Arnau Albert, Alex Varela, Luis Cardenete, Mar Takiguchi

Pregunta 1) Importació del Dataset

1. Explicació del context. Què son aquestes dades? Posar referències.

Aquestes dades són registros naºcionals de espanya on es poden veure els anys quan van ser diagnosticats

2. Explicar les columnes que usareu (no cal totes).

A. “Año de diagnóstico”
   a. Tipus data
B. “Sexo”
   a. Tipus string
C. “Edad años”
   a. Tipus int
D. “Edad meses”
   a. Tipus int
E. “provincia”
   a. Tipus string
F. “Pais de origen”
   a. Tipus string
G. “Código Grupo de riesgo”
   a. Tipus int
H. “Grupo de riesgo”
   a. Tipus string

```
import numpy as np
import pandas as pd

sida = pd.read_csv('registro-regional-de-sida.csv', sep=";")
sida.dtypes
```

```
Año de diagnóstico        float64
Sexo                       object
Edad años                 float64
Edad meses                float64
provincia                  object
País de origen             object
Código Grupo de riesgo      int64
Grupo de riesgo            object
dtype: object
```

c) Per a què serveix, si no queda clar amb el nom.

3. Quantes files hi ha?

```
import numpy as np
import pandas as pd

sida = pd.read_csv('registro-regional-de-sida.csv', sep=";")
print('Row count is:', len(sida.index))


Row count is: 2702
```

4. Hi ha NAs? A on?

```
import pandas as pd

sida: pd.DataFrame = pd.read_csv("registro-regional-de-sida.csv",
sep=";")

# sida.isnull().sum()
nan_rows = sida[sida.isnull().any(1)]
display(nan_rows)
```

```
import pandas as pd

sida: pd.DataFrame = pd.read_csv("registro-regional-de-sida.csv", sep=";")

# sida.isnull().sum()
nan_rows = sida[sida.isnull().any(1)]
display(nan_rows)
```

[18] ✓ 0.1s

	Año de diagnóstico	Sexo	Edad años	Edad meses	provincia	País de origen	Código Grupo de riesgo	Grupo de riesgo
63	NaN	H	28.0	0.0	Salamanca	España	20	Personas que se inyectan drogas
962	NaN	H	64.0	0.0	Salamanca	España	80	Grupo de riesgo desconocido
964	NaN	H	27.0	0.0	Salamanca	España	20	Personas que se inyectan drogas
983	NaN	H	30.0	0.0	Segovia	España	80	Grupo de riesgo desconocido
1130	NaN	H	28.0	0.0	Zamora	España	20	Personas que se inyectan drogas

5. Resultat final, fitxer Jupyter Notebook amb:

a) Text responent les anteriors observacions

b) Codi font que permeti carregar el CSV en un dataframe i mostri les primeres línies.

Pregunta 2) Arreglar el Dataset.

1. El dataset està en format «tidy»? Justifiqueu la vostra resposta.

Si, perque cada fila es una observació, cada columna es una variable i com es pot veure cada cela nomes conte una dada

2. Si no ho està, poseu-lo en aquest format utilitzant Pandas.

3. Resultat final, completar el fitxer Jupyter Notebook amb la resposta, i el codi en Pandas que heu usat, si us ha fet falta.

Pregunta 3) Tractament de valors no disponibles, NaN.

1. Si el fitxer no té valors NaN crea algunes files amb alguns valors NaN.

Si tenemos filas con valores NaN

2. Ara, aplica una d'aquestes dues operacions i justifica el motiu:

a) Substituïr el valor dels NaN d'una columna per un altre valor. (operació fillna)

b) Eliminar les files que tinguin algún valor NaN concret. (operació dropna)

```
import numpy as np
import pandas as pd

sida = pd.read_csv('registro-regional-de-sida.csv', sep=";")
sida.dropna(inplace=True)
```

3. Resultat final, completar el fitxer Jupyter Notebook amb la resposta, i el codi en Pandas que heu usat.

```
import numpy as np
import pandas as pd

sida = pd.read_csv('registro-regional-de-sida.csv', sep=";")
sida.dropna(inplace=True)
```

Pregunta 4) Consulta que filtri resultats.

1. Que mostri només algunes de les columnes del dataframe.

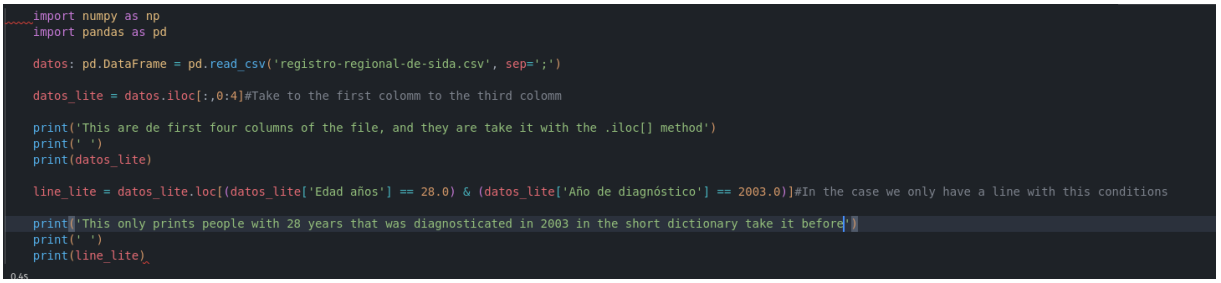
```
import numpy as np
import pandas as pd

datos: pd.DataFrame = pd.read_csv('registro-regional-de-sida.csv',
sep=';')

datos_lite = datos.iloc[:,0:4]#Take to the first colomm to the third
colomm

print('This are de first four columns of the file, and they are take it
with the .iloc[] method')
print(' ')
print(datos_lite)
```

2. Que filtri algunes de les files per un o més criteris.

```
import numpy as np
import pandas as pd

datos: pd.DataFrame = pd.read_csv('registro-regional-de-sida.csv',
sep=';')


line_lite = datos_lite.loc[(datos_lite['Edad años'] == 28.0) &
(datos_lite['Año de diagnóstico'] == 2003.0)]#In the case we only have
a line with this conditions

print('This only prints people with 28 years that was diagnosticated in
2003 in the short dictionary take it before')
```

```
print(' ')
print(line_lite)
```

3. Resultat final, Jupyter Notebook o projecte Python amb el codi.



```
import numpy as np
import pandas as pd

datos: pd.DataFrame = pd.read_csv('registro-regional-de-sida.csv', sep=';')

datos_lite = datos.iloc[:,0:4]#Take to the first colomm to the third colomm

print('This are de first four columns of the file, and they are take it with the .iloc[] method')
print(' ')
print(datos_lite)

line_lite = datos_lite.loc[(datos_lite['Edad años'] == 28.0) & (datos_lite['Año de diagnóstico'] == 2003.0)]#In the case we only have a line with this conditions

print('This only prints people with 28 years that was diagnosticated in 2003 in the short dictionary take it before')
print(' ')
print(line_lite)
```

```
This are de first four columns of the file, and they are take it with
the .iloc[] method

      Año de diagnóstico Sexo  Edad años  Edad meses
0                 1985.0    H       28.0         0.0
1                 1982.0    M       28.0         0.0
2                 1987.0    H       10.0         0.0
3                 1987.0    H       30.0         0.0
4                 1987.0    M       25.0         0.0
...                  ...  ...        ...         ...
2697              2020.0    H       42.0         0.0
2698              2020.0    H       27.0         0.0
2699              2020.0    H       54.0         0.0
2700              2021.0    H       31.0         0.0
2701              2021.0    H       60.0         0.0

[2702 rows x 4 columns]
This only prints persoan with 28 years that was diagnosticated in 2003

     Año de diagnóstico Sexo  Edad años  Edad meses
642              2003.0    H       28.0         0.0
```

Pregunta 5) Consulta que crei un rànquing.

1. És a dir, que ordeni els valors d'una columna i mostri els primers per pantalla.

Hem creat un ranking per any de diagnòstic en ordre ascendent.

```
import numpy as np
import pandas as pd

sida = pd.read_csv('registro-regional-de-sida.csv', sep=";")
sida.dropna(inplace=True)
sida.sort_values(by='Año de diagnóstico', ascending=True)
```

	Año de diagnóstico	Sexo	Edad años	Edad meses	provincia	País de origen	Código Grupo de riesgo	Grupo de riesgo
1	1982.0	M	28.0	0.0	Valladolid	España	20	Personas que se inyectan drogas
1810	1984.0	H	34.0	0.0	Valladolid	España	80	Grupo de riesgo desconocido
894	1984.0	H	56.0	0.0	Valladolid	España	41	Receptores de hemoderivados
893	1984.0	H	11.0	0.0	Salamanca	España	40	Receptores de hemoderivados
0	1985.0	H	28.0	0.0	Salamanca	España	20	Personas que se inyectan drogas
...	...	...	...	...	...	...	...	...
2700	2021.0	H	31.0	0.0	Palencia	España	75	Relaciones Heterosexuales
892	2021.0	H	42.0	0.0	Avila	Ecuador	80	Grupo de riesgo desconocido
891	2021.0	H	42.0	0.0	Segovia	España	75	Relaciones Heterosexuales
888	2021.0	M	44.0	0.0	Burgos	Brasil	80	Grupo de riesgo desconocido
2701	2021.0	H	60.0	0.0	Salamanca	España	70	Relaciones Heterosexuales

2. També heu de mostrar un gràfic.

3. Resultat final, Jupyter Notebook o projecte Python amb el codi i el gràfic.

import numpy as np
import pandas as pd

sida = pd.read_csv('registro-regional-de-sida.csv', sep=";")
sida.dropna(inplace=True)
sida.sort_values(by='Año de diagnóstico', ascending=True)

Pregunta 6) Consulta que crei almenys una columna calculada.

1. És a dir, que la consulta crei un nou camp depenent d'un altre camp, o calculat a partir d'altres columnes.

```
import numpy as np
import pandas as pd

datos: pd.DataFrame = pd.read_csv('registro-regional-de-sida.csv',
sep=';')

datos_lite = datos.iloc[:,:5]#Take to the first colomm to the third
colomm



datos.loc[:, 'born'] = datos.loc[:, 'Año de diagnóstico'] -
datos.loc[:, 'Edad años']
print('This is the exercise 6, the new columns is born, and it is (Año
de diagnóstico)- (Edad años) and it return the born year')
print('')
datos
```

2. Exemples:

a) camp Apte/NoApte depenent de les notes d'alumnes

b) càlcul imc a partir del pes i l'alçada.

3. Resultat final, Jupyter Notebook o projecte Python amb el codi.

En el nostre cas, crearem una columna calculada en base a l'any de diagnostic i la seva edad en aquell moment, el que retornara es l'any que van neixer aquestes persones

```
This is the exercise 6, the new columns is born, and it is (Año de diagnóstico)- (Edad años) and it return the born year
```

	Año de diagnóstico	Sexo	Edad años	Edad meses	provincia	País de origen	Código Grupo de riesgo	Grupo de riesgo	born
0	1985.0	H	28.0	0.0	Salamanca	España	20	Personas que se inyectan drogas	1957.0
1	1982.0	M	28.0	0.0	Valladolid	España	20	Personas que se inyectan drogas	1954.0
2	1987.0	H	10.0	0.0	Palencia	España	40	Receptores de hemoderivados	1977.0
3	1987.0	H	30.0	0.0	Salamanca	España	20	Personas que se inyectan drogas	1957.0
4	1987.0	M	25.0	0.0	Burgos	España	77	Relaciones Heterosexuales	1962.0
...	...	...	...	...	...	...	...	...	...
2697	2020.0	H	42.0	0.0	Burgos	España	80	Grupo de riesgo desconocido	1978.0
2698	2020.0	H	27.0	0.0	León	Mali	80	Grupo de riesgo desconocido	1993.0
2699	2020.0	H	54.0	0.0	León	España	10	Varones homosexuales / bisexuales	1966.0
2700	2021.0	H	31.0	0.0	Palencia	España	75	Relaciones Heterosexuales	1990.0
2701	2021.0	H	60.0	0.0	Salamanca	España	70	Relaciones Heterosexuales	1961.0

Pregunta 7) Consulta amb dades agrupades per un camp de tipus categòric.

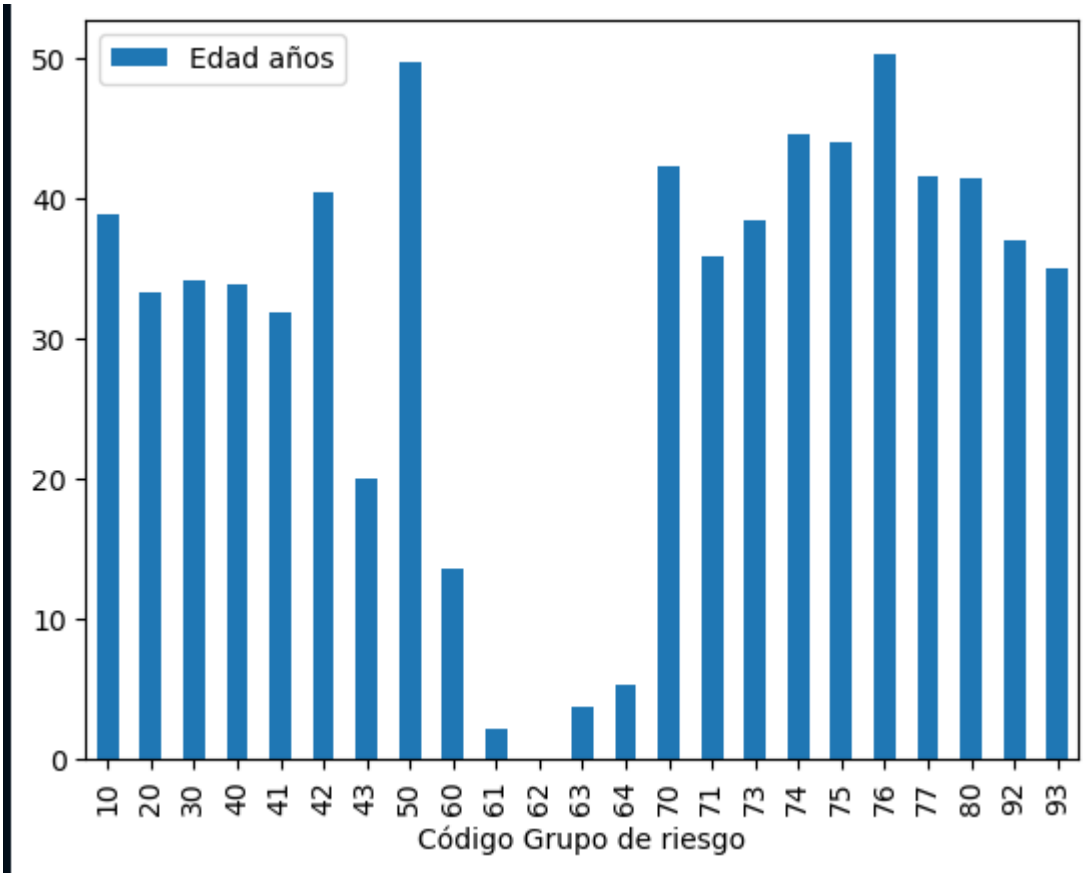
1. Si no teniu un camp que es pugui convertir a tipus categòric, haureu de crear-ne un.

```
import numpy as np
import pandas as pd
sida: pd.DataFrame = pd.read_csv("registro-regional-de-sida.csv",
sep=";")

media = sida.groupby(by = "Código Grupo de riesgo").mean()

media_edad = media.loc[:, ["Edad años"]]
media_edad
```

Código Grupo de riesgo	Edad años
10	38.941176
20	33.415426
30	34.151515
40	33.891892
41	31.941176
42	40.500000
43	20.000000
50	49.760000
60	13.666667
61	2.238095
62	0.000000
63	3.800000
64	5.333333
70	42.327869
71	35.886792
73	38.500000
74	44.602740
75	44.094444
76	50.285714
77	41.707317
80	41.491018
92	37.000000
93	35.000000

2. També heu de mostrar un gràfic de totes les categories.

```
import numpy as np
import pandas as pd
sida: pd.DataFrame = pd.read_csv("registro-regional-de-sida.csv",
sep=";")
sida["Código Grupo de riesgo"]
media = sida.groupby(by = "Código Grupo de riesgo").mean()

media_edad = media.loc[:, ["Edad años"]]
media_edad.plot(kind="bar")
```



3. Resultat final, Jupyter Notebook o projecte Python amb el codi i el gràfic.

Pregunta 8) Consulta amb dades agrupades per data.

1. És a dir, que si les dades no estan agrupades les haureu d'agrupar per data; ja sigui per any, per mes o per dia.

```
import numpy as np
import pandas as pd



sida: pd.DataFrame = pd.read_csv("registro-regional-de-sida.csv",
sep=";")
sida["Año de diagnóstico"]
media = sida.groupby(by = "Año de diagnóstico").mean()
media.iloc[:,0:1]
```

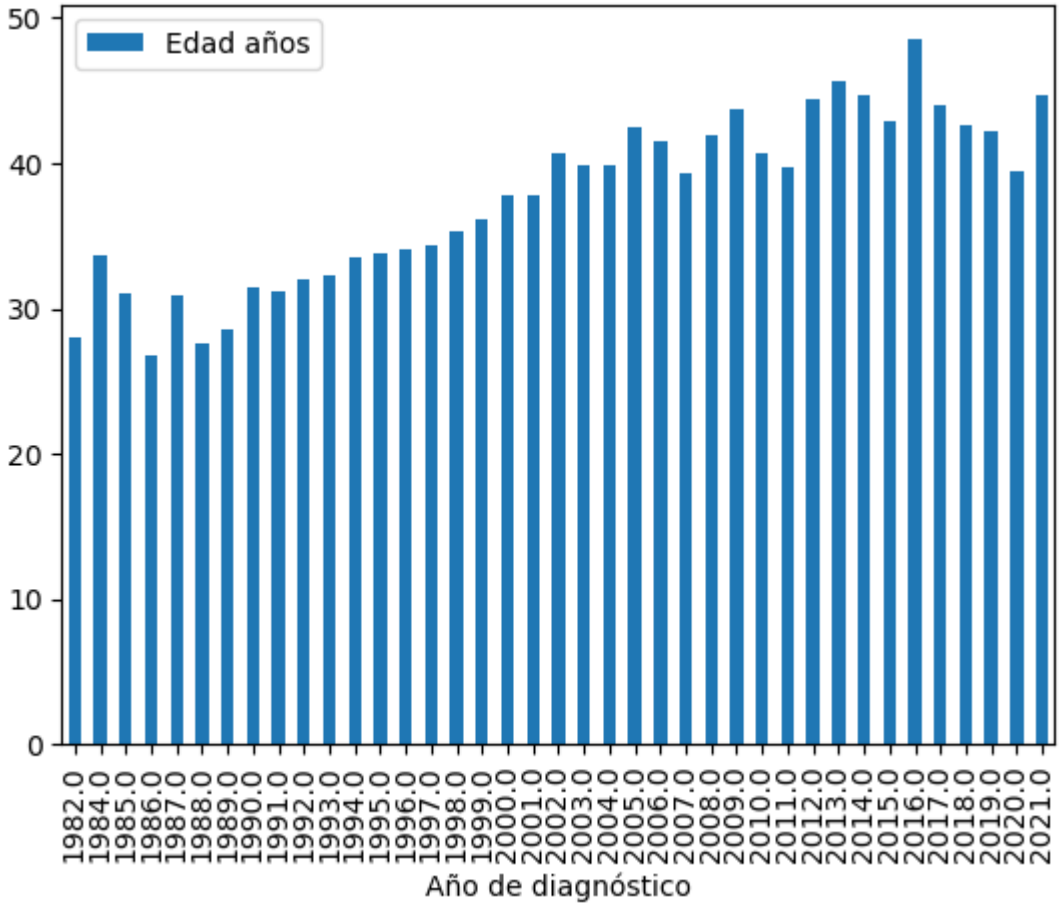
```python
import numpy as np
import pandas as pd


sida: pd.DataFrame = pd.read_csv("registro-regional-de-sida.csv", sep=";")
sida["Año de diagnóstico"]
media = sida.groupby(by = "Año de diagnóstico").mean()
media.iloc[:,0:1]
```

✓ 0.1s

Año de diagnóstico	Edad años
1982.0	28.000000
1984.0	33.666667
1985.0	31.000000
1986.0	26.800000
1987.0	30.842105
1988.0	27.576271
1989.0	28.500000
1990.0	31.427184
1991.0	31.180952
1992.0	32.016529
1993.0	32.226667
1994.0	33.520619
1995.0	33.762712

2. També heu de mostrar un gràfic.

```python
import numpy as np
import pandas as pd
sida: pd.DataFrame = pd.read_csv("registro-regional-de-sida.csv",
sep=";")
sida["Año de diagnóstico"]
media = sida.groupby(by = "Año de diagnóstico").mean()
grafico = media.iloc[:,0:1]

grafico.plot(kind="bar")
```



3. Resultat final, Jupyter Notebook o projecte Python amb el codi i el gràfic.

Pregunta 9) Separació i fusió de datasets.

1. Tria una de les 2 operacions:

a) Fes una còpia del dataSet, aconsegueix crear 2 dataSet amb camps i files separats però que comparteixin un camp comú, i després fes el merge.

b) Si el teu dataSet està desactualitzat o falten dades d'alguns anys i les trobes dades per altres fonts, crea un nou conjunt de dades amb algunes files i/o alguna columna nova.

```
import numpy as np

import pandas as pd

import copy



sida: pd.DataFrame = pd.read_csv("registro-regional-de-sida.csv",
sep=";")



genero: pd.DataFrame = (copy.deepcopy(sida).drop(columns=["Código Grupo
de riesgo","provincia","Edad años","Edad meses","País de origen"]))



cgdr : pd.DataFrame =
(copy.deepcopy(sida).drop(columns=["Sexo","provincia","Edad años","Edad
meses","País de origen"]))

# Código Grupo de riesgo



join : pd.DataFrame = pd.merge(genero,cgdr)



join
```

```python
import numpy as np
import pandas as pd
import copy

sida: pd.DataFrame = pd.read_csv("registro-regional-de-sida.csv", sep=";")

genero: pd.DataFrame = (copy.deepcopy(sida).drop(columns=["Código Grupo de riesgo","provincia","Edad años","Edad meses","País de origen"]))

cgdr : pd.DataFrame = (copy.deepcopy(sida).drop(columns=["Sexo","provincia","Edad años","Edad meses","País de origen"]))
# Código Grupo de riesgo

join : pd.DataFrame = pd.merge(genero,cgdr)

join
```

[19] ✓ 0.1s Python

	Año de diagnóstico	Sexo	Grupo de riesgo	Código Grupo de riesgo
0	1985.0	H	Personas que se inyectan drogas	20
1	1985.0	H	Personas que se inyectan drogas	20
2	1985.0	H	Personas que se inyectan drogas	20
3	1985.0	H	Personas que se inyectan drogas	20
4	1982.0	M	Personas que se inyectan drogas	20
...	...	...	...	...
129591	2004.0	M	Hijo de madre a riesgo	61
129592	2009.0	H	Receptores de hemoderivados	40
129593	2015.0	H	Hijo de madre a riesgo	60
129594	2017.0	H	Hijo de madre a riesgo	60
129595	2020.0	M	Personas que se inyectan drogas	20

129596 rows × 4 columns

	Año de diagnóstico	Sexo	Grupo de riesgo	Código Grupo de riesgo
0	1985.0	H	Personas que se inyectan drogas	20
1	1985.0	H	Personas que se inyectan drogas	20
2	1985.0	H	Personas que se inyectan drogas	20
3	1985.0	H	Personas que se inyectan drogas	20
4	1982.0	M	Personas que se inyectan drogas	20
...	...	...	...	...
129591	2004.0	M	Hijo de madre a riesgo	61
129592	2009.0	H	Receptores de hemoderivados	40
129593	2015.0	H	Hijo de madre a riesgo	60
129594	2017.0	H	Hijo de madre a riesgo	60
129595	2020.0	M	Personas que se inyectan drogas	20

129596 rows × 4 columns

Finalment, fes el merge.

2. Resultat final, Jupyter Notebook o projecte Python