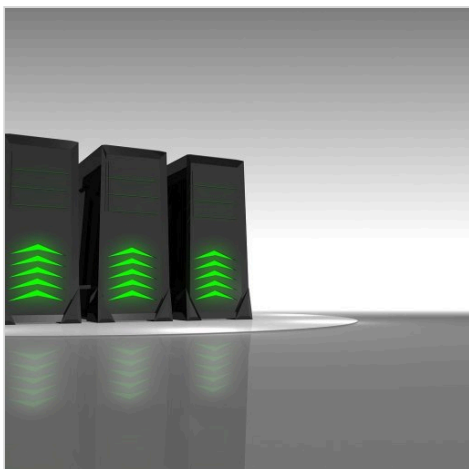


L U M B R E



Filtrado, groupby y agregación

Grado: 2

Materia: Programación con Python para Datos · Programación

11/10/2026

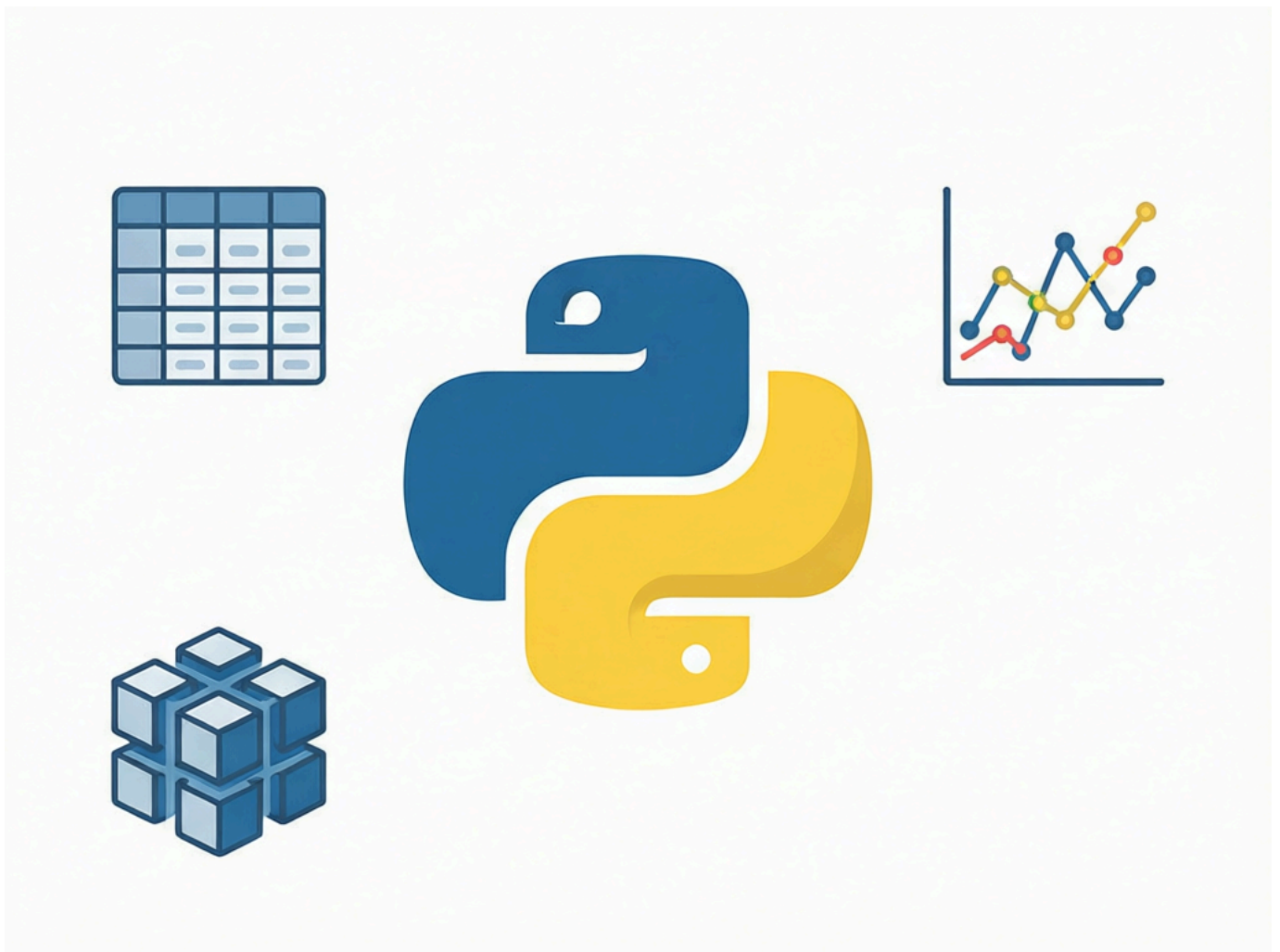
Antes de empezar

El corazón del análisis es preguntar a los datos: «¿cuánto vendió cada región?». En Pandas eso es agrupar por una columna y aplicar una función de resumen a las demás: groupby.

QUÉ APRENDERÁS

- Segmentar datos con filtros, agrupar por categorías y calcular estadísticos agregados con el patrón split-apply-combine.

Filtrado, groupby y agregación



Python para ciencia de datos: NumPy, pandas y manipulación de datos.

Mapa conceptual

- **Python para Datos**
 - Entorno
 - pip

- venv
- Reproducibilidad
- Funcional
 - Comprensiones
 - Lambda
 - map filter
- Calidad
 - try except
 - pytest
 - Asserts
- NumPy
 - Arrays
 - Broadcasting
 - Indexacion
- Pandas
 - DataFrame
 - Groupby
 - Merge Join

El mapa mental interactivo es interactivo en la versión online.



Dividir, aplicar, combinar

Groupby es la generalización de la tabla dinámica de una hoja de cálculo: divide los datos en grupos, aplica un cálculo a cada grupo y combina los resultados. Es la operación más rentable de Pandas para responder preguntas de negocio.

Filtrar filas

Filtrar es quedarse con las filas que cumplen una condición, usando máscaras booleanas como en NumPy: `df[df.ventas > 100]`. Se combinan condiciones con `and` y `or` y paréntesis. También existe `df.query(ventas > 100)`, más legible para expresiones largas.

Seleccionar columnas

Antes o después de filtrar, eliges columnas: `df[[region, ventas]]` te trae solo lo necesario. Filtrar filas y seleccionar columnas suele ser el primer paso de cualquier análisis puntual.

La idea de groupby

Imagina agrupar las ventas por región: todas las filas de una misma región forman un grupo. `df.groupby(region)` crea el objeto de grupo; sobre él llamas a una función de resumen y Pandas la aplica a cada grupo por separado, devolviendo una fila de resultados por grupo.

Split-apply-combine

- Split: parte el DataFrame en grupos según una o varias columnas clave.
- Apply: ejecuta una función de resumen sobre cada grupo de forma independiente.
- Combine: vuelve a juntar los resultados en un nuevo DataFrame indexado por grupo.

Funciones de agregación

Las más comunes son `sum`, `mean`, `count`, `min`, `max` y `std`. `df.groupby(region)[ventas].sum()` responde «cuánto vendió cada región» en una línea. Se pueden agregar varias columnas a la vez pidiendo un subconjunto antes de `agg`.

agg: varias funciones

Con `.agg([sum, mean, count])` obtienes varios estadísticos por grupo en una sola pasada, y con un diccionario puedes pedir funciones distintas por columna: `ventas: sum`, `cliente: nunique`. Es la forma expresiva de construir tablas resumen.

El patrón groupby-agg cubre la mayoría de resúmenes tabulares.

Pregunta	Expresión (idea)	Resultado
Total por región	<code>df.groupby(region)[ventas].sum()</code>	una fila por región
Media por dos claves	<code>df.groupby([region, año])...</code>	índice multinivel
Recuento distinto	<code>.agg(nunique)</code>	valor único por grupo
Varias métricas	<code>.agg([sum, mean])</code>	una columna por métrica

1. ¿Qué hace exactamente `df.groupby(region)[ventas].mean()`?

- ☐ Ordena las ventas por región
- ☐ Calcula la media de ventas dentro de cada región, una fila por región
- ☐ Filtra solo la región mayor
- ☐ Suma todas las ventas del DataFrame

Filtrar tras agrupar

A veces quieres quedarte con grupos que cumplen algo: `.filter(lambda g: len(g) > 5)` conserva solo los grupos con suficientes filas. Y con `.transform` aplicas un cálculo de grupo devolviendo el tamaño original, útil para restar la media de cada grupo fila a fila.

Índices multinivel

Al agrupar por varias columnas, el resultado tiene un índice multinivel (varios niveles). `.reset_index()` lo convierte de nuevo en columnas normales, facilitando exportar o seguir operando.

value_counts

Para contar categorías rápidas sin un `groupby` completo, `value_counts()` da la frecuencia de cada valor ordenada de mayor a menor. Es el resumen de una columna categórica en un gesto.

Ordenar resultados

Un resumen casi siempre se ordena para interpretarse: `.sort_values(ventas, ascending=False)` muestra primero lo importante. Agrupar, agregar y ordenar es la trinidad de las preguntas de negocio.

Errores frecuentes

- Aplicar la agregación a todas las columnas por descuido y obtener resultados sin sentido en las textuales.
- Olvidar `reset_index` y pelearse con un índice multinivel al exportar.
- Confundir `count` (no nulos) con `nunique` (distintos) al contar clientes.

2. Tras agrupar por varias columnas, `reset_index()` convierte los niveles del índice de nuevo en columnas.

- ☐ Verdadero
- ☐ Falso

De la pregunta al código

El valor de groupby es traduce preguntas de lenguaje natural a una línea: «media de notas por materia», «total gastado por cliente». Practicar esa traducción es volverte rápido interrogando datos.

3. Ordena las etapas del patrón split-apply-combine:

Elementos:

- groupby separa en grupos
- una función resume cada grupo
- Pandas junta los resultados
- sort_values ordena para leer

Dividir	. . .
Aplicar	. . .
Combinar	. . .
Presentar	. . .

Pandas Docs https://pandas.pydata.org/docs/user_guide/groupby.html Guía oficial de groupby y agregación.

VIDEO**Video complementario**

Recurso audiovisual para reforzar los conceptos.



Escanea para verlo en Lumbre

<https://www.youtube.com/watch?v=rfscVS0vtbw>

4. Con un DataFrame de transacciones (fecha, cliente, producto, importe), escribe las expresiones para: importe total por cliente, y media de importe por producto y mes. ¿Qué estadístico usarías para saber qué cliente compra artículos más distintos?

Documentación pandas <https://pandas.pydata.org/docs/Guia de referencia para manipulacion de datos en Python>.

Referencia curricular

Grado:	2
Materia:	Programación con Python para Datos · Programación

Versión interactiva completa (con actividades, videos y autoevaluación):



<https://univ.lumbre.lat/object/cdia-pyth-008-filtrado-groupby-y-agregacion>

Material educativo generado desde Lumbre · lumbre.lat