«Talento Tech»

Data Analytics

con Python

Clase 06

Buenos Aires Ciudad

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Clase N° 6 | Transformación de datos

Temario:

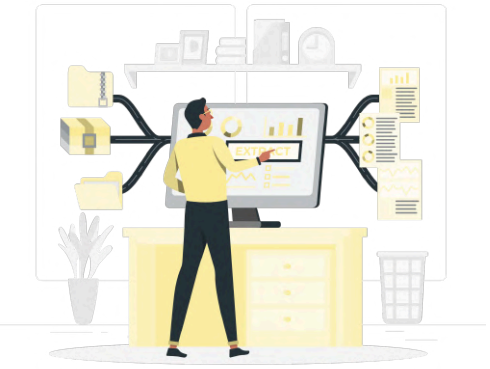
- Introducción a operaciones de transformación
- Filtros, selecciones y transformaciones con Pandas.

Objetivos de la clase:

- Comprender la importancia de la transformación de datos en la analítica.
- Aprender a utilizar la biblioteca Pandas de Python para realizar transformaciones.
- Aplicar filtros y selecciones para preparar los datos para análisis avanzado.
- Familiarizarse con las operaciones básicas de transformación en Pandas.

Buenos Aires aprende
Agencia de Habilidades para el Futuro

¿Qué son las selecciones y filtros de datos?



Las selecciones y filtros de datos son **técnicas utilizadas para extraer subconjuntos específicos de información de un conjunto de datos más grande**. Estas operaciones permiten a los analistas enfocarse en las dimensiones y métricas que son relevantes para su análisis.

- **Selecciones**: Se refieren a la elección de columnas o filas particulares de un DataFrame.
- **Filtros**: Implican aplicar condiciones para obtener sólo aquellos registros que cumplen ciertos criterios, permitiéndonos restringir la visualización de datos.

Ejemplo

Si tenemos un DataFrame con datos de ventas, podemos filtrar para mostrar sólo las ventas realizadas en una región específica o seleccionar sólo las columnas que nos interesan, como el nombre del vendedor y la cantidad vendida.

¿Qué son las transformaciones de datos?

Las transformaciones de datos son **operaciones que modifican la estructura o el contenido de un conjunto de datos**. Esto puede incluir la creación de nuevas columnas, la eliminación de información no relevante o la reestructuración de los datos para facilitar su análisis.

Ejemplo

Si deseamos analizar el rendimiento de un producto, podemos crear una nueva columna que calcule la tasa de crecimiento de las ventas a partir de datos existentes.

Buenos Aires aprende
Agencia de Habilidades para el Futuro

¿Para qué se utilizan las selecciones, filtros y transformaciones de datos en Data Analytics?

Estas operaciones son fundamentales en Data Analytics por varias razones:

Exploración de datos	Permiten a los analistas comprender la estructura y el contenido de los datos, ayudando a identificar patrones y anomalías.
Preparación de datos	Facilitan la limpieza y estructuración de datos, asegurando que el análisis se realice sobre información relevante y precisa.
Análisis específico	Ayudan a focalizar el análisis en áreas de interés, permitiendo una toma de decisiones informada basada en resultados concretos.

1. Filtros y selecciones con Pandas

Obtener los nombres de las columnas y los índices

```
# Crear un DataFrame simple
data = {
    'Producto': ['A', 'B', 'C', 'D'],
    'Ventas': [100, 150, 200, 250],
    'Fecha': ['01/01/2023', '02/01/2023', '03/01/2023',
'04/01/2023']
}
```

Buenos Aires aprende
Agencia de Habilidades para el Futuro

```
df = pd.DataFrame(data)

# Nombres de las columnas
columnas = df.columns
print("Nombres de las columnas:\n", columnas)

# Índices del DataFrame
indices = df.index
print("Índices del DataFrame:\n", indices)
```

Obtener datos de una columna

```
# Obtener datos de la columna 'Ventas'
ventas = df['Ventas']
print("Datos de la columna Ventas:\n", ventas)
```

Obtener los datos de una fila

```
# Obtener datos de la fila con índice 1
fila_1 = df.loc[1]
print("Datos de la fila 1:\n", fila_1)
```

Seleccionar dos o más columnas

```
# Seleccionar columnas 'Producto' y 'Ventas'
sub_df = df[['Producto', 'Ventas']]
print("Selección de columnas Producto y Ventas:\n", sub_df)
```

Seleccionar un subconjunto de filas y columnas

```
# Seleccionar filas 0 a 2 y columnas 'Producto' y 'Ventas'
subset = df.loc[0:2, ['Producto', 'Ventas']]
print("Subconjunto de filas y columnas:\n", subset)
```

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Selección condicional

```
# Selección condicional: productos con ventas mayores a 150
filtro_condicional = df[df['Ventas'] > 150]
print("Productos con ventas mayores a 150:\n",
filtro_condicional)
```

Búsqueda condicional usando `query()`

```
# Usar query para obtener productos con ventas mayores a 150
query_result = df.query('Ventas > 150')
print("Resultados de la búsqueda condicional:\n",
query_result)
```

2. Transformaciones con Pandas

Crear una nueva columna

```
# Crear una nueva columna con el 10% de aumento en ventas
df['Ventas_Aumento'] = df['Ventas'] * 1.1
print("DataFrame con nueva columna de ventas aumentadas:\n",
df)
```

Eliminar una columna

```
# Eliminar la columna 'Fecha'
df_dropped = df.drop(columns=['Fecha'])
print("DataFrame después de eliminar la columna Fecha:\n",
df_dropped)
```

Eliminar una fila

```
# Eliminar la fila con índice 0
df_dropped_row = df.drop(index=0)
print("DataFrame después de eliminar la fila con índice 0:\n",
df_dropped_row)
```

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Reindexación

```
# Reindexar el DataFrame
df_reindexed = df.reset_index(drop=True)
print("DataFrame reindexado:\n", df_reindexed)
```

Transposición

```
# Transponer el Dataframe
df_transposed  = df.transpose()
print("DataFrame transpuesto:\n", df_transposed)
```

Reflexión final

La práctica de selecciones, filtros y transformaciones en pandas permiten **manipular y preparar datos de manera eficiente para obtener insights significativos y tomar decisiones fundadas**. Al dominar estas técnicas, se desarrollarán las habilidades necesarias para enfrentar con éxito proyectos analíticos complejos.

Materiales y recursos adicionales

- Documentación de Pandas

Próximos pasos

- Uso de groupby y pivot_table en Pandas.

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Ejercicios prácticos

SYNTHDATA
SINTETIZA CONOCIMIENTO

Actividad 1: Análisis de Supervivencia

Contexto

En esta primera actividad, trabajarás en un proyecto fuera de lo común para un historiador que necesita incorporar algunas estadísticas a su nuevo trabajo. Se trata de una base de datos que contiene información sobre los pasajeros del Titanic. Tu objetivo será desentrañar quiénes fueron los que sobrevivieron y evaluar cómo factores como el género, la edad y el precio del ticket influyeron en su supervivencia. Con la guía de Luis, aprenderás a filtrar y seleccionar la información relevante para tu análisis.

Objetivos

Aprender a:

- Aplicar filtros sobre los registros de un DataFrame.
- Obtener bases de datos que contengan sólo la información relevante para el análisis.
- Modificar las estructuras de los DataFrames.

Ejercicio práctico

1. Filtrar pasajeros sobrevivientes: Tu primer paso será identificar y extraer a los pasajeros que lograron sobrevivir al desastre.

2. Seleccionar columnas relevantes: De los sobrevivientes, deberás extraer únicamente las columnas que te interesan para el análisis: 'sex', 'age' y 'fare'.
3. Crear una nueva columna: Vas a agregar una nueva columna al DataFrame que clasifique a cada pasajero según la clase en la que viajaban, etiquetándola como una categoría.

Sets de datos

Utilizarás el conjunto de datos titanic.csv que se encuentra en la librería Seaborn. Para cargar el conjunto de datos, puedes usar el siguiente código:

```
import pandas as pd
import seaborn as sns

# Cargar el dataset de Titanic
df_titanic = sns.load_dataset('titanic')
```

¿Por qué importa esto en SynthData?

El análisis de supervivencia es fundamental para la comprensión y detección de patrones y las dinámicas de eventos críticos en conjuntos de datos. En SynthData, al aprender a analizar datos históricos como los del Titanic, estarás desarrollando habilidades que se pueden aplicar en una variedad de situaciones del mundo real, desde investigaciones de mercado hasta estudios de comportamiento del consumidor.

Actividad 2: Manipulación de Datos

Contexto

En esta actividad podrás practicar técnicas fundamentales en la manipulación de DataFrames, obtener información, modificar estructuras y manipular índices, gracias a la orientación de Matías.

Objetivos

Practicar técnicas de manipulación de columnas, índices y obtención de información de los DataFrames.

Buenos Aires aprende
Agencia de Habilidades para el Futuro

Ejercicio práctico

Practicar la manipulación de datos haciendo selecciones y transformaciones sobre el conjunto de datos de Titanic:

1. Obtener los nombres de las columnas del DataFrame.
2. Eliminar la columna 'deck'. (tener en cuenta que puede contener valores nulos y la función puede dar advertencias).
3. Reindexar el DataFrame después de eliminar la columna.

¿Por qué importa esto en SynthData?

La manipulación de datos es un paso esencial en el proceso de análisis, ya que te permite trabajar con la información de manera eficiente y efectiva. En SynthData, aprender a seleccionar y transformar datos de conjuntos complejos como el del Titanic te enseñará a preparar la información para análisis más profundos y a comunicar hallazgos significativos.

⛔ **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