Ejercicio Masajeo de Datos I

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

In [ ]:
```
cwd = os.getcwd()
cwd
```

Out[ ]:
```
'e:\\WORK IN PROGRESS\\Data Analytics course\\parte 2 python\\week 16'
```

In [ ]:
```
os.chdir('e:\\WORK IN PROGRESS\\Data Analytics course\\parte 2 python\\week 16')
```

In [ ]:
```
cwd = os.getcwd()
cwd
```

Out[ ]:
```
'e:\\WORK IN PROGRESS\\Data Analytics course\\parte 2 python\\week 16'
```

In [ ]:
```
df = pd.read_csv('fifa_eda.csv')
```

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

Out[ ]:
```
(18207, 18)
```

1. Mostrar las primeras 20 filas del archivo, las últimas 5 y un sample de 10

In [ ]:
```
# Las primeras 20 filas del archivo
df.head(20)
```

Out[ ]:

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	International Reputation	Skill Moves	Position	Joined	Contract Valid Until
0	158023	L. Messi	31	Argentina	94	94	FC Barcelona	110500.0	565.0	Left	5.0	4.0	RF	2004	2021-01-01
1	20801	Cristiano Ronaldo	33	Portugal	94	94	Juventus	77000.0	405.0	Right	5.0	5.0	ST	2018	2022-01-01
2	190871	Neymar Jr	26	Brazil	92	93	Paris Saint-Germain	118500.0	290.0	Right	5.0	5.0	LW	2017	2022-01-01
3	193080	De Gea	27	Spain	91	93	Manchester United	72000.0	260.0	Right	4.0	1.0	GK	2011	2020-01-01
4	192985	K. De Bruyne	27	Belgium	91	92	Manchester City	102000.0	355.0	Right	4.0	4.0	RCM	2015	2023-01-01
5	183277	E. Hazard	27	Belgium	91	91	Chelsea	93000.0	340.0	Right	4.0	4.0	LF	2012	2020-01-01
6	177003	L. Modrić	32	Croatia	91	91	Real Madrid	67000.0	420.0	Right	4.0	4.0	RCM	2012	2020-01-01
7	176580	L. Suárez	31	Uruguay	91	91	FC Barcelona	80000.0	455.0	Right	5.0	3.0	RS	2014	2021-01-01
8	155862	Sergio Ramos	32	Spain	91	91	Real Madrid	51000.0	380.0	Right	4.0	3.0	RCB	2005	2020-01-01
9	200389	J. Oblak	25	Slovenia	90	93	Atlético Madrid	68000.0	94.0	Right	3.0	1.0	GK	2014	2021-01-01
10	188545	R. Lewandowski	29	Poland	90	90	FC Bayern München	77000.0	205.0	Right	4.0	4.0	ST	2014	2021-01-01
11	182521	T. Kroos	28	Germany	90	90	Real Madrid	76500.0	355.0	Right	4.0	3.0	LCM	2014	2022-01-01
12	182493	D. Godín	32	Uruguay	90	90	Atlético Madrid	44000.0	125.0	Right	3.0	2.0	CB	2010	2019-01-01
13	168542	David Silva	32	Spain	90	90	Manchester City	60000.0	285.0	Left	4.0	4.0	LCM	2010	2020-01-01
14	215914	N. Kanté	27	France	89	90	Chelsea	63000.0	225.0	Right	3.0	2.0	LDM	2016	2023-01-

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	International Reputation	Skill Moves	Position	Joined	Contract Valid Until
															01
15	211110	P. Dybala	24	Argentina	89	94	Juventus	89000.0	205.0	Left	3.0	4.0	LF	2015	2022-01-01
16	202126	H. Kane	24	England	89	91	Tottenham Hotspur	83500.0	205.0	Right	3.0	3.0	ST	2010	2024-01-01
17	194765	A. Griezmann	27	France	89	90	Atlético Madrid	78000.0	145.0	Left	4.0	4.0	CAM	2014	2023-01-01
18	192448	M. ter Stegen	26	Germany	89	92	FC Barcelona	58000.0	240.0	Right	3.0	1.0	GK	2014	2022-01-01
19	192119	T. Courtois	26	Belgium	89	90	Real Madrid	53500.0	240.0	Left	4.0	1.0	GK	2018	2024-01-01

In [ ]:
```
# Las últimas 5 filas del archivo
df.tail(5)
```

Out[ ]:

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	International Reputation	Skill Moves	Position	Joined	Contrac Vali Unt
18202	238813	J. Lundstram	19	England	47	65	Crewe Alexandra	60.0	1.0	Right	1.0	2.0	CM	2017	2019-01 0
18203	243165	N. Christoffersson	19	Sweden	47	63	Trelleborgs FF	60.0	1.0	Right	1.0	2.0	ST	2018	2020-01 0
18204	241638	B. Worman	16	England	47	67	Cambridge United	60.0	1.0	Right	1.0	2.0	ST	2017	2021-01 0
18205	246268	D. Walker-Rice	17	England	47	66	Tranmere Rovers	60.0	1.0	Right	1.0	2.0	RW	2018	2019-01 0
18206	246269	G. Nugent	16	England	46	66	Tranmere Rovers	60.0	1.0	Right	1.0	2.0	CM	2018	2019-01 0

In [ ]:
```
# 10 muestras del archivo
df.sample(10)
```

Out[ ]:

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	International Reputation	Skill Moves	Position	Joined	Contract Valid Until
9421	232847	Y. Akimoto	30	Japan	66	66	Shonan Bellmare	425.0	1.0	Right	1.0	1.0	GK	2017	2019-01-01
9832	208111	S. Sandberg	24	Sweden	66	71	Hammarby IF	725.0	2.0	Right	1.0	2.0	RB	2018	2019-01-01
12148	53778	M. Richards	35	England	63	63	Swindon Town	170.0	3.0	Right	1.0	2.0	ST	2018	2019-01-01
16527	239781	T. Arndal	20	Denmark	57	68	AC Horsens	180.0	1.0	Right	1.0	2.0	LM	2017	2021-01-01
16551	245173	O. Mireles	19	Mexico	57	70	Club León	170.0	2.0	Right	1.0	2.0	CB	2018	2021-01-01
2394	204699	L. Melgarejo	27	Paraguay	74	74	Spartak Moscow	6000.0	1.0	Left	1.0	4.0	LM	2016	2020-01-01
4971	210236	M. Hefele	27	Germany	70	72	Nottingham Forest	1700.0	24.0	Right	1.0	2.0	CB	2018	2021-01-01
12047	238843	T. Lebeau	19	France	64	74	Chamois Niortais Football Club	700.0	1.0	Right	1.0	2.0	CAM	2017	2021-01-01
15044	246427	M. Berden	20	Netherlands	60	70	Vitesse	325.0	2.0	Right	1.0	3.0	RW	2018	2019-01-01
6362	216514	B. Sagredo	29	Chile	69	69	San Luis de Quillota	1000.0	3.0	Left	1.0	3.0	LM	2015	2019-01-01

2. Generar data estadística con .describe() y además los tipos de datos del dataset

In [ ]:
```
# Tipos de datos del dataset
df.dtypes
```

Out[ ]:
```
ID                          int64
Name                       object
Age                         int64
Nationality                object
Overall                     int64
Potential                   int64
Club                       object
Value                     float64
Wage                      float64
Preferred Foot             object
International Reputation  float64
Skill Moves               float64
Position                   object
Joined                      int64
Contract Valid Until       object
Height                    float64
Weight                    float64
Release Clause            float64
dtype: object
```

In [ ]:
```
# Estadìstica descriptiva del dataset
df.describe()
```

Out[ ]:

	ID	Age	Overall	Potential	Value	Wage	International Reputation	Skill Moves	Joined	Height
count	18207.000000	18207.000000	18207.000000	18207.000000	17955.000000	18207.000000	18159.000000	18159.000000	18207.000000	18207.000000
mean	214298.338606	25.122206	66.238699	71.307299	2444.530214	9.731312	1.113222	2.361308	2016.420607	5.946771
std	29965.244204	4.669943	6.908930	6.136496	5626.715434	21.999290	0.394031	0.756164	2.018194	0.220514
min	16.000000	16.000000	46.000000	48.000000	10.000000	0.000000	1.000000	1.000000	1991.000000	5.083333
25%	200315.500000	21.000000	62.000000	67.000000	325.000000	1.000000	1.000000	2.000000	2016.000000	5.750000
50%	221759.000000	25.000000	66.000000	71.000000	700.000000	3.000000	1.000000	2.000000	2017.000000	5.916667
75%	236529.500000	28.000000	71.000000	75.000000	2100.000000	9.000000	1.000000	3.000000	2018.000000	6.083333
max	246620.000000	45.000000	94.000000	95.000000	118500.000000	565.000000	5.000000	5.000000	2018.000000	6.750000

3. Si es necesario, pasar a numéricas por lo menos 2 columnas que contengan números para

incluirlas en .describe(). Aplicar las técnicas aprendidas.

In [ ]:
```
# Al revisar el tipo de cada columna del dataset, Constate que solo se le podìa
# hacer el cambio de tipo de datos a la columna "Contract Valid Until" pero en
# formato numèrico sino a formato fecha. Por lo cual procedì usando # la funciòn
# to_datetime con el formato AA-MM-DD asì como estaba en el dataset en formato
#  object.

# No considerè necesario realizar otros cambios.

df['Contract Valid Until'] = pd.to_datetime(df['Contract Valid Until'], format = "%Y-%m-%d")
```

In [ ]:
```
# Aqui se puede observar que los datos de la columna 'Contract Valid Until'
# tiene otro formato.

df.dtypes
```

Out[ ]:
```
ID                                 int64
Name                              object
Age                                int64
Nationality                       object
Overall                            int64
Potential                          int64
Club                              object
Value                            float64
Wage                             float64
Preferred Foot                    object
International Reputation         float64
Skill Moves                      float64
Position                          object
Joined                             int64
Contract Valid Until        datetime64[ns]
Height                           float64
Weight                           float64
Release Clause                   float64
dtype: object
```

3. Añadir una columna “Years Playing”, que calcule el año actual menos la columna “Joined”.

In [ ]:
```
# Para este punto, como el archivo no esta actualizado (por ejemplo Messi
# actualmente esta jugando en el Paris Saint Germain), # lo que hice fue
# primero hallar una columna binaria que dijerà "si" si el contrato que el
# jugador firmò cuando ingresò al club en el cual jugarìa todavìa era valido.
# Es decir, si dicho contrato se acaba despuès del año 2023. De lo contrariò
# tendrìa que decir "no".

# De esta manera no incurro en el error de, por ejemplo en el caso de Messi,
# contarle años adicionales que èl no ha jugado en el Barcelona.

# Asì, solo calculo la columna Years Playing para aquellos jugadores que,
# en teorìa, para la fecha en la cual se hizo este data set estarìan jugando
# actualmente en sus respectivos cubles.

# Procedo generando una nueva columna con la palabra "no".
# Despuès, cambio el tipo de datos de esa columna. La convierto a String.
```

```
df['expired contract'] = 'no'
df['expired contract'] = df['expired contract'].astype(pd.StringDtype())
```

In [ ]:
```
# Utilizo la funciòn count para determinar si los valores de la columna
# "Contract Valid Until" corresponde con el total de registros.
# Encuentro que hay una difencia. Lo que quiere decir que hay valores nulos
# en esa columna.
df.count()
```

Out[ ]:
```
ID                        18207
Name                      18207
Age                       18207
Nationality               18207
Overall                   18207
Potential                 18207
Club                      17966
Value                     17955
Wage                      18207
Preferred Foot            18207
International Reputation  18159
Skill Moves               18159
Position                  18207
Joined                    18207
Contract Valid Until      17918
Height                    18207
Weight                    18207
Release Clause            18207
expired contract          18207
dtype: int64
```

In [ ]:
```
# Para saber exactamente cuantos valores nulos hay en dicha columna utilizo las
# funciones isnull() y sum()
df['Contract Valid Until'].isnull().sum()
```

Out[ ]:
```
289
```

In [ ]:
```
# Para solucionar este problema eliminè aquellos registros que tenian un valor
# nulo en la columna 'Contract Valid Until'.

df.dropna(axis = 0, subset='Contract Valid Until', inplace = True)
```

In [ ]:
```
# De esta manera se eliminaron 289 registros. Por lo cual, las dimensiones del
# data set cambiaron:
```

```
df.shape
```

Out[ ]:
```
(17918, 19)
```

In [ ]:
```
# Ahora procedo con la construcciòn de la columna "expired contract". Cambio el
# "no" por el "si" para aquellos jugadores
# que tuvieron un contrato vigente para antes del 2023.

df.loc[df["Contract Valid Until"].dt.year < 2023, "expired contract"] = 'si'
```

In [ ]:
```
df.sample(10)
```

Out[ ]:

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	International Reputation	Skill Moves	Position	Joined	Contract Valid Until
6756	208422	R. Sanusi	26	Belgium	68	71	Grenoble Foot 38	925.0	3.0	Right	1.0	3.0	CDM	2018	2020-01-01
1870	152211	R. Özcan	34	Austria	75	75	Bayer 04 Leverkusen	2700.0	36.0	Left	1.0	1.0	GK	2016	2019-01-01
10525	223378	A. Çeviker	28	Turkey	65	66	Akhisar Belediyespor	500.0	4.0	Right	1.0	2.0	RDM	2015	2021-01-01
14162	241485	Chen Binbin	20	China PR	61	74	Shanghai SIPG FC	475.0	2.0	Right	1.0	2.0	LW	2017	2019-01-01
4069	209698	B. Touré	26	Mali	71	75	AJ Auxerre	2500.0	5.0	Right	1.0	2.0	RDM	2017	2020-01-01
13425	238182	A. Amaya	17	Colombia	62	82	Atlético Huila	600.0	1.0	Left	1.0	3.0	RM	2017	2021-01-01
10310	244316	Luís Maximiano	19	Portugal	65	80	Sporting CP	950.0	1.0	Right	1.0	1.0	GK	2017	2019-01-01
4735	169978	R. Beerens	30	Netherlands	71	71	Vitesse	2200.0	11.0	Right	1.0	3.0	RW	2018	2021-01-01
11979	244971	D. Furtado	21	France	64	76	Rio Ave FC	800.0	2.0	Right	1.0	3.0	LW	2018	2021-01-01
8438	205750	I. Koné	27	France	67	69	KSV Cercle Brugge	750.0	5.0	Right	1.0	2.0	CDM	2017	2019-01-01

In [ ]:
```
# Ahora creo la columna 'Years Playing'. Inicialmente, la creo con un valor 0
# para todos los campos.
# Adiconalmente, cambio el formato de los datos y lo  paso a numero enteros.
df['Years Playing'] = 0
df['Years Playing'] = df['Years Playing'].astype(pd.Int64Dtype())
```

In [ ]:
```
# Verifico que la columna haya sido añadida.
df.shape
```

Out[ ]: (17918, 20)

In [ ]:
```
# Finalmente, para aquellos jugadores que aùn tienen vigente su contrato realizo
# la resta entre el 2023 y el año en el cual el jugador fue contratado por el club.

df.loc[df["expired contract"] == 'no', "Years Playing"] = 2023-df['Joined']
```

In [ ]:
```
# Para los jugadores que ya se les acabò el contrato tienen como valor en la
# columna "Years Playing" 0 porque no tenemos ninguna certeza de que el contrato
# haya sido renovado o de que los jugadores hayan cambiado de club.

df.sample(5)
```

Out[ ]:

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	International Reputation	Skill Moves	Position	Joined	Contract Valid Until	
3721	225701	P. Gallese	28	Peru	72	73	Tiburones Rojos de Veracruz	2500.0	5.0	Right	1.0	1.0	GK	2016	2019-01-01	6.
1277	193839	Rômulo	31	Italy	76	76	Genoa	6500.0	17.0	Right	1.0	4.0	RM	2018	2020-01-01	5.
18005	229615	J. Heaton	21	England	51	63	St. Mirren	60.0	1.0	Right	1.0	2.0	CB	2018	2021-01-01	6.
17207	236470	C. Sepúlveda	21	Colombia	55	67	Jaguares de Córdoba	140.0	1.0	Right	1.0	2.0	CDM	2018	2021-01-01	5.
9272	224624	Fu Huan	24	China PR	66	71	Shanghai SIPG FC	725.0	4.0	Right	1.0	2.0	RB	2012	2020-01-01	6.

4. Buscar y mostrar a todos los jugadores de Colombia.

In [ ]:
```
# Utilice la funciòn loc para encontrar todos los jugadores de Colombia.
df.loc[df['Nationality']=='Colombia']
```

Out[ ]:

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	International Reputation	Skill Moves	Position	Joined	Contra Vali Unt
28	198710	J. Rodríguez	26	Colombia	88	89	FC Bayern München	69500.0	315.0	Left	4.0	4.0	LAM	2016	2019-0 3
110	220793	D. Sánchez	22	Colombia	84	88	Tottenham Hotspur	34000.0	105.0	Right	2.0	2.0	RCB	2017	2024-0
129	193082	J. Cuadrado	30	Colombia	84	84	Juventus	29500.0	150.0	Right	3.0	5.0	RAM	2015	2020-0
148	167397	Falcao	32	Colombia	84	84	AS Monaco	25000.0	115.0	Right	3.0	3.0	RS	2013	2020-0
346	207664	C. Bacca	31	Colombia	81	81	Villarreal CF	16000.0	45.0	Right	3.0	3.0	LS	2018	2022-0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
18044	246109	K. Lara	16	Colombia	50	74	Atlético Huila	60.0	1.0	Right	1.0	2.0	RB	2018	2021-0
18054	238468	C. Mesa	20	Colombia	50	67	América de Cali	60.0	1.0	Right	1.0	2.0	CB	2018	2021-0
18110	243434	G. Tegue	18	Colombia	50	68	Independiente Medellín	50.0	1.0	Left	1.0	2.0	CB	2018	2021-0
18152	245755	J. Yabur	19	Colombia	49	62	Atlético Nacional	50.0	1.0	Right	1.0	2.0	CM	2018	2021-0
18182	246001	Y. Góez	18	Colombia	48	65	Atlético Nacional	50.0	1.0	Right	1.0	2.0	CDM	2018	2021-0

616 rows × 20 columns

5. Ordenar y mostrar los datos por la columna ReleaseClause (sueldo).

In [ ]:
```
# Lo hice con el nuevo dataset del punto anterior, para los jugadores de Colombia.
```

```
df.loc[df['Nationality']=='Colombia'].sort_values(by='Release Clause',ascending= False)
```

Out[ ]:

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	International Reputation	Skill Moves	Position	Joined	Contract Valid Until	
371	193165	J. Corona	25	Mexico	81	83	FC Porto	21500.0	18.0	Right	3.0	5.0	RM	2015	2020-01-01	5
306	171897	A. Guardado	31	Mexico	82	82	Real Betis	19000.0	35.0	Left	3.0	4.0	CM	2017	2020-01-01	5
329	221992	H. Lozano	22	Mexico	81	86	PSV	24000.0	22.0	Right	3.0	4.0	LS	2017	2023-01-01	5
406	156519	H. Herrera	28	Mexico	81	81	FC Porto	17500.0	20.0	Right	3.0	3.0	CM	2013	2019-01-01	6
397	169416	C. Vela	29	Mexico	81	81	Los Angeles FC	17500.0	15.0	Left	3.0	4.0	RW	2018	2022-01-01	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18037	246089	C. Landa	19	Mexico	50	60	Tiburones Rojos de Veracruz	50.0	1.0	Left	1.0	2.0	CM	2018	2021-01-01	6
18068	240286	J. García	20	Mexico	50	62	Santos Laguna	40.0	1.0	Right	1.0	1.0	GK	2017	2021-01-01	6
18113	237045	R. Pasquel	22	Mexico	50	60	Deportivo Toluca	40.0	2.0	Right	1.0	1.0	GK	2017	2021-01-01	5
9689	139213	L. Michel	38	Mexico	66	66	Club Tijuana	40.0	3.0	Right	1.0	1.0	GK	2018	2018-01-01	6
14311	140164	Y. Gutiérrez	37	Mexico	61	61	Club Necaxa	20.0	1.0	Right	1.0	1.0	GK	2017	2018-01-01	6

365 rows × 20 columns

6. Generar un nuevo dataset que contenga el año (joined) y el número de jugadores (groupby).

In [ ]:
```
df_temp = df.groupby('Joined').size().rename('Jugadores').reset_index()

df_temp
```

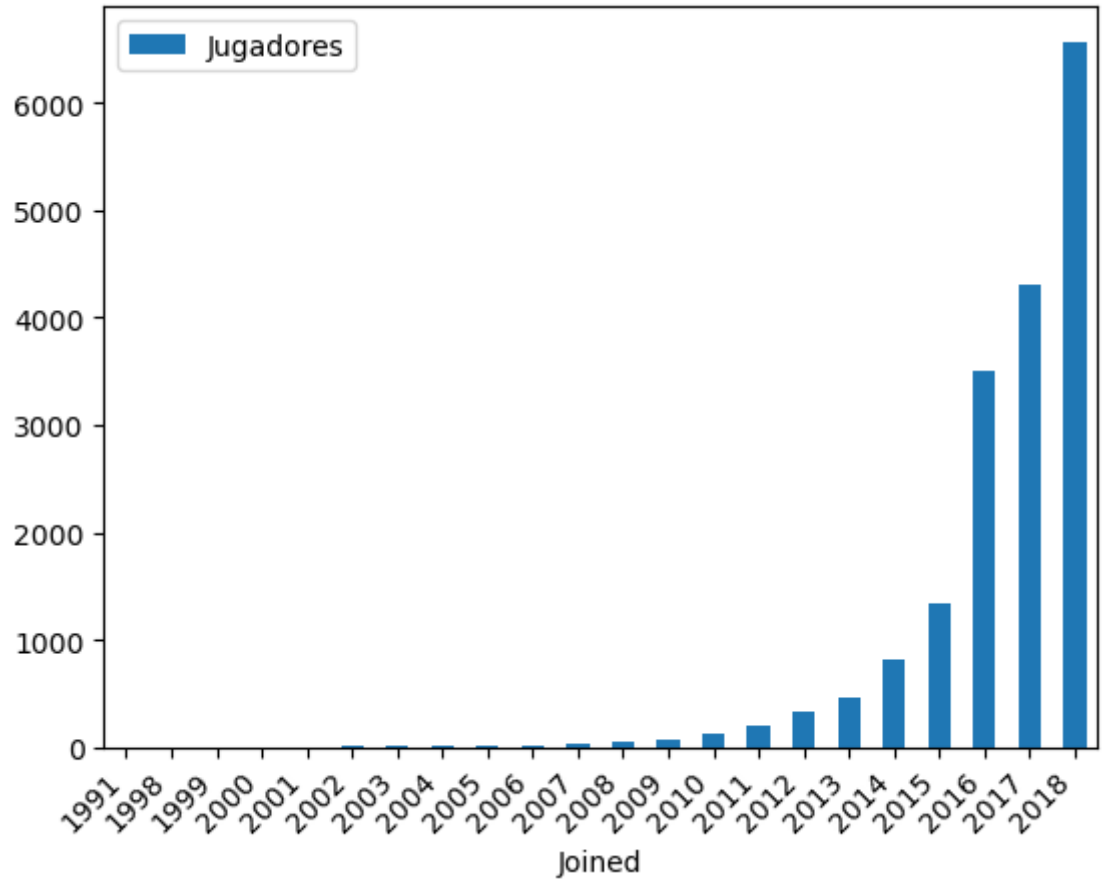
Out[ ]:

	Joined	Jugadores
0	1991	1
1	1998	3
2	1999	1
3	2000	2
4	2001	2
5	2002	10
6	2003	13
7	2004	12
8	2005	17
9	2006	18
10	2007	38
11	2008	53
12	2009	78
13	2010	131
14	2011	201
15	2012	340
16	2013	458
17	2014	818
18	2015	1336
19	2016	3510
20	2017	4307
21	2018	6569

7. Opcional: Generar un gráfico que contenga, por año, el número de jugadores.

In [ ]:
```
# Lo hice con la informacion extraìda en el apartado anterior.

import matplotlib.pyplot as plt
ax = df_temp.plot.bar(x='Joined',y='Jugadores',rot=0)
plt.xticks(rotation=45, ha='right')
plt.show()
```



In [ ]:
```
df.to_csv('fifa_eda_2.csv')
```