S05 T01: Tasca mètodes de mostreig

Descripció Aprèn a realitzar mostreig de les dades amb Python.

NIVELL 1

Exercici 1

Agafa un conjunt de dades de tema esportiu que t'agradi. Realitza un mostreig de les dades generant una mostra aleatòria simple i una mostra sistemàtica.

He escollit un conjunt de dades relacionades amb el mundial del futbol femení del 2019 que he trobat al web https://data.world/, on es veuen les jugadores de cada equip i diversa informació de cada una

In [1]:
```
# Crido a les llibreries que necessito
# Faig entrar l'arxiu CSV gràcies a pandas

import pandas as pd
import numpy as np
import random
import matplotlib.pyplot as plt
import seaborn as sns
import imblearn

women = pd.read_csv("C:\\Users\\Anna\DataScience\\SPRINTS\\SPRINT 5\\Womens Squads.csv", e

display(women)
```

	Squad no.	Country	Pos.	Player	DOB	Age	Caps	Goals	Club
0	1	US	GK	Alyssa Naeher	20-abr-88	31	43.0	0.0	Chicago Red Stars
1	2	US	FW	Mallory Pugh	29-abr-98	21	50.0	15.0	Washington Spirit
2	3	US	MF	Sam Mewis	09-oct-92	26	47.0	9.0	North Carolina Courage
3	4	US	DF	Becky Sauerbrunn	06-jun-85	34	155.0	0.0	Utah Royals
4	5	US	DF	Kelley O'Hara	04-ago-88	30	115.0	2.0	Utah Royals
...	...	...	...	...	...	...	...	...	...
547	19	France	DF	Griedge Mbock Bathy	26-feb-95	24	49.0	4.0	Lyon
548	20	France	FW	Delphine Cascarino	05-feb-97	22	11.0	1.0	Lyon
549	21	France	GK	Pauline Peyraud-Magnin	17-mar-92	27	1.0	0.0	Arsenal
550	22	France	DF	Julie Debever	18-abr-88	31	2.0	0.0	Guingamp
551	23	France	MF	Maéva Clémaron	10-nov-92	26	3.0	0.0	Fleury

552 rows × 9 columns

In [2]:
```
# Tipus de dades que trobem a cada columna
women.dtypes
```

Out[2]:
```
Squad no.      int64
Country       object
```

```
Pos.           object
Player         object
DOB            object
Age             int64
Caps          float64
Goals         float64
Club           object
dtype: object
```

In [3]:
```
# Petita descripció de cada columna
women.describe()
```

Out[3]:

	Squad no.	Age	Caps	Goals
count	552.000000	552.000000	520.000000	520.000000
mean	12.000000	26.050725	43.661538	7.348077
std	6.639266	4.060920	43.674846	15.541727
min	1.000000	16.000000	0.000000	0.000000
25%	6.000000	23.000000	11.750000	0.000000
50%	12.000000	26.000000	29.500000	1.500000
75%	18.000000	29.000000	62.000000	8.250000
max	23.000000	41.000000	282.000000	181.000000

MOSTRA ALEATÒRIA SIMPLE

Utilitzo el mètode "sample" i em treu una mostra aleatòria simple de 100 jugadores.

La mostra aleatòria simple es caracteritza perquè tots els membres de la població tenen les mateixes possibilitats de ser seleccionats.

In [4]:
```
mostra_simple = women.sample(100)

display(mostra_simple)
```

	Squad no.	Country	Pos.	Player	DOB	Age	Caps	Goals	Club
511	6	Nigeria	MF	Evelyn Nwabuoku	14-nov-85	33	42.0	3.0	Rivers Angels
17	18	US	GK	Ashlyn Harris	19-oct-85	33	21.0	0.0	Orlando Pride
76	8	Chile	MF	Karen Araya	16-oct-90	28	20.0	6.0	Sevilla
291	16	Jamaica	DF	Dominique Bond-Flasza	11-sep-96	22	16.0	2.0	PSV
537	9	France	FW	Eugénie Le Sommer	18-may-89	30	159.0	73.0	Lyon
...	...	...	...	...	...	...	...	...	...
311	13	Italy	DF	Elisa Bartoli	07-may-91	28	45.0	1.0	Roma
62	17	Sweden	MF	Caroline Seger	19-mar-85	34	193.0	27.0	Rosengård
205	22	Scotland	FW	Erin Cuthbert	19-jul-98	20	29.0	9.0	Chelsea
262	10	Argentina	MF	Estefanía Banini	21-jun-90	28	32.0	9.0	Levante
44	22	Thailand	GK	Tiffany Sornpao	22-may-98	21	1.0	0.0	Kennesaw State Owls

100 rows × 9 columns

MOSTRA ALEATÒRIA SISTEMÀTICA

És similar a la simple però en comptes de generar números aleatoris, els individus s'escullen a intervals regulars a partir d'un primer número aleatori.

La fòrmula és k = N\n, on K serà l'interval, N majúscula és la mida total de la població i n minúscula és la mida de la mostra

In [5]:
```
# Definim la mida de la mostra
print(women.shape)

n = 100               # mida de la mostra que volem
N = len(women)        # mida de la població
k = int(N/n)          # càlcul de l'interval

print("L'interval per fer el càlcul de la mostra serà de ", k)
```

```
(552, 9)
L'interval per fer el càlcul de la mostra serà de  5
```

In [6]:
```
# Eleció del número aleatori
# Tot aquest apartat m'ha quedat inutilitzat pel problema que m'he trobat i que explico er

randomNum = random.randint(1,552)

print("El número des d'on s'iniciarà l'elecció dels valors de la mostra és el", randomNum)
```

```
El número des d'on s'iniciarà l'elecció dels valors de la mostra és el 459
```

In [7]:
```
# Definim la funció de la mostra aleatòria sistemàtica

def systematic (women, step):
    index = np.arange(k, len(women), step=step)
    mostra_sistematica = women.iloc[index]
    return mostra_sistematica

# Obtenim el restultat de la funció amb els valors que volem

mostra_sistematica = systematic(women, k)

display(mostra_sistematica)
print(mostra_sistematica.shape)

# La mostra surt de 110 en comptes de 100 perquè he convertit la k en int, per tant al ari


# He intentat fer que el primer número fos random, però al arribar al final del dataframe
# index = np.arange(randomNum, len(women), step=step)
```

	Squad no.	Country	Pos.	Player	DOB	Age	Caps	Goals	Club
5	6	US	MF	Morgan Brian	26-feb-93	26	82.0	6.0	Chicago Red Stars
10	11	US	DF	Ali Krieger	28-jul-84	34	99.0	1.0	Orlando Pride
15	16	US	MF	Rose Lavelle	14-may-95	24	24.0	6.0	Washington Spirit
20	21	US	GK	Adrianna Franch	12-nov-90	28	1.0	0.0	Portland Thorns
25	3	Thailand	DF	Natthakarn Chinwong	15-mar-92	27	11.0	0.0	Bundit Asia

	Squad no.	Country	Pos.	Player	DOB	Age	Caps	Goals	Club
...	...	...	...	...	...	...	...	...	...
530	2	France	DF	Ève Périsset	24-dic-94	24	13.0	0.0	Paris Saint-Germain
535	7	France	DF	Sakina Karchaoui	26-ene-96	23	23.0	0.0	Montpellier
540	12	France	FW	Emelyne Laurent	04-nov-98	20	3.0	0.0	Guingamp
545	17	France	MF	Gaëtane Thiney	28-oct-85	33	154.0	58.0	Paris FC
550	22	France	DF	Julie Debever	18-abr-88	31	2.0	0.0	Guingamp

110 rows × 9 columns

```
(110, 9)
```

NIVELL 2

Exercici 2

Continua amb el conjunt de dades de tema esportiu i genera una mostra estratificada i una mostra utilitzant SMOTE (Synthetic Minority Oversampling Technique).

MOSTRA ESTRATIFICADA

Consisteix en dividir la població en subpoblacions que poden diferir de manera important. Permet extreure conclusions més precises assegurant que cada subgrup (o estrat) està representat correctament a la mostra. Per utilitzar-lo s'ha de dividir la població en subgrups, calcular quantes persones s'haurien de prendre mostres de cada subgrup i utilitzar el mostreig aleatori o sistemàtic per seleccionar la mostra de cada subgrup

In [8]:

```
pd.set_option('max_rows', None)             #ho modificio per veure tota la mostra

# Primer decideixo com vull dividir la població, i en aquest cas ho faig per país de proce

def mostra_estratificada(women,n,numClusters):            # funció princiapl
    N = len(women)
    k = int(N/n)
    numGrups = int(N/k)

    def pes_subgrups (women, numGrups):                   # funció per saber quants valors

        def mostra_subgrup (x):                           # funció per l'elecció de la pobl
            n_x = int(np.rint(k*len(x)/len(women)))
            mostra_x = x.sample(n_x)
            return(mostra_x)

        pes_mostra = women.groupby("Country").apply(mostra_subgrup)
        return(pes_mostra)

    estrats = None
    for k in range (numGrups):                #bucle for perquè vagi recorrent i trobant ca
        pes_mostra_k = pes_subgrups(women,k).reset_index(drop = True)
        pes_mostra_k["cluster"] = np.repeat(k, len(pes_mostra_k))
        estrats = pd.concat([estrats, pes_mostra_k], axis=0)
        women.drop(index = pes_mostra_k.index)
    clusters_seleccionats = np.random.randint(k,numGrups,size= numClusters)
    mostres_estratificades = estrats[estrats.cluster.isin(clusters_seleccionats)]
    return (mostres_estratificades)

mostra = mostra_estratificada(women = women,n=100,numClusters=20)
```

```
display(mostra)

#Com que totes els equips tenen el mateix número de jugadores, els clústers tenen tots el
```

	Squad no.	Country	Pos.	Player	DOB	Age	Caps	Goals	Club	cluster
0	10	Argentina	MF	Estefanía Banini	21-jun-90	28	32.0	9.0	Levante	109
1	15	Argentina	FW	Belén Potassa	12-dic-88	30	28.0	7.0	UAI Urquiza	109
2	6	Argentina	DF	Aldana Cometti	03-mar-96	23	32.0	3.0	Sevilla	109
3	16	Argentina	MF	Lorena Benítez	03-dic-98	20	2.0	0.0	Boca Juniors	109
4	21	Argentina	DF	Natalie Juncos	28-dic-90	28	6.0	0.0	Unattached	109
5	3	Australia	MF	Aivi Luik	18-mar-85	34	21.0	0.0	Levante	109
6	6	Australia	MF	Chloe Logarzo	22-dic-94	24	37.0	6.0	Washington Spirit	109
7	14	Australia	DF	Alanna Kennedy	21-ene-95	24	77.0	7.0	Orlando Pride	109
8	15	Australia	FW	Emily Gielnik	13-may-92	27	28.0	7.0	Melbourne Victory	109
9	22	Australia	MF	Amy Harrison	21-abr-96	23	10.0	0.0	Washington Spirit	109
10	1	Brazil	GK	Bárbara	04-jul-88	30	41.0	0.0	Kindermann	109
11	3	Brazil	DF	Érika	04-feb-88	31	65.0	13.0	Corinthians	109
12	18	Brazil	MF	Luana	02-may-93	26	6.0	0.0	Hwacheon KSPO	109
13	9	Brazil	FW	Debinha	20-oct-91	27	45.0	14.0	North Carolina Courage	109
14	8	Brazil	MF	Formiga	03-mar-78	41	160.0	23.0	Paris Saint-Germain	109
15	21	Cameroon	FW	Alexanda Takounda	07-jul-00	18	NaN	NaN	Eclair	109
16	23	Cameroon	GK	Marthe Ongmahan	12-jun-92	26	0.0	0.0	AWA Yaoundé	109
17	1	Cameroon	GK	Annette Ngo Ndom	03-jun-85	34	50.0	0.0	Amazone FAP	109
18	6	Cameroon	DF	Estelle Johnson	21-jul-88	30	NaN	NaN	Sky Blue FC	109
19	9	Cameroon	FW	Madeleine Ngono Mani	16-oct-83	35	128.0	49.0	Ambilly [fr]	109
20	22	Canada	DF	Lindsay Agnew	31-mar-95	24	11.0	0.0	Houston Dash	109

	Squad no.	Country	Pos.	Player	DOB	Age	Caps	Goals	Club	cluster
21	8	Canada	DF	Jayde Riviere	22-ene-01	18	5.0	0.0	Vancouver Whitecaps	109
22	13	Canada	MF	Sophie Schmidt	28-jun-88	30	184.0	19.0	Houston Dash	109
23	23	Canada	DF	Jenna Hellstrom	02-abr-95	24	4.0	0.0	KIF Örebro	109
24	16	Canada	FW	Janine Beckie	20-ago-94	24	56.0	25.0	Manchester City	109
25	7	Chile	FW	María José Rojas	17-dic-87	31	20.0	12.0	Slavia Praha	109
26	15	Chile	DF	Su Helen Galaz	27-may-91	28	10.0	0.0	Zaragoza	109
27	12	Chile	GK	Natalia Campos	12-ene-92	27	3.0	0.0	Universidad Católica	109
28	1	Chile	GK	Christiane Endler	23-jul-91	27	20.0	0.0	Paris Saint-Germain	109
29	4	Chile	MF	Francisca Lara	29-jul-90	28	22.0	9.0	Sevilla	109
30	14	China PR	DF	Wang Ying	18-nov-97	21	NaN	NaN	Wuhan Jianghan University [zh]	109
31	18	China PR	GK	Bi Xiaolin	18-sep-89	29	NaN	NaN	Dalian Quanjian	109
32	10	China PR	FW	Li Ying	07-ene-93	26	92.0	24.0	Guangdong Huijun	109
33	11	China PR	FW	Wang Shanshan	27-ene-90	29	77.0	10.0	Dalian Quanjian	109
34	2	China PR	DF	Liu Shanshan	16-mar-92	27	58.0	0.0	Beijing Phoenix	109
35	10	England	FW	Fran Kirby	29-jun-93	25	37.0	12.0	Chelsea	109
36	14	England	DF	Leah Williamson	29-mar-97	22	6.0	0.0	Arsenal	109
37	22	England	FW	Beth Mead	09-may-95	24	12.0	5.0	Arsenal	109
38	20	England	MF	Karen Carney	01-ago-87	31	138.0	32.0	Chelsea	109
39	19	England	MF	Georgia Stanway	03-ene-99	20	6.0	1.0	Manchester City	109
40	13	France	FW	Valérie Gauvin	01-jun-96	23	17.0	9.0	Montpellier	109
41	11	France	FW	Kadidiatou Diani	01-abr-95	24	45.0	4.0	Paris Saint-Germain	109
42	4	France	DF	Marion Torrent	17-abr-92	27	20.0	0.0	Montpellier	109
43	2	France	DF	Ève Périsset	24-dic-94	24	13.0	0.0	Paris Saint-Germain	109

	Squad no.	Country	Pos.	Player	DOB	Age	Caps	Goals	Club	cluster
44	9	France	FW	Eugénie Le Sommer	18-may-89	30	159.0	73.0	Lyon	109
45	5	Germany	DF	Marina Hegering	17-abr-90	29	2.0	0.0	SGS Essen	109
46	4	Germany	DF	Leonie Maier	29-sep-92	26	69.0	10.0	Bayern Munich	109
47	3	Germany	DF	Kathrin Hendrich	06-abr-92	27	29.0	4.0	Bayern Munich	109
48	9	Germany	MF	Svenja Huth	25-ene-91	28	43.0	7.0	Turbine Potsdam	109
49	2	Germany	DF	Carolin Simon	24-nov-92	26	15.0	2.0	Lyon	109
50	7	Italy	DF	Alia Guagni	01-oct-87	31	61.0	5.0	Fiorentina	109
51	21	Italy	MF	Valentina Cernoia	22-jun-91	27	30.0	6.0	Juventus	109
52	5	Italy	DF	Elena Linari	15-abr-94	25	28.0	0.0	Atlético Madrid	109
53	6	Italy	MF	Martina Rosucci	09-may-92	27	35.0	1.0	Juventus	109
54	18	Italy	FW	Ilaria Mauro	22-may-88	31	24.0	8.0	Fiorentina	109
55	7	Jamaica	MF	Chinyelu Asher	20-may-93	26	22.0	3.0	Stabæk	109
56	2	Jamaica	DF	Lauren Silver	22-mar-93	26	19.0	1.0	Trondheims-Ørn	109
57	12	Jamaica	MF	Sashana Campbell	02-mar-91	28	8.0	2.0	Maccabi Kishronot Hadera	109
58	14	Jamaica	DF	Deneisha Blackwood	07-mar-97	22	13.0	2.0	West Florida Argonauts	109
59	8	Jamaica	MF	Ashleigh Shim	11-nov-93	25	6.0	1.0	Unattached	109
60	11	Japan	FW	Rikako Kobayashi	21-jul-97	21	5.0	2.0	Nippon TV Beleza	109
61	10	Japan	MF	Mizuho Sakaguchi	15-oct-87	31	124.0	29.0	Nippon TV Beleza	109
62	19	Japan	FW	Jun Endo	24-may-00	19	4.0	0.0	Nippon TV Beleza	109
63	3	Japan	DF	Aya Sameshima	16-jun-87	31	108.0	5.0	INAC Kobe Leonessa	109
64	2	Japan	DF	Rumi Utsugi	05-dic-88	30	112.0	6.0	Reign FC	109
65	17	Netherlands	FW	Ellen Jansen	06-oct-92	26	14.0	1.0	Ajax	109
66	14	Netherlands	MF	Jackie Groenen	17-dic-94	24	45.0	2.0	1. FFC Frankfurt	109

	Squad no.	Country	Pos.	Player	DOB	Age	Caps	Goals	Club	cluster
67	3	Netherlands	DF	Stefanie van der Gragt	16-ago-92	26	55.0	7.0	Barcelona	109
68	19	Netherlands	MF	Jill Roord	22-abr-97	22	40.0	3.0	Bayern Munich	109
69	9	Netherlands	FW	Vivianne Miedema	15-jul-96	22	74.0	57.0	Arsenal	109
70	12	New Zealand	MF	Betsy Hassett	04-ago-90	28	111.0	13.0	KR Reykjavik	109
71	1	New Zealand	GK	Erin Nayler	17-abr-92	27	61.0	0.0	Bordeaux	109
72	7	New Zealand	DF	Ali Riley	30-oct-87	31	123.0	1.0	Chelsea	109
73	20	New Zealand	MF	Daisy Cleverley	30-abr-97	22	8.0	2.0	California Golden Bears	109
74	4	New Zealand	DF	CJ Bott	22-abr-95	24	16.0	1.0	Vittsjö	109
75	2	Nigeria	MF	Amarachi Okoronkwo	12-dic-92	26	NaN	NaN	Nasarawa Amazons	109
76	5	Nigeria	DF	Onome Ebi	08-may-83	36	81.0	0.0	Henan Huishang	109
77	15	Nigeria	FW	Rasheedat Ajibade	08-dic-99	19	NaN	NaN	Avaldsnes	109
78	9	Nigeria	FW	Desire Oparanozie	17-dic-93	25	35.0	22.0	Guingamp	109
79	6	Nigeria	MF	Evelyn Nwabuoku	14-nov-85	33	42.0	3.0	Rivers Angels	109
80	23	Norway	GK	Oda Bogstad	24-abr-96	23	0.0	0.0	Arna-Bjørnar	109
81	21	Norway	MF	Karina Sævik	24-mar-96	23	3.0	1.0	Kolbotn	109
82	20	Norway	FW	Emilie Haavi	16-jun-92	26	81.0	16.0	LSK Kvinner	109
83	1	Norway	GK	Ingrid Hjelmseth	10-abr-80	39	131.0	0.0	Stabæk	109
84	13	Norway	FW	Therese Åsland	26-ago-95	23	5.0	1.0	LSK Kvinner	109
85	5	Scotland	DF	Jennifer Beattie	13-may-91	28	123.0	22.0	Manchester City	109
86	9	Scotland	MF	Caroline Weir	20-jun-95	23	62.0	7.0	Manchester City	109
87	12	Scotland	GK	Shannon Lynn	22-oct-85	33	30.0	0.0	Vittsjö	109
88	6	Scotland	MF	Joanne Love	06-dic-85	33	191.0	13.0	Glasgow City	109
89	15	Scotland	DF	Sophie Howard	17-sep-93	25	13.0	0.0	Reading	109

	Squad no.	Country	Pos.	Player	DOB	Age	Caps	Goals	Club	cluster
90	1	South Africa	GK	Mapaseka Mpuru	09-abr-98	21	0.0	0.0	University of Pretoria	109
91	14	South Africa	DF	Tiisetso Makhubela	24-abr-97	22	2.0	0.0	Mamelodi Sundowns	109
92	16	South Africa	GK	Andile Dlamini	02-sep-92	26	37.0	0.0	Mamelodi Sundowns	109
93	21	South Africa	MF	Busisiwe Ndimeni	25-jun-91	27	26.0	0.0	Tshwane University of Technology	109
94	13	South Africa	DF	Bambanani Mbane	12-mar-90	29	43.0	0.0	Bloemfontein Celtic	109
95	5	South Korea	DF	Kim Do-yeon	07-dic-88	30	79.0	1.0	Incheon Hyundai Steel Red Angels	109
96	16	South Korea	DF	Jang Sel-gi	31-may-94	25	54.0	11.0	Incheon Hyundai Steel Red Angels	109
97	2	South Korea	DF	Lee Eun-mi	18-ago-88	30	87.0	14.0	Suwon UDC	109
98	21	South Korea	GK	Jung Bo-ram	22-jul-91	27	3.0	0.0	Hwacheon KSPO	109
99	6	South Korea	DF	Lim Seon-joo	27-nov-90	28	76.0	5.0	Incheon Hyundai Steel Red Angels	109
100	7	Spain	MF	Marta Corredera	08-ago-91	27	66.0	5.0	Levante	109
101	18	Spain	MF	Aitana Bonmatí	18-ene-98	21	12.0	1.0	Barcelona	109
102	8	Spain	DF	Marta Torrejón	27-feb-90	29	85.0	9.0	Barcelona	109
103	4	Spain	DF	Irene Paredes	04-jul-91	27	61.0	8.0	Paris Saint-Germain	109
104	2	Spain	DF	Celia Jiménez	20-jun-95	23	22.0	0.0	Reign FC	109
105	6	Sweden	DF	Magdalena Eriksson	08-sep-93	25	48.0	5.0	Chelsea	109
106	23	Sweden	MF	Elin Rubensson	11-may-93	26	62.0	2.0	Kopparbergs/Göteborg	109
107	1	Sweden	GK	Hedvig Lindahl	29-abr-83	36	157.0	0.0	Chelsea	109
108	8	Sweden	MF	Lina Hurtig	15-sep-95	23	17.0	3.0	Linköping	109
109	7	Sweden	FW	Madelen Janogy	12-nov-95	23	3.0	0.0	Piteå	109
110	11	Thailand	MF	Sudarat Chucheun	19-jun-97	21	NaN	NaN	Sisaket	109
111	14	Thailand	FW	Saowalak Pengngam	30-nov-96	22	NaN	NaN	Chonburi Sriprathum	109
112	8	Thailand	FW	Suchawadee Nildhamrong	01-abr-97	22	17.0	12.0	California Golden Bears	109

	Squad no.	Country	Pos.	Player	DOB	Age	Caps	Goals	Club	cluster
113	9	Thailand	DF	Warunee Phetwiset	13-dic-90	28	44.0	0.0	Chonburi Sriprathum	109
114	18	Thailand	GK	Sukanya Chor Charoenying	24-nov-87	31	8.0	0.0	Air Force United	109
115	22	US	FW	Jessica McDonald	28-feb-88	31	7.0	2.0	North Carolina Courage	109
116	11	US	DF	Ali Krieger	28-jul-84	34	99.0	1.0	Orlando Pride	109
117	16	US	MF	Rose Lavelle	14-may-95	24	24.0	6.0	Washington Spirit	109
118	18	US	GK	Ashlyn Harris	19-oct-85	33	21.0	0.0	Orlando Pride	109
119	7	US	DF	Abby Dahlkemper	13-may-93	26	37.0	0.0	North Carolina Courage	109

MOSTRA UTILITZANT SMOTE (Synthetic Minority Oversampling Technique).

Tècnica de mostreig que ens permet generar mostres sintètiques per a les categories minoritàries d'un grup

Per veure si tenim alguna classe minoritària, farem una mica d'estudi bàsic de les dades per tenir una idea del que ens podem trobar

In [9]:
```
women.describe()
```

Out[9]:

	Squad no.	Age	Caps	Goals
count	552.000000	552.000000	520.000000	520.000000
mean	12.000000	26.050725	43.661538	7.348077
std	6.639266	4.060920	43.674846	15.541727
min	1.000000	16.000000	0.000000	0.000000
25%	6.000000	23.000000	11.750000	0.000000
50%	12.000000	26.000000	29.500000	1.500000
75%	18.000000	29.000000	62.000000	8.250000
max	23.000000	41.000000	282.000000	181.000000

In [10]:
```
women.count()
```

Out[10]:
```
Squad no.    552
Country      552
Pos.         552
Player       552
DOB          552
Age          552
Caps         520
Goals        520
Club         552
dtype: int64
```

In [11]:
```
# Com que en l'anterior punt veiem que les columnes "Caps" i "Goals" tenen menys quantitat

print(women.isnull())

print("_____________________________________________________________________________")

print(women.count())

print("_____________________________________________________________________________")

print(women.isnull().sum())
```

	Squad no.	Country	Pos.	Player	DOB	Age	Caps	Goals	Club
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
5	False	False	False	False	False	False	False	False	False
6	False	False	False	False	False	False	False	False	False
7	False	False	False	False	False	False	False	False	False
8	False	False	False	False	False	False	False	False	False
9	False	False	False	False	False	False	False	False	False
10	False	False	False	False	False	False	False	False	False
11	False	False	False	False	False	False	False	False	False
12	False	False	False	False	False	False	False	False	False
13	False	False	False	False	False	False	False	False	False
14	False	False	False	False	False	False	False	False	False
15	False	False	False	False	False	False	False	False	False
16	False	False	False	False	False	False	False	False	False
17	False	False	False	False	False	False	False	False	False
18	False	False	False	False	False	False	False	False	False
19	False	False	False	False	False	False	False	False	False
20	False	False	False	False	False	False	False	False	False
21	False	False	False	False	False	False	False	False	False
22	False	False	False	False	False	False	False	False	False
23	False	False	False	False	False	False	False	False	False
24	False	False	False	False	False	False	True	True	False
25	False	False	False	False	False	False	False	False	False
26	False	False	False	False	False	False	False	False	False
27	False	False	False	False	False	False	False	False	False
28	False	False	False	False	False	False	False	False	False
29	False	False	False	False	False	False	False	False	False
30	False	False	False	False	False	False	False	False	False
31	False	False	False	False	False	False	False	False	False
32	False	False	False	False	False	False	False	False	False
33	False	False	False	False	False	False	True	True	False
34	False	False	False	False	False	False	False	False	False
35	False	False	False	False	False	False	False	False	False
36	False	False	False	False	False	False	True	True	False
37	False	False	False	False	False	False	True	True	False
38	False	False	False	False	False	False	False	False	False
39	False	False	False	False	False	False	False	False	False
40	False	False	False	False	False	False	False	False	False
41	False	False	False	False	False	False	True	True	False
42	False	False	False	False	False	False	False	False	False
43	False	False	False	False	False	False	False	False	False
44	False	False	False	False	False	False	False	False	False
45	False	False	False	False	False	False	True	True	False
46	False	False	False	False	False	False	False	False	False
47	False	False	False	False	False	False	False	False	False
48	False	False	False	False	False	False	False	False	False
49	False	False	False	False	False	False	False	False	False
50	False	False	False	False	False	False	False	False	False
51	False	False	False	False	False	False	False	False	False

52	False	False	False	False	False	False	False	False	False
53	False	False	False	False	False	False	False	False	False
54	False	False	False	False	False	False	False	False	False
55	False	False	False	False	False	False	False	False	False
56	False	False	False	False	False	False	False	False	False
57	False	False	False	False	False	False	False	False	False
58	False	False	False	False	False	False	False	False	False
59	False	False	False	False	False	False	False	False	False
60	False	False	False	False	False	False	False	False	False
61	False	False	False	False	False	False	False	False	False
62	False	False	False	False	False	False	False	False	False
63	False	False	False	False	False	False	False	False	False
64	False	False	False	False	False	False	False	False	False
65	False	False	False	False	False	False	False	False	False
66	False	False	False	False	False	False	False	False	False
67	False	False	False	False	False	False	False	False	False
68	False	False	False	False	False	False	False	False	False
69	False	False	False	False	False	False	False	False	False
70	False	False	False	False	False	False	False	False	False
71	False	False	False	False	False	False	False	False	False
72	False	False	False	False	False	False	False	False	False
73	False	False	False	False	False	False	False	False	False
74	False	False	False	False	False	False	False	False	False
75	False	False	False	False	False	False	False	False	False
76	False	False	False	False	False	False	False	False	False
77	False	False	False	False	False	False	False	False	False
78	False	False	False	False	False	False	False	False	False
79	False	False	False	False	False	False	False	False	False
80	False	False	False	False	False	False	False	False	False
81	False	False	False	False	False	False	False	False	False
82	False	False	False	False	False	False	False	False	False
83	False	False	False	False	False	False	False	False	False
84	False	False	False	False	False	False	False	False	False
85	False	False	False	False	False	False	False	False	False
86	False	False	False	False	False	False	False	False	False
87	False	False	False	False	False	False	False	False	False
88	False	False	False	False	False	False	False	False	False
89	False	False	False	False	False	False	False	False	False
90	False	False	False	False	False	False	False	False	False
91	False	False	False	False	False	False	False	False	False
92	False	False	False	False	False	False	False	False	False
93	False	False	False	False	False	False	False	False	False
94	False	False	False	False	False	False	False	False	False
95	False	False	False	False	False	False	False	False	False
96	False	False	False	False	False	False	False	False	False
97	False	False	False	False	False	False	False	False	False
98	False	False	False	False	False	False	False	False	False
99	False	False	False	False	False	False	False	False	False
100	False	False	False	False	False	False	False	False	False
101	False	False	False	False	False	False	False	False	False
102	False	False	False	False	False	False	False	False	False
103	False	False	False	False	False	False	False	False	False
104	False	False	False	False	False	False	False	False	False
105	False	False	False	False	False	False	False	False	False
106	False	False	False	False	False	False	False	False	False
107	False	False	False	False	False	False	False	False	False
108	False	False	False	False	False	False	False	False	False
109	False	False	False	False	False	False	False	False	False
110	False	False	False	False	False	False	False	False	False
111	False	False	False	False	False	False	False	False	False
112	False	False	False	False	False	False	False	False	False
113	False	False	False	False	False	False	False	False	False
114	False	False	False	False	False	False	False	False	False
115	False	False	False	False	False	False	False	False	False
116	False	False	False	False	False	False	False	False	False
117	False	False	False	False	False	False	False	False	False

```
118      False    False  False   False  False  False  False  False  False
119      False    False  False   False  False  False  False  False  False
120      False    False  False   False  False  False  False  False  False
121      False    False  False   False  False  False  False  False  False
122      False    False  False   False  False  False  False  False  False
123      False    False  False   False  False  False  False  False  False
124      False    False  False   False  False  False  False  False  False
125      False    False  False   False  False  False  False  False  False
126      False    False  False   False  False  False  False  False  False
127      False    False  False   False  False  False  False  False  False
128      False    False  False   False  False  False  False  False  False
129      False    False  False   False  False  False  False  False  False
130      False    False  False   False  False  False  False  False  False
131      False    False  False   False  False  False  False  False  False
132      False    False  False   False  False  False  False  False  False
133      False    False  False   False  False  False  False  False  False
134      F
```

limit_output extension: Maximum message size of 10000 exceeded with 41399 characters

In [12]:

```
# Eliminem les files que contenen algun valor nul i comprobem que s'han eliminat, de manei

women = women.dropna(subset=["Caps", "Goals"])

women.count()
```

Out[12]:

```
Squad no.    520
Country      520
Pos.         520
Player       520
DOB          520
Age          520
Caps         520
Goals        520
Club         520
dtype: int64
```

In [13]:

```
women.columns
```

Out[13]:

```
Index(['Squad no.', 'Country', 'Pos.', 'Player', 'DOB', 'Age', 'Caps', 'Goals',
       'Club'],
      dtype='object')
```

In [14]:

```
# Utilitzo la fòrmula següent per tenir una idea de si tinc alguna categoria minoritària a

numSamarreta = pd.value_counts(women['Squad no.'], sort = True)

display(numSamarreta)
```

```
12    24
17    24
4     24
5     24
8     24
9     24
10    24
3     23
6     23
7     23
20    23
18    23
22    22
1     22
13    22
```

```
2     22
11    22
23    22
15    21
16    21
14    21
19    21
21    21
Name: Squad no., dtype: int64
```

In [15]:

```
pais = pd.value_counts(women['Country'], sort = True)

display(pais)
```

```
US              23
Argentina       23
Norway          23
South Korea     23
Germany         23
South Africa    23
Spain           23
Australia       23
Brazil          23
Italy           23
Jamaica         23
England         23
Japan           23
Scotland        23
Canada          23
Netherlands     23
New Zealand     23
Chile           23
Sweden          23
France          23
Thailand        17
Cameroon        15
China PR        14
Nigeria         14
Name: Country, dtype: int64
```

In [16]:

```
posicio = pd.value_counts(women['Pos.'], sort = True)

display(posicio)
```

```
DF    166
MF    156
FW    132
GK     66
Name: Pos., dtype: int64
```

In [17]:

```
edat = pd.value_counts(women['Age'], sort = True)

display(edat)
```

```
26    59
25    52
28    49
27    47
24    44
23    37
29    33
22    30
31    30
```

```
30    30
21    26
20    17
33    15
32    10
34    10
18     8
19     7
17     5
35     4
36     3
37     1
41     1
16     1
39     1
Name: Age, dtype: int64
```

In [18]:

```
partitsJugats = pd.value_counts(women['Caps'], sort = True)

display(partitsJugats)
```

```
3.0      18
20.0     17
2.0      15
6.0      13
0.0      12
4.0      12
13.0     12
1.0      11
8.0      11
45.0     10
21.0     10
14.0      9
7.0       9
12.0      9
11.0      8
40.0      8
9.0       8
26.0      8
16.0      8
18.0      7
34.0      7
5.0       7
43.0      7
19.0      7
31.0      7
24.0      7
22.0      6
35.0      6
37.0      6
42.0      6
17.0      6
10.0      6
61.0      6
62.0      5
28.0      5
15.0      5
77.0      5
71.0      5
50.0      5
48.0      4
30.0      4
123.0     4
59.0      4
49.0      4
```

66.0	4
29.0	4
25.0	4
32.0	4
23.0	4
95.0	3
36.0	3
58.0	3
63.0	3
104.0	3
74.0	3
60.0	3
69.0	3
39.0	3
64.0	3
76.0	3
65.0	3
108.0	3
38.0	3
52.0	3
57.0	3
99.0	3
46.0	3
87.0	3
44.0	3
115.0	3
96.0	2
47.0	2
80.0	2
79.0	2
160.0	2
85.0	2
53.0	2
147.0	2
83.0	2
135.0	2
27.0	2
56.0	2
113.0	2
139.0	2
72.0	2
81.0	2
126.0	2
88.0	2
89.0	1
78.0	1
51.0	1
41.0	1
133.0	1
117.0	1
116.0	1
106.0	1
125.0	1
73.0	1
86.0	1
131.0	1
129.0	1
152.0	1
54.0	1
166.0	1
120.0	1
186.0	1
155.0	1
92.0	1
159.0	1
93.0	1

```
33.0      1
122.0     1
138.0     1
82.0      1
70.0      1
55.0      1
91.0      1
161.0     1
193.0     1
102.0     1
100.0     1
175.0     1
109.0     1
157.0     1
143.0     1
282.0     1
184.0     1
128.0     1
191.0     1
98.0      1
132.0     1
150.0     1
112.0     1
103.0     1
124.0     1
111.0     1
75.0      1
271.0     1
134.0     1
154.0     1
Name: Caps, dtype: int64
```

In [19]:

```
gols = pd.value_counts(women['Goals'], sort = True)

display(gols)
```

```
0.0      202
1.0       58
3.0       31
2.0       28
6.0       17
5.0       17
9.0       14
8.0       14
7.0       12
10.0      11
4.0       11
11.0       9
17.0       9
12.0       7
14.0       6
13.0       5
16.0       5
15.0       5
25.0       4
20.0       4
22.0       4
27.0       3
18.0       3
28.0       3
19.0       3
24.0       3
23.0       3
32.0       3
58.0       2
```

```
29.0       2
57.0       2
47.0       2
30.0       2
110.0      1
21.0       1
54.0       1
45.0       1
61.0       1
31.0       1
83.0       1
33.0       1
42.0       1
44.0       1
53.0       1
101.0      1
49.0       1
107.0      1
181.0      1
73.0       1
Name: Goals, dtype: int64
```

In [20]:

```
club = pd.value_counts(women['Club'], sort = True)

display(club)
```

```
Barcelona                          15
Lyon                               14
Manchester City                    12
Chelsea                            12
Unattached                         11
Incheon Hyundai Steel Red Angels   11
Nippon TV Beleza                   10
Bayern Munich                      10
Arsenal                             9
Portland Thorns                     9
Paris Saint-Germain                 9
LSK Kvinner                         8
Juventus                            8
Reign FC                            8
Orlando Pride                       8
Atlético Madrid                     8
North Carolina Courage              7
Milan                               7
Houston Dash                        7
VfL Wolfsburg                       7
Bundit Asia                         7
Montpellier                         7
Kopparbergs/Göteborg                6
UAI Urquiza                         6
INAC Kobe Leonessa                  6
Utah Royals                         6
Guingamp                            5
Glasgow City                        5
Washington Spirit                   5
SGS Essen                           5
Roma                                5
Chicago Red Stars                   5
Linköping                           5
Rosengård                           5
Arna-Bjørnar                        4
Beijing Phoenix                     4
Bangkok                             4
Ajax                                4
West Ham United                     4
```

SC Freiburg	4
Vittsjö	4
Levante	4
Hwacheon KSPO	4
Fiorentina	4
Sandviken	3
Urawa Red Diamonds	3
Mamelodi Sundowns	3
Rayo Vallecano	3
Birmingham City	3
Paris FC	3
Manchester United	3
Eskilstuna United	3
Turbine Potsdam	3
Sevilla	3
Guangdong Huijun	3
Colo-Colo	3
Sporting Huelva	3
Suwon UDC	3
Corinthians	3
Melbourne Victory	3
Chonburi Sriprathum	3
Stabæk	3
Bordeaux	3
Klepp	2
Sky Blue FC	2
JVW	2
Gintra Universitetas	2
Benfica	2
Piteå	2
Reading	2
Boca Juniors	2
Hibernian	2
River Plate	2
Rivers Angels	2
MaIndies	2
Real Sociedad	2
Rosario Central	2
Brisbane Roar	2
California Golden Bears	2
Dijon	2
Granadilla	2
Papakura City	2
Djurgården	2
Western Springs	2
Jiangsu Suning [zh]	2
Miramar Rangers	2
Shanghai [zh]	2
Everton	2
Kolbotn	2
Vålerenga	2
Cáceres	2
Avaldsnes	2
Zaragoza	2
Henan Huishang	2
Växjö	2
Vancouver Whitecaps	2
UCLA Bruins	2
Santiago Morning	2
Amazone FAP	2
Gyeongju KHNP	2
Dalian Quanjian	2
Internacional	1
Kindermann	1
Gumi Sportstoto	1
Fortuna Hjørring	1

São Paulo	1
ChievoVerona Valpo [it]	1
SC Sand	1
Florentia	1
Changchun Zhuoyue [zh]	1
Málaga	1
Newcastle Jets	1
Sydney FC	1
Bankstown City	1
Changnyeong	1
Athletic Bilbao	1
Southeastern Fire	1
Wuhan Jianghan University [zh]	1
University of Pretoria	1
Kristianstads	1
Tshwane University of Technology	1
Golden Stars	1
University of KwaZulu-Natal	1
University of the Western Cape	1
Bloemfontein Celtic	1
University of Johannesburg	1
Sophakama Ladies/HPC	1
Liverpool	1
United Soccer Alliance	1
Pink Sport Time	1
KIF Örebro	1
Florida State Seminoles	1
Texas Longhorns	1
Florida Gators	1
Bristol City	1
1. FFC Frankfurt	1
Twente	1
ADO Den Haag	1
Real Betis	1
Three Kings United	1
Onehunga Sports	1
KR Reykjavik	1
MSV Duisburg	1
Universidad de Chile	1
Curicó Unido	1
Universidad Católica	1
Valencia	1
3B da Amazônia [pt]	1
Slavia Praha	1
Santos	1
Wolfsburg	1
Kennesaw State Owls	1
Air Force United	1
Nancy	1
Strasbourg	1
CSKA Moscow	1
Trondheims-Ørn	1
PSV	1
West Florida Argonauts	1
Sion Swifts	1
Maccabi Kishronot Hadera	1
Tennessee Volunteers	1
Montverde Academy	1
Szent Mihály	1
UCF Knights	1
Rutgers Scarlet Knights	1
Memphis Tigers	1
UNC Wilmington Seahawks	1
Ambilly [fr]	1
Granada	1
Tacón	1

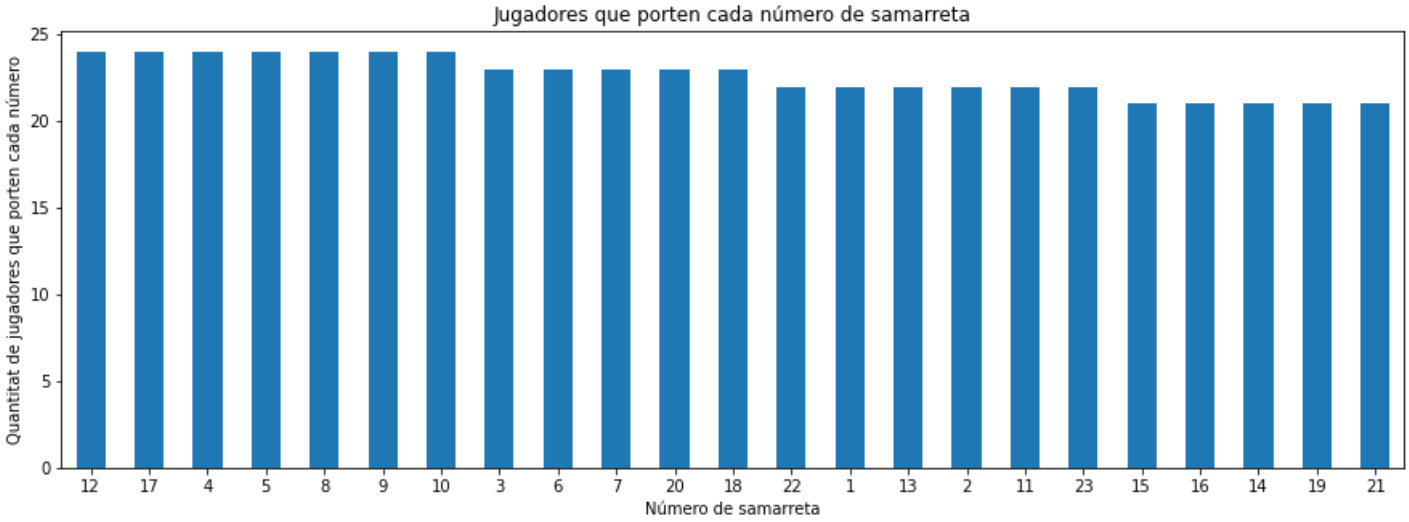
```
Logroño                          1
Madrid CFF                       1
Albirex Niigata                  1
AC Nagano Parceiro               1
Vegalta Sendai                   1
AWA Yaoundé                      1
Be?ikta?                         1
Saint-Malo                       1
Fleury                           1
Name: Club, dtype: int64
```

In [21]:

```
# I per fer els resultats anteriors més visuals els converteixo en gràfics
plt.figure(figsize=(15,5))

numSamarreta.plot(kind = "bar", rot=0)

plt.title("Jugadores que porten cada número de samarreta")
plt.xlabel("Número de samarreta")
plt.ylabel("Quantitat de jugadores que porten cada número");
```

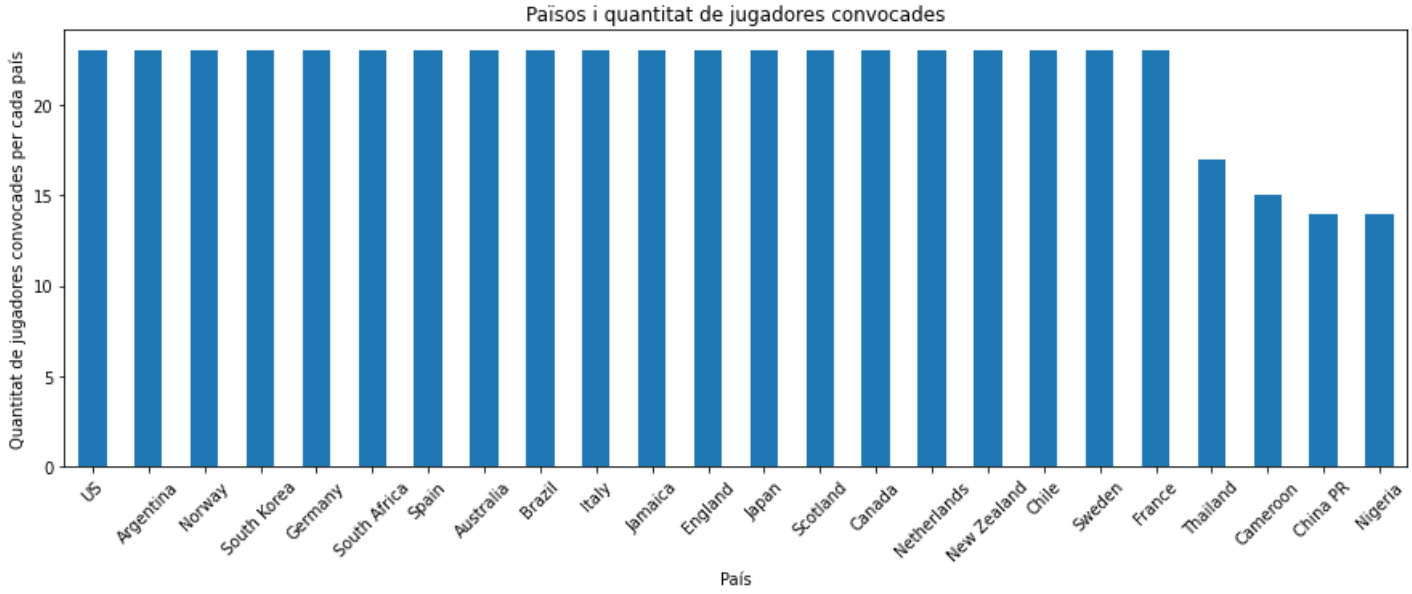


In [22]:

```
plt.figure(figsize=(15,5))

pais.plot(kind = "bar", rot=0)

plt.xticks(rotation=45)
plt.title("Països i quantitat de jugadores convocades")
plt.xlabel("País")
plt.ylabel("Quantitat de jugadores convocades per cada país");
```

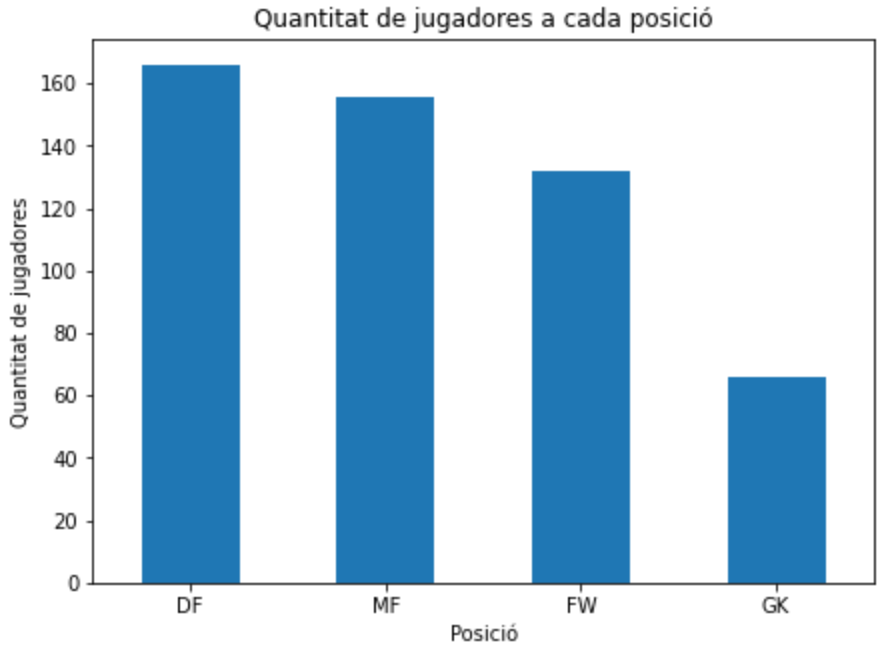


In [23]:

```
plt.figure(figsize=(7,5))

posicio.plot(kind = "bar", rot=0)

plt.title("Quantitat de jugadores a cada posició")
plt.xlabel("Posició")
plt.ylabel("Quantitat de jugadores");
```

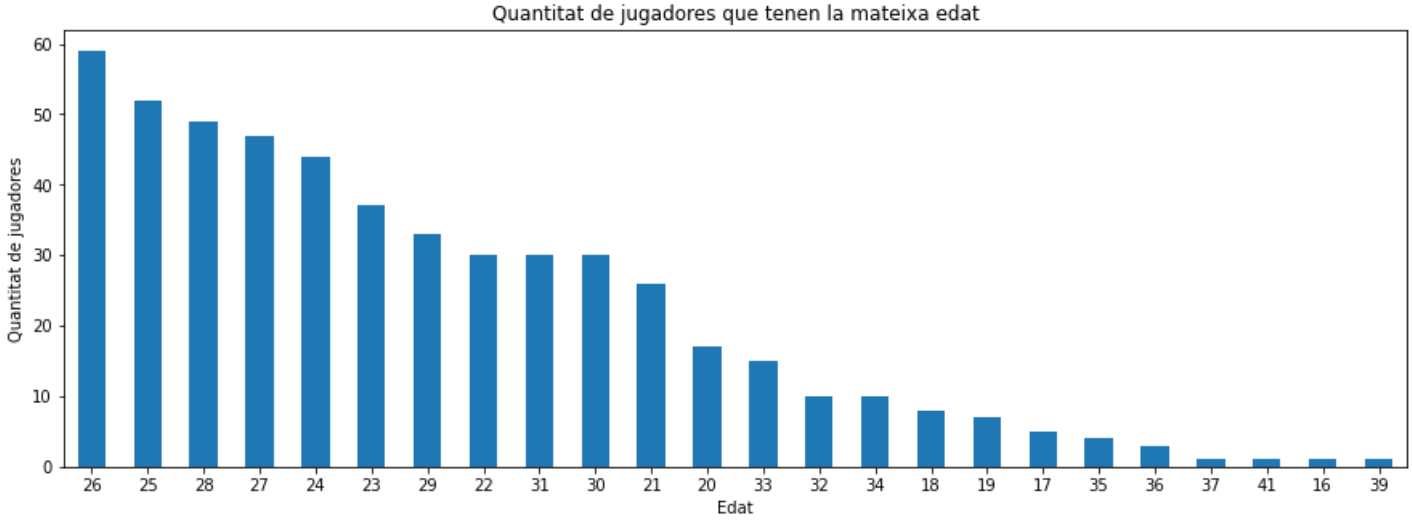


In [24]:

```
plt.figure(figsize=(15,5))

edat.plot(kind = "bar", rot=0)

plt.title("Quantitat de jugadores que tenen la mateixa edat")
plt.xlabel("Edat")
plt.ylabel("Quantitat de jugadores");
```

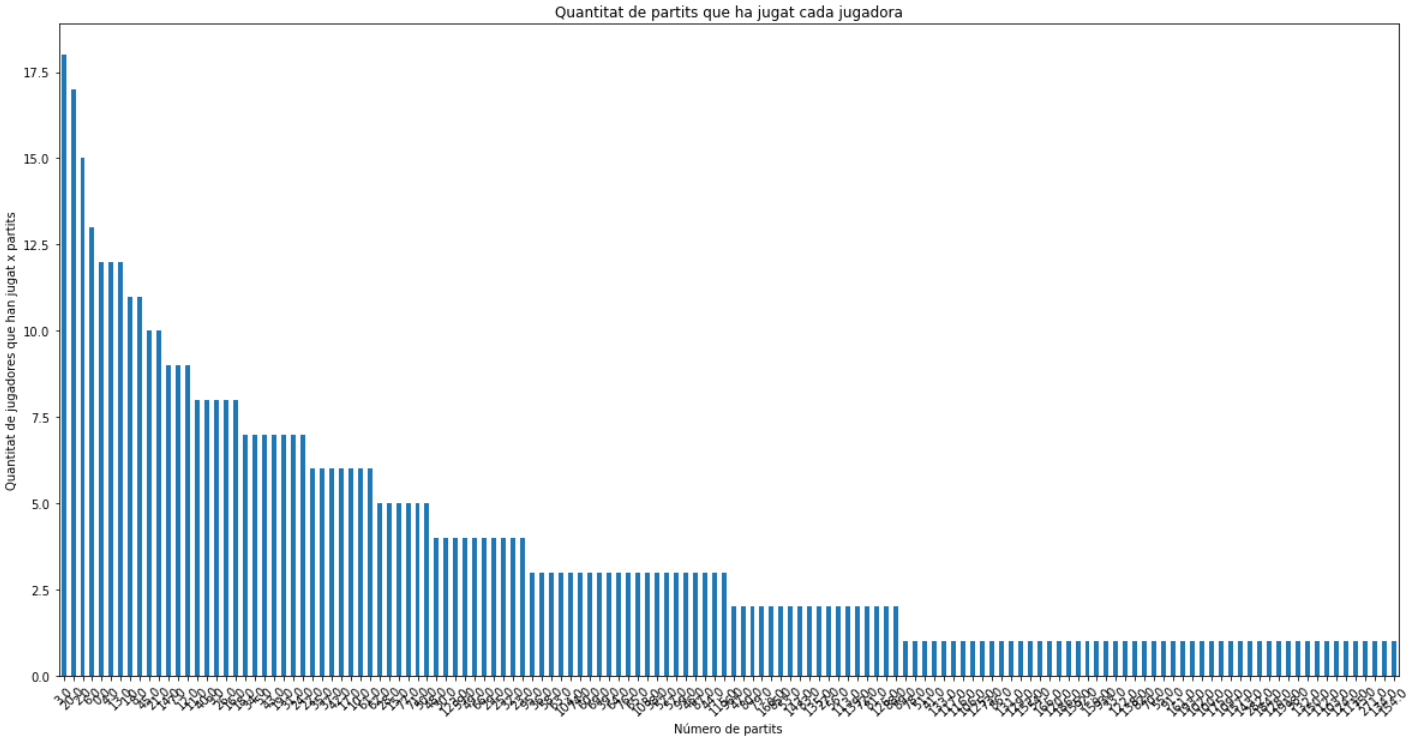


In [25]:

```
plt.figure(figsize=(20,10))

partitsJugats.plot(kind = "bar", rot=0)

plt.xticks(rotation=45)
plt.title("Quantitat de partits que ha jugat cada jugadora")
plt.xlabel("Número de partits")
plt.ylabel("Quantitat de jugadores que han jugat x partits");
```

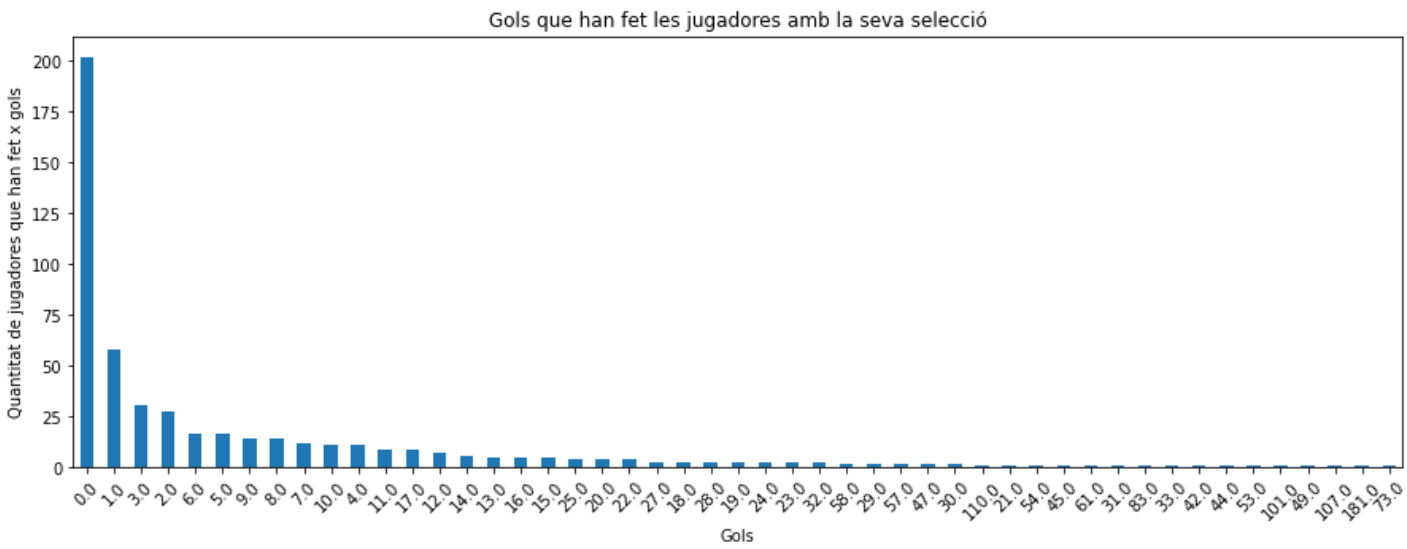


In [26]:

```
plt.figure(figsize=(15,5))

gols.plot(kind = "bar", rot=0)

plt.xticks(rotation=45)
plt.title("Gols que han fet les jugadores amb la seva selecció")
plt.xlabel("Gols")
plt.ylabel("Quantitat de jugadores que han fet x gols");
```

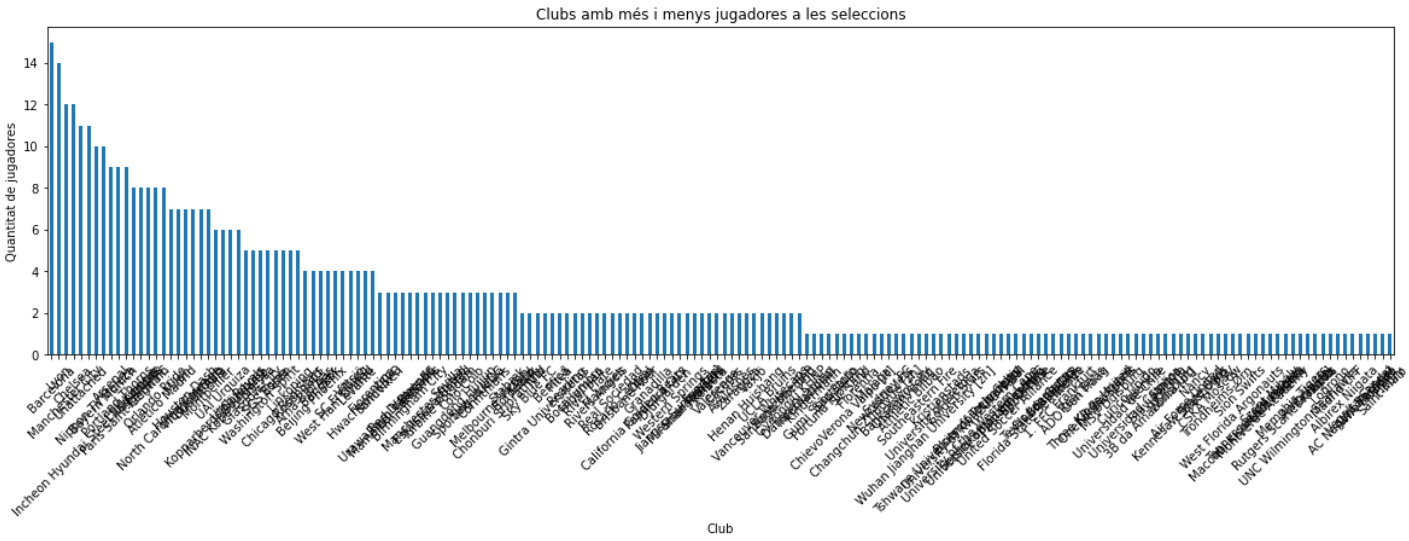


In [27]:

```
plt.figure(figsize=(20,5))

club.plot(kind = "bar", rot=0)

plt.xticks(rotation=45)
plt.title("Clubs amb més i menys jugadores a les seleccions")
plt.xlabel("Club")
plt.ylabel("Quantitat de jugadores");
```



Ara divideixo el dataset original en dos, un que serà el df amb les variables explicatives (x) i l'altre amb la variable objectiu (Y)

In [28]:

```
# Ho he provat tot i no hi ha manera d'aconseguir que em fucioni, l'error clau és el següe
#   ValueError: could not convert string to float: 'US'

'''
from imblearn.over_sampling import SMOTE

X = women[["Squad no.", "Country", "Pos.", 'Player', "DOB", "Age", "Caps", "Club"]]
y = women["Goals"]

sm = SMOTE()
X_res, y_res = sm.fit_resample(X,y)

df_over = X_res
df_over["Goals"] = y_res
'''
```

Out[28]:
```
'\nfrom imblearn.over_sampling import SMOTE\n\nX = women[["Squad no.", "Country", "Pos.",
\'Player\', "DOB", "Age", "Caps", "Club"]]\ny = women["Goals"]\n\nsm = SMOTE()\nX_res, y_r
es = sm.fit_resample(X,y)\n\ndf_over = X_res\ndf_over["Goals"] = y_res\n'
```

NIVELL 3

Exercici 3

Continua amb el conjunt de dades de tema esportiu i genera una mostra utilitzant el mètode Reservoir sampling.

El RESERVOIR SAMPLING s'utilitza en la mineria de dades per obtenir una mostra de mida n a partir d'un flux de dades de longitud desconeguda

In [35]:
```
# Em surt l'error següent i no sé entendre què vol dir:
# KeyError: 0

'''
# Preparo les dades
n = len(women)
k = 100

mostra_reservoir = []  #creació de la llista buida on s'hi aniran posant els valors aleato

# Inicialització de mostra_reservoir
for i in range (k):
    mostra_reservoir.append(women[i])

# Iteració pel dataset fins a n-1
for j in range (k,n):
    index = random.randint(0,j)
    if index < k:
        mostra_reservoir[index] = women[j]
print("Input array:")
print(women)
print("Output array:")
print(mostra_reservoir)
'''
```

Out[35]:
```
'\n# Preparo les dades\nn = len(women)\nk = 100\n\nmostra_reservoir = []  #creació de la l
lista buida on s\'hi aniran posant els valors aleatoris que vagin sortint per la mostra\n
\n# Inicialització de mostra_reservoir\nfor i in range (k):\n    mostra_reservoir.append(w
omen[i])\n\n# Iteració pel dataset fins a n-1    \nfor j in range (k,n):\n    index = rand
om.randint(0,j)\n    if index < k:\n        mostra_reservoir[index] = women[j]\nprint("Inp
ut array:")\nprint(women)\nprint("Output array:")\nprint(mostra_reservoir)\n'
```

In [39]:
```
# He provat de fer-ho d'una altra manera i em surt el mateix error:
# KeyError: 0

'''

# Creo funció per imprimir la array
def imprimirArray(women,n):
    for i in range(n):
        print(women[i],end=" ")
    print()

# Creo funció per seleccionar aleatòriament els elements que vull afegir a la mostra
def seleccioMostra(women, n, k):
    i = 0
    reservoir = [0]*k  # creació i inicialització d'on es guardarà la mostra
```

```
    for i in range (k):
        reservoir[i] = women[i]

    # iteració per tot el dataset
    while (i<n):
        j = random.randint(i+1)
        if (j<k):   # si el número aleatòri es més petit que k, s'agafa el corresponent de
            reservoir[j] = women[i]
        i+=1
    print(women, k)

# Codi amb la info que necessitem perquè funcioni tot lo anterior
if __name__ == "__main__":
    women = women;
    n = len(women);
    k = 100;
    seleccioMostra(women, n, k);

'''
```

Out[39]:

```
'\n\n# Creo funció per imprimir la array\ndef imprimirArray(women,n):\n    for i in range
(n):\n        print(women[i],end=" ")\n    print()\n\n# Creo funció per seleccionar aleatò
riament els elements que vull afegir a la mostra \ndef seleccioMostra(women, n, k):\n    i
= 0\n    reservoir = [0]*k  # creació i inicialització d\'on es guardarà la mostra\n    fo
r i in range (k):\n        reservoir[i] = women[i]\n    \n    # iteració per tot el dataset
\n    while (i<n):\n        j = random.randint(i+1)\n        if (j<k):   # si el número al
eatòri es més petit que k, s\'agafa el corresponent del dataset i s\'afegeix a la mostra\n
            reservoir[j] = women[i]\n        i+=1\n    print(women, k)\n    \n# Codi amb l
a info que necessitem perquè funcioni tot lo anterior    \nif __name__ == "__main__":\n
women = women;\n    n = len(women);\n    k = 100;\n    seleccioMostra(women, n, k);\n\n'
```

In [ ]: