



# ENGIN 604 INTRODUCCIÓN A PYTHON PARA LAS FINANZAS

## ANÁLISIS DE DATOS CON PANDAS

**Profesor:** *Gabriel E. Cabrera*

**Ayudante:** *Alex Den Braber*



La librería Pandas está pensada para análisis de datos del tipo tabular. No solo provee clases y funciones útiles, también permite aplicar funciones desde otras librerías como NumPy.

Existen dos tipos de estructuras fundamentales en Pandas:

Cuadro 1: Estructuras Fundamentales

Tipo de objeto	Descripción	Usado para
<b>DataFrame</b>	Objeto de 2-dimensiones con índice ( <b>index</b> )	Datos tabulares organizados en columnas
<b>Series</b>	Objeto de 1-dimensión con índice ( <b>index</b> )	Serie (de tiempo) de datos única

Para importar Pandas:

```
# se importa pandas
import pandas as pd
```

## 1. Series

Una Series es un *array* de una dimensión que contiene una secuencia de valores (como en NumPy) y una etiqueta (*label*) denominada índice (**index**). Para crear una Series:

```
# se crea Series
obj = pd.Series([4, 7, -5, 3],                # se define los datos
                index = ['a', 'b', 'c', 'd']) # se especifica el índice
```

```
# se verifica el obj
obj
```

```
## a    4
## b    7
## c   -5
## d    3
## dtype: int64
```

## 2. DataFrames

Un DataFrame es una estructura de 2-dimensiones con datos etiquetados (*labels*), índice en las filas como en las columnas. La columna potencialmente puede contener diferentes tipos de datos. Para crear un objeto **DataFrame**:

```
# se crea DataFrame
df = pd.DataFrame([10, 20, 30, 40],          # se define los datos
                  columns = ['numbers'],      # nombre de columna
                  index = ['a', 'b', 'c', 'd']) # se especifica el índice

# se verifica df
df
```

```
##      numbers
## a         10
## b         20
## c         30
## d         40
```

Es importante notar que:

- Los datos están organizados en columna (puede tener nombres personalizados)
- Hay un índice que puede tomar diferentes formatos (e.g números, *strings*, etc).

Para acceder al índice:

```
# extrae columna
df.columns
```

```
## Index(['numbers'], dtype='object')
```

Se puede pasar una Series a un DataFrame:

```
# serie a df
obj_to_df = obj.to_frame()

# se le asigna el nombre de la columna
obj_to_df.columns = ['numbers']

# verifica obj_to_df
obj_to_df
```

```
##      numbers
## a          4
## b          7
## c         -5
## d          3
```

En resumen la estructura de DataFrame:

The diagram illustrates the structure of a DataFrame with the following components:

- Column names:** Name, Team, Number, Position, Age, Height, Weight, College, Salary.
- Index label:** 0, 1, 2, 3, 4, 5, 6.
- Data:** The values within the cells of the DataFrame.
- Missing value:** Indicated by 'NaN' in the 'Number' column for index 3 and 'NaN' in the 'Age' column for index 5.

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston University	NaN
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0	6-8	235.0	LSU	1170960.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN	6-9	260.0	Ohio State	2569260.0
6	Evan Turner	Boston Celtics	11.0	SG	27.0	6-7	220.0	Ohio State	3425510.0

Para ver la documentación de la librería Pandas ir a: <https://pandas.pydata.org/docs/>

### 3. Aplicación

1. El archivo `gapminder.xlsx` y `gapminder.dta` contiene un extracto del proyecto Gapminder sobre expectativa de vida (`lifeExp`), PIB per cápita (`gdpPercap`) y población (`pop`), según país (`continent`). Utilizando la librería Pandas, cargue a su espacio de trabajo ambas bases de datos. Nombre uno de los dos **DataFrame** como `gapminder`.
2. Muestre las 10 primeras y últimas observaciones de `gapminder`.
3. Genere un nuevo **DataFrame** que contenga solo los países del continente americano (`americas`) en el año 2007.
4. A partir del **DataFrame** generado en (3), muestre:
  - El país con mayor PIB per cápita en el año 2007
  - El país con menor PIB per cápita en el año 2007¿Qué observa en el índice?
5. Reinicie el índice del **DataFrame** generado en (3). Luego elimine las variables `country` y `continent`.
6. Utilizando una *list comprehension*, renombre las columnas con su nombre original en minúscula.
7. Genere una variable que contenga la expectativa de vida (`lifeexp`) en meses. Realice definiendo una función y utilizando una función anónima.
8. Utilizando la base de datos original:
  - a. Trabaje solo con los países Europeos.
  - b. Genere el crecimiento del PIB per cápita por país.
  - c. Elimine los **NAs**.
  - d. Construya una breve estadística descriptiva por país.
  - e. Guarde las estadísticas descriptivas en un archivo con extensión `.xlsx` (excel).