«Talento Tech»

Data Analytics

con Python

Clase 05

Buenos Aires Ciudad

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Clase N° 5 | Limpieza de datos con Pandas

Temario:

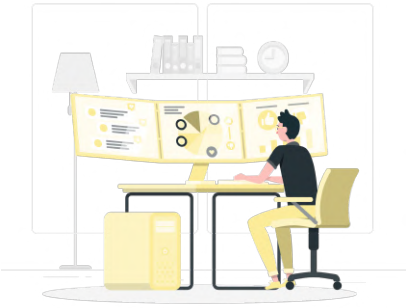
- Los desafíos de la raw data.
- Técnicas para limpiar datos: eliminación de duplicados, caracteres no deseados, corrección en los tipos de datos.

Objetivos de la clase:

- Conocer la importancia de la limpieza de datos en el campo de la analítica.
- Familiarizarse con las técnicas de limpieza de datos de Pandas.

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Limpieza de Datos con Pandas en Data Analytics



La limpieza de datos es uno de los pasos más críticos en el proceso de análisis de datos. Utilizando **Python** y la **biblioteca Pandas**, los analistas pueden transformar datos crudos en información útil y significativa.

Desafíos de los Datos Crudos

Los ***raw data*** o datos crudos suelen presentar varios problemas que dificultan su análisis. Algunos de los desafíos más comunes incluyen:

1. **Inconsistencias:** Los datos pueden contener formatos diferentes, como fechas escritas en distintos estilos (`dd/mm/yyyy` vs. `mm/dd/yyyy`).
2. **Datos Faltantes:** Es común que algunos valores se encuentren ausentes en el conjunto de datos, lo que puede afectar el análisis.
3. **Errores:** Los datos pueden incluir errores tipográficos o valores que no tienen sentido (por ejemplo, una edad negativa).
4. **Duplicados:** Puede haber entradas repetidas que influyen en los resultados del análisis.
5. **Valores No Deseados:** Caracteres especiales o etiquetas que no son relevantes para el análisis pueden estar presentes.

Importancia de la Limpieza de Datos en Data Analytics

La limpieza de datos es fundamental por varias razones:

- **Precisión:** Los análisis con datos sucios pueden llevar a conclusiones erróneas. Limpiarlos asegura que el análisis es preciso.

- **Eficiencia:** La limpieza de datos antes de realizar análisis ahorra tiempo en el proceso, ya que evita trabajar con información irrelevante o incorrecta.

- **Mejor Toma de Decisiones:** Unos datos bien formateados y precisos facilitan una mejor interpretación y análisis, lo que a su vez lleva a decisiones más informadas.

Técnicas para Limpiar Datos con Pandas

Pandas ofrece una variedad de herramientas y métodos para limpiar datos. A continuación, se presentan algunas técnicas junto con ejemplos.

1. Eliminación de Duplicados

Los duplicados pueden distorsionar los resultados del análisis. Para eliminarlos, se puede utilizar el método `drop_duplicates()`.

```
# Crear un DataFrame de ejemplo
data = {'Nombre': ['Juan', 'Ana', 'Juan'],
        'Edad': [28, 22, 28]}
df = pd.DataFrame(data)

# Eliminar duplicados
df_limpio = df.drop_duplicates()
print(df_limpio)
```

2. Eliminación de Caracteres No Deseados

A menudo, los datos pueden contener caracteres no deseados. El método `str.replace()` permite limpiar cadenas de texto.

```
# Supongamos que tenemos un DataFrame con nombres que
contienen caracteres especiales
df['Nombre'] = df['Nombre'].str.replace('ñ', 'n')
```

3. Corrección de Tipos de Datos

Es importante que los datos estén en el tipo correcto, ya que afecta las operaciones que se pueden realizar. Para cambiar el tipo de una columna, se utiliza `astype()`.

```
# Cambiar la columna 'Edad' a tipo entero
df['Edad'] = df['Edad'].astype(int)
```

4. Manejo de Datos Faltantes

Pandas ofrece métodos como `fillna()` para sustituir valores nulos o `dropna()` para eliminarlos.

```
# Sustituir valores nulos
df['Edad'] = df['Edad'].fillna(df['Edad'].mean())

# O, eliminar filas con datos faltantes
df = df.dropna()
```

5. Normalización de Datos

La normalización puede incluir la conversión de todas las entradas de texto a minúsculas para asegurar consistencia.

```
# Convertir todos los nombres a minúsculas
df['Nombre'] = df['Nombre'].str.lower()
```

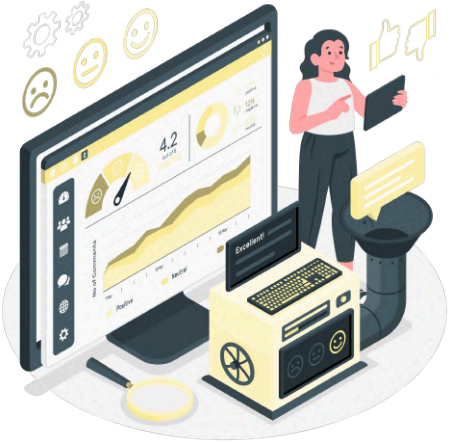
6. Filtrado de Datos

A veces, es útil filtrar datos según ciertas condiciones. Esto se puede lograr con querys que permiten trabajar solo con datos relevantes.

```
# Filtrar para obtener sólo las filas donde la edad es mayor a
25
df_filtrado = df[df['Edad'] > 25]
```

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Reflexión final



La **limpieza de datos** es un paso esencial en el proceso de análisis de datos que no debe ser pasado por alto. Utilizando herramientas como Pandas, es posible transformar datos crudos en un conjunto de datos limpio y estructurado que permita análisis precisos y relevantes. Con las técnicas presentadas, podremos abordar los desafíos comunes en tus proyectos de análisis de datos y mejorar nuestros resultados.

Materiales y recursos adicionales

- Documentación de Pandas

Próximos pasos

- Introducción a operaciones de transformación.
- Filtros, selecciones y transformaciones con Pandas.

Ejercicios prácticos

SYNTHDATA
SINTETIZA CONOCIMIENTO

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Actividad 1: Eliminación de Duplicados y Tratamiento de Datos Faltantes

Contexto

En esta etapa, te enfrentarás a un proyecto real. Tu mentor para esta actividad será Matías, nuestro Data Analyst, quien te guiará mientras integrás tus conocimientos técnicos con las decisiones analíticas.
En esta actividad, trabajarás con un conjunto de datos ficticios de una tienda en línea que incluye información sobre clientes, como nombre, edad y correo electrónico. Tu objetivo será limpiar y preparar estos datos para un análisis más profundo. Esto implica eliminar registros duplicados y decidir cómo manejar los datos faltantes en la columna de edad, lo cual es fundamental para segmentar adecuadamente a nuestros clientes.

Objetivos

- Cargar el conjunto de datos y familiarizarte con su estructura.
- Aplicar métodos para identificar y eliminar las filas duplicadas.
- Evaluar diferentes estrategias para manejar los datos faltantes.

Ejercicio práctico

- Cargar el conjunto de datos.
- Identificar y eliminar las filas duplicadas: Usarás técnicas específicas para encontrar y eliminar cualquier registro duplicado que pueda afectar la calidad de tu análisis. No olvides escribir un reporte de los hallazgos y modificaciones.
- Manejar los datos faltantes en la columna de edad, evaluando cuál es la mejor decisión, considerando que vamos a realizar un análisis enfocado en grupos etarios de los clientes. Evaluar el impacto sobre el análisis si:
    - Se eliminan las filas que no contienen la edad.
    - Se completa el dato con la media de la columna.

Sets de datos

data_clientes Conjunto de datos sintéticos que contiene información de clientes de un comercio en línea.

¿Por qué importa esto en SynthData?

La eliminación de duplicados y el tratamiento de datos faltantes son procesos cruciales para asegurar la calidad de nuestros análisis. La información precisa nos permite a nosotros, en SynthData, crear personas efectivas que describan a nuestros clientes y desarrollar estrategias basadas en insights confiables. Matías te ayudará a comprender la relevancia de cada paso en esta tarea, asegurando que tu aprendizaje se ajuste a las exigencias del mercado.

Actividad 2: Corrección de Tipos de Datos y Normalización

Contexto

En esta actividad, tendrás acceso a un conjunto de datos que contiene información sobre productos disponibles en un comercio, incluyendo el nombre del producto, su precio y la categoría. Tu objetivo será cargar este conjunto de datos y aplicar las transformaciones necesarias para garantizar que todos los registros sean consistentes y estén listos para un análisis más profundo. Tu mentor para esta tarea será Sabrina, nuestra Data Engineer, quien te guiará a través de los pasos necesarios para corregir tipos de datos y normalizar la información en un conjunto de datos sobre productos.

Objetivos

Aplicar los criterios y métodos apropiados para obtener tipos de datos consistentes y normalizados.

Ejercicio práctico

1. Cargar el conjunto de datos.
2. Corregir el tipo de datos de la columna de precio a un tipo numérico.
3. Normalizar el nombre del producto para que todas las entradas estén en minúsculas y sin caracteres especiales.

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Sets de datos

productos

¿Por qué importa esto en SynthData?
La corrección de tipos de datos y la normalización son procesos esenciales para asegurar que cada dato se maneje de manera correcta y estandarizada. En SynthData, trabajamos con grandes volúmenes de información, y es crucial que los datos sean precisos y consistentes para obtener insights valiosos. Sabrina te mostrará la relevancia de estas tareas y cómo afectan la calidad de la información que utilizamos para la toma de decisiones.

⛔ **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.**
Recordá 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