

Data Analysis con Python

Juan Felipe Acevedo Pérez
Monitor Junior

Curso libre:
Data Analysis con
Python

Monitor encargado:
Juan Felipe Acevedo
Pérez

Correo: uniic_bog@unal.edu.co

Tel: 3165000 **Ext:** 12301

4

Sesión 4

Limpieza y preparación de datos- Pandas 2

Correo: uniic_bog@unal.edu.co

Teléfono: 3165000 ext 12301

Contenido



**Aplicación y
mapeo de
funciones**



**Manejo de datos
faltantes**







**Datos
duplicados**

Apply, Map & Applymap

Apply, Map y Applymap son métodos que permiten *aplicar funciones* a un DataFrame, se diferencian principalmente según los elementos a los que se aplica.

- **Apply:** Se aplica tanto a series como a DataFrames.
- **Map:** Solo se aplica a series.
- **Applymap:** Solo se aplica a Dataframes.

Nota: En términos de eficiencia Map > Applymap > Apply (No siempre se cumple).

	DataFrame	Series
apply		
map		
applymap		

Datos faltantes

NaN None

- La diferencia entre los datos que se encuentran en muchos tutoriales y los datos del mundo real es que los **datos del mundo real rara vez son limpios y homogéneos**. En particular, a muchos conjuntos de datos interesantes les faltará cierta cantidad de datos. Para complicar aún más las cosas, diferentes fuentes de datos pueden indicar datos faltantes de diferentes maneras.

Valores nulos de Python

NaN

- NaN (acrónimo de **Not a Number**), es un valor float reconocido por todos los sistemas que usan la representación de punto flotante estándar IEEE.

None

- Debido a que *None* es un objeto de Python, no se puede usar en una matriz NumPy/Pandas arbitraria, sino solo en matrices con el tipo de datos '**objeto**'.

Manejo de None y NaN

- Pandas trata a None y NaN como esencialmente intercambiables para indicar valores faltantes o nulos. Para facilitar esta convención, existen varios métodos útiles para detectar, eliminar y reemplazar valores nulos en las estructuras de datos de Pandas:
 - **isnull()**: Genere una máscara booleana que indique los valores faltantes.
 - **notnull()**: Opuesto a isnull().
 - **dropna()**: Devuelve una versión filtrada de los datos.
 - **fillna()**: Devolver una copia de los datos con los valores faltantes completados o imputados.

Datos duplicados



Se pueden encontrar filas duplicadas en un DataFrame por varias razones. La duplicidad puede ser del registro completo o solamente de unos elementos.

Saber cómo eliminar estos registros duplicados es imprescindible para evitar *posibles errores* en los análisis posteriores