«Talento Tech»

Data Analytics

con Python

Clase 07

Buenos Aires Ciudad

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Clase N° 07 | Funciones de Agregación y Agrupamiento de Datos

Temario:

- Agregación, agrupamiento y tablas dinámicas.
- Paralelismo entre las funciones de agregación y agrupamiento de SQL y Pandas

Objetivos de la clase:

- Repasar los conceptos y la importancia de la agregación y agrupamiento de datos.
- Familiarizarse con las funciones de Pandas.
- Relacionar y comparar con las funciones de SQL.

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Funciones de Agregación y Agrupamiento de Datos

En el ámbito del data analytics, manejar grandes volúmenes de datos se ha vuelto indispensable. Existen técnicas clave que nos permiten sintetizar y analizar información de manera efectiva, entre ellas la **agregación**, el **agrupamiento de datos** y las **tablas dinámicas**. A lo largo de esta clase, se explorarán estos conceptos aplicados en Python a través de la biblioteca Pandas, y nos apoyaremos en sus similitudes con las funciones de SQL.

Agregación de Datos

La agregación implica **resumir varios registros en un único valor**. Este procedimiento permite obtener métricas básicas que ayudan a comprender patrones dentro de un conjunto de datos. En SQL, este concepto también se aplica mediante las funciones de agregación como `SUM()`, `AVG()`, o `COUNT()`, donde se pueden obtener totales o promedios por registros.
Por ejemplo, en un conjunto de datos con ventas, podríamos querer saber el total de ingresos generados. En Pandas, esto se logra fácilmente usando la función `sum()`.

Ejemplo de Agregación en Pandas:

```
# Creando un DataFrame de ejemplo
datos = {
    'Producto': ['A', 'B', 'A', 'B', 'C'],
    'Ventas': [200, 150, 300, 200, 400]
}
df = pd.DataFrame(datos)

# Agregando las ventas totales
ventas_totales = df['Ventas'].sum()
print(f'Ventas totales: {ventas_totales}')
```

El resultado nos mostrará el total de ventas, similar a lo que obtendríamos en SQL al ejecutar `SELECT SUM(Ventas) FROM tabla;`

Agrupamiento de Datos

El agrupamiento de datos permite **segmentar la información en grupos basados en características específicas**, lo que facilita el análisis comparativo. Esto se puede lograr en Pandas con la función `groupby()`, que agrupa el DataFrame por una o más columnas y permite aplicar funciones de agregación en estos grupos.
SQL tiene un uso similar, donde podemos emplear `GROUP BY` para segmentar resultados.

Ejemplo de Agrupamiento en Pandas:

```
# Agrupando las ventas por producto
grupo_ventas = df.groupby('Producto')['Ventas'].sum()
print(grupo_ventas)
```

Este código genera un resumen de las ventas totales por producto, semejante a la instrucción SQL `SELECT Producto, SUM(Ventas) FROM tabla GROUP BY Producto;`.

Tablas Dinámicas

Las tablas dinámicas permiten realizar **análisis interactivos en datasets**. En Pandas, utilizamos `pivot_table()` para crear estas tablas, las cuales permiten resumir datos basados en categorías y ajustar la visualización según sea necesario.
En SQL, podríamos lograr una funcionalidad similar utilizando instrucciones que combinan agrupación y agregación, pero sin la interactividad que ofrecen las tablas dinámicas.

Ejemplo de Tabla Dinámica en Pandas:

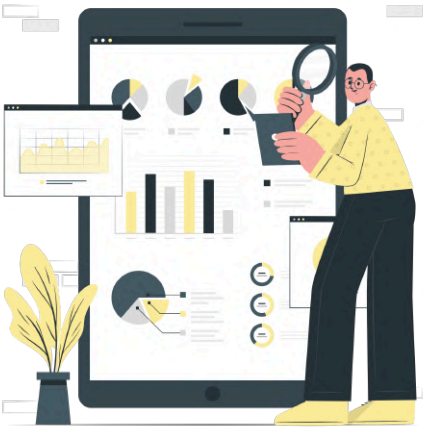
```
# Creando una tabla dinámica
tabla_dinamica = df.pivot_table(index='Producto',
values='Ventas', aggfunc='sum')
print(tabla_dinamica)
```

Este resultado nos proporciona un resumen claro y organizado de las ventas por producto.

Reflexión final

A lo largo de esta clase, hemos revisado los conceptos de **agregación**, **agrupamiento de datos** y **tablas dinámicas**, mostrándolos en un contexto práctico, utilizando **Python** y la biblioteca **Pandas**. Además, hemos comparado estas funcionalidades con sus equivalentes en SQL, resaltando la flexibilidad y eficiencia que ofrecen ambas herramientas en el análisis de datos.

Dominar estos conceptos es fundamental para un analista de datos y se traduce en una mayor capacidad para realizar análisis profundos.



Buenos Aires aprende
Agencia de Habilidades para el Futuro

Materiales y recursos adicionales

Documentación oficial de Pandas:

- Algunas funciones de agregación
- Agrupamiento
- Tablas dinámicas

Próximos pasos

- Combinación de diferentes fuentes de datos
- `merge()`
- `join()`

Ejercicios prácticos:

SYNTHDATA
SINTETIZA CONOCIMIENTO

Actividad 1: Agrupando Datos para Análisis de Ventas

Contexto

Hoy te enfrentás a la tarea de analizar un conjunto de datos sobre las ventas de productos en la última campaña. Silvia, tu mentora y Project Manager, te solicitó que agrupes los datos para obtener un informe claro sobre el rendimiento de cada producto. Tu trabajo es esencial, ya que estos informes serán el fundamento para decidir el enfoque de ventas en el próximo trimestre.

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Objetivos

- Aplicar la función `groupby()` de Pandas para segmentar los datos de ventas.
- Calcular el total de ventas por producto.
- Presentar los resultados de manera clara y concisa.

Ejercicio práctico

1. Cargar el conjunto de datos llamado `ventas.csv`, que contiene información sobre las ventas de diferentes productos en distintos meses.
2. Agrupar las ventas por producto y sumar las ventas totales para cada uno.

3. Generar un nuevo DataFrame que contenga el total de ventas por producto y mostrarlo en pantalla.

Set de datos

- ventas.csv

¿Por qué importa esto en SynthData?

Agrupar datos es una tarea crítica en el análisis de datos, ya que permite a los analistas identificar qué productos están liderando las ventas y cuáles necesitan estrategias de mejora. La capacidad de segmentar datos te proporcionará una abanico de insights, fundamentando las decisiones estratégicas de la empresa. Sintetizar la información destacará no sólo tu habilidad técnica, sino también tu capacidad para influir en el rendimiento del negocio.

Actividad 2: Creación de Tablas Dinámicas para Resumir Datos

Contexto

Después de finalizar el análisis de ventas, ahora es el turno de Matías, tu mentor Data Analyst, de evaluar el rendimiento de las campañas de marketing. Matías te pide que utilices tablas dinámicas para resumir información sobre las ventas y el rendimiento de las campañas.

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Tu contribución será importante a la hora de entender qué estrategias funcionan mejor y dónde se pueden hacer mejoras.

Objetivos

- Aplicar la función `pivot_table()` de Pandas para resumir información clave de los datos de ventas.
- Crear un resumen que permita analizar las ventas mensuales por producto.
- Interpretar los resultados y prepararlos para una presentación.

Ejercicio práctico

1. Cargar el conjunto de datos `ventas.csv` de la actividad anterior.
2. Utilizar la función `pivot_table()` para crear una tabla dinámica que muestre las ventas mensuales por producto.
3. Asegurarse de que la tabla muestre los productos en las filas y los meses en las columnas, con las ventas totales como valores dentro de la tabla.
4. Mostrar la tabla dinámica en pantalla.

¿Por qué importa esto en SynthData?

El uso de tablas dinámicas es una habilidad poderosa que permite resumir grandes volúmenes de datos de manera eficiente. En SynthData, ser capaz de visualizar estos resúmenes facilitará la comprensión de patrones complejos y responderá preguntas críticas sobre el rendimiento de las campañas. Este análisis es fundamental para que el equipo de marketing ajuste su estrategia y maximice el retorno sobre la inversión.

⛔ **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