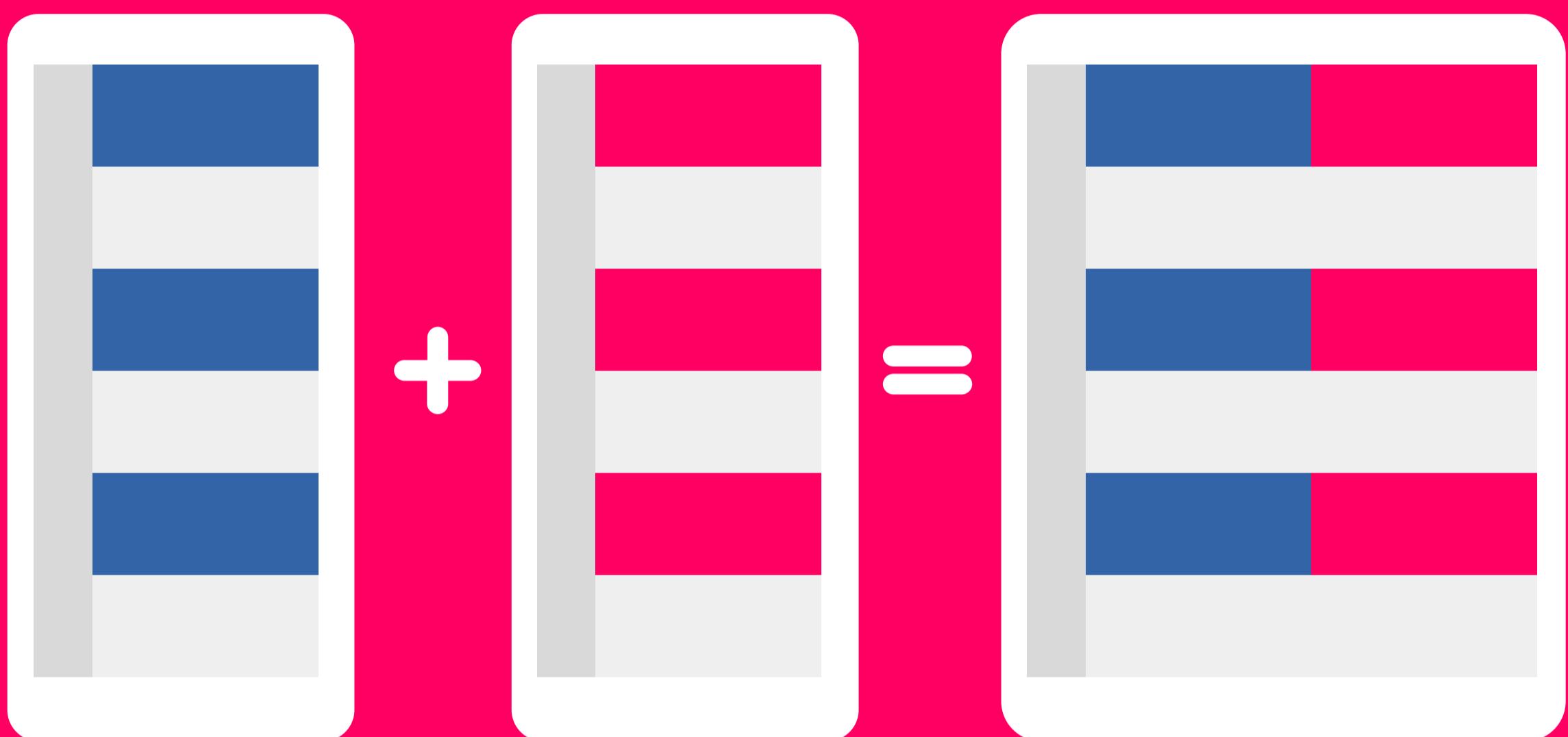




El futuro digital  
es de todos

MinTIC



# Introducción a los Dataframes



Universidad de Caldas

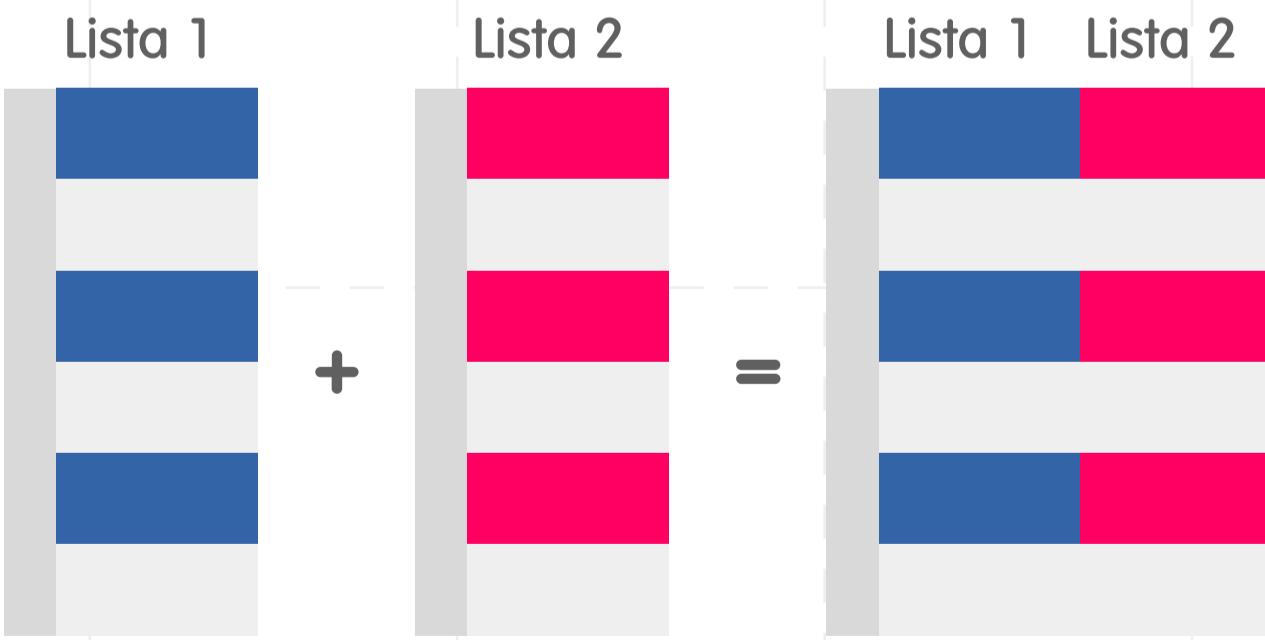
# Hola:

Además del manejo de imágenes en Python, esta semana vamos a ver algunas estructuras de datos bastante interesantes que nos va a permitir manipular información de una manera más fácil.

La nueva estructura se llama *Dataframes*. Ya hemos manejado el concepto de listas o series, que son un conjunto de elementos que están de manera secuencial. Supongamos que tenemos dos (2) listas para manejar información y que estas dos (2) listas se encuentran relacionadas, en el sentido de que el elemento en la posición cero (0) de la lista uno (1) corresponde a la misma entidad general del primer elemento de la lista dos (2) y así sucesivamente. Podemos unir estas dos (2) listas y obtener un *Dataframe*, es decir, vamos a tener una sola estructura de datos donde cada una de sus filas hace referencia a una identidad única que se forma de estas dos (2) listas. Esta es una manera de entenderlo, pero podemos crear *Dataframes* no solo sumando o adicionando listas.

**Series**

**Dataframes**



Un Dataframe es una estructura de datos bidimensional, es decir, que se parece mucho a una matriz y que incluso tiene columnas y filas que están relacionadas entre sí que va a permitirnos utilizar una librería para hacer una manipulación más directa de la información que se encuentra allí.

### ¿Cómo lo definimos en Python?

En Python definimos esto de la siguiente manera: lo primero que tenemos que hacer es importar una librería que se llama **pandas** que tiene muchas funciones para el manejo de datos y gráficos.

```
import pandas as pd

datos = {
    "alimentos": ["Manzana", "Piña", "Pera", "Arándano", "Fresa"],
    "calorías": [52, 55, 55, 35, 32]
}

# Se "cargan" en un Dataframe
df = pd.DataFrame(datos)

print(df)
```

Tenemos un diccionario de alimentos con la cantidad de calorías de cada alimento con el que podemos con una instrucción cargar esos datos en un *Dataframe* que al imprimirlo nos queda de la siguiente manera:

```
import pandas as pd

datos = {
    "alimentos": ["Manzana", "Piña", "Pera", "Arándano", "Fresa"],
    "calorías": [52, 55, 55, 35, 32]
}

# Se "cargan" en un Dataframe
df = pd.DataFrame(datos)

print(df)
```

	alimentos	calorías
0	Manzana	52
1	Piña	55
2	Pera	55
3	Arándano	35
4	Fresa	32

Al imprimir el Dataframe, nos da una estructura matricial con columnas y filas relacionadas uno a uno.

Cuando tenemos librerías, sabemos que son funciones que nos van a servir para cosas como el siguiente ejemplo:

```
>>> df.columns  
Index(['alimentos', 'calorias'], dtype='object')  
  
print(df["calorias"])  
  
0    52  
1    55  
2    55  
3    35  
4    32  
Name: calorias, dtype: int64
```

Si llamamos al *Dataframe* y le decimos “Columns”, solo nos mostrará las columnas del *Dataframe*, en este caso son alimentos y calorías. Además nos dice que el tipo de datos es objeto, esto será un tema para el próximo curso de Java.

Generalmente un *Dataframe* sirve para manejar muchos datos y volúmenes de información, sin embargo, aquí tenemos ejemplos sencillos para una mejor comprensión.

Si decimos por ejemplo, `.head` el *Dataframe* nos muestra los dos (2) primeros elementos de la lista.

```
print(df.head(2))
```

	alimentos	calorias
0	Manzana	52
1	Piña	55

.tail, nos muestra los últimos elementos de la lista.

```
df.tail(2)
```

	alimentos	calorías
3	Arándano	35
4	Fresa	32

Con.info podemos obtener la información del Dataframe.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype  
---  --  
 0   alimentos    5 non-null      object 
 1   calorías     5 non-null      int64  
dtypes: int64(1), object(1)
memory usage: 208.0+ bytes
```

También podemos utilizar otra función que es .describe, la cual hace cálculos estadísticos de la columna.

```
df.describe()
```

	calorías
count	5.000000
mean	45.800000
std	11.344602
min	32.000000
25%	35.000000
50%	52.000000
75%	55.000000
max	55.000000

Lo que nos interesa de esto y que es lo que queremos mostrar es que nos ahorra mucho trabajo porque no tenemos que hacer las funciones que me calculen estas estadísticas, sino que ya están explícitamente escritas allí.

Podemos convertirlos en listas con `list` y ubicarnos en la primera posición del Dataframe con la instrucción `.loc`, y con la instrucción `.max` nos muestra el elemento máximo que en este caso son las calorías.

```
list(df["alimentos"])
['Manzana', 'Piña', 'Pera', 'Arándano', 'Fresa']

df.loc[0]
   alimentos    Manzana
   calorias        52
   Name: 0, dtype: object

df.max()
   alimentos    Piña
   calorias        55
   dtype: object
```

Existen muchas funciones de *Dataframe* y si queremos aprender una estructura de estas lo mejor es buscar en la documentación de Python o en el w3school y allí vamos a encontrar más funciones, lo importante es entender el concepto de que un *Dataframe* es una estructura matricial que me va a permitir manejar varios volúmenes de información.

En resumen, vimos los *Dataframe* que es una formación o estructura de datos matricial que nos permite almacenar en una sola estructura elementos que están relacionados por filas, entidades, objetos, personas. Vimos que como toda librería y sobre todo **pandas**, tiene muchas funciones. Además vimos que se puede manejar información al respecto de todos los cálculos que podemos hacer con esas funciones y muchos de los elementos que tiene esas funciones, que dependiendo del contexto nos permite hacer análisis.



**Mision  
TIC 2022**



Universidad de Caldas