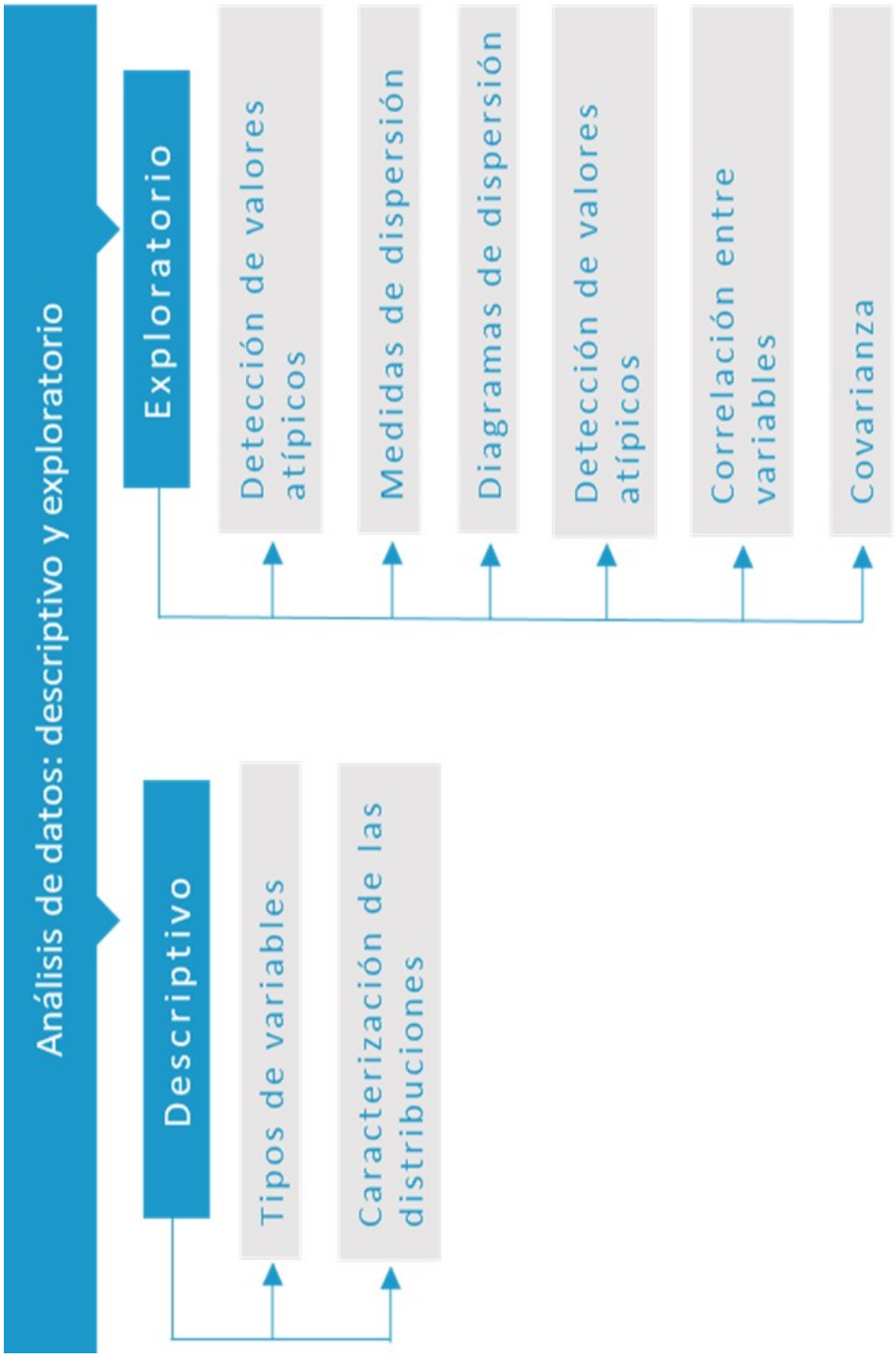
Técnicas de Aprendizaje Automático

Tema 2. Análisis de datos. Descriptivo y exploratorio

Índice

Esquema



2.1. Introducción y objetivos

El presente tema tiene como propósito conocer los principios básicos de un análisis descriptivo de los datos, esto implica recolectar, organizar, presentar y describir un conjunto de datos. Una vez se describen los datos se continúa con el análisis exploratorio de los mismos EDA (Exploratory Data Analysis, por sus siglas en inglés), utilizando la visualización y la transformación para explorar los datos de manera sistemática.

- Conocer diferentes procesos de organización y presentación de los datos a través de herramientas como los histogramas.
- Calcular las medidas de dispersión para un conjunto de datos.
- Interpretar la representación de los datos a través de diagramas de caja.
- Calcular el coeficiente de Pearson para dos atributos de un conjunto de datos y su interpretación.
- Estimar la matriz de covarianza para un conjunto de atributos.

2.2. Tipos de variables

Una de las primeras etapas del análisis descriptivo es reconocer los diferentes tipos de datos, ya que estas observaciones son una representación del mundo que nos rodea a través de las denominadas variables. Una variable se define como «característica o atributo de un objeto que puede ser observable, siendo susceptible de variación, debido a que puede medirse, asumiendo diferentes valores» (Dicovskiy, 2008). Otros autores definen variable como los diferentes valores que puede tomar una característica (Rustom, 2012).

La representación de una variable puede darse de diferentes formas, no solo numérica (datos discretos y continuos), sino también categóricos y nominales como se observa en la Figura 1.

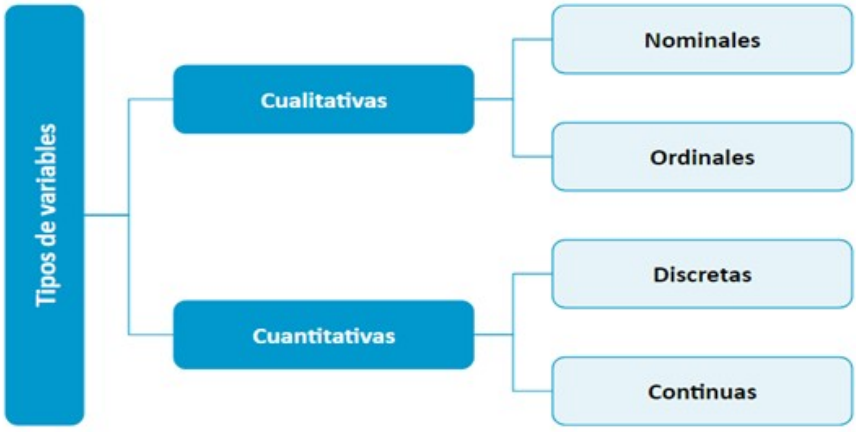


Figura 1. Clasificación de los tipos de variables. Fuente: elaboración propia.

A continuación, se realizará una descripción de cada uno de los tipos de variable:

Variable cualitativa

Es aquella que expresa una característica de forma no numérica, es decir, representa una cualidad; por ejemplo, color, género, grupo sanguíneo, estado civil. A su vez, puede ser de dos tipos (Gorgas García *et al.*, 2011):

- **Nominal:** es aquella cuyos valores son códigos o nombres que no se pueden ordenar; por ejemplo, nacionalidad, religión, color de piel.
- **Ordinal:** es aquella cuyos valores conllevan a un ordenamiento de mayor a menor o de mejor a peor; por ejemplo, intensidad del dolor (ausente, leve, moderado, severo, muy severo), calificación (excelente, bueno, regular, malo).

Variable cuantitativa

Se expresa mediante un número. Puede ser de dos tipos:

- **Discreta:** cuando solo admite tomar una cantidad de valores numéricos y usualmente son valores enteros; por ejemplo, número de hijos por hogar, número de computadores en un aula, el número de electrones de un átomo (Gorgas García *et al.*, 2011).
- **Continua:** entre dos valores distintos de la variable, puede haber infinitos posibles valores y pueden ser datos enteros o fraccionarios. Estos datos o variables continuas representan mediciones, por lo que sus posibles valores no pueden enumerarse o ser contados, debido a esto admite cualquier valor dentro de un rango o intervalo en la recta de los números reales; por ejemplo, peso (Kg), estatura, edad, el tiempo que demora un atleta en recorrer 100 metros; cualquier medida que tomemos del mundo real, siempre pueden darse pequeñas o grandes variaciones (Hérnandez, 2012; López Briega, 2016).

Tipos de datos usados en Python

Cuando se realiza un análisis de datos categóricos en el lenguaje de programación Python (a partir del cual se generarán una serie de gráficos ejemplo a lo largo de este módulo), es necesario, en primer lugar, conocer sobre los tipos de datos básicos, dado que son los que establecen qué valores puede tomar una variable y qué operaciones se pueden realizar sobre la misma (ver Figura 2).

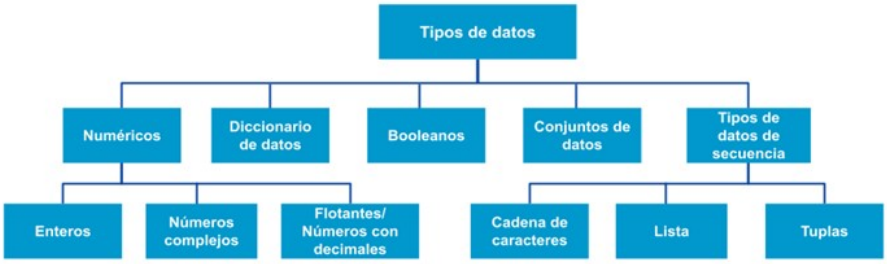


Figura 2. Tipos de datos en Python. Fuente: adaptado de Geeksforgeeks, 2024.

En la Figura 3 se pueden apreciar en la columna izquierda los cuatro componentes numerados desde 0 (cero) y frente a cada uno (en la columna de la derecha), el tipo de dato (Dtype) *float*. Así, conocemos que el total de las características corresponden a datos *float* con decimales. Se destaca, además, DataFrame, la cual es una estructura de datos de etiquetado bidimensional con columnas de tipos potencialmente diferentes, siendo necesario la biblioteca de pandas para su ejecución:

Visualización de los tipos del dataframe

```
import numpy as np

import pandas as pd

df = pd.read_csv('wine.data', header=None, usecols=[1,2,3,4]])

df.info()
```

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Alcohol           178 non-null    float64
 1   Malic acid        178 non-null    float64
 2   Ash               178 non-null    float64
 3   Alcalinity of ash 178 non-null    float64
dtypes: float64(4)
memory usage: 5.7 KB
```

Figura 3. Visualización de los tipos del dataframe. Fuente: elaboración propia.

2.3. Caracterización de las distribuciones de las variables

Histograma

Es la manera más común de representar gráficamente la distribución de frecuencia de los datos agrupados tomando como entrada solo una variable numérica, por tanto, es comúnmente utilizado para variables cuantitativas continuas. Las clases o categorías de estas distribuciones están formadas mediante intervalos (Figura 4).

Gráficamente, está compuesto por una serie de rectángulos adyacentes, cada uno de los cuales representa a una categoría, con la condición de que el área de cada uno de ellos es igual o proporcional a la frecuencia de la categoría que representa (Salazar y Castillo, 2018). Los valores de los elementos se dividen en contenedores (*bins*) y se crean gráficos de barra por cada contenedor (Bravo Márquez, 2009); su base corresponde a cada intervalo de clase y la altura de cada barra indica el número de elementos o frecuencia del contenedor. La altura se puede calcular como el cociente entre la frecuencia (absoluta o relativa) y la amplitud del intervalo (Gorgas García *et al*., 2011); por lo tanto, en el eje vertical (Y) se representan las frecuencias de los datos, considerando que la frecuencia de un intervalo corresponde al número de datos que se encuentran en él; y en el eje horizontal (X) se representan los valores de las variables, normalmente señalando las marcas de clase, las cuales indican la mitad del intervalo en el que están agrupados los datos, por lo que efectivamente el centro de la base de cada rectángulo o barra corresponde a una marca de clase (Rodríguez Ojeda, 2007). Teniendo en cuenta la información proporcionada por la base de datos Wine, en donde se realiza el conteo de los datos por tipo de vino, se presenta la Figura 4, la cual contiene un histograma agrupado correspondiente a los cuatro primeros componentes (alcohol, ácido málico, ceniza, alcalinidad de la ceniza) de los trece que en total integran cada uno de los tres vinos:

Representación de un histograma

```
histogram = df.plot.hist(bins= 8, edgecolor='black', color = colors )

print(histogram)

plt.show()
```

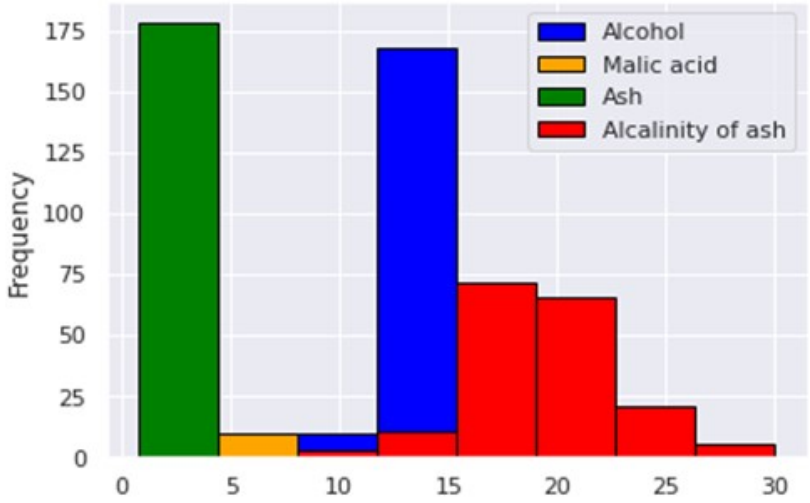


Figura 4. Representación de un histograma. Fuente: elaboración propia.

En la Figura 4 se puede identificar que los datos se han dividido, agrupándolos en intervalos consecutivos. Cada una de las barras corresponde a un intervalo o clase diferente y representan características de los vinos (diferenciados por colores con su respectiva nomenclatura); en el eje X (horizontal) se encuentran las cantidades en que se presentan dichos componentes y en el eje Y (vertical) se pueden ver las frecuencias absolutas. Es decir, las veces que dichos componentes se encuentran en los vinos. Para la construcción del histograma, se toman los valores de rango y valor de K estimado (Bernal, 2017).

Tipos de simetría

El histograma permite una primera visualización del tipo de distribución de datos (Figura 5). Su forma depende del número de intervalos, y la simetría, es decir la manera en la que los valores se reparten de la misma forma a uno y otro lado del centro, puede variar; cuanto más similares sean, más simétrica será la distribución; cuanto más distintos, más asimétrica (Rodríguez Ojeda, 2007):

- Si la altura o eje y de las barras son similares unas a otras, se tendría una distribución uniforme.
- Si las alturas son mayores en la zona central, se forma una «campana», la cual puede ser simétrica o asimétrica, hacia el lado positivo (a la derecha) o negativo (hacia la izquierda)
- Si hay barras alejadas en el grupo, se considera que son datos atípicos, los cuales probablemente se deban a errores de medición y pueden ser descartados, ya que no pertenecen al grupo que se desea caracterizar (Rodríguez Ojeda, 2007).

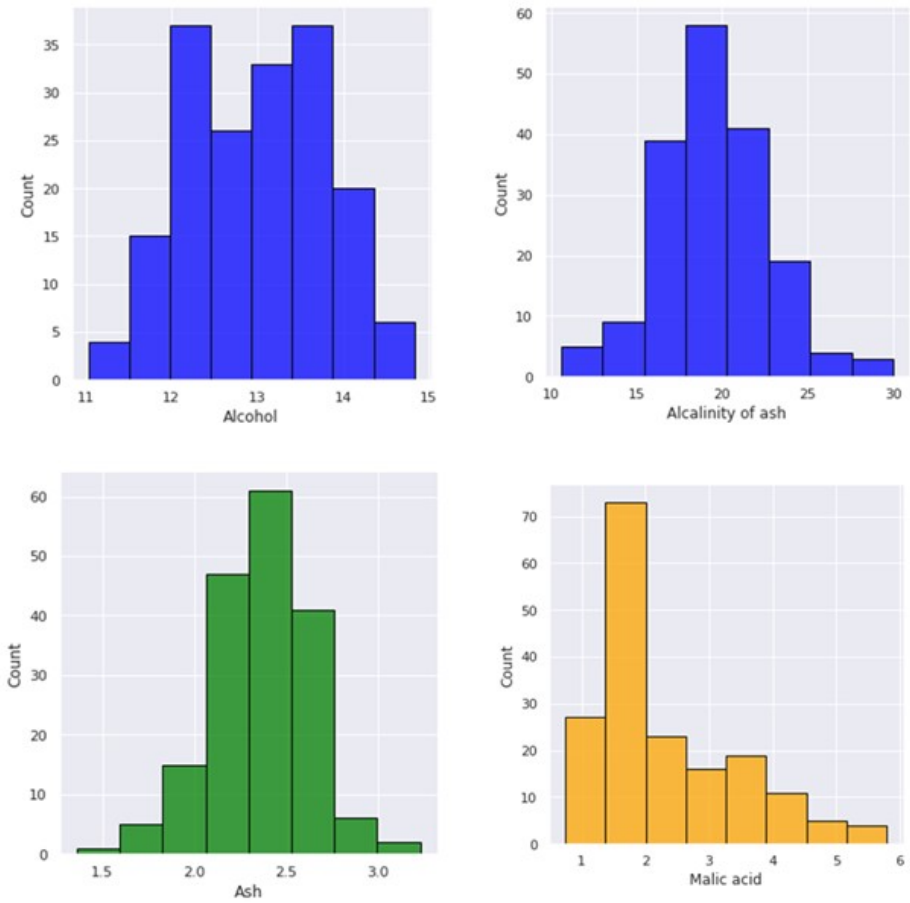


Figura 5. Representación de cuatro histogramas correspondientes a los cuatro primeros componentes de los vinos. Fuente: elaboración propia.

Como observamos en la figura, el histograma para alcohol tiene tendencia a una distribución simétrica unimodal, el histograma de ácido málico tiene tendencia a una distribución asimétrica a la derecha.

2.4. Medidas de dispersión

En un conjunto de datos existe la tendencia a agruparse hacia el centro de la distribución de frecuencias. Sin embargo, los datos extremos pueden estar bastante alejados de esa tendencia central. En el tratamiento estadístico de datos, medir esa distancia respecto a los promedios es un cálculo importante. A estas medidas se les denomina **de dispersión o de variación**.

Rango

Se trata de la diferencia entre el límite superior y el límite inferior de un conjunto de datos (Witte y Witte, 2017). Para su cálculo solo se requiere que los datos brutos estén ordenados. Como medida de dispersión, su utilidad es limitada, pues se deja afectar fácilmente por los valores extremos de poca frecuencia. Además, el valor del rango tiende a aumentar a medida que el número de datos es mayor. La influencia de los extremos en el cálculo del rango debe ser suprimida, y, para ello, es común usar el rango intercuartílico que consiste en calcular la diferencia entre el tercer cuartil y el primero.

Rango de una distribución

	Alcohol	Malic acid	Ash	Alcalinity of ash
count	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944
std	0.811827	1.117146	0.274344	3.339564
min	11.030000	0.740000	1.360000	10.600000
25%	12.362500	1.602500	2.210000	17.200000
50%	13.050000	1.865000	2.360000	19.500000
75%	13.677500	3.082500	2.557500	21.500000
max	14.830000	5.800000	3.230000	30.000000

Figura 6. Medidas de tendencia central y de dispersión. Fuente: elaboración propia.

Mediante el comando df.describe en Python, se detallan las medidas de tendencia central y de dispersión de los cuatro conjuntos de datos correspondientes a los atributos *alcohol*, *malic acid*, *ash* y *alkalinity of ash*. Al tener los valores máximos y mínimos de cada uno de los atributos, el rango es simplemente la diferencia entre tales valores:

Rango_alcohol=14,83-11,03=3,80

Rango_(malic acid)=5,80-0,74=5,06

Rango_ash=3,23-1,36=1,87

Rango_(alkalinity of ash)=30,00-10,60=19,40

Cuartiles y Rango Intercuartílico

El rango intercuartílico (RIC) es simplemente el rango para el 50 % de los datos que se encuentra en el medio de la distribución. Los cuartiles se hallan dividiendo la distribución en cuatro partes iguales, de modo que cada una contenga el 25 % de las observaciones. Los tres puntos de separación de los valores son los cuartiles. El cuartil inferior (Q1) representa el 25 % de las observaciones y es superado por el 75 % restante. El segundo cuartil (Q2) corresponde a la mediana de la distribución. El tercer cuartil (Q3) representa el 75 % y es superado por el 25 % restante de las observaciones.

Específicamente, el RIC equivale a la distancia entre el tercer cuartil (percentil 75) y el primer cuartil (percentil 25); es decir, es el resultado de remover los cuartos superiores (25 % más alto) e inferior (el 25 % más bajo) del conjunto original de datos (Witte y Witte, 2017). Dado que la mayoría de las distribuciones están más

dispersas en las extremidades que en el medio, el RIC tiende a ser menor que la mitad del rango. Una propiedad clave del RIC es su resistencia al efecto de distorsión de los valores extremos o atípicos.

Rango Intercuartílico

Del mismo modo que en el ejemplo previo, el rango intercuartílico para cada atributo es la diferencia de dos valores representados en la tabla generada en Python, los cuartiles terceros y primero:

RIC_alcohol=13,6775-12,3625=1,315

RIC_(malic acid)=3,0825-1,6025=1,48

RIC_ash=2,5575-2,2100=0,3475

RIC_(alkalinity of ash)=21,5-17,2=4,3

En Python se calcularía utilizando la librería `numpy` y la clase `np`, función percentil:

```
q3, q1= np.percentil(dataframe, [75,25])
```

```
iqr=q3-q1.
```

Varianza

Aunque el rango y su derivado, el rango intercuartílico, son medidas de dispersión válidas, una de las medidas más usadas en estadística es la varianza, que a su vez da origen a la **desviación estándar**, mucho más significativa. Es definida como el promedio de los cuadrados de las desviaciones de cada dato respecto a la media aritmética. El promedio de las desviaciones no es una medida útil, dado que su sumatoria va a resultar cero, pues las desviaciones negativas cancelan las positivas. Por esta razón, se elevan al cuadrado las desviaciones, eliminando como resultado

los signos negativos. Los símbolos s^2 y σ^2 representan la varianza muestral y la varianza poblacional, respectivamente (Spiegel y Stephen, 2009). Dos distribuciones pueden ser comparadas en cuanto a su variabilidad absoluta, teniendo en cuenta sus varianzas, de manera que el resultado indique cuál de ellas es más homogénea o heterogénea. Sin embargo, la varianza tiene la desventaja de presentar valores con unidades cuadradas cuya interpretación es compleja.

Desviación típica

Para resolver el problema de las unidades de medida paradójicas se calcula la raíz cuadrada de la varianza, **tomando siempre el valor positivo**. Esto produce una nueva medida, conocida como **desviación típica o estándar** que describe la variabilidad en las unidades de medida originales. Como las anteriores medidas, la desviación típica debe distinguirse simbolizada por s en la muestra y σ en la población.

2.5. Detección de valores atípicos

Un valor atípico es una observación numéricamente distante del resto de los datos. Los valores atípicos en un conjunto de datos pueden causar problemas en los análisis estadísticos al generar conclusiones engañosas. Usualmente, un valor atípico puede indicar un error de medición o una distribución de cola larga en la población, aunque pueden ocurrir de manera fortuita.

Diagrama de caja

Es una representación gráfica de una serie de datos numéricos a través de sus cuartiles. Las cajas dispuestas en este método estandarizado contienen los siguientes elementos: **la mediana (Q2), el rango intercuartil y los cuartiles (Q1 y Q3)**. Los diagramas de caja también tienen líneas que se extienden desde las cajas (bigotes) que indican variabilidad fuera de los cuartiles superior e inferior, de ahí el término **diagrama de caja** y **bigotes**. Según el tipo de representación, los bigotes pueden llegar hasta el valor mínimo en un extremo y el valor máximo en el otro, representando el rango completo. Para la identificación de valores atípicos, John Tukey propone unos bigotes cuya longitud es 1,5 veces el RIC (Tukey, 1977), es decir, a partir de las bisagras (Q1 y Q3) se suma o se resta respectivamente 1,5 veces el IRC, luego, los valores más allá de estos bigotes se consideran atípicos y se representan como puntos individuales.

Diagrama de caja

Para la construcción de las cajas en cada conjunto de datos se sigue el siguiente procedimiento en Python:

```
ax = sns.boxplot(data=df, x=name_components[0], color=colors[0] )

plt.show()

ay= sns.boxplot(data=df, x=name_components[1], color=colors[1] )
```

```
plt.show()

az= sns.boxplot(data=df, x=name_components[2], color=colors[2] )

plt.show()

aw= sns.boxplot(data=df, x=name_components[3], color=colors[3] )

plt.show()
```

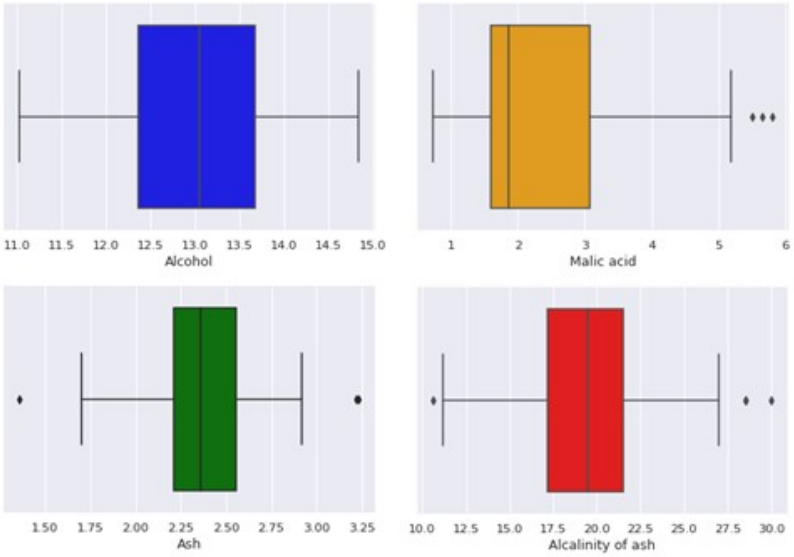


Figura 7. Diagrama de caja. Fuente: elaboración propia.

Para construir los diagramas de caja se utilizan la mediana, los cuartiles primero y tercero, y el rango intercuartílico de cada atributo. Después, para los bigotes, se suma arriba y se resta abajo de la caja 1,5 veces el RIC. Finalmente, se toma el último valor dentro de esos límites y es ahí hasta donde se grafican los bigotes.

2.6. Diagramas de dispersión

Las representaciones gráficas más útiles para describir el comportamiento de un conjunto de variables son los diagramas de dispersión, también conocidos como diagramas de puntos o *scatter plots*. Estas gráficas muestran la relación existente entre las variables numéricas (Ribecca, 2018), además, permiten descubrir y confirmar las relaciones anticipadas entre dos conjuntos asociados de datos (Hernández Medrano, 2017). Las variables numéricas están representadas por coordenadas en el espacio euclidiano de dimensión n.

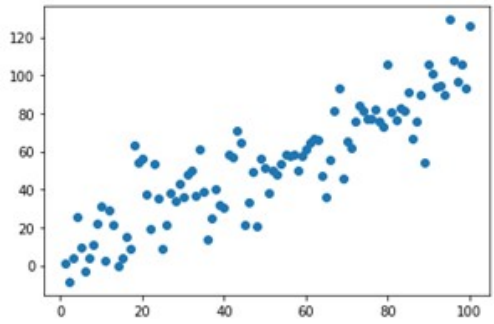
Diagramas de dispersión bidimensional

Para el caso de dos variables, los conjuntos de datos deberán tener la misma longitud, una para el eje X (horizontal) y otra para el eje Y (vertical). En Python se puede obtener un gráfico de dispersión de dos variables numéricas X y Y usando el comando `plot ( X,Y)` de la biblioteca `matplotlib`. En el siguiente ejemplo se generan valores aleatorios entre [0 y 101] para X y entre [-20 y 130] para `Yalmacenados` en un dataframe (Bravo Márquez, 2013).

Ejemplo de diagrama de dispersión en Python

```
import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

import seaborn as sns

from matplotlib.colors import ListedColormap

df=pd.DataFrame({'x_values':range(1,101),'y_values':np.random.randn(100)*15+range(1,10
1) })
```

```
# plot

plt.plot( 'x_values', 'y_values', data=df, linestyle='none', marker='o')

plt.show()
```



La representación de cada variable en cada eje permite identificar la relación de incremento o disminución de los valores (Figura 8), de las variables dependiente e independiente (Ribecca, 2018).

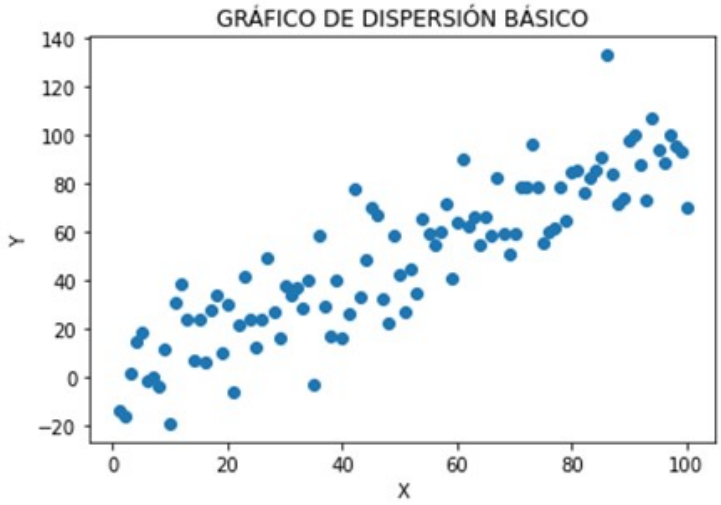


Figura 8. Gráfico de dispersión básico. Fuente: elaborado a partir de The Python Graph Gallery, s.f.

Los diagramas de dispersión son adecuados cuando se tienen datos numéricos emparejados y se desea ver si una variable afecta a la otra. Sin embargo, es necesario entender que la correlación no es causal y otra variable inadvertida puede

influir en los resultados (Ribecca, 2018). En el caso de una base de datos, las variables numéricas corresponden a los atributos de un conjunto de datos y estas se usan para definir aspectos de una instancia, como tamaño, forma y color (Bravo Márquez, 2013). A continuación, se genera una representación tipo tabla (Figura 9) a partir del dataset iris con Python, donde se observan los atributos, tamaño y una de las primeras instancias usando dataframe pandas con el siguiente código:

Lectura del dataset iris en Python

```
df = pd.read_csv("https://archive.ics.uci.edu/ml/iris.data")

df.head()
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Figura 9. Codificación inicial del dataset iris. Fuente: elaborado a partir de InteractiveChaos, s.f.

Adicionalmente, a partir de las variables longitud del sépalo [cm] y ancho del sépalo [cm], se genera un gráfico de dispersión a través de la función `plt.scatter`. Se puede indicar el color según la clase a la que pertenece (Burrueco, s.f.):

Gráfico de dispersión dataset iris en Python

```
fig, ax = plt.subplots()

for species in set(data.species):
```

```
ax.scatter(df.sepal_length[df.species == species],

df.sepal_width[df.species == species], s = 30, colors[species],

label = species)

plt.legend()

plt.show()
```

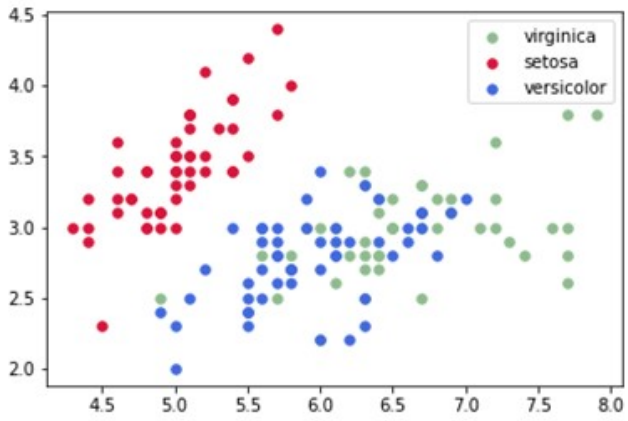


Figura 10. Gráfico de dispersión para variables longitud de pétalo vs ancho de pétalo de los tres tipos de especie de Iris. Fuente: elaborado a partir de InteractiveChaos, s.f.

En la Figura 10, se aprecia el conjunto de atributos seleccionados, segmentados por las clases: setosa (rojo), versicolor (azul) y virginica (verde).

Matriz de gráficos de dispersión

La matriz de gráficos de dispersión (SPLOM, por sus siglas en inglés — *scatterplot matrix*) presenta múltiples gráficos de dispersión adyacentes para todas las comparaciones de variables en una sola pantalla (Figura 11). En cierta medida, logra solventar uno de los problemas geométricos para conjuntos de datos que tienen más de tres atributos, en su lugar, se presentan dos o más parejas de variables en un mismo gráfico, lo cual resulta muy pertinente para identificar dependencias entre los datos (Wilkinson *et al.*, 2006).

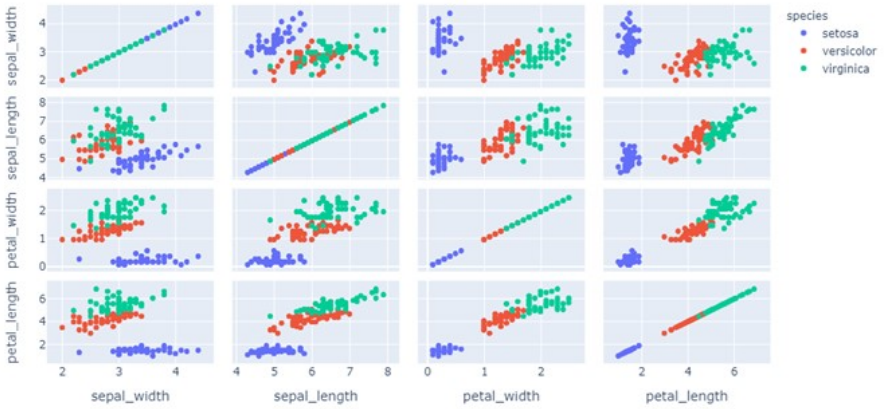


Figura 11. Diagramas de dispersión de las cuatro variables de la flor Iris con un color y carácter distinto para cada especie. Fuente: Bravo Márquez, 2013.

Los círculos azules representan a las variables de la especie Setosa; los círculos naranjas, a las variables de la especie Versicolor y los verdes a las variables de la especie Virginica.

Una de las desventajas es que SPLOM requiere una gran cantidad de espacio en la pantalla y la formación de asociaciones multivariante sigue siendo un desafío. Además, se requiere incorporar medidas estadísticas adicionales para organizar SPLOM y guiar al observador a través de un análisis exploratorio de los conjuntos de datos de alta dimensión (Wilkinson *et al.*, 2006).

2.7. Correlación entre variables

Correlación lineal

La correlación se conoce como una medida estadística que cuantifica el grado de variación conjunta que existe entre dos variables y, en específico, evalúa la tendencia creciente o decreciente de los datos. Los tipos de correlación se pueden observar en la Tabla 1.

Tipos de correlación	
Correlación positiva	Los valores de las variables aumentan juntos, dado que un aumento de "Y" (variable dependiente) depende de un aumento de "X" (variable independiente) (Figura 2).
Correlación negativa	Un valor disminuye a medida que el otro aumenta; por lo tanto, un aumento de los valores de la variable "X" causará una disminución de los valores de la variable "Y".
Nulo o sin correlación	No existe ninguna relación entre las variables, por lo tanto, las variables son independientes. La gráfica no sigue ninguna tendencia, ocasionando que los puntos se encuentren totalmente dispersos.
Lineal	Existe una relación entre las variables, la cual es lineal.
No lineal	Si bien existe una relación entre las variables, no es lineal. Puede ser exponencial, en forma de U, etc.

Tabla 1. Tipos de correlación en un diagrama de dispersión. Fuente: elaborado a partir de The Data Visualisation Catalogue y Aprendiendo Calidad y ADR (2017).

En la Figura 12 se muestran algunos diagramas de dispersión; en los gráficos superiores a), b) y c) se pueden observar las relaciones positivo, negativo y nulo, respectivamente; en los gráficos inferiores d), e) y f) se pueden observar las relaciones lineal, exponencial y tipo U, respectivamente.

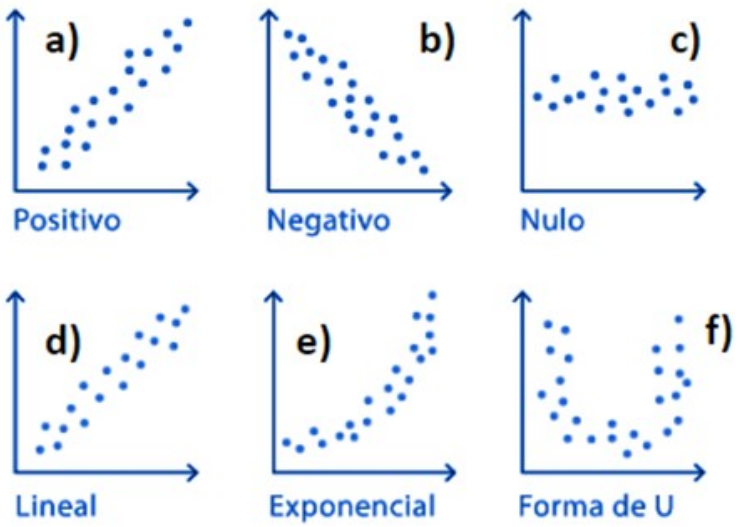


Figura 12. Tipos de correlación en un diagrama de dispersión. Fuente: Catálogo de Visualización de Datos, s.f.

Dada la subjetividad para la interpretación de estos diagramas de dispersión, se hace necesario el uso de un coeficiente que permita medir el grado de dependencia entre variables. El coeficiente de correlación lineal representa el comportamiento de una variable dependiente Y respecto a una variable independiente X (De Corso *et al*, 2017).

La fuerza de la correlación está dada por la proximidad de los puntos entre sí en el gráfico. Si los puntos se encuentran más juntos o cercanos unos de otros, la fuerza será mayor; si no se encuentran tan concentrados, la fuerza será débil, y si se observa un patrón de puntos muy dispersos entre sí, la fuerza de correlación será nula (Figura 13).

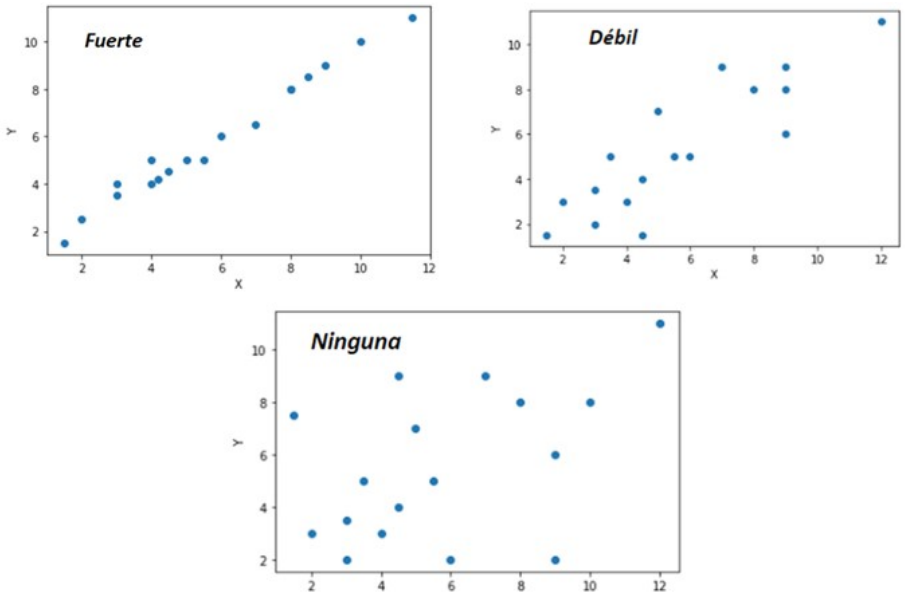


Figura 13. Fuerza de correlación en un diagrama de dispersión. Fuente: elaborado a partir de Catálogo de Visualización de Datos, s.f.

Los puntos que se encuentran muy lejanos del conjunto general se conocen como **valores atípicos,** debido a que son numéricamente distantes del resto de los datos. Las líneas o curvas se ajustan dentro del diagrama para ayudar en el análisis; se trazan tan cerca de todos los puntos como sea posible con el fin de visualizar cómo se concentran todos los puntos en una sola línea (Figura 13) (Ribecca, 2018).

Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es un estadístico cuyos valores varían entre -1 y 1 ($-1 \leq P \leq 1$); una correlación cercana a 1 indica una asociación positiva donde las dos variables crecen y disminuyen simultáneamente; una correlación próxima a -1 indica una asociación negativa debido a que si una variable crece, la otra decrece, y viceversa (Balzarini *et al.*, 2011); si el coeficiente es 0 (cero) o está muy cercano a este valor, se considera que no existe relación lineal, aunque puede existir otro tipo de relación no lineal. Este coeficiente solo puede utilizarse para comparar variables cuantitativas y continuas (Flores, 2020).

El coeficiente de correlación se calcula dividiendo la covarianza entre el producto obtenido de las desviaciones estándar tanto de la variable X como Y, así:

```
P=Cov(X,Y)/Sx*Sy
```

Donde P es el coeficiente de correlación lineal de Pearson, `Cov(X,Y)` corresponde a la covarianza entre `X` y `Y`, `Sx` es la desviación estándar de la variable `X`, y `Sy` es la desviación estándar de la variable Y. La importancia de este coeficiente radica en que le permite al investigador decidir acerca de la existencia o inexistencia de una asociación lineal entre las variables, así como de su intensidad (González Támara, 2018).

Con lo anterior y como lo veremos en el próximo apartado de esta unidad, el signo del coeficiente de correlación se interpreta igual que el de la covarianza; si se presenta una covarianza positiva, se entiende que la correlación es directa; si la covarianza resultante es negativa, la correlación será inversa, y si la covarianza es nula, no existe correlación como tal (Berrendero, s.f.). El coeficiente de correlación de Pearson, además de que permite identificar en qué cuadrantes están la mayoría de los puntos de un gráfico de dispersión, también mide qué tan próximos están los puntos de una línea recta (la cual refleja la tendencia de los datos).

Tomando como referencia nuestra base de datos Iris, a continuación, se muestra un ejemplo sobre coeficiente de correlación de Pearson en Python:

Cálculo de correlación entre dos variables dataset iris

```
import pandas as pd

iris=pd.read_csv("https://raw.githubusercontent.com/toneloy/data/master/iris.csv")

iris.head()
```

```
corr_mat.loc["petal_length", "sepal_width"]

-0.42844010433054003
```

Figura 14. Cálculo de la relación entre "petal_length" (largo del pétalo) y "sepal_width" (ancho del sépalo) de las especies de flor Iris). Fuente: elaborado a partir de Odio la Estadística, 2020.

En la Figura 14 se puede apreciar que la relación lineal entre las variables petal_length (largo del pétalo) y sepal_width (ancho del sépalo) es inversa porque la correlación es negativa (-0,428440).

En el lenguaje de programación de Python se emplea el comando .corr() para calcular la matriz de correlación al realizar el cruce de todas las variables de estudio (Figura 12):

Cálculo de correlación de todas las variables dataset iris en Python

```
corr_mat = iris.corr()

corr_mat
```

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	-0.117570	0.871754	0.817941
sepal_width	-0.117570	1.000000	-0.428440	-0.366126
petal_length	0.871754	-0.428440	1.000000	0.962865
petal_width	0.817941	-0.366126	0.962865	1.000000

Figura 15. Matriz de correlación considerando las cuatro variables (ancho y largo del pétalo y ancho y largo del sépalo de las especies de flor Iris). Fuente: elaborado a partir de Odio la Estadística, 2020.

2.8. Matriz de covarianza

A través de la estadística no solamente podemos entender cada variable de manera independiente, sino que también podemos entender las relaciones entre dos o más variables simultáneamente, a través de gráficos y estadísticos. Si la covarianza es positiva, la asociación lineal entre las variables también será positiva (Tabla 2). Por otra parte, si valores pequeños de X tienden a valores grandes de Y, los valores grandes de X tienden a valores pequeños de Y. Entonces, se obtienen productos mayoritariamente negativos, lo cual implica que, si la covarianza es negativa, la asociación lineal entre las variables también será negativa (González Támara, 2018).

Tipos de relaciones lineales	
Si Cov (X,Y)=0	La relación lineal entre X y Y es inexistente
Si Cov (X,Y)>0	Existe una relación lineal positiva (directa) entre "X" y "Y", por lo cual, a mayores valores de "X" , en promedio se tendrán mayores valores de "Y" y así al contrario.
Si Cov (X,Y)<0	Existe una relación lineal inversa o negativa entre "X" y "Y", por lo cual, a mayores valores de "X", en promedio tendremos menores valores de "Y" y viceversa.

Tabla 2. Covarianza y tipos de relaciones lineales. Fuente: elaborado a partir de Odio la Estadística, 2020.

En la literatura se define a la matriz de covarianza como una matriz de forma cuadrada cuyo número de filas es igual a su número de columnas y contiene tanto a las varianzas en la diagonal principal (observadas en la Figura 13 y resaltadas en un recuadro rojo) como a las covarianzas, estas últimas se ubican por fuera de la mencionada diagonal (Figura 13). La matriz de varianzas y covarianzas generalmente es simétrica, por lo tanto, la covarianza entre X e Y será igual a la covarianza entre Y y X; dado lo anterior, la covarianza para cada pareja de variables se muestra en la matriz dos veces (Minitab, 2019).

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	0.685694	-0.042434	1.274315	0.516271
sepal_width	-0.042434	0.189979	-0.329656	-0.121639
petal_length	1.274315	-0.329656	3.116278	1.295609
petal_width	0.516271	-0.121639	1.295609	0.581006

Figura 16. Matriz de covarianza de las cuatro variables (ancho y largo del pétalo y sépalo de las especies de flor Iris) generada en Python; se observa en los recuadros rojos a las varianzas. Fuente: Odio la Estadística, 2020.

Con Python podemos hacer uso del comando `cov` para realizar el cálculo de la matriz de varianzas y covarianzas con DataFrame, importando `pandas` :

Matriz de covarianza dataset iris en Python

```
import pandas as pd

iris=pd.read_csv("https://raw.githubusercontent.com/toneloy/data/master/iris.csv")

iris.head()

cov_mat = iris.cov()

cov_mat
```

En Python se puede emplear el comando `.loc` cuando se requiere conocer una covarianza entre dos variables específicas de la matriz, por ejemplo, “ petal_width ” y “ sepal_width ”:

```
cov_mat.loc["petal_width", "sepal_width"]
```

La matriz de covarianza estimada para los cuatro atributos de la base de datos Iris se encuentra definida por la Figura 17:

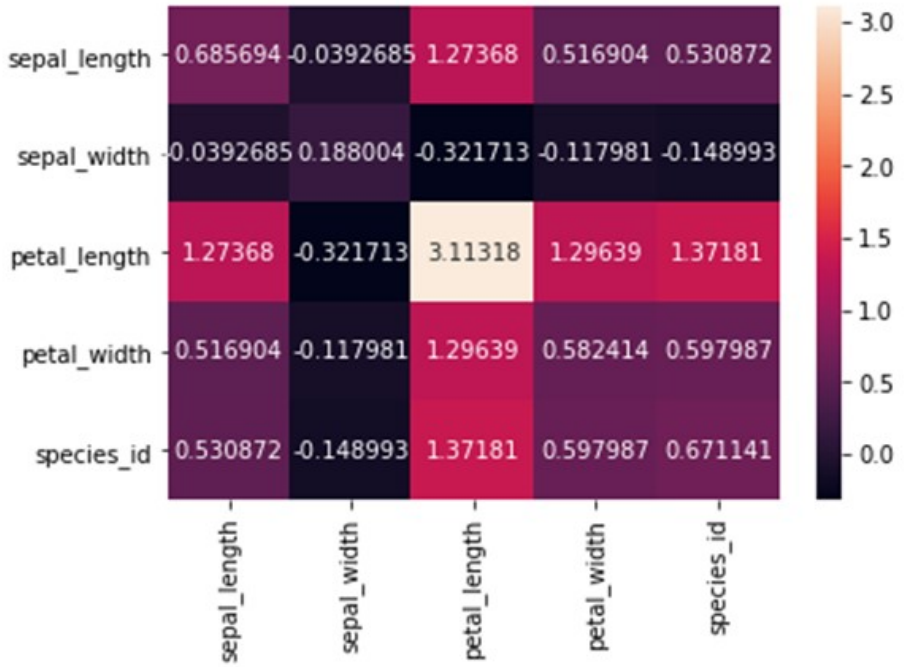


Figura 17. Representación gráfica de la matriz de covarianza de las cuatro variables (ancho y largo del pétalo y ancho y largo del sépalo de las especies de flor Iris). Fuente: elaborado a partir de Economipedia, 2020.

De la Figura 17 podemos inferir que la covarianza entre las variables largo del pétalo y ancho del sépalo es negativa, por lo cual la relación lineal entre ellas es inversa. Por su parte, la covarianza entre las variables ancho del pétalo y largo del sépalo es positiva, presentando una relación lineal directa entre ellas, como ya se había ilustrado anteriormente en la Tabla 3.

2.9. Cuaderno de ejercicios

- **Conjunto de datos generado con un 10% de outliers:** genera el conjunto de datos con el siguiente código:

```
# define the x co-ordinates

X = np.random.normal(mu, sigma, (395, 1))

# define the y co-ordinates

Y = np.random.normal(mu * 2, sigma * 3, (395, 1))
```

- Genera el diagrama de dispersión y comenta los resultados.

Inmuebles: con el siguiente conjunto de datos, https://raw.githubusercontent.com/jbrownlee/Datasets/master/housing.csv y su descripción, https://raw.githubusercontent.com/jbrownlee/Datasets/master/housing.names

- Crea la matriz de correlación aplicando Pearson y comenta los resultados.
- Con el conjunto de datos de los inmuebles genera la matriz de covarianza y compara los datos obtenidos con la matriz de correlaciones construida en el punto anterior.
- Utilizando la biblioteca de matplotlib para crear un diagrama de dispersión básico, genera datos aleatoriamente mediante la función: np.random.rand() .

SOLUCIÓN

```
import matplotlib.pyplot as plt

import numpy as np
```

```
# Generar datos aleatorios

np.random.seed(42)

x = np.random.rand(50)

y = np.random.rand(50)

# Crear el diagrama de dispersión

plt.scatter(x, y)

# Configurar el título y las etiquetas de los ejes

plt.title("Diagrama de dispersión")

plt.xlabel("Eje X")

plt.ylabel("Eje Y")

# Mostrar el diagrama de dispersión

plt.show()
```

- Modifica algunos valores de los datos del ejercicio anterior, cambia los colores o agrega leyendas según tus preferencias. Además, explica qué representa cada eje y cómo interpreta este diagrama.

2.10. Referencias bibliográficas

Balzarini, M., Di Rienzo, J., Tablada, M., González, L., Bruno, C., Córdoba, M., Robledo, W., & Casanoves, F. (2011). *Estadística y Biometría: Ilustraciones del Uso de InfoStat en Problemas de Agronomía* (1ª ed.). Editorial Brujas. https://www.researchgate.net/profile/Fernando-Casanoves/publication/283506487_Estadistica_y_biometria_Ilustraciones_del_uso_de_Infostat_en_Problemas_de_Agronomia/links/57f9195508ae280dd0bdc87d/Estadistica-y-biometria-Ilustraciones-del-uso-de-Infostat-en-Problemas-de-Agronomia.pdf

Bernal Villanueva, J. A. (2017). *Cómo hacer un histograma*. Ingenio Empresa. https://ingenioempresa.com/histograma/

Berrendero, J. R. (s.f.). *Tema 1: Descripción de datos*. Departamento de Matemáticas Universidad Autónoma de Madrid. https://verso.mat.uam.es/~joser.berrendero/cursos/est1/est1-descripcion-21.html#2

Burrueco, D. (s.f.). Gráficos de dispersión. Interactive Chaos. https://www.interactivechaos.com/es/manual/tutorial-de-matplotlib/graficos-de-dispersion

Bravo Márquez, F. J. (2013). *Análisis exploratorio de datos en R*. https://cupdf.com/document/analisis-de-datos-con-r.html

De Corso Sicilia, G. B., Pinilla Rivera, M. y Jaime, G. N. (2017). Métodos gráficos de análisis exploratorio de datos espaciales con variables espacialmente distribuidas. *Cuadernos Latinoamericanos de Administración, 25*, 92-104. Redalyc.org. https://www.redalyc.org/articulo.oa?id=409655122009

Ideas clave

Dicovskiy Riobóo, L. M. (2008). *Estadística Básica*. Universidad Nacional De Ingeniería . https://www.studocu.com/latam/document/universidad-nacional-de-ingenieria-nicaragua/estadistica-i/08-estadistica-basica-autor-luis-maria-dicovskiy-rioboo/26829838

González Támara, L. (2018). *Análisis Exploratorio de Datos*. Editorial Universidad Jorge Tadeo Lozano. https://www.researchgate.net/publication/340208750_Analisis_Exploratorio_de_Datos

Gorgas García, J., Cardiel López, N. y Zamorano Calvo, J. (2011). Estadística básica para estudiantes de Ciencias. Departamento de Astrofísica y Ciencias de la Atmósfera. Facultad de Ciencias Físicas. Universidad Complutense de Madrid. https://guaix.fis.ucm.es/~ncl/homepage/teaching/libro_GCZ2009.pdf

Hernández Martín, Z. (2012). *Métodos de análisis de datos: apuntes*. Universidad de la Rioja. Servicio de publicaciones, ed. Recuperado de: https://www.unirioja.es/cu/zehernan/docencia/MAD_710/Lib489791.pdf

Hernández Medrano, G. (2017). *Diagrama de dispersión*. Calidad y ADR. https://aprendiendocalidadyadr.com/diagrama-de-dispersion/

López Briega, R. E. (2016). *Análisis de datos categóricos con Python*. Matemáticas, análisis de datos y Python. https://relopezbriega.github.io/blog/2016/02/29/analisis-de-datos-categoricos-con-python/

Ribecca, S. (s.f.). *Diagrama de Dispersión*. Catálogo de Visualización de Datos. https://datavizcatalogue.com/ES/metodos/diagrama_de_dispersion.html#:%7E:text=Se%20pueden%20interpretar%20varios%20tipos,y%20en%20forma%20de%20U

Rodríguez Ojeda, L. (2007). Probabilidad y estadística básica para ingenieros. Instituto de Ciencias Matemáticas. Escuela Superior Técnica del Litoral, ESPOL. https://archuto.files.wordpress.com/2011/02/probabilidad_y_estadistica_basica.pdf

Rustom, J, A. (2012). *Estadística descriptiva, probabilidad e inferencia: Una visión conceptual y aplicada*. Departamento de Economía Agraria. Facultad de Ciencias Agronómicas. Universidad de Chile. http://repositorio.uchile.cl/bitstream/handle/2250/120284/Rustom_Antonio_Estadistica_descriptiva.pdf?sequence=1

Salazar, P., C. y Del Castillo G., S. (2018). *Fundamentos Básicos de Estadística* (1ª ed). http://www.dspace.uce.edu.ec/bitstream/25000/13720/3/Fundamentos%20B%C3%A1sicos%20de%20Estad%C3%ADstica-Libro.pdf

Spiegel, M. R. y Stephens, L. J. (2009). *Estadística: Serie Schaum* (4ª ed.). McGraw-Hill/Interamericana. https://www.cimat.mx/ciencia_para_jovenes/bachillerato/libros/[Spiegel]Probabilidad_y_Estadistica.pdf

Wilkinson, L., Anand, A. y Grossman, R. (2006). High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics, 12*, 1366-1372. https://ieeexplore.ieee.org/document/1703359

Witte, R. S. y Witte, J. S. (2017). *Statistics* (11ª ed.). Wiley Editorial. https://www.wiley.com/en-us/Statistics%2C+11th+Edition-p-9781119254515

Estadística Descriptiva

Rendón-Macías, Mario Enrique y Villasís-Keeve, Miguel Ángel y Miranda-Novales, María Guadalupe (2016). Estadística descriptiva. *Revista Alergia México, 63*(4), 397-407. https://www.redalyc.org/pdf/4867/486755026009.pdf

El anterior recurso contiene información con la que, a través de ejemplos y conceptos, el estudiante puede profundizar más sobre lo relacionado con la estadística descriptiva, su importancia y aplicación, así como sobre la caracterización de las variables, sobre medidas de dispersión y detección de valores atípicos a través de los diagramas de cajas.

Aplicación de la estadística descriptiva en la ingeniería

Ochoa, J. (2020). *Aplicación de la estadística descriptiva en la ingeniería de la confiabilidad, parte I*. Medium. Disponible en: https://medium.com/@javier8amoreno/aplicaci%C3%B3n-de-la-estad%C3%ADstica-descriptiva-en-la-ingenier%C3%ADa-de-la-confiabilidad-parte-i-bdd3feb8b593

A través de este recurso, el estudiante puede conocer y realizar ejercicios prácticos a partir de la estadística descriptiva apoyado en el uso de Python para su ejecución, como complemento de lo ya tratado durante la segunda unidad.

Adivina la correlación

Wagih, O. (s. f.). *Guess the Correlation*. http://guessthecorrelation.com/

Se trata de un juego online gratuito cuyo objetivo se enfoca en que el participante adivine qué tan correlacionadas se encuentran dos variables en un diagrama de dispersión; la respuesta del participante debe estar entre cero (0) y uno (1), donde cero significa no correlación y uno es correlación perfecta. Para este juego no se emplean correlaciones negativas. Este recurso, al ser didáctico y educativo, permite al estudiante aprender, reforzar y repasar conceptos a través del juego.

Galería gráfica de Python

Página de The Python Graph Gallery (https://python-graph-gallery.com/).

A través de este útil recurso interactivo, el estudiante podrá conocer múltiples gráficos, como los tratados en la unidad de estadística exploratoria basados en datos manejados en Python. En esta página podrás acceder a los ejemplos, así como a la codificación en Python para la realización de los gráficos que necesites.

Test

1. Es una característica de los diagramas de dispersión.

   A. Muestran una relación existente entre variables numéricas.

   B. Permiten descubrir relaciones.

   C. Permiten identificar relación de incremento o disminución de valores entre variables.

   D. Todas son características de los diagramas de dispersión.

2. La matriz de gráficos de dispersión facilita:

   A. Graficar tres atributos o variables en un mismo gráfico.

   B. Emparejar datos numéricos y saber si una variable afecta a la otra.

   C. La comparación de las variables en una sola pantalla e identificar dependencias.

   D. En un espacio pequeño asociar múltiples variables.

3. El coeficiente de correlación de Pearson permite comparar:

   A. Variables cuantitativas y continuas.

   B. Solamente variables cuantitativas.

   C. Solamente variables continuas.

   D. Variables cualitativas.

4. Es una buena interpretación de coeficiente de correlación:

   A. Si el resultado es -1, la correlación es positiva perfecta.

   B. Si el resultado es -1, la correlación es negativa perfecta.

   C. Si el resultado es 1, la correlación es positiva moderada.

   D. Si el resultado es 1, la correlación es fuerte y positiva.

5. Es una afirmación falsa sobre el coeficiente de correlación de Pearson:

   A. El coeficiente se interpreta exactamente igual que el de la covarianza.

   B. Si se presenta covarianza positiva, la correlación es directa.

   C. Si se presenta covarianza negativa, la correlación es inversa.

   D. Todas las anteriores son verdaderas.

6. Las medidas de dispersión tienen como objetivo:

   A. Definir la frecuencia relativa y absoluta.

   B. Describir la concentración de los datos alrededor de cierto número.

   C. Construir diagramas para contextualizar los datos.

   D. Ninguna de las anteriores.

7. Son tipos de variables:

   A. Nominales.

   B. Ordinales.

   C. Discretas.

   D. Todas las anteriores.

8. No es una característica del diagrama de cajas:

   A. El Q2 representa la mediana.

   B. Los bigotes que representan el rango intercuartil.

   C. Los bigotes que se extienden desde el valor mínimo al valor máximo de los datos.

   D. Los cuartiles Q1 y Q3.

9. Es un ejemplo de un valor muy atípico:

A. Temperatura por debajo de 0 ºC en invierno.

B. Temperatura corporal de 38 ºC.

C. Temperatura por debajo de 0 ºC en verano.

D. Temperatura corporal de 36 ºC.

10. No es una afirmación válida para la varianza:

A. La varianza puede indicar la homogeneidad o heterogeneidad de dos distribuciones.

B. Es una medida de dispersión.

C. Es el promedio de los cuadrados de las desviaciones de cada dato con respecto a la mediana.

D. El promedio de las desviaciones no es una medida útil, ya que la sumatoria es cero.