Open in Colab

- Raúl Correa Ocañas
- A01722401
- ICI - IDM

Operaciones con dataframes -

Referencia: https://aprendeconalf.es/docencia/python/manual/pandas/

Reshape: https://pandas.pydata.org/docs/user_guide/reshaping.html

Lectura de Datos

In [ ]:
```
import pandas as pd
import numpy as np
```

In [ ]:
```
from google.colab import drive
drive.mount('/content/gdrive')
```

```
Mounted at /content/gdrive
```

In [ ]:
```
df = pd.read_csv('/content/gdrive/MyDrive/Colab Notebooks/TC2004B.101/data/colester
```

Descripción General

In [ ]:
```
df.head()
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol
0	José Luis Martínez Izquierdo	18	H	85.0	179	182.0
1	Rosa Díaz Díaz	32	M	65.0	173	232.0
2	Javier García Sánchez	24	H	NaN	181	191.0
3	Carmen López Pinzón	35	M	65.0	170	200.0
4	Marisa López Collado	46	M	51.0	158	148.0

In [ ]:
```
df.shape
```

Out[ ]:
```
(14, 6)
```

In [ ]:
```
df.size
```

Out[ ]:
```
84
```

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14 entries, 0 to 13
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   nombre      14 non-null     object 
 1   edad        14 non-null     int64  
 2   sexo        14 non-null     object 
 3   peso        13 non-null     float64
 4   altura      14 non-null     int64  
 5   colesterol  13 non-null     float64
dtypes: float64(2), int64(2), object(2)
memory usage: 800.0+ bytes
```

In [ ]:
```
#Mostrar las columnas
df.columns
```

Out[ ]:
```
Index(['nombre', 'edad', 'sexo', 'peso', 'altura', 'colesterol'], dtype='object')
```

In [ ]:
```
df.index
```

Out[ ]:
```
RangeIndex(start=0, stop=14, step=1)
```

In [ ]:
```
df.dtypes
```

Out[ ]:
```
nombre         object
edad            int64
sexo           object
peso          float64
altura          int64
colesterol    float64
dtype: object
```

Acceso a elementos

Acceso por posición

In [ ]:
```
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol
0	José Luis Martínez Izquierdo	18	H	85.0	179	182.0
1	Rosa Díaz Díaz	32	M	65.0	173	232.0
2	Javier García Sánchez	24	H	NaN	181	191.0
3	Carmen López Pinzón	35	M	65.0	170	200.0
4	Marisa López Collado	46	M	51.0	158	148.0
5	Antonio Ruiz Cruz	68	H	66.0	174	249.0
6	Antonio Fernández Ocaña	51	H	62.0	172	276.0
7	Pilar Martín González	22	M	60.0	166	NaN
8	Pedro Gálvez Tenorio	35	H	90.0	194	241.0
9	Santiago Reillo Manzano	46	H	75.0	185	280.0
10	Macarena Álvarez Luna	53	M	55.0	162	262.0
11	José María de la Guía Sanz	58	H	78.0	187	198.0
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	198	210.0
13	Carolina Rubio Moreno	20	M	61.0	177	194.0

In [ ]:
```
#Peso de Rosa Díaz
df.iloc[1, 3]  #filas,columnas
```

Out[ ]:
```
65.0
```

In [ ]:
```
#Peso de los dos primeros
df.iloc[:2, [0,3]]
```

Out[ ]:

	nombre	peso
0	José Luis Martínez Izquierdo	85.0
1	Rosa Díaz Díaz	65.0

Acceso por nombre

In [ ]:
```
df.loc[2, 'colesterol']
```

Out[ ]:
```
191.0
```

In [ ]:
```
df.loc[:3, ('nombre','colesterol')]
```

Out[ ]:

	nombre	colesterol
0	José Luis Martínez Izquierdo	182.0
1	Rosa Díaz Díaz	232.0
2	Javier García Sánchez	191.0
3	Carmen López Pinzón	200.0

In [ ]:
```
df.describe()
```

Out[ ]:

	edad	peso	altura	colesterol
count	14.000000	13.000000	14.000000	13.000000
mean	38.214286	70.923077	176.857143	220.230769
std	15.621379	16.126901	11.501553	39.847948
min	18.000000	51.000000	158.000000	148.000000
25%	24.750000	61.000000	170.500000	194.000000
50%	35.000000	65.000000	175.500000	210.000000
75%	49.750000	78.000000	184.000000	249.000000
max	68.000000	109.000000	198.000000	280.000000

In [ ]:
```
df.head(14)
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol
0	José Luis Martínez Izquierdo	18	H	85.0	179	182.0
1	Rosa Díaz Díaz	32	M	65.0	173	232.0
2	Javier García Sánchez	24	H	NaN	181	191.0
3	Carmen López Pinzón	35	M	65.0	170	200.0
4	Marisa López Collado	46	M	51.0	158	148.0
5	Antonio Ruiz Cruz	68	H	66.0	174	249.0
6	Antonio Fernández Ocaña	51	H	62.0	172	276.0
7	Pilar Martín González	22	M	60.0	166	NaN
8	Pedro Gálvez Tenorio	35	H	90.0	194	241.0
9	Santiago Reillo Manzano	46	H	75.0	185	280.0
10	Macarena Álvarez Luna	53	M	55.0	162	262.0
11	José María de la Guía Sanz	58	H	78.0	187	198.0
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	198	210.0
13	Carolina Rubio Moreno	20	M	61.0	177	194.0

Operaciones con columnas

Agregar columnas al data frame

In [ ]:
```
df['diabetes']=pd.Series([False, False, True, False, True])
df['fecha_nac']=pd.Series(['05-03-2000', '20-05-2001', '10-12-1999'])
df
```

```
/usr/local/lib/python3.10/dist-packages/lida/components/summarizer.py:74: UserWarnin
g: Parsing dates in DD/MM/YYYY format when dayfirst=False (the default) was specifie
d. This may lead to inconsistently parsed dates! Specify a format to ensure consiste
nt parsing.
  cast_date_col = pd.to_datetime(df[column], errors='coerce')
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	diabetes	fecha_nac
0	José Luis Martínez Izquierdo	18	H	85.0	179	182.0	False	05-03-2000
1	Rosa Díaz Díaz	32	M	65.0	173	232.0	False	20-05-2001
2	Javier García Sánchez	24	H	NaN	181	191.0	True	10-12-1999
3	Carmen López Pinzón	35	M	65.0	170	200.0	False	NaN
4	Marisa López Collado	46	M	51.0	158	148.0	True	NaN
5	Antonio Ruiz Cruz	68	H	66.0	174	249.0	NaN	NaN
6	Antonio Fernández Ocaña	51	H	62.0	172	276.0	NaN	NaN
7	Pilar Martín González	22	M	60.0	166	NaN	NaN	NaN
8	Pedro Gálvez Tenorio	35	H	90.0	194	241.0	NaN	NaN
9	Santiago Reillo Manzano	46	H	75.0	185	280.0	NaN	NaN
10	Macarena Álvarez Luna	53	M	55.0	162	262.0	NaN	NaN
11	José María de la Guía Sanz	58	H	78.0	187	198.0	NaN	NaN
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	198	210.0	NaN	NaN
13	Carolina Rubio Moreno	20	M	61.0	177	194.0	NaN	NaN

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14 entries, 0 to 13
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   nombre      14 non-null     object 
 1   edad        14 non-null     int64  
 2   sexo        14 non-null     object 
 3   peso        13 non-null     float64
 4   altura      14 non-null     int64  
 5   colesterol  13 non-null     float64
 6   diabetes    5 non-null      object 
 7   fecha_nac   3 non-null      object 
dtypes: float64(2), int64(2), object(4)
memory usage: 1.0+ KB
```

Cambiar tipo de dato de columna a datetime

In [ ]:
```
df['fecha_nac'] = pd.to_datetime(df.fecha_nac, format = '%d-%m-%Y')
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	diabetes	fecha_nac
0	José Luis Martínez Izquierdo	18	H	85.0	179	182.0	False	2000-03-05
1	Rosa Díaz Díaz	32	M	65.0	173	232.0	False	2001-05-20
2	Javier García Sánchez	24	H	NaN	181	191.0	True	1999-12-10
3	Carmen López Pinzón	35	M	65.0	170	200.0	False	NaT
4	Marisa López Collado	46	M	51.0	158	148.0	True	NaT
5	Antonio Ruiz Cruz	68	H	66.0	174	249.0	NaN	NaT
6	Antonio Fernández Ocaña	51	H	62.0	172	276.0	NaN	NaT
7	Pilar Martín González	22	M	60.0	166	NaN	NaN	NaT
8	Pedro Gálvez Tenorio	35	H	90.0	194	241.0	NaN	NaT
9	Santiago Reillo Manzano	46	H	75.0	185	280.0	NaN	NaT
10	Macarena Álvarez Luna	53	M	55.0	162	262.0	NaN	NaT
11	José María de la Guía Sanz	58	H	78.0	187	198.0	NaN	NaT
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	198	210.0	NaN	NaT
13	Carolina Rubio Moreno	20	M	61.0	177	194.0	NaN	NaT

In [ ]:
```
df.dtypes
```

Out[ ]:
```
nombre                  object
edad                     int64
sexo                    object
peso                   float64
altura                   int64
colesterol             float64
diabetes                object
fecha_nac       datetime64[ns]
dtype: object
```

Operación sobre una columna

Dividir la columna entre un valor

In [ ]:
```
#Mostrar altura en metros
df['altura']/100
```

Out[ ]:
```
0     1.79
1     1.73
2     1.81
3     1.70
4     1.58
5     1.74
6     1.72
7     1.66
8     1.94
9     1.85
10    1.62
11    1.87
12    1.98
13    1.77
Name: altura, dtype: float64
```

In [ ]:
```
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	diabetes	fecha_nac
0	José Luis Martínez Izquierdo	18	H	85.0	179	182.0	False	2000-03-05
1	Rosa Díaz Díaz	32	M	65.0	173	232.0	False	2001-05-20
2	Javier García Sánchez	24	H	NaN	181	191.0	True	1999-12-10
3	Carmen López Pinzón	35	M	65.0	170	200.0	False	NaT
4	Marisa López Collado	46	M	51.0	158	148.0	True	NaT
5	Antonio Ruiz Cruz	68	H	66.0	174	249.0	NaN	NaT
6	Antonio Fernández Ocaña	51	H	62.0	172	276.0	NaN	NaT
7	Pilar Martín González	22	M	60.0	166	NaN	NaN	NaT
8	Pedro Gálvez Tenorio	35	H	90.0	194	241.0	NaN	NaT
9	Santiago Reillo Manzano	46	H	75.0	185	280.0	NaN	NaT
10	Macarena Álvarez Luna	53	M	55.0	162	262.0	NaN	NaT
11	José María de la Guía Sanz	58	H	78.0	187	198.0	NaN	NaT
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	198	210.0	NaN	NaT
13	Carolina Rubio Moreno	20	M	61.0	177	194.0	NaN	NaT

In [ ]:
```
df['altura']=df['altura']/100
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	diabetes	fecha_nac
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	False	2000-03-05
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	False	2001-05-20
2	Javier García Sánchez	24	H	NaN	1.81	191.0	True	1999-12-10
3	Carmen López Pinzón	35	M	65.0	1.70	200.0	False	NaT
4	Marisa López Collado	46	M	51.0	1.58	148.0	True	NaT
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaN	NaT
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaN	NaT
7	Pilar Martín González	22	M	60.0	1.66	NaN	NaN	NaT
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaN	NaT
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaN	NaT
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaN	NaT
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaN	NaT
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaN	NaT
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaN	NaT

Aplicar funciones a una columna

In [ ]:
```
df['altura2']=df['altura'].apply(np.square)
```

In [ ]:
```
df['imc']=df['peso']/df['altura2']
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	diabetes	fecha_nac	altura2	im
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	False	2000-03-05	3.2041	26.52851
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	False	2001-05-20	2.9929	21.71806
2	Javier García Sánchez	24	H	NaN	1.81	191.0	True	1999-12-10	3.2761	Na
3	Carmen López Pinzón	35	M	65.0	1.70	200.0	False	NaT	2.8900	22.49134
4	Marisa López Collado	46	M	51.0	1.58	148.0	True	NaT	2.4964	20.42941
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaN	NaT	3.0276	21.79944
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaN	NaT	2.9584	20.95727
7	Pilar Martín González	22	M	60.0	1.66	NaN	NaN	NaT	2.7556	21.77384
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaN	NaT	3.7636	23.91327
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaN	NaT	3.4225	21.91380
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaN	NaT	2.6244	20.95717
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaN	NaT	3.4969	22.30547
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaN	NaT	3.9204	27.80328
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaN	NaT	3.1329	19.47077

Renombrar columnas si es necesario

- Usar el método rename
- Usar inplace=True para que los cambios tengan efecto en el mismo dataframe

df.rename(columns={'nombre_actual': 'nombre_nuevo', 'nombre_actual': 'nombre_nuevo'}, inplace=True)

In [ ]:
```
df.rename(columns={'diabetes': 'diabetes_mellitus'}, inplace=True)
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	diabetes_mellitus	fecha_nac	altura2
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	False	2000-03-05	3.2041
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	False	2001-05-20	2.9929
2	Javier García Sánchez	24	H	NaN	1.81	191.0	True	1999-12-10	3.2761
3	Carmen López Pinzón	35	M	65.0	1.70	200.0	False	NaT	2.8900
4	Marisa López Collado	46	M	51.0	1.58	148.0	True	NaT	2.4964
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaN	NaT	3.0276
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaN	NaT	2.9584
7	Pilar Martín González	22	M	60.0	1.66	NaN	NaN	NaT	2.7556
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaN	NaT	3.7636
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaN	NaT	3.4225
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaN	NaT	2.6244
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaN	NaT	3.4969
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaN	NaT	3.9204
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaN	NaT	3.1329

Seleccionar ciertas columnas de un dataframe

In [ ]:
```
#Se crea un nuevo dataframe con las columnas seleccionadas
df2=df[['nombre', 'edad']]
df2
```

Out[ ]:

	nombre	edad
0	José Luis Martínez Izquierdo	18
1	Rosa Díaz Díaz	32
2	Javier García Sánchez	24
3	Carmen López Pinzón	35
4	Marisa López Collado	46
5	Antonio Ruiz Cruz	68
6	Antonio Fernández Ocaña	51
7	Pilar Martín González	22
8	Pedro Gálvez Tenorio	35
9	Santiago Reillo Manzano	46
10	Macarena Álvarez Luna	53
11	José María de la Guía Sanz	58
12	Miguel Angel Cuadrado Gutiérrez	27
13	Carolina Rubio Moreno	20

Eliminar columnas de un dataframe

del d[nombre] : Elimina la columna indicada del DataFrame df.

df.pop(nombre) : Elimina la columna indicada del DataFrame df y la devuelve como una serie.

In [ ]:
```
del(df['diabetes_mellitus'])
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.528510
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	2001-05-20	2.9929	21.718066
2	Javier García Sánchez	24	H	NaN	1.81	191.0	1999-12-10	3.2761	NaN
3	Carmen López Pinzón	35	M	65.0	1.70	200.0	NaT	2.8900	22.491349
4	Marisa López Collado	46	M	51.0	1.58	148.0	NaT	2.4964	20.429418
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaT	3.0276	21.799445
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaT	2.9584	20.957274
7	Pilar Martín González	22	M	60.0	1.66	NaN	NaT	2.7556	21.773842
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaT	3.4225	21.913806
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaT	2.6244	20.957171
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaT	3.4969	22.305471
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaT	3.1329	19.470778

Operaciones con Filas/Renglones

Añadir una fila a un dataframe

In [ ]:
```
#df.append(pd.Series(['Carlos Rivas', 28, 'H', 89.0, 1.78, 245.0], index=['nombre',
#Append deprecado, usar concat.
s2 = pd.Series(['Carlos Rivas', 28, 'H', 89.0, 1.78, 245.0], index=['nombre','edad'
s2
```

Out[ ]:
```
nombre        Carlos Rivas
edad                    28
sexo                     H
peso                  89.0
altura                1.78
colesterol           245.0
dtype: object
```

In [ ]:
```
#Convertir a dataframe y aplicar la transpuesta
s2.to_frame().T
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol
0	Carlos Rivas	28	H	89.0	1.78	245.0

In [ ]:
```
pd.concat([df,s2.to_frame().T], ignore_index=True)
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.528510
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	2001-05-20	2.9929	21.718066
2	Javier García Sánchez	24	H	NaN	1.81	191.0	1999-12-10	3.2761	NaN
3	Carmen López Pinzón	35	M	65.0	1.7	200.0	NaT	2.8900	22.491349
4	Marisa López Collado	46	M	51.0	1.58	148.0	NaT	2.4964	20.429418
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaT	3.0276	21.799445
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaT	2.9584	20.957274
7	Pilar Martín González	22	M	60.0	1.66	NaN	NaT	2.7556	21.773842
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaT	3.4225	21.913806
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaT	2.6244	20.957171
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaT	3.4969	22.305471
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaT	3.1329	19.470778
14	Carlos Rivas	28	H	89.0	1.78	245.0	NaT	NaN	NaN

In [ ]:
```
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.528510
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	2001-05-20	2.9929	21.718066
2	Javier García Sánchez	24	H	NaN	1.81	191.0	1999-12-10	3.2761	NaN
3	Carmen López Pinzón	35	M	65.0	1.70	200.0	NaT	2.8900	22.491349
4	Marisa López Collado	46	M	51.0	1.58	148.0	NaT	2.4964	20.429418
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaT	3.0276	21.799445
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaT	2.9584	20.957274
7	Pilar Martín González	22	M	60.0	1.66	NaN	NaT	2.7556	21.773842
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaT	3.4225	21.913806
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaT	2.6244	20.957171
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaT	3.4969	22.305471
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaT	3.1329	19.470778

Es necesario guardarlo en el dataframe

In [ ]:
```
#df=df.append(pd.Series(['Carlos Rivas', 28, 'H', 89.0, 1.78, 245.0], index=['nombr
s2 = pd.Series(['Carlos Rivas', 28, 'H', 89.0, 1.78, 245.0], index=['nombre','edad'
```

```
df = pd.concat([df,s2.to_frame().T], ignore_index=True)
df.tail()
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaT	2.6244	20.957171
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaT	3.4969	22.305471
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaT	3.1329	19.470778
14	Carlos Rivas	28	H	89.0	1.78	245.0	NaT	NaN	NaN

Seleccionar filas de un dataframe

select the rows of the dataframe for which float column is larger than 0.15 Select the rows for which float column is larger than 0.1 and integer column is larger than 2. Change 'and' by 'or' Select the rows for which string column is not 'a'

In [ ]:
```
df[df.peso>80]
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.528510
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
14	Carlos Rivas	28	H	89.0	1.78	245.0	NaT	NaN	NaN

In [ ]:
```
df.loc[df['peso'] > 80]
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.528510
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
14	Carlos Rivas	28	H	89.0	1.78	245.0	NaT	NaN	NaN

In [ ]:
```
df[(df.peso>80) & (df.colesterol>200)]
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
14	Carlos Rivas	28	H	89.0	1.78	245.0	NaT	NaN	NaN

In [ ]:
```
df[(df.peso > 80) | (df.colesterol>200)]
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.528510
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	2001-05-20	2.9929	21.718066
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaT	3.0276	21.799445
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaT	2.9584	20.957274
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaT	3.4225	21.913806
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaT	2.6244	20.957171
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
14	Carlos Rivas	28	H	89.0	1.78	245.0	NaT	NaN	NaN

In [ ]:
```
df[(df.edad > 18)]
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	2001-05-20	2.9929	21.718066
2	Javier García Sánchez	24	H	NaN	1.81	191.0	1999-12-10	3.2761	NaN
3	Carmen López Pinzón	35	M	65.0	1.7	200.0	NaT	2.8900	22.491349
4	Marisa López Collado	46	M	51.0	1.58	148.0	NaT	2.4964	20.429418
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaT	3.0276	21.799445
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaT	2.9584	20.957274
7	Pilar Martín González	22	M	60.0	1.66	NaN	NaT	2.7556	21.773842
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaT	3.4225	21.913806
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaT	2.6244	20.957171
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaT	3.4969	22.305471
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaT	3.1329	19.470778
14	Carlos Rivas	28	H	89.0	1.78	245.0	NaT	NaN	NaN

In [ ]:
```
df[~(df.edad > 18)]
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.52851

In [ ]:
```
df.loc[df['nombre'] != 'Carlos Rivas']
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.528510
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	2001-05-20	2.9929	21.718066
2	Javier García Sánchez	24	H	NaN	1.81	191.0	1999-12-10	3.2761	NaN
3	Carmen López Pinzón	35	M	65.0	1.7	200.0	NaT	2.8900	22.491349
4	Marisa López Collado	46	M	51.0	1.58	148.0	NaT	2.4964	20.429418
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaT	3.0276	21.799445
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaT	2.9584	20.957274
7	Pilar Martín González	22	M	60.0	1.66	NaN	NaT	2.7556	21.773842
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaT	3.4225	21.913806
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaT	2.6244	20.957171
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaT	3.4969	22.305471
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaT	3.1329	19.470778

Eliminar filas de un dataframe

In [ ]:
```
#Drop: Elimina los renglones con los indices indcados
df.drop([1,3])
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.528510
2	Javier García Sánchez	24	H	NaN	1.81	191.0	1999-12-10	3.2761	NaN
4	Marisa López Collado	46	M	51.0	1.58	148.0	NaT	2.4964	20.429418
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaT	3.0276	21.799445
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaT	2.9584	20.957274
7	Pilar Martín González	22	M	60.0	1.66	NaN	NaT	2.7556	21.773842
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaT	3.4225	21.913806
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaT	2.6244	20.957171
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaT	3.4969	22.305471
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaT	3.1329	19.470778
14	Carlos Rivas	28	H	89.0	1.78	245.0	NaT	NaN	NaN

In [ ]:
```
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.528510
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	2001-05-20	2.9929	21.718066
2	Javier García Sánchez	24	H	NaN	1.81	191.0	1999-12-10	3.2761	NaN
3	Carmen López Pinzón	35	M	65.0	1.7	200.0	NaT	2.8900	22.491349
4	Marisa López Collado	46	M	51.0	1.58	148.0	NaT	2.4964	20.429418
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaT	3.0276	21.799445
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaT	2.9584	20.957274
7	Pilar Martín González	22	M	60.0	1.66	NaN	NaT	2.7556	21.773842
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaT	3.4225	21.913806
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaT	2.6244	20.957171
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaT	3.4969	22.305471
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaT	3.1329	19.470778
14	Carlos Rivas	28	H	89.0	1.78	245.0	NaT	NaN	NaN

In [ ]:
```
#Reasignarlo al dataframe
df = df.drop([1,3])
```

```
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.528510
2	Javier García Sánchez	24	H	NaN	1.81	191.0	1999-12-10	3.2761	NaN
4	Marisa López Collado	46	M	51.0	1.58	148.0	NaT	2.4964	20.429418
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	NaT	3.0276	21.799445
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	NaT	2.9584	20.957274
7	Pilar Martín González	22	M	60.0	1.66	NaN	NaT	2.7556	21.773842
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	NaT	3.7636	23.913275
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	NaT	3.4225	21.913806
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	NaT	2.6244	20.957171
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	NaT	3.4969	22.305471
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	NaT	3.9204	27.803285
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	NaT	3.1329	19.470778
14	Carlos Rivas	28	H	89.0	1.78	245.0	NaT	NaN	NaN

Eliminar filas que tienen algún dato desconocido

In [ ]:
```
#Se eliminarán a los renglones 2 y 7 que tienen NA
df=df.dropna()
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol	fecha_nac	altura2	imc
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	2000-03-05	3.2041	26.52851

Agrupación, Ordenamiento y Agregación de datos

In [ ]:
```
#Agrupación de datos

df.groupby('sexo').get_group('M')
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol
1	Rosa Díaz Díaz	32	M	65.0	173	232.0
3	Carmen López Pinzón	35	M	65.0	170	200.0
4	Marisa López Collado	46	M	51.0	158	148.0
7	Pilar Martín González	22	M	60.0	166	NaN
10	Macarena Álvarez Luna	53	M	55.0	162	262.0
13	Carolina Rubio Moreno	20	M	61.0	177	194.0

In [ ]:
```
dfh = df.groupby('sexo').get_group('H')
dfh
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol
0	José Luis Martínez Izquierdo	18	H	85.0	179	182.0
2	Javier García Sánchez	24	H	NaN	181	191.0
5	Antonio Ruiz Cruz	68	H	66.0	174	249.0
6	Antonio Fernández Ocaña	51	H	62.0	172	276.0
8	Pedro Gálvez Tenorio	35	H	90.0	194	241.0
9	Santiago Reillo Manzano	46	H	75.0	185	280.0
11	José María de la Guía Sanz	58	H	78.0	187	198.0
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	198	210.0

In [ ]:
```
dfh.sort_values('colesterol')
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol
0	José Luis Martínez Izquierdo	18	H	85.0	179	182.0
2	Javier García Sánchez	24	H	NaN	181	191.0
11	José María de la Guía Sanz	58	H	78.0	187	198.0
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	198	210.0
8	Pedro Gálvez Tenorio	35	H	90.0	194	241.0
5	Antonio Ruiz Cruz	68	H	66.0	174	249.0
6	Antonio Fernández Ocaña	51	H	62.0	172	276.0
9	Santiago Reillo Manzano	46	H	75.0	185	280.0

In [ ]:
```
#Obtener el peso mínimo
dfh['peso'].min()
```

Out[ ]:
```
62.0
```

In [ ]:
```
df.groupby('sexo').mean()
```

```
<ipython-input-52-bde78877453e>:1: FutureWarning: The default value of numeric_only
in DataFrameGroupBy.mean is deprecated. In a future version, numeric_only will defau
lt to False. Either specify numeric_only or select only columns which should be vali
d for the function.
  df.groupby('sexo').mean()
```

Out[ ]:

sexo	edad	peso	altura	colesterol
H	40.875000	80.714286	183.750000	228.375
M	34.666667	59.500000	167.666667	207.200

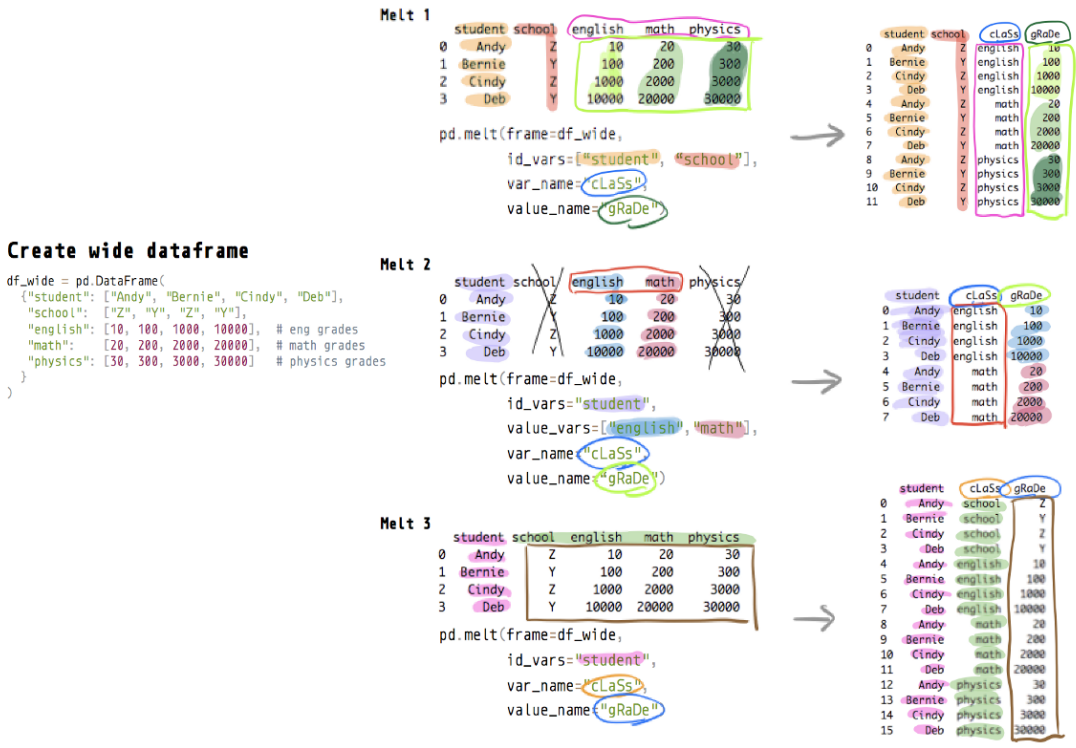
In [ ]:
```

```

Reformateo del dataframe

*melt: de ancho a largo (de columnas a filas) Convierte una dataframe a otro formato donde ciertas columnas se definen como id, y las otras columnas se consideran variables a medir, quitándolas del eje del renglón

Imagen de: https://towardsdatascience.com/reshape-pandas-dataframe-with-melt-in-python-tutorial-and-visualization-29ec1450bb02

*pivot: de largo a ancho (de filas a columnas)



In [ ]:
```
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol
0	José Luis Martínez Izquierdo	18	H	85.0	179	182.0
1	Rosa Díaz Díaz	32	M	65.0	173	232.0
2	Javier García Sánchez	24	H	NaN	181	191.0
3	Carmen López Pinzón	35	M	65.0	170	200.0
4	Marisa López Collado	46	M	51.0	158	148.0
5	Antonio Ruiz Cruz	68	H	66.0	174	249.0
6	Antonio Fernández Ocaña	51	H	62.0	172	276.0
7	Pilar Martín González	22	M	60.0	166	NaN
8	Pedro Gálvez Tenorio	35	H	90.0	194	241.0
9	Santiago Reillo Manzano	46	H	75.0	185	280.0
10	Macarena Álvarez Luna	53	M	55.0	162	262.0
11	José María de la Guía Sanz	58	H	78.0	187	198.0
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	198	210.0
13	Carolina Rubio Moreno	20	M	61.0	177	194.0

In [ ]:
```
df_reshape= df.melt(id_vars=['nombre','edad'])
df_reshape
```

Out[ ]:

	nombre	edad	variable	value
0	José Luis Martínez Izquierdo	18	sexo	H
1	Rosa Díaz Díaz	32	sexo	M
2	Javier García Sánchez	24	sexo	H
3	Carmen López Pinzón	35	sexo	M
4	Marisa López Collado	46	sexo	M
5	Antonio Ruiz Cruz	68	sexo	H
6	Antonio Fernández Ocaña	51	sexo	H
7	Pilar Martín González	22	sexo	M
8	Pedro Gálvez Tenorio	35	sexo	H
9	Santiago Reillo Manzano	46	sexo	H
10	Macarena Álvarez Luna	53	sexo	M
11	José María de la Guía Sanz	58	sexo	H
12	Miguel Angel Cuadrado Gutiérrez	27	sexo	H
13	Carolina Rubio Moreno	20	sexo	M
14	José Luis Martínez Izquierdo	18	peso	85.0
15	Rosa Díaz Díaz	32	peso	65.0
16	Javier García Sánchez	24	peso	NaN
17	Carmen López Pinzón	35	peso	65.0
18	Marisa López Collado	46	peso	51.0
19	Antonio Ruiz Cruz	68	peso	66.0
20	Antonio Fernández Ocaña	51	peso	62.0
21	Pilar Martín González	22	peso	60.0
22	Pedro Gálvez Tenorio	35	peso	90.0
23	Santiago Reillo Manzano	46	peso	75.0
24	Macarena Álvarez Luna	53	peso	55.0
25	José María de la Guía Sanz	58	peso	78.0
26	Miguel Angel Cuadrado Gutiérrez	27	peso	109.0
27	Carolina Rubio Moreno	20	peso	61.0
28	José Luis Martínez Izquierdo	18	altura	179
29	Rosa Díaz Díaz	32	altura	173

	nombre	edad	variable	value
30	Javier García Sánchez	24	altura	181
31	Carmen López Pinzón	35	altura	170
32	Marisa López Collado	46	altura	158
33	Antonio Ruiz Cruz	68	altura	174
34	Antonio Fernández Ocaña	51	altura	172
35	Pilar Martín González	22	altura	166
36	Pedro Gálvez Tenorio	35	altura	194
37	Santiago Reillo Manzano	46	altura	185
38	Macarena Álvarez Luna	53	altura	162
39	José María de la Guía Sanz	58	altura	187
40	Miguel Angel Cuadrado Gutiérrez	27	altura	198
41	Carolina Rubio Moreno	20	altura	177
42	José Luis Martínez Izquierdo	18	colesterol	182.0
43	Rosa Díaz Díaz	32	colesterol	232.0
44	Javier García Sánchez	24	colesterol	191.0
45	Carmen López Pinzón	35	colesterol	200.0
46	Marisa López Collado	46	colesterol	148.0
47	Antonio Ruiz Cruz	68	colesterol	249.0
48	Antonio Fernández Ocaña	51	colesterol	276.0
49	Pilar Martín González	22	colesterol	NaN
50	Pedro Gálvez Tenorio	35	colesterol	241.0
51	Santiago Reillo Manzano	46	colesterol	280.0
52	Macarena Álvarez Luna	53	colesterol	262.0
53	José María de la Guía Sanz	58	colesterol	198.0
54	Miguel Angel Cuadrado Gutiérrez	27	colesterol	210.0
55	Carolina Rubio Moreno	20	colesterol	194.0

In [ ]:
```
# unmelting using pivot()
# https://www.journaldev.com/33398/pandas-melt-unmelt-pivot-function

df_unmelted = df_reshape.pivot(index=['nombre', 'edad'], columns='variable')
df_unmelted = df_unmelted['value'].reset_index()
```

```
df_unmelted.columns.name = None
df_unmelted
```

Out[ ]:

	nombre	edad	altura	colesterol	peso	sexo
0	Antonio Fernández Ocaña	51	172	276.0	62.0	H
1	Antonio Ruiz Cruz	68	174	249.0	66.0	H
2	Carmen López Pinzón	35	170	200.0	65.0	M
3	Carolina Rubio Moreno	20	177	194.0	61.0	M
4	Javier García Sánchez	24	181	191.0	NaN	H
5	José Luis Martínez Izquierdo	18	179	182.0	85.0	H
6	José María de la Guía Sanz	58	187	198.0	78.0	H
7	Macarena Álvarez Luna	53	162	262.0	55.0	M
8	Marisa López Collado	46	158	148.0	51.0	M
9	Miguel Angel Cuadrado Gutiérrez	27	198	210.0	109.0	H
10	Pedro Gálvez Tenorio	35	194	241.0	90.0	H
11	Pilar Martín González	22	166	NaN	60.0	M
12	Rosa Díaz Díaz	32	173	232.0	65.0	M
13	Santiago Reillo Manzano	46	185	280.0	75.0	H

Combinar dataframes

*Concatenación: Combinación de varios DataFrames concatenando sus filas o columnas.
*Mezcla: Combinación de varios DataFrames usando columnas o índices comunes.

Concat

In [ ]:
```
df1 = pd.read_csv('/content/gdrive/MyDrive/Colab Notebooks/TC2004B.101/data/Cars1.c
df2 = pd.read_csv('/content/gdrive/MyDrive/Colab Notebooks/TC2004B.101/data/Cars2.c
```

In [ ]:
```
df1.head(20)
```

Out[ ]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	origin	
0	18.0	8	307	130	3504	12	70	1	che ch r
1	15.0	8	350	165	3693	11.5	70	1	skyla
2	18.0	8	318	150	3436	11	70	1	ply s
3	16.0	8	304	150	3433	12	70	1	amc
4	17.0	8	302	140	3449	10.5	70	1	ford
5	15.0	8	429	198	4341	10	70	1	ford g
6	14.0	8	454	220	4354	9	70	1	che i
7	14.0	8	440	215	4312	8.5	70	1	ply
8	14.0	8	455	225	4425	10	70	1	p c
9	15.0	8	390	190	3850	8.5	70	1	amba
10	15.0	8	383	170	3563	10	70	1	chal
11	14.0	8	340	160	3609	8	70	1	ply 'cu
12	15.0	8	400	150	3761	9.5	70	1	che mont
13	14.0	8	455	225	3086	10	70	1	buick wagc
14	24.0	4	113	95	2372	15	70	3	r
15	22.0	6	198	95	2833	15.5	70	1	ply
16	18.0	6	199	97	2774	15.5	70	1	amc
17	21.0	6	200	85	2587	16	70	1	ma
18	27.0	4	97	88	2130	14.5	70	3	

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	origin	
19	26.0	4	97	46	1835	20.5	70	2	volks 1131

In [ ]:
```
df2.head()
```

Out[ ]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	origin	ca
0	33.0	4	91	53	1795	17.4	76	3	hond civi
1	20.0	6	225	100	3651	17.7	76	1	dodg aspe s
2	18.0	6	250	78	3574	21	76	1	for granad ghi
3	18.5	6	250	110	3645	16.2	76	1	pontia ventur
4	17.5	6	258	95	3193	17.8	76	1	am pace d,

In [ ]:
```
del(df1['data1'])
del(df1['data2'])
df1
```

Out[ ]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	origin	
0	18.0	8	307	130	3504	12	70	1	chev che m
1	15.0	8	350	165	3693	11.5	70	1	sk
2	18.0	8	318	150	3436	11	70	1	plym sat
3	16.0	8	304	150	3433	12	70	1	reb
4	17.0	8	302	140	3449	10.5	70	1	t
...	...	...	...	...	...	...	...	...	
193	24.0	6	200	81	3012	17.6	76	1	mav
194	22.5	6	232	90	3085	17.6	76	1	h
195	29.0	4	85	52	2035	22.2	76	1	chev che
196	24.5	4	98	60	2164	22.1	76	1	chev w
197	29.0	4	90	70	1937	14.2	76	2	vw r

198 rows × 9 columns

In [ ]:
```
print(df1.shape)
print(df2.shape)
```

```
(198, 9)
(200, 9)
```

In [ ]:
```
total_cars = pd.concat([df1,df2])
total_cars
```

Out[ ]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	origin	
0	18.0	8	307	130	3504	12	70	1	chev che m
1	15.0	8	350	165	3693	11.5	70	1	sk
2	18.0	8	318	150	3436	11	70	1	plym sat
3	16.0	8	304	150	3433	12	70	1	reb
4	17.0	8	302	140	3449	10.5	70	1	t
...	...	...	...	...	...	...	...	...	
195	27.0	4	140	86	2790	15.6	82	1	mus
196	44.0	4	97	52	2130	24.6	82	2	p
197	32.0	4	135	84	2295	11.6	82	1	d ram
198	28.0	4	120	79	2625	18.6	82	1	ra
199	31.0	4	119	82	2720	19.4	82	1	che

398 rows × 9 columns

Merge - (join)

In this exercise, we'll merge the details of students from two datasets, namely student.csv and marks.csv. The student dataset contains columns such as Age, Gender, Grade, and Employed. The marks.csv dataset contains columns such as Mark and City. The Student_id column is common between the two datasets. Follow these steps to complete this exercise.
Reference: Data Science with Python By Rohan Chopra, Aaron England, Mohamed Noordeen Alaudeen July 2019
https://subscription.packtpub.com/book/data/9781838552862/1/ch01lvl1sec06/data-integration

In [ ]:
```
df1 = pd.read_csv('/content/gdrive/MyDrive/Colab Notebooks/TC2004B.101/data/mark.cs
```

```
df2 = pd.read_csv('/content/gdrive/MyDrive/Colab Notebooks/TC2004B.101/data/student
```

In [ ]:
```
df1.head()
```

Out[ ]:

	Student_id	Mark	City
0	1	95	Chennai
1	2	70	Delhi
2	3	98	Mumbai
3	4	75	Pune
4	5	89	Kochi

In [ ]:
```
df2.head()
```

Out[ ]:

	Student_id	Age	Gender	Grade	Employed
0	1	19	Male	1st Class	yes
1	2	20	Female	2nd Class	no
2	3	18	Male	1st Class	no
3	4	21	Female	2nd Class	no
4	5	19	Male	1st Class	no

In [ ]:
```
df_completo = pd.merge(df1, df2, on = 'Student_id')
df_completo.head()
```

Out[ ]:

	Student_id	Mark	City	Age	Gender	Grade	Employed
0	1	95	Chennai	19	Male	1st Class	yes
1	2	70	Delhi	20	Female	2nd Class	no
2	3	98	Mumbai	18	Male	1st Class	no
3	4	75	Pune	21	Female	2nd Class	no
4	5	89	Kochi	19	Male	1st Class	no

Ejemplos:

1. Identificar que variables son números y crear un DataFrame temporal en donde solo se tengan datos de ese tipo.
2. Usando los datos de 1, crear dos subconjuntos basados en género y muestra sus respectivas medias de colesterol.

In [ ]:
```
df = pd.read_csv('/content/gdrive/MyDrive/Colab Notebooks/TC2004B.101/data/colester
```

In [ ]:
```
df
```

Out[ ]:

	nombre	edad	sexo	peso	altura	colesterol
0	José Luis Martínez Izquierdo	18	H	85.0	179	182.0
1	Rosa Díaz Díaz	32	M	65.0	173	232.0
2	Javier García Sánchez	24	H	NaN	181	191.0
3	Carmen López Pinzón	35	M	65.0	170	200.0
4	Marisa López Collado	46	M	51.0	158	148.0
5	Antonio Ruiz Cruz	68	H	66.0	174	249.0
6	Antonio Fernández Ocaña	51	H	62.0	172	276.0
7	Pilar Martín González	22	M	60.0	166	NaN
8	Pedro Gálvez Tenorio	35	H	90.0	194	241.0
9	Santiago Reillo Manzano	46	H	75.0	185	280.0
10	Macarena Álvarez Luna	53	M	55.0	162	262.0
11	José María de la Guía Sanz	58	H	78.0	187	198.0
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	198	210.0
13	Carolina Rubio Moreno	20	M	61.0	177	194.0

In [ ]:
```
# Variables numericas unicamente
temp = df.select_dtypes(include=['float64', 'int64'])

# ahora agrupamos por sexo utilizando df.sexo
temp = temp.groupby(df.sexo)
# mostramos la estadistica de medias
temp.mean()
```

Out[ ]:

sexo	edad	peso	altura	colesterol
H	40.875000	80.714286	183.750000	228.375
M	34.666667	59.500000	167.666667	207.200

Función	Ejemplo
Lectura de Datos	`pd.read_csv('/dir/file.csv')`
Descripción General	`df.head()` , `df.shape` , `df.size` , `df.info()`

Función	Ejemplo
Estadísticas Descriptivas	`df.describe()`, `df.mode()`, `df['col'].value_counts()`
Limpieza de Datos	`df.dropna()`, `df.drop_duplicates()`
Acceso a elementos	`df.iloc[1, 3]`, `df.loc[2, 'colesterol']`
Operaciones con columnas	`df['new_col'] = valores`, `df.drop(columns=['col'])`
Operaciones con filas/renglones	`df.append(row)`, `df[df['condition']]`, `df.drop(index)`
Agrupación, Ordenamiento y Agregación de datos	`df.groupby('col').mean()`, `df.sort_values('col')`
Reformateo del dataframe	`df.melt(id_vars=['id'])`, `df.pivot(index, columns)`
Combinar dataframes	`pd.concat([df1, df2])`, `pd.merge(df1, df2, on='key')`

In [ ]: