Módulo 3

Sesión N° 2

ACTIVIDAD:

Selección, Filtrado, Sumarización y Análisis con Pandas

Objetivo: Manipular y analizar datos tabulares en Pandas aplicando selección y filtrado avanzado, agregaciones, agrupaciones, creación de variables categóricas, manejo de datos faltantes y visualización exploratoria para interpretar resultados y tomar decisiones.

Contexto

Como analista de datos de una empresa ficticia, recibirás un conjunto de datos de empleados. Deberás aplicar técnicas de manipulación, exploración, filtrado, creación de variables, análisis y visualización básica con Pandas para responder preguntas de negocio y preparar la información para informes ejecutivos.

En este contexto, la gerencia quiere comprender mejor la distribución de edad, ciudad e ingresos de los empleados para planificar una estrategia de compensaciones y detectar posibles desigualdades. Además, el equipo de datos desea practicar la preparación de información antes de usar técnicas de machine learning.

Datos iniciales
(nombre, edad, ciudad, ingresos, área, años_experiencia)

- nombre: ['Ana', 'Luis', 'Sofía', 'Pedro', 'Marta', 'Juan', 'Lucía', 'Miguel']
- edad: [23, 35, 28, 40, 29, 35, 31, 27]
- ciudad: ['Madrid', 'Barcelona', 'Valencia', 'Bilbao', 'Madrid', 'Barcelona', 'Valencia', 'Bilbao']
- ingresos: [2500, 3200, 2900, 3600, 2700, 3100, 3300, 2750]
- área: ['IT', 'Ventas', 'IT', 'Dirección', 'Recursos Humanos', 'Ventas', 'IT', 'Recursos Humanos']
- años_experiencia: [2, 8, 5, 15, 3, 9, 6, 4]

Agrega al menos un valor nulo (None) o atípico a propósito en “ingresos” y “edad” para practicar limpieza de datos.

Requerimientos técnicos
- Python 3.x
- Pandas
- Matplotlib o Seaborn (opcional para visualización)
- Jupyter Notebook, Google Colab o entorno similar

Requerimientos:

1. Carga y limpieza de datos
   - Crea un DataFrame con los datos entregados, agregando intencionalmente al menos un valor faltante (`None` o `np.nan`) en “ingresos” y otro atípico en “edad” (por ejemplo, un valor negativo o superior a 100).
   - Identifica y muestra cuántos valores faltantes o anómalos hay en cada columna.
2. Exploración básica
   - Usa los métodos `.head()`, `.info()`, `.describe()`, `.value_counts()` sobre al menos dos columnas categóricas.
   - Comenta los resultados: ¿Qué distribuciones, rangos y datos inusuales identificas?
3. Limpieza de datos
   - Reemplaza los valores atípicos detectados en “edad” por el valor mediano de la columna.
   - Imputa los valores faltantes en “ingresos” usando la media por “área”.
4. Filtrado avanzado y selección múltiple
   - Selecciona todos los empleados que:
     - Trabajen en Madrid o Barcelona,
     - Tengan más de 30 años,
     - Y más de 5 años de experiencia.
   - ¿Cuántos cumplen todos estos criterios? ¿Qué características comparten?
5. Agrupación y resumen
   - Calcula el ingreso promedio y la edad máxima por “área” y por “ciudad”.
   - ¿En qué ciudad y área se encuentran los ingresos más altos?
6. Creación y transformación de variables
   - Crea una columna “nivel_ingresos”:
     - “alto” si ingresos > 3200,
     - “medio” si ingresos entre 2800 y 3200 (inclusive),
     - “bajo” si ingresos < 2800.
   - Crea una columna “senioridad”:
     - “junior” si años_experiencia < 5,
     - “semisenior” si entre 5 y 10,
     - “senior” si > 10.
7. Visualización exploratoria
   - Grafica un boxplot de ingresos por ciudad.
   - Grafica un conteo de empleados por área y senioridad (puedes usar pandas o seaborn/matplotlib).
8. Interpretación y propuesta
   - Resume en 4-6 líneas los principales hallazgos sobre desigualdades, grupos destacados, anomalías corregidas y recomendaciones para la gerencia.
   - Propón al menos una pregunta o análisis adicional que podría hacerse con este dataset.

Instrucciones de Desarrollo:
Modalidad grupal.
Tiempo: 75 minutos.
Entrega: Notebook ejecutado y comentado + gráficos explicados.