«Talento Tech»

Data Analytics

con Python

Clase 11

Buenos Aires Ciudad

Clase N° 11 | Análisis de Correlación

Temario:

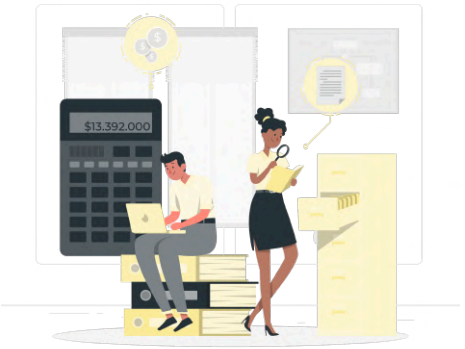
- Introducción a la correlación y su importancia en el análisis de datos.
- Calcular y visualizar la correlación usando Pandas y Seaborn.

Objetivos de la Clase:

- **Comprender la Correlación**: Definir la correlación y su importancia en el análisis de datos.
- **Calcular y Medir la Correlación**: Aprender a utilizar las medidas de correlación, y aplicar el método DataFrame.corr() de Pandas para calcular la matriz de correlación.
- **Manipular y Visualizar Datos**: crear visualizaciones de correlaciones utilizando Seaborn, como mapas de calor y diagramas de dispersión.

Análisis de Correlación

Introducción a la Correlación



La **correlación** se refiere a una medida estadística que describe la relación entre dos o más variables. Cuando decimos que dos variables están correlacionadas, esto significa que **un cambio en una variable tiende a estar acompañado por un cambio en la otra variable**. Por ejemplo, el aumento en la cantidad de horas estudiadas está generalmente relacionado con un aumento en el rendimiento académico.

Analizar la correlación entre variables nos permite identificar patrones y tendencias en nuestros datos.

Repaso de Conceptos Básicos de Estadística

Antes de adentrarnos en el análisis de correlación, es fundamental repasar algunos conceptos estadísticos básicos:

- **Media, Mediana y Moda**: La media es el promedio aritmético de un conjunto de valores; la mediana es el valor central cuando los números están ordenados, y la moda es el valor que más se repite en un conjunto.
- **Varianza y Desviación Estándar**: La varianza mide la dispersión de un conjunto de datos con respecto a su media, mientras que la desviación estándar es la raíz cuadrada de la varianza, proporcionando una medida de dispersión en las mismas unidades que los datos.

Tipos de Correlación

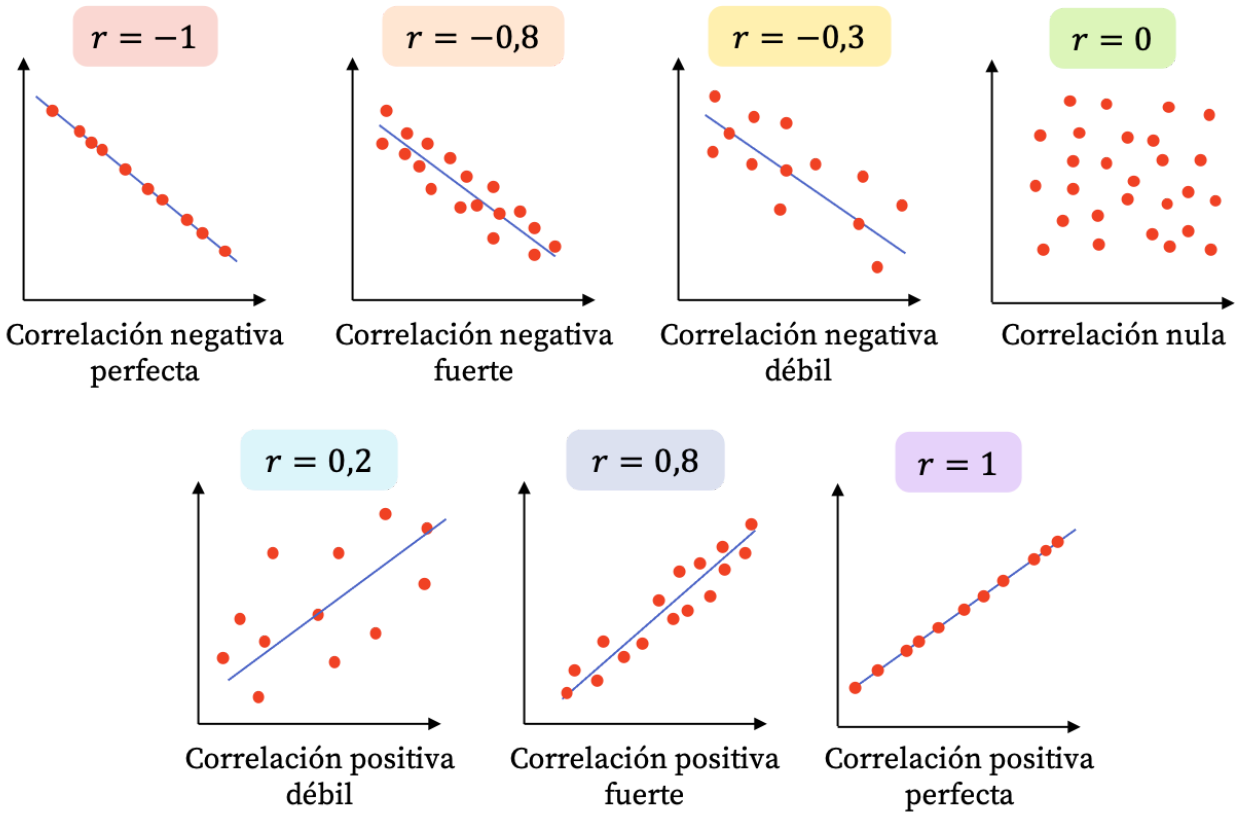
Existen varios tipos de correlación que debemos tener en cuenta:

1. **Correlación Positiva**: Indica que, a medida que una variable aumenta, la otra también lo hace.
2. **Correlación Negativa**: Significa que cuando una variable aumenta, la otra disminuye.
3. **Correlación Nula**: No hay una relación discernible entre las variables.

Además, podemos distinguir entre correlaciones **lineales** y **no lineales**. La correlación lineal implica que los datos son alineados en una línea recta al graficarlos, mientras que la correlación no lineal indica que puede haber una relación más compleja entre las variables. En este curso trabajaremos con correlaciones lineales.

Medidas de Correlación

Para cuantificar la correlación, utilizamos el **coeficiente de correlación de Pearson,** que mide la fuerza y la dirección de la relación lineal entre dos variables continuas.



Debemos resaltar el hecho de que dos variables tengan una alta correlación, no implica que el cambio del valor de una de las variables, sea la causa del cambio en la otra. Sólo indica que ambas variables cambian en la misma proporción.

¡Correlación no implica causalidad!

Cálculo de la Correlación con Pandas

Para calcular la correlación entre nuestras variables, utilizamos el método `DataFrame.corr()` en Pandas. Este método devuelve una matriz de correlación.

```
correlation_matrix = mi_dataframe.corr()
print(correlation_matrix)
```

Es importante interpretar correctamente esta matriz: valores cercanos a +1 indican una fuerte correlación positiva, mientras que valores cerca de -1 indican una fuerte correlación negativa. Los valores cercanos a 0 indican que no existe correlación entre las variables.

Visualización de Correlaciones con Seaborn

La visualización es de gran ayuda para entender las correlaciones. **Seaborn** ofrece herramientas prácticas para crear gráficos como el mapa de calor, ideal para representar la matriz de correlación:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Crear un mapa de calor
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.show()
```

Además, podemos utilizar diagramas de dispersión para observar la relación entre dos variables específicas:

```
sns.scatterplot(x='Variable1', y='Variable2', data=data)
plt.show()
```

Seaborn facilita la personalización de las gráficas, como modificar títulos, etiquetas e incluso colores para mejorar la presentación.

Ejemplo Práctico

Ahora, apliquemos todo lo aprendido. Utilizaremos un conjunto de datos real. Por ejemplo, un archivo que contenga información sobre las dimensiones de pétalos y sépalos de algunas especies de flores.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# Cargar el set de datos
df = sns.load_dataset("iris")
```

El dataframe obtenido tiene esta forma:

	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
..	...	...	...	...	...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

[150 rows x 5 columns]

Observamos que hay más de una especie en el dataframe, por lo que vamos a obtener los valores numéricos de sólo una de las especies de flores.

```
# Exploramos la columna de especies
df['species'].unique()

# seleccionamos sólo los datos de una especie
setosa = df[df['species'] == 'setosa']

#obtenemos las variables numéricas
setosa.drop(['species'], axis='columns', inplace=True)
```

El nuevo dataframe tiene el siguiente aspecto:

	sepal_length	sepal_width	petal_length	petal_width
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
5	5.4	3.9	1.7	0.4
6	4.6	3.4	1.4	0.3
7	5.0	3.4	1.5	0.2
8	4.4	2.9	1.4	0.2
9	4.9	3.1	1.5	0.1
10	5.4	3.7	1.5	0.2
11	4.8	3.4	1.6	0.2

Calcularemos la matriz de correlación y visualizaremos los resultados.

```
# Crear la matriz de correlación
correlation_matrix = setosa.corr()
print(correlation_matrix)
```

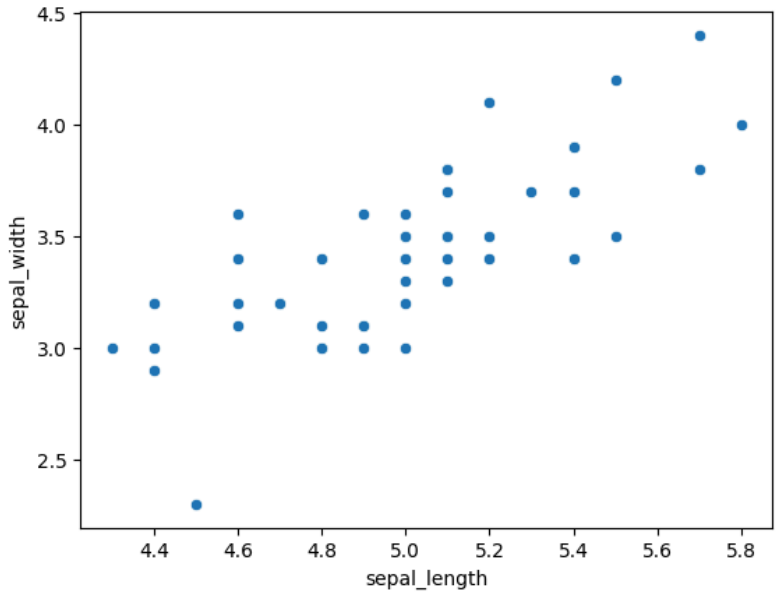
	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	0.742547	0.267176	0.278098
sepal_width	0.742547	1.000000	0.177700	0.232752
petal_length	0.267176	0.177700	1.000000	0.331630
petal_width	0.278098	0.232752	0.331630	1.000000

Podemos explorar el valor de correlación entre dos variables fácilmente:

```
# Investigar la correlación entre el largo y el ancho de
sépalo
correlation_matrix['sepal_length']['sepal_width']
# 0.7425466856651594
```

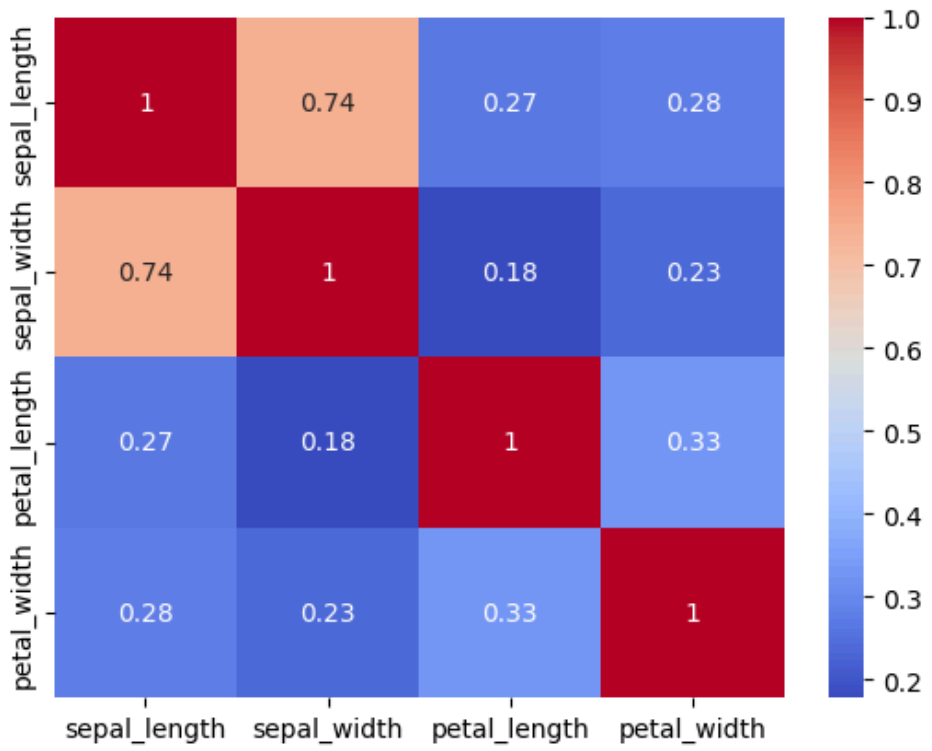
Visualmente, la mejor opción es el gráfico de dispersión:

```
sns.scatterplot(x='sepal_length', y='sepal_width', data=setosa)
plt.show()
```

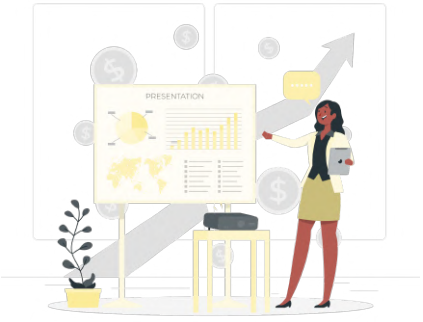


Mejor opción para explorar visualmente el mapa de correlaciones es el mapa de calor:

```
# Crear el mapa de calor
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.show()
```



Reflexión final



El análisis de correlación es una herramienta poderosa que nos permite identificar relaciones entre variables, pero es vital recordar que **correlación no implica causalidad**.

Utilizando librerías como Pandas y Seaborn, podemos calcular y visualizar estas correlaciones de manera intuitiva, mejorando nuestra comprensión de los datos.

Materiales y recursos adicionales

- Documentación Seaborn

Próximos Pasos

- Técnicas para consolidar datos y preparar para el análisis final.
- Consolidar, transformar y preparar un conjunto de datos para análisis posterior.

Ejercicios Prácticos

Actividad 1: Análisis de Correlación

Contexto

Durante esta semana de tu pasantía en SynthData, vas a trabajar con un conjunto de datos que contiene información sobre las características de diferentes diamantes. Silvia, tu mentora y Project Manager, te ha asignado esta tarea para que puedas aplicar conceptos de correlación y análisis de datos.

Objetivos

El objetivo de esta actividad es que aprendas a identificar relaciones significativas entre las características de los diamantes, como el precio y quilates. Esto te permitirá desarrollar una comprensión más profunda de cómo los diferentes factores influyen en el valor de los diamantes en el mercado.

Ejercicio práctico

- Calcular la correlación entre precio y quilates de los diamantes color “F” y corte Premium.
- Visualizar los resultados mediante gráfico de dispersión.

Dataset

```
#Set de datos seaborn
df = sns.load_dataset("iris")
```

¿Por qué importa esto en SynthData?

Entender las correlaciones entre dos variables ayudará a las decisiones comerciales y a la estrategia de precios. Esta habilidad te permitirá utilizarla en otros contextos para, por ejemplo, eliminar datos que puedan resultar redundantes.

Actividad 2: Análisis de Estadístico

Contexto

Matías, el Data Analyst de SynthData, te invita a analizar un conjunto de datos sobre pingüinos. Este conjunto contiene características físicas de diferentes especies. Se busca descubrir posibles patrones que puedan influir en su conservación. Este ejercicio te permitirá explorar diversas variables numéricas, su relación entre sí, y su distribución estadística.

Objetivos

El objetivo de esta actividad es que realices un análisis de distribución y correlaciones para identificar patrones que describan a las especies de pingüinos. Los insights obtenidos podrán ayudar en la elaboración de estrategias de conservación.

Ejercicio práctico

A continuación, deberás calcular la matriz de correlación, visualizar sus resultados y realizar gráficos que muestren la distribución de las principales variables.

1. **Observar cómo se distribuyen los pesos de los pingüinos según su especie.**
2. **Visualizar la distribución de características clave:**
   a) **Longitud del pìco**
   b) **Longitud de la aleta**
3. **Desglosar por ubicación geográfica: M**ostrar para cada ubicación, analítica y gráficamente la distribución de:
   a) Ejemplares por sexo
   b) Ejemplares por especie

    c) Desafio: por especie y sexo
4. **Calcular la matriz de correlación y visualizarla**
5. **Explorar, interpretar y mostrar gráficamente las variables que tienen**
    a) **Mayor correlación.**
    b) **Menor correlación.**

Dataset

```
#Set de datos seaborn
penguins = sns.load_dataset("penguins")
```

¿Por qué importa esto en SynthData?

Este análisis proporcionará información valiosa sobre cómo las diferentes características físicas y de comportamiento de los pingüinos están interrelacionadas y distribuidas. Estos insights son la base para definir estrategias que estén alineadas con la conservación de las diferentes especies. Al entender la variabilidad y relaciones entre variables, podrás contribuir a la sostenibilidad y el bienestar de estas especies.

⛔ **Estos ejercicios son una simulación de cómo se podría resolver el problema en este contexto específico. Las soluciones encontradas no aplican de ninguna manera a todos los casos.**
Recuerda que las soluciones dependen de los sets de datos, el contexto y los requerimientos específicos de los stakeholders y las organizaciones.

Buenos Aires aprende

Agencia de Habilidades para el Futuro

Buenos Aires Ciudad