DataFramesOperations

February 15, 2024

Ricardo Kaleb Flores Alfonso, A01198716, IDM

1 Operaciones con dataframes -

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

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

#Lectura de Datos

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

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

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call
drive.mount("/content/gdrive", force_remount=True).
```

[ ]:
```
df = pd.read_csv('/content/gdrive/MyDrive/Pandas/Act 1/colesterol.csv')
original_df=df
```

#Descripción General

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

[ ]:

	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

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

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

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

[ ]:
```
84
```

[ ]:
```
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
```

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

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

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

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

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

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

#Acceso a elementos

1.1 Acceso por posición

[ ]:
```
df
```

[ ]:

	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
```

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

[ ]:
```
65.0
```

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

[ ]:
```
                        nombre  peso
0  José Luis Martínez Izquierdo  85.0
1               Rosa Díaz Díaz  65.0
```

##Acceso por nombre

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

[ ]:
```
191.0
```

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

[ ]:
```
                        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
```

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

[ ]:
```
            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
```

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

[ ]:

	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

1.2 Agregar columnas al data frame

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

[ ]:

	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	

	diabetes
0	False
1	False
2	True
3	False
4	True
5	NaN
6	NaN

```
7       NaN
8       NaN
9       NaN
10      NaN
11      NaN
12      NaN
13      NaN
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14 entries, 0 to 13
Data columns (total 7 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
dtypes: float64(2), int64(2), object(3)
memory usage: 912.0+ bytes
```

1.3 Cambiar tipo de dato de columna a datetime

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

[ ]:

	nombre	edad	sexo	peso	altura	colesterol	diabetes
0	José Luis Martínez Izquierdo	18	H	85.0	179	182.0	False
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	

```
1     False
2      True
3     False
4      True
5       NaN
6       NaN
7       NaN
8       NaN
9       NaN
10      NaN
11      NaN
12      NaN
13      NaN
```

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

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

1.4 Operación sobre una columna

Dividir la columna entre un valor

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

[ ]:
```
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
```

[ ]:
```
df
```

[ ]:

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

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

[ ]:

	nombre	edad	sexo	peso	altura	colesterol
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0
2	Javier García Sánchez	24	H	NaN	1.81	191.0
3	Carmen López Pinzón	35	M	65.0	1.70	200.0
4	Marisa López Collado	46	M	51.0	1.58	148.0
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0
7	Pilar Martín González	22	M	60.0	1.66	NaN
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0

```
9           Santiago Reillo Manzano    46    H   75.0    1.85       280.0
10            Macarena Álvarez Luna    53    M   55.0    1.62       262.0
11       José María de la Guía Sanz    58    H   78.0    1.87       198.0
12  Miguel Angel Cuadrado Gutiérrez    27    H  109.0    1.98       210.0
13            Carolina Rubio Moreno    20    M   61.0    1.77       194.0

   diabetes
0     False
1     False
2      True
3     False
4      True
5       NaN
6       NaN
7       NaN
8       NaN
9       NaN
10      NaN
11      NaN
12      NaN
13      NaN
```

##Aplicar funciones a una columna

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

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

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

   diabetes  altura2        imc
0     False   3.2041  26.528510
1     False   2.9929  21.718066
```

```
2      True   3.2761        NaN
3     False   2.8900  22.491349
4      True   2.4964  20.429418
5       NaN   3.0276  21.799445
6       NaN   2.9584  20.957274
7       NaN   2.7556  21.773842
8       NaN   3.7636  23.913275
9       NaN   3.4225  21.913806
10      NaN   2.6244  20.957171
11      NaN   3.4969  22.305471
12      NaN   3.9204  27.803285
13      NaN   3.1329  19.470778
```

1.5 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)

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

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

   diabetes_mellitus  altura2        imc
0              False   3.2041  26.528510
1              False   2.9929  21.718066
2               True   3.2761        NaN
3              False   2.8900  22.491349
4               True   2.4964  20.429418
5                NaN   3.0276  21.799445
6                NaN   2.9584  20.957274
```

```
7                NaN   2.7556  21.773842
8                NaN   3.7636  23.913275
9                NaN   3.4225  21.913806
10               NaN   2.6244  20.957171
11               NaN   3.4969  22.305471
12               NaN   3.9204  27.803285
13               NaN   3.1329  19.470778
```

1.6 Seleccionar ciertas columnas de un dataframe

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

[ ]:
```
                          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
```

1.7 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.

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

[ ]:
```
                          nombre  edad sexo   peso  altura  colesterol  \
0      José Luis Martínez Izquierdo    18    H   85.0    1.79       182.0
1                  Rosa Díaz Díaz    32    M   65.0    1.73       232.0
2            Javier García Sánchez    24    H    NaN    1.81       191.0
3              Carmen López Pinzón    35    M   65.0    1.70       200.0
4             Marisa López Collado    46    M   51.0    1.58       148.0
5                Antonio Ruiz Cruz    68    H   66.0    1.74       249.0
6          Antonio Fernández Ocaña    51    H   62.0    1.72       276.0
7            Pilar Martín González    22    M   60.0    1.66         NaN
```

```
8              Pedro Gálvez Tenorio    35    H   90.0    1.94       241.0
9           Santiago Reillo Manzano    46    H   75.0    1.85       280.0
10            Macarena Álvarez Luna    53    M   55.0    1.62       262.0
11       José María de la Guía Sanz    58    H   78.0    1.87       198.0
12  Miguel Angel Cuadrado Gutiérrez    27    H  109.0    1.98       210.0
13            Carolina Rubio Moreno    20    M   61.0    1.77       194.0
```

```
    altura2        imc
0    3.2041  26.528510
1    2.9929  21.718066
2    3.2761        NaN
3    2.8900  22.491349
4    2.4964  20.429418
5    3.0276  21.799445
6    2.9584  20.957274
7    2.7556  21.773842
8    3.7636  23.913275
9    3.4225  21.913806
10   2.6244  20.957171
11   3.4969  22.305471
12   3.9204  27.803285
13   3.1329  19.470778
```

#Operaciones con Filas/Renglones

1.8 Añadir una fila a un dataframe

[ ]:
```
#df.append(pd.Series(['Carlos Rivas', 28, 'H', 89.0, 1.78, 245.0],␣
 ↪index=['nombre','edad','sexo','peso','altura','colesterol']),␣
 ↪ignore_index=True)
#Append deprecado, usar concat.
s2 = pd.Series(['Carlos Rivas', 28, 'H', 89.0, 1.78, 245.0],␣
 ↪index=['nombre','edad','peso','sexo','colesterol','altura'])
s2
```

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

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

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

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

[ ]:

	nombre	edad	sexo	peso	altura	colesterol	\
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	
2	Javier García Sánchez	24	H	NaN	1.81	191.0	
3	Carmen López Pinzón	35	M	65.0	1.7	200.0	
4	Marisa López Collado	46	M	51.0	1.58	148.0	
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	
7	Pilar Martín González	22	M	60.0	1.66	NaN	
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0	
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0	
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0	
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0	
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0	
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0	
14	Carlos Rivas	28	89.0	H	245.0	1.78	

	altura2	imc
0	3.2041	26.528510
1	2.9929	21.718066
2	3.2761	NaN
3	2.8900	22.491349
4	2.4964	20.429418
5	3.0276	21.799445
6	2.9584	20.957274
7	2.7556	21.773842
8	3.7636	23.913275
9	3.4225	21.913806
10	2.6244	20.957171
11	3.4969	22.305471
12	3.9204	27.803285
13	3.1329	19.470778
14	NaN	NaN

[ ]:
```
df
```

[ ]:

	nombre	edad	sexo	peso	altura	colesterol	\
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0	
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0	
2	Javier García Sánchez	24	H	NaN	1.81	191.0	
3	Carmen López Pinzón	35	M	65.0	1.70	200.0	
4	Marisa López Collado	46	M	51.0	1.58	148.0	
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0	
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0	
7	Pilar Martín González	22	M	60.0	1.66	NaN	

```
8              Pedro Gálvez Tenorio    35    H   90.0    1.94       241.0
9           Santiago Reillo Manzano    46    H   75.0    1.85       280.0
10            Macarena Álvarez Luna    53    M   55.0    1.62       262.0
11       José María de la Guía Sanz    58    H   78.0    1.87       198.0
12  Miguel Angel Cuadrado Gutiérrez    27    H  109.0    1.98       210.0
13            Carolina Rubio Moreno    20    M   61.0    1.77       194.0

    altura2        imc
0    3.2041  26.528510
1    2.9929  21.718066
2    3.2761        NaN
3    2.8900  22.491349
4    2.4964  20.429418
5    3.0276  21.799445
6    2.9584  20.957274
7    2.7556  21.773842
8    3.7636  23.913275
9    3.4225  21.913806
10   2.6244  20.957171
11   3.4969  22.305471
12   3.9204  27.803285
13   3.1329  19.470778
```

Es necesario guardarlo en el dataframe

[ ]:
```
#df=df.append(pd.Series(['Carlos Rivas', 28, 'H', 89.0, 1.78, 245.0],␣
 ↪index=['nombre','edad','sexo','peso','altura','colesterol']),␣
 ↪ignore_index=True)
s2 = pd.Series(['Carlos Rivas', 28, 'H', 89.0, 1.78, 245.0],␣
 ↪index=['nombre','edad','sexo','peso','altura','colesterol'])
df = pd.concat([df,s2.to_frame().T], ignore_index=True)
df.tail()
```

[ ]:
```
                          nombre edad sexo   peso altura colesterol  \
10            Macarena Álvarez Luna   53    M   55.0   1.62      262.0
11       José María de la Guía Sanz   58    H   78.0   1.87      198.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0
13            Carolina Rubio Moreno   20    M   61.0   1.77      194.0
14                     Carlos Rivas   28    H   89.0   1.78      245.0

    altura2        imc
10   2.6244  20.957171
11   3.4969  22.305471
12   3.9204  27.803285
13   3.1329  19.470778
14      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'

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

[ ]:
```
                            nombre edad sexo   peso altura colesterol  \
0      José Luis Martínez Izquierdo   18    H   85.0   1.79      182.0
8              Pedro Gálvez Tenorio   35    H   90.0   1.94      241.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0
14                     Carlos Rivas   28    H   89.0   1.78      245.0

    altura2        imc
0    3.2041  26.528510
8    3.7636  23.913275
12   3.9204  27.803285
14      NaN        NaN
```

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

[ ]:
```
                            nombre edad sexo   peso altura colesterol  \
0      José Luis Martínez Izquierdo   18    H   85.0   1.79      182.0
8              Pedro Gálvez Tenorio   35    H   90.0   1.94      241.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0
14                     Carlos Rivas   28    H   89.0   1.78      245.0

    altura2        imc
0    3.2041  26.528510
8    3.7636  23.913275
12   3.9204  27.803285
14      NaN        NaN
```

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

[ ]:
```
                            nombre edad sexo   peso altura colesterol  \
8              Pedro Gálvez Tenorio   35    H   90.0   1.94      241.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0
14                     Carlos Rivas   28    H   89.0   1.78      245.0

    altura2        imc
8    3.7636  23.913275
12   3.9204  27.803285
14      NaN        NaN
```

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

[ ]:
```
                            nombre edad sexo   peso altura colesterol  \
0      José Luis Martínez Izquierdo   18    H   85.0   1.79      182.0
1                    Rosa Díaz Díaz   32    M   65.0   1.73      232.0
```

```
5                 Antonio Ruiz Cruz   68    H   66.0   1.74      249.0
6           Antonio Fernández Ocaña   51    H   62.0   1.72      276.0
8              Pedro Gálvez Tenorio   35    H   90.0   1.94      241.0
9           Santiago Reillo Manzano   46    H   75.0   1.85      280.0
10            Macarena Álvarez Luna   53    M   55.0   1.62      262.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0
14                     Carlos Rivas   28    H   89.0   1.78      245.0

    altura2        imc
0    3.2041  26.528510
1    2.9929  21.718066
5    3.0276  21.799445
6    2.9584  20.957274
8    3.7636  23.913275
9    3.4225  21.913806
10   2.6244  20.957171
12   3.9204  27.803285
14      NaN        NaN
```

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

[ ]:
```
                          nombre edad sexo   peso altura colesterol  \
1                  Rosa Díaz Díaz   32    M   65.0   1.73      232.0
2             Javier García Sánchez   24    H    NaN   1.81      191.0
3               Carmen López Pinzón   35    M   65.0    1.7      200.0
4              Marisa López Collado   46    M   51.0   1.58      148.0
5                 Antonio Ruiz Cruz   68    H   66.0   1.74      249.0
6           Antonio Fernández Ocaña   51    H   62.0   1.72      276.0
7             Pilar Martín González   22    M   60.0   1.66        NaN
8              Pedro Gálvez Tenorio   35    H   90.0   1.94      241.0
9           Santiago Reillo Manzano   46    H   75.0   1.85      280.0
10            Macarena Álvarez Luna   53    M   55.0   1.62      262.0
11      José María de la Guía Sanz   58    H   78.0   1.87      198.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0
13            Carolina Rubio Moreno   20    M   61.0   1.77      194.0
14                     Carlos Rivas   28    H   89.0   1.78      245.0

    altura2        imc
1    2.9929  21.718066
2    3.2761        NaN
3    2.8900  22.491349
4    2.4964  20.429418
5    3.0276  21.799445
6    2.9584  20.957274
7    2.7556  21.773842
8    3.7636  23.913275
9    3.4225  21.913806
```

```
10   2.6244  20.957171
11   3.4969  22.305471
12   3.9204  27.803285
13   3.1329  19.470778
14      NaN        NaN
```

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

[ ]:
```
                        nombre edad sexo  peso altura colesterol  altura2  \
0  José Luis Martínez Izquierdo   18    H  85.0   1.79      182.0   3.2041

        imc
0  26.52851
```

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

[ ]:
```
                            nombre edad sexo   peso altura colesterol  \
0      José Luis Martínez Izquierdo   18    H   85.0   1.79      182.0
1                   Rosa Díaz Díaz   32    M   65.0   1.73      232.0
2            Javier García Sánchez   24    H    NaN   1.81      191.0
3              Carmen López Pinzón   35    M   65.0    1.7      200.0
4             Marisa López Collado   46    M   51.0   1.58      148.0
5                Antonio Ruiz Cruz   68    H   66.0   1.74      249.0
6          Antonio Fernández Ocaña   51    H   62.0   1.72      276.0
7            Pilar Martín González   22    M   60.0   1.66        NaN
8             Pedro Gálvez Tenorio   35    H   90.0   1.94      241.0
9          Santiago Reillo Manzano   46    H   75.0   1.85      280.0
10           Macarena Álvarez Luna   53    M   55.0   1.62      262.0
11      José María de la Guía Sanz   58    H   78.0   1.87      198.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0
13           Carolina Rubio Moreno   20    M   61.0   1.77      194.0

    altura2        imc
0    3.2041  26.528510
1    2.9929  21.718066
2    3.2761        NaN
3    2.8900  22.491349
4    2.4964  20.429418
5    3.0276  21.799445
6    2.9584  20.957274
7    2.7556  21.773842
8    3.7636  23.913275
9    3.4225  21.913806
10   2.6244  20.957171
11   3.4969  22.305471
12   3.9204  27.803285
13   3.1329  19.470778
```

1.9 Eliminar filas de un dataframe

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

[ ]:

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

[ ]:
```
df
```

[ ]:

	nombre	edad	sexo	peso	altura	colesterol
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0
1	Rosa Díaz Díaz	32	M	65.0	1.73	232.0
2	Javier García Sánchez	24	H	NaN	1.81	191.0
3	Carmen López Pinzón	35	M	65.0	1.7	200.0
4	Marisa López Collado	46	M	51.0	1.58	148.0
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0
7	Pilar Martín González	22	M	60.0	1.66	NaN
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0

```
10            Macarena Álvarez Luna   53    M   55.0   1.62      262.0
11       José María de la Guía Sanz   58    H   78.0   1.87      198.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0
13            Carolina Rubio Moreno   20    M   61.0   1.77      194.0
14                    Carlos Rivas   28    H   89.0   1.78      245.0

    altura2        imc
0    3.2041  26.528510
1    2.9929  21.718066
2    3.2761        NaN
3    2.8900  22.491349
4    2.4964  20.429418
5    3.0276  21.799445
6    2.9584  20.957274
7    2.7556  21.773842
8    3.7636  23.913275
9    3.4225  21.913806
10   2.6244  20.957171
11   3.4969  22.305471
12   3.9204  27.803285
13   3.1329  19.470778
14      NaN        NaN
```

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

[ ]:
```
                          nombre edad sexo   peso altura colesterol  \
0      José Luis Martínez Izquierdo   18    H   85.0   1.79      182.0
2             Javier García Sánchez   24    H    NaN   1.81      191.0
4              Marisa López Collado   46    M   51.0   1.58      148.0
5                 Antonio Ruiz Cruz   68    H   66.0   1.74      249.0
6           Antonio Fernández Ocaña   51    H   62.0   1.72      276.0
7             Pilar Martín González   22    M   60.0   1.66        NaN
8              Pedro Gálvez Tenorio   35    H   90.0   1.94      241.0
9           Santiago Reillo Manzano   46    H   75.0   1.85      280.0
10            Macarena Álvarez Luna   53    M   55.0   1.62      262.0
11       José María de la Guía Sanz   58    H   78.0   1.87      198.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0
13            Carolina Rubio Moreno   20    M   61.0   1.77      194.0
14                    Carlos Rivas   28    H   89.0   1.78      245.0

    altura2        imc
0    3.2041  26.528510
2    3.2761        NaN
4    2.4964  20.429418
5    3.0276  21.799445
```

```
6    2.9584  20.957274
7    2.7556  21.773842
8    3.7636  23.913275
9    3.4225  21.913806
10   2.6244  20.957171
11   3.4969  22.305471
12   3.9204  27.803285
13   3.1329  19.470778
14      NaN        NaN
```

1.10 Eliminar filas que tienen algún dato desconocido

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

[ ]:
```
                          nombre edad sexo   peso altura colesterol  \
0      José Luis Martínez Izquierdo   18    H   85.0   1.79      182.0
4              Marisa López Collado   46    M   51.0   1.58      148.0
5                 Antonio Ruiz Cruz   68    H   66.0   1.74      249.0
6           Antonio Fernández Ocaña   51    H   62.0   1.72      276.0
8              Pedro Gálvez Tenorio   35    H   90.0   1.94      241.0
9           Santiago Reillo Manzano   46    H   75.0   1.85      280.0
10            Macarena Álvarez Luna   53    M   55.0   1.62      262.0
11       José María de la Guía Sanz   58    H   78.0   1.87      198.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0
13            Carolina Rubio Moreno   20    M   61.0   1.77      194.0

    altura2        imc
0    3.2041  26.528510
4    2.4964  20.429418
5    3.0276  21.799445
6    2.9584  20.957274
8    3.7636  23.913275
9    3.4225  21.913806
10   2.6244  20.957171
11   3.4969  22.305471
12   3.9204  27.803285
13   3.1329  19.470778
```

#Agrupación, Ordenamiento y Agregación de datos

[ ]:
```
#Agrupación de datos
df.groupby('sexo').get_group('M')
```

[ ]:
```
                  nombre edad sexo  peso altura colesterol  altura2  \
4    Marisa López Collado   46    M  51.0   1.58      148.0   2.4964
10  Macarena Álvarez Luna   53    M  55.0   1.62      262.0   2.6244
```

```
13  Carolina Rubio Moreno   20    M  61.0   1.77      194.0   3.1329

          imc
4   20.429418
10  20.957171
13  19.470778
```

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

[ ]:
```
                            nombre edad sexo   peso altura colesterol  \
0      José Luis Martínez Izquierdo   18    H   85.0   1.79      182.0
5                 Antonio Ruiz Cruz   68    H   66.0   1.74      249.0
6            Antonio Fernández Ocaña   51    H   62.0   1.72      276.0
8               Pedro Gálvez Tenorio   35    H   90.0   1.94      241.0
9            Santiago Reillo Manzano   46    H   75.0   1.85      280.0
11        José María de la Guía Sanz   58    H   78.0   1.87      198.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0

    altura2        imc
0    3.2041  26.528510
5    3.0276  21.799445
6    2.9584  20.957274
8    3.7636  23.913275
9    3.4225  21.913806
11   3.4969  22.305471
12   3.9204  27.803285
```

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

[ ]:
```
                            nombre edad sexo   peso altura colesterol  \
0      José Luis Martínez Izquierdo   18    H   85.0   1.79      182.0
11        José María de la Guía Sanz   58    H   78.0   1.87      198.0
12  Miguel Angel Cuadrado Gutiérrez   27    H  109.0   1.98      210.0
8               Pedro Gálvez Tenorio   35    H   90.0   1.94      241.0
5                 Antonio Ruiz Cruz   68    H   66.0   1.74      249.0
6            Antonio Fernández Ocaña   51    H   62.0   1.72      276.0
9            Santiago Reillo Manzano   46    H   75.0   1.85      280.0

    altura2        imc
0    3.2041  26.528510
11   3.4969  22.305471
12   3.9204  27.803285
8    3.7636  23.913275
5    3.0276  21.799445
6    2.9584  20.957274
9    3.4225  21.913806
```

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

[ ]:
```
62.0
```

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

```
<ipython-input-427-bde78877453e>:1: FutureWarning: The default value of
numeric_only in DataFrameGroupBy.mean is deprecated. In a future version,
numeric_only will default to False. Either specify numeric_only or select only
columns which should be valid for the function.
  df.groupby('sexo').mean()
```

[ ]:

sexo	altura2	imc
H	3.399071	23.603009
M	2.751233	20.285789

[ ]:
```

```

2 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)

[ ]:
```
df
```

[ ]:

	nombre	edad	sexo	peso	altura	colesterol
0	José Luis Martínez Izquierdo	18	H	85.0	1.79	182.0
4	Marisa López Collado	46	M	51.0	1.58	148.0
5	Antonio Ruiz Cruz	68	H	66.0	1.74	249.0
6	Antonio Fernández Ocaña	51	H	62.0	1.72	276.0
8	Pedro Gálvez Tenorio	35	H	90.0	1.94	241.0
9	Santiago Reillo Manzano	46	H	75.0	1.85	280.0
10	Macarena Álvarez Luna	53	M	55.0	1.62	262.0
11	José María de la Guía Sanz	58	H	78.0	1.87	198.0
12	Miguel Angel Cuadrado Gutiérrez	27	H	109.0	1.98	210.0
13	Carolina Rubio Moreno	20	M	61.0	1.77	194.0

	altura2	imc
0	3.2041	26.528510

```
4    2.4964  20.429418
5    3.0276  21.799445
6    2.9584  20.957274
8    3.7636  23.913275
9    3.4225  21.913806
10   2.6244  20.957171
11   3.4969  22.305471
12   3.9204  27.803285
13   3.1329  19.470778
```

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

[ ]:

	nombre	edad	variable	value
0	José Luis Martínez Izquierdo	18	sexo	H
1	Marisa López Collado	46	sexo	M
2	Antonio Ruiz Cruz	68	sexo	H
3	Antonio Fernández Ocaña	51	sexo	H
4	Pedro Gálvez Tenorio	35	sexo	H
5	Santiago Reillo Manzano	46	sexo	H
6	Macarena Álvarez Luna	53	sexo	M
7	José María de la Guía Sanz	58	sexo	H
8	Miguel Angel Cuadrado Gutiérrez	27	sexo	H
9	Carolina Rubio Moreno	20	sexo	M
10	José Luis Martínez Izquierdo	18	peso	85.0
11	Marisa López Collado	46	peso	51.0
12	Antonio Ruiz Cruz	68	peso	66.0
13	Antonio Fernández Ocaña	51	peso	62.0
14	Pedro Gálvez Tenorio	35	peso	90.0
15	Santiago Reillo Manzano	46	peso	75.0
16	Macarena Álvarez Luna	53	peso	55.0
17	José María de la Guía Sanz	58	peso	78.0
18	Miguel Angel Cuadrado Gutiérrez	27	peso	109.0
19	Carolina Rubio Moreno	20	peso	61.0
20	José Luis Martínez Izquierdo	18	altura	1.79
21	Marisa López Collado	46	altura	1.58
22	Antonio Ruiz Cruz	68	altura	1.74
23	Antonio Fernández Ocaña	51	altura	1.72
24	Pedro Gálvez Tenorio	35	altura	1.94
25	Santiago Reillo Manzano	46	altura	1.85
26	Macarena Álvarez Luna	53	altura	1.62
27	José María de la Guía Sanz	58	altura	1.87
28	Miguel Angel Cuadrado Gutiérrez	27	altura	1.98
29	Carolina Rubio Moreno	20	altura	1.77
30	José Luis Martínez Izquierdo	18	colesterol	182.0
31	Marisa López Collado	46	colesterol	148.0
32	Antonio Ruiz Cruz	68	colesterol	249.0

```
33          Antonio Fernández Ocaña   51  colesterol      276.0
34             Pedro Gálvez Tenorio   35  colesterol      241.0
35          Santiago Reillo Manzano   46  colesterol      280.0
36            Macarena Álvarez Luna   53  colesterol      262.0
37       José María de la Guía Sanz   58  colesterol      198.0
38  Miguel Angel Cuadrado Gutiérrez   27  colesterol      210.0
39            Carolina Rubio Moreno   20  colesterol      194.0
40     José Luis Martínez Izquierdo   18     altura2     3.2041
41             Marisa López Collado   46     altura2     2.4964
42                Antonio Ruiz Cruz   68     altura2     3.0276
43          Antonio Fernández Ocaña   51     altura2     2.9584
44             Pedro Gálvez Tenorio   35     altura2     3.7636
45          Santiago Reillo Manzano   46     altura2     3.4225
46            Macarena Álvarez Luna   53     altura2     2.6244
47       José María de la Guía Sanz   58     altura2     3.4969
48  Miguel Angel Cuadrado Gutiérrez   27     altura2     3.9204
49            Carolina Rubio Moreno   20     altura2     3.1329
50     José Luis Martínez Izquierdo   18         imc   26.52851
51             Marisa López Collado   46         imc  20.429418
52                Antonio Ruiz Cruz   68         imc  21.799445
53          Antonio Fernández Ocaña   51         imc  20.957274
54             Pedro Gálvez Tenorio   35         imc  23.913275
55          Santiago Reillo Manzano   46         imc  21.913806
56            Macarena Álvarez Luna   53         imc  20.957171
57       José María de la Guía Sanz   58         imc  22.305471
58  Miguel Angel Cuadrado Gutiérrez   27         imc  27.803285
59            Carolina Rubio Moreno   20         imc  19.470778
```

[ ]:
```python
# 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
```

[ ]:
```
                          nombre  edad altura altura2 colesterol        imc  \
0          Antonio Fernández Ocaña    51   1.72  2.9584      276.0  20.957274
1                Antonio Ruiz Cruz    68   1.74  3.0276      249.0  21.799445
2            Carolina Rubio Moreno    20   1.77  3.1329      194.0  19.470778
3     José Luis Martínez Izquierdo    18   1.79  3.2041      182.0   26.52851
4       José María de la Guía Sanz    58   1.87  3.4969      198.0  22.305471
5            Macarena Álvarez Luna    53   1.62  2.6244      262.0  20.957171
6              Marisa López Collado    46   1.58  2.4964      148.0  20.429418
7  Miguel Angel Cuadrado Gutiérrez    27   1.98  3.9204      210.0  27.803285
8              Pedro Gálvez Tenorio    35   1.94  3.7636      241.0  23.913275
9          Santiago Reillo Manzano    46   1.85  3.4225      280.0  21.913806
```

```
    peso sexo
0   62.0    H
1   66.0    H
2   61.0    M
3   85.0    H
4   78.0    H
5   55.0    M
6   51.0    M
7  109.0    H
8   90.0    H
9   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

[ ]:
```
df1 = pd.read_csv('/content/gdrive/MyDrive/Pandas/Act 1/Cars1.csv')
df2 = pd.read_csv('/content/gdrive/MyDrive/Pandas/Act 1/Cars2.csv')
```

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

[ ]:
```
     mpg  cylinders  displacement horsepower  weight acceleration  model  \
0   18.0          8           307        130    3504           12     70
1   15.0          8           350        165    3693         11.5     70
2   18.0          8           318        150    3436           11     70
3   16.0          8           304        150    3433           12     70
4   17.0          8           302        140    3449         10.5     70
5   15.0          8           429        198    4341           10     70
6   14.0          8           454        220    4354            9     70
7   14.0          8           440        215    4312          8.5     70
8   14.0          8           455        225    4425           10     70
9   15.0          8           390        190    3850          8.5     70
10  15.0          8           383        170    3563           10     70
11  14.0          8           340        160    3609            8     70
12  15.0          8           400        150    3761          9.5     70
13  14.0          8           455        225    3086           10     70
14  24.0          4           113         95    2372           15     70
15  22.0          6           198         95    2833         15.5     70
16  18.0          6           199         97    2774         15.5     70
17  21.0          6           200         85    2587           16     70
18  27.0          4            97         88    2130         14.5     70
19  26.0          4            97         46    1835         20.5     70

    origin                          car  data1  data2
0        1     chevrolet chevelle malibu    NaN    NaN
```

```
1        1             buick skylark 320    NaN    NaN
2        1            plymouth satellite    NaN    NaN
3        1                 amc rebel sst    NaN    NaN
4        1                   ford torino    NaN    NaN
5        1              ford galaxie 500    NaN    NaN
6        1              chevrolet impala    NaN    NaN
7        1             plymouth fury iii    NaN    NaN
8        1              pontiac catalina    NaN    NaN
9        1            amc ambassador dpl    NaN    NaN
10       1           dodge challenger se    NaN    NaN
11       1            plymouth 'cuda 340    NaN    NaN
12       1         chevrolet monte carlo    NaN    NaN
13       1       buick estate wagon (sw)    NaN    NaN
14       3         toyota corona mark ii    NaN    NaN
15       1               plymouth duster    NaN    NaN
16       1                    amc hornet    NaN    NaN
17       1                ford maverick    NaN    NaN
18       3                 datsun pl510    NaN    NaN
19       2  volkswagen 1131 deluxe sedan    NaN    NaN
```

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

[ ]:
```
    mpg  cylinders  displacement horsepower  weight acceleration  model  \
0  33.0          4            91         53    1795         17.4     76
1  20.0          6           225        100    3651         17.7     76
2  18.0          6           250         78    3574           21     76
3  18.5          6           250        110    3645         16.2     76
4  17.5          6           258         95    3193         17.8     76

   origin                 car
0       3         honda civic
1       1      dodge aspen se
2       1   ford granada ghia
3       1  pontiac ventura sj
4       1       amc pacer d/l
```

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

[ ]:
```
      mpg  cylinders  displacement horsepower  weight acceleration  model  \
0    18.0          8           307        130    3504           12     70
1    15.0          8           350        165    3693         11.5     70
2    18.0          8           318        150    3436           11     70
3    16.0          8           304        150    3433           12     70
4    17.0          8           302        140    3449         10.5     70
..    …          …           …        …      …          …     …
```

```
193  24.0          6           200         81    3012         17.6     76
194  22.5          6           232         90    3085         17.6     76
195  29.0          4            85         52    2035         22.2     76
196  24.5          4            98         60    2164         22.1     76
197  29.0          4            90         70    1937         14.2     76

     origin                       car
0         1  chevrolet chevelle malibu
1         1          buick skylark 320
2         1         plymouth satellite
3         1              amc rebel sst
4         1                ford torino
..      …                        …
193       1              ford maverick
194       1                 amc hornet
195       1          chevrolet chevette
196       1             chevrolet woody
197       2                   vw rabbit

[198 rows x 9 columns]
```

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

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

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

[ ]:
```
      mpg  cylinders  displacement horsepower  weight acceleration  model  \
0    18.0          8           307        130    3504           12     70
1    15.0          8           350        165    3693         11.5     70
2    18.0          8           318        150    3436           11     70
3    16.0          8           304        150    3433           12     70
4    17.0          8           302        140    3449         10.5     70
..    …          …             …          …      …            …     …
195  27.0          4           140         86    2790         15.6     82
196  44.0          4            97         52    2130         24.6     82
197  32.0          4           135         84    2295         11.6     82
198  28.0          4           120         79    2625         18.6     82
199  31.0          4           119         82    2720         19.4     82

     origin                       car
0         1  chevrolet chevelle malibu
1         1          buick skylark 320
2         1         plymouth satellite
3         1              amc rebel sst
```

```
4         1                ford torino
..      …                         …
195       1            ford mustang gl
196       2                  vw pickup
197       1              dodge rampage
198       1                ford ranger
199       1                 chevy s-10

[398 rows x 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

[ ]:
```
df1 = pd.read_csv('/content/gdrive/MyDrive/Pandas/Act 1/mark.csv')
df2 = pd.read_csv('/content/gdrive/MyDrive/Pandas/Act 1/student.csv')
```

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

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

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

[ ]:
```
   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
```

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

[ ]:
```
   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

[ ]:
```
total_cars
```

[ ]:
```
      mpg  cylinders  displacement horsepower  weight acceleration  model  \
0    18.0          8           307        130    3504           12     70
1    15.0          8           350        165    3693         11.5     70
2    18.0          8           318        150    3436           11     70
3    16.0          8           304        150    3433           12     70
4    17.0          8           302        140    3449         10.5     70
..    …          …           …          …      …            …      …
195  27.0          4           140         86    2790         15.6     82
196  44.0          4            97         52    2130         24.6     82
197  32.0          4           135         84    2295         11.6     82
198  28.0          4           120         79    2625         18.6     82
199  31.0          4           119         82    2720         19.4     82

     origin                       car
0         1  chevrolet chevelle malibu
1         1          buick skylark 320
2         1         plymouth satellite
3         1              amc rebel sst
4         1                ford torino
..      …                          …
195       1            ford mustang gl
196       2                  vw pickup
197       1              dodge rampage
198       1                ford ranger
199       1                 chevy s-10

[398 rows x 9 columns]
```

[ ]:
```
ndf=total_cars.groupby("cylinders").mean()
ndf
```

```
<ipython-input-442-0fe79ee3c086>:1: FutureWarning: The default value of
numeric_only in DataFrameGroupBy.mean is deprecated. In a future version,
numeric_only will default to False. Either specify numeric_only or select only
columns which should be valid for the function.
  ndf=total_cars.groupby("cylinders").mean()
```

[ ]:
```
                mpg  displacement       weight      model    origin
cylinders
3         20.550000     72.500000  2398.500000  75.500000  3.000000
4         29.286765    109.799020  2308.127451  77.073529  1.985294
5         27.366667    145.000000  3103.333333  79.000000  2.000000
```

```
6           19.985714    218.142857  3198.226190  75.928571  1.190476
8           14.963107    345.009709  4114.718447  73.902913  1.000000
```

[ ]:
```
ndf.drop(["model"],axis=1)
ndf.drop(["origin"],axis=1)
```

[ ]:
```
                mpg  displacement       weight      model
cylinders
3          20.550000     72.500000  2398.500000  75.500000
4          29.286765    109.799020  2308.127451  77.073529
5          27.366667    145.000000  3103.333333  79.000000
6          19.985714    218.142857  3198.226190  75.928571
8          14.963107    345.009709  4114.718447  73.902913
```

[ ]:
```
ndf = ndf.sort_values(by = "mpg", ascending=False)
```

[ ]:
```
ndf
```

[ ]:
```
                mpg  displacement       weight      model    origin
cylinders
4          29.286765    109.799020  2308.127451  77.073529  1.985294
5          27.366667    145.000000  3103.333333  79.000000  2.000000
3          20.550000     72.500000  2398.500000  75.500000  3.000000
6          19.985714    218.142857  3198.226190  75.928571  1.190476
8          14.963107    345.009709  4114.718447  73.902913  1.000000
```

4 cilindros es el carro que más ahorra