«Talento Tech»

Data Analytics

con Python

Clase 12

Buenos Aires Ciudad

Clase N° 12 | Consolidación de Datos para Análisis

Temario:

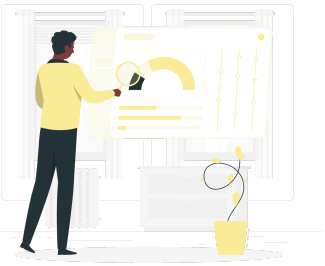
- Integración de conocimientos previos.
- Técnicas para Consolidar Datos y Preparar para el Análisis Final

---

Objetivos de la Clase:

- Integrar, profundizar y reforzar conocimientos imprescindibles para enfrentar el módulo de visualización de datos
- Comprender la importancia de la consolidación de datos.
- Dominar las herramientas y técnicas de manipulación de datos.

Consolidación de Datos para Análisis en Python con NumPy y Pandas



La **consolidación de datos** es el paso previo al análisis de datos, que implica reunir información de diferentes fuentes y transformarla en un formato que sea adecuado para su posterior análisis. En este documento, repasaremos todos lo aprendido, exploraremos cómo utilizar Python, junto con las **bibliotecas NumPy y Pandas**, integrando todas estas herramientas, para consolidar, transformar y preparar conjuntos de datos.

¿Qué es la Consolidación de Datos (ETL)?

La consolidación de datos (**ETL: "extraer, transformar, cargar"**) se refiere al proceso de combinar datos dispersos, a menudo provenientes de diferentes sistemas o formatos, en una única representación coherente. Esto permite a los analistas obtener una visión más clara de la información y facilita la identificación de patrones, tendencias y relaciones dentro de los datos. Un conjunto de datos bien consolidado es esencial para llevar a cabo un análisis efectivo y significativo.

Pasos para consolidar datos y prepararlos para el análisis final

1. Importación de Bibliotecas

Para comenzar, es fundamental importar las bibliotecas necesarias en Python. Usaremos Pandas para la manipulación de los datos y NumPy para operaciones numéricas:

```
import pandas as pd
import numpy as np
```

2. Carga de Conjuntos de Datos

Los conjuntos de datos pueden ser importados desde diversos formatos, como CSV, Excel o bases de datos SQL. La función read_csv de Pandas es muy común para cargar archivos CSV:

```
df = pd.read_csv('nombre_del_archivo.csv')
```

3. Exploración Inicial de los Datos

Antes de consolidar, es importante explorar los datos para entender su estructura y contenido. Podemos usar métodos como head(), info() y describe() para obtener una visión general:

```
print(df.head())  # Muestra las primeras filas del DataFrame
print(df.info())  # Información sobre el DataFrame
print(df.describe())  # Estadísticas del DataFrame
```

4. Limpieza de Datos

La limpieza de datos es otro paso crítico en la consolidación. Esto puede incluir la eliminación de duplicados, el manejo de valores faltantes, y la corrección de formatos inconsistentes:

```
df.drop_duplicates(inplace=True)  # Elimina duplicados
df.fillna(method='ffill', inplace=True)   # Rellena valores
faltantes
```

5. Transformación de Datos

Una vez que los datos están limpios, podemos proceder a transformarlos. Esto incluye operaciones como la creación de nuevas columnas, la conversión de tipos de datos y la normalización de los valores:

```
df['nueva_columna'] = df['columna_existente'] * 2   # Crear
una nueva columna
df['fecha'] = pd.to_datetime(df['fecha'])  # Convertir a tipo
de dato fecha
```

6. Consolidación de Datos

La consolidación puede implicar combinar múltiples DataFrames. Utilizaremos operaciones como merge, concat y join para combinar datos de diferentes fuentes:

```
otros_datos = pd.read_csv('otros_datos.csv')
df_consolidado = pd.merge(df, otros_datos, on='clave_comun',
how='inner')  # Combina DataFrames
```

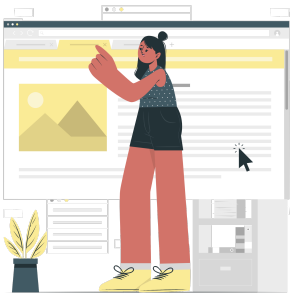
7. Agrupación y Agregación

La agrupación de datos es especialmente útil para realizar análisis descriptivos. Podemos usar el método groupby para agrupar datos según ciertas características y aplicar funciones de agregación:

```
agrupado = df.groupby('categoría')['valor'].sum()  # Suma los valores por categoría
```

Reflexión final

La consolidación de datos es un proceso multifacético que sienta las bases para realizar análisis más profundos. Saber cómo consolidar, transformar y preparar datos de manera eficiente es parte fundamental en el campo de la ciencia de datos. Ya estamos listos para adentrarnos de lleno en el campo de la visualización de datos.



Materiales y recursos adicionales

- Numpy.org
- Pandas.pydata.org

Próximos Pasos

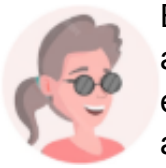
- Matplotlib: Repaso sobre tipos de gráficos básicos.
- Profundización de gráficos básicos: líneas, barras y dispersión utilizando Matplotlib.
- Personalización.

Ejercicios Prácticos

SYNTHDATA
SINTETIZA CONOCIMIENTO

Actividad 1: Análisis de Ventas y Productos

Contexto



En esta semana de tu pasantía en SynthData, te asignaron la tarea de analizar la información de ventas de un cliente especializado en e-commerce. Silvia, la Project Manager y Data Scientist, te ofreció su apoyo para esta tarea. Esta análisis no solo es importante para entender las ventas pasadas, sino también para proponer mejoras y optimizaciones en las estrategias de comercialización.

Objetivos

- Cargar y explorar múltiples conjuntos de datos que contengan información sobre las ventas, productos y clientes.
- Limpiar y transformar los datos para garantizar su calidad y utilidad.
- Realizar un análisis descriptivo.

Ejercicio práctico

1. Cargá los conjuntos de datos que contienen la información sobre las ventas, productos y clientes..
2. Explorá cada DataFrame.
3. Limpiá los datos, eliminando duplicados y manejando valores nulos. Asegurate de que las columnas de fecha estén en el formato adecuado.
4. Ideá una manera de calcular el total gastado por cada cliente.
5. En cada paso, documentá tus hallazgos y las modificaciones que vas aplicando.
6. Realizá un análisis de las ventas por categoría de producto y visualiza los resultados utilizando gráficos de barras.

Datasets

- ventas.csv
- productos.csv
- clientes.csv

¿Por qué importa esto en SynthData?

Entender las ventas y el comportamiento de los consumidores proporciona información sobre cómo los cambios en la estrategia pueden influir en el rendimiento de ventas. Este ejercicio te permitirá adquirir competencias valiosas en análisis de datos.

Actividad 2: Análisis de Asistencia y Resultados en Talleres de Capacitación

Contexto

Después de realizar un excelente trabajo en la primera actividad, ahora Matías, el Data Analyst en SynthData, te asignó una nueva tarea. Esta vez, vas a analizar la participación y el rendimiento de los asistentes en una serie de talleres de capacitación organizados por una ONG asociada. Tu trabajo es evaluar la efectividad de estos talleres para ayudar a la organización a tomar decisiones para futuras ediciones.

Objetivos

- Cargar y combinar datos sobre la asistencia a talleres, los participantes y su rendimiento en los talleres.
- Limpiar y transformar los datos, asegurando su integridad y aplicabilidad.
- Realizar un análisis descriptivo para resumir los resultados y la participación en los talleres.

Ejercicio práctico

1. Cargá los conjuntos de datos.
2. Explorá cada DataFrame para familiarizarte con los datos.
3. Limpiá los datos.
4. Combiná los tres conjuntos de datos en un único DataFrame utilizando las claves apropiadas.
5. Realizá un análisis para calcular la asistencia promedio a los talleres y el puntaje promedio obtenido por los participantes*, visualizando los resultados en gráficos*.

Datasets:

- talleres.csv,
- participantes.csv
- resultados_taller.csv.

¿Por qué importa esto en SynthData?

Evaluar la asistencia y los resultados de los talleres de capacitación es fundamental para SynthData, ya que les permite realizar mejor sus informes y ayudar a las organizaciones a medir el impacto de sus programas. Esta actividad fortalecerá tu capacidad para trabajar con datos de múltiples fuentes y te preparará para abordar problemas complejos en el ámbito de la ciencia de datos.

⛔ **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