In [ ]:

```
#S06 T01: Tasca mètodes de mostreig
```

In [ ]:

```
#Nivell 1
```

In [498]:

```
#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.

#Llibreries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import random
import seaborn as sns
import warnings

#Prenem el conjunt de dades de tema esportiu de la pàgina web
#https://www.kaggle.com/abecklas/fifa-world-cup/version/5
```

In [470]:

```
#Dataset FIFA World Cup
#Fitxer: Python/WorldCups.csv

wcupplay_df = pd.read_csv('Python/WorldCupPlayers.csv', engine="python", error_bad_lines=Fa
wcupplay_df.head(5)
```

Out[470]:

	RoundID	MatchID	Team Initials	Coach Name	Line-up	Shirt Number	Player Name	Position	Event
0	201	1096	FRA	CAUDRON Raoul (FRA)	S	0	Alex THEPOT	GK	NaN
1	201	1096	MEX	LUQUE Juan (MEX)	S	0	Oscar BONFIGLIO	GK	NaN
2	201	1096	FRA	CAUDRON Raoul (FRA)	S	0	Marcel LANGILLER	NaN	G40'
3	201	1096	MEX	LUQUE Juan (MEX)	S	0	Juan CARRENO	NaN	G70'
4	201	1096	FRA	CAUDRON Raoul (FRA)	S	0	Ernest LIBERATI	NaN	NaN

In [471]:

```
#Dataset FIFA World Cup
#Fitxer Python/WorldCups.csv

wcup_df = pd.read_csv('Python/WorldCups.csv', engine="python", error_bad_lines=False, warn_
wcup_df.head(5)
```

Out[471]:

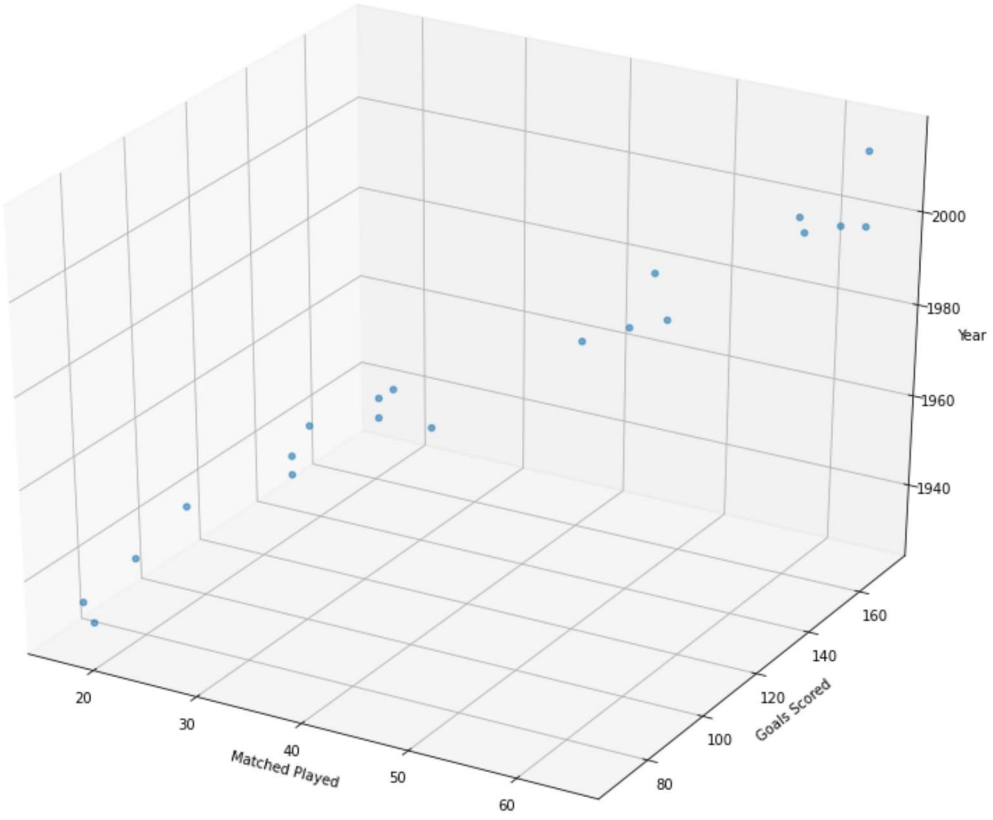
	Year	Country	Winner	Runners-Up	Third	Fourth	GoalsScored	QualifiedTeam
0	1930	Uruguay	Uruguay	Argentina	USA	Yugoslavia	70	1
1	1934	Italy	Italy	Czechoslovakia	Germany	Austria	70	1
2	1938	France	Italy	Hungary	Brazil	Sweden	84	1
3	1950	Brazil	Uruguay	Brazil	Sweden	Spain	88	1
4	1954	Switzerland	Germany FR	Hungary	Austria	Uruguay	140	1

In [636]:

```
#Graphic World Cup:
fig = plt.figure(figsize=(15,12))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(wcup_df['MatchesPlayed'],
           wcup_df['GoalsScored'],
           wcup_df['Year'],alpha=.6)

ax.set_xlabel('Matched Played')
ax.set_ylabel('Goals Scored')
ax.set_zlabel('Year')

plt.show()
```



In [472]:

```
#Dataset FIFA World Cup
#Fitxer Python/WorldCupMatches.csv

wcupmatch_df = pd.read_csv('Python/WorldCupMatches.csv', engine="python", error_bad_lines=F
wcupmatch_df.head(5)
```

Out[472]:

	Year	Datetime	Stage	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	Win conditions
0	1930	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4	1	Mexico	
1	1930	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3	0	Belgium	
2	1930	14 Jul 1930 - 12:45	Group 2	Parque Central	Montevideo	Yugoslavia	2	1	Brazil	
3	1930	14 Jul 1930 - 14:50	Group 3	Pocitos	Montevideo	Romania	3	1	Peru	
4	1930	15 Jul 1930 - 16:00	Group 1	Parque Central	Montevideo	Argentina	1	0	France	

In [473]:

```
wcup_df.dtypes
```

Out[473]:

```
Year              int64
Country          object
Winner           object
Runners-Up       object
Third            object
Fourth           object
GoalsScored       int64
QualifiedTeams    int64
MatchesPlayed     int64
Attendance       object
dtype: object
```

In [474]:

```
wcupplay_df.dtypes
```

Out[474]:

```
RoundID           int64
MatchID           int64
Team Initials    object
Coach Name       object
Line-up          object
Shirt Number      int64
Player Name      object
Position         object
Event            object
dtype: object
```

In [475]:

```
wcupmatch_df.dtypes
```

Out[475]:

```
Year                      int64
Datetime                 object
Stage                    object
Stadium                  object
City                     object
Home Team Name           object
Home Team Goals           int64
Away Team Goals           int64
Away Team Name           object
Win conditions           object
Attendance              float64
Half-time Home Goals      int64
Half-time Away Goals      int64
Referee                  object
Assistant 1              object
Assistant 2              object
RoundID                   int64
MatchID                   int64
Home Team Initials       object
Away Team Initials       object
dtype: object
```

In [476]:

```
wcup_df.columns.values
```

Out[476]:

```
array(['Year', 'Country', 'Winner', 'Runners-Up', 'Third', 'Fourth',
       'GoalsScored', 'QualifiedTeams', 'MatchesPlayed', 'Attendance'],
      dtype=object)
```

In [477]:

```
wcupplay_df.columns.values
```

Out[477]:

```
array(['RoundID', 'MatchID', 'Team Initials', 'Coach Name', 'Line-up',
       'Shirt Number', 'Player Name', 'Position', 'Event'], dtype=object)
```

In [478]:

```
wcupmatch_df.columns.values
```

Out[478]:

```
array(['Year', 'Datetime', 'Stage', 'Stadium', 'City', 'Home Team Name',
       'Home Team Goals', 'Away Team Goals', 'Away Team Name',
       'Win conditions', 'Attendance', 'Half-time Home Goals',
       'Half-time Away Goals', 'Referee', 'Assistant 1', 'Assistant 2',
       'RoundID', 'MatchID', 'Home Team Initials', 'Away Team Initials'],
      dtype=object)
```

In [479]:

```
#Mètodes de mostreig: 1.mostra aleatòria simple
```

In [480]:

```
#Aquest procés implica definir el conjunt d'individus de la població d'estudi,
#la grandària de la població (N), i un cop definida a partir d'aquest registre
#seleccionar-los aleatòriament. Aquest mostreig pot ser difícil de realitzar
#en aquells casos que la població és difícil de reclutar o els registres existents
#són poc exhaustius, així doncs definir la grandària i els individus que componen
#la població és essencial.
```

In [481]:

```
#Població d'estudi: jugadors de World Cup
#Grandària: 37784 individus - wcupplay_df['Player Name'].unique
mostra = wcupplay_df.sample(n=60).shape
mostra[0] # files 60
mostra[1] # columnes 9
```

Out[481]:

```
9
```

In [482]:

```
#Mostra de 10 jugadors aleatoris del dataframe
wcupplay_df.sample(n=10, random_state=0)
```

Out[482]:

	RoundID	MatchID	Team Initials	Coach Name	Line-up	Shirt Number	Player Name	Position	Event
32012	249722	300061457	GRE	REHHAGEL Otto (GER)	N	13	SIFAKIS	GK	NaN
21334	337	3076	GER	VOGTS Berti (GER)	S	20	Stefan EFFENBERG	NaN	Y44' O75'
12128	278	2197	ARG	MENOTTI Cesar Luis (ARG)	S	9	Rene HOUSEMAN	NaN	NaN
14045	293	774	BEL	THYS Guy (BEL)	N	17	Rene VERHEYEN	NaN	NaN
3393	211	1300	SUI	RAPPAN Karl (AUT)	S	20	Eugen MEIER	NaN	NaN
12292	278	2408	SCO	MacLEOD Alistair (SCO)	N	20	Bobby CLARK	NaN	NaN
16798	308	389	BUL	VUTSOV Ivan (BUL)	S	15	Georgi YORDANOV	NaN	NaN
37728	255957	300186502	BRA	SCOLARI Luiz Felipe (BRA)	N	13	DANTE	NaN	NaN
34357	255931	300186489	CRC	PINTO Jorge Luis (COL)	N	18	PEMBERTON P.	GK	NaN
2951	209	1231	URU	LOPEZ Juan (URU)	S	0	Matias GONZALEZ	NaN	NaN

In [483]:

```
#Visualització de paisos on juguen 8 jugadors escollits aleatoriament.
wcupplay_df.sample(n=8, random_state=1)[['Player Name','Team Initials']]
```

Out[483]:

	Player Name	Team Initials
7536	Nestor COMBIN	FRA
25498	GALLARDO	ARG
11150	Harry VOS	NED
14225	MIGUEL ANGEL	ESP
26566	M. VIDRIO	MEX
4589	ZOZIMO	BRA
13244	EDINHO	BRA
8284	Jose SASIA	URU

In [ ]:

```
from matplotlib import pyplot as plt
import seaborn as sns

airlines_df[["ActualElapsedTime","CRSElapsedTime","UniqueCarrier","TaxiIn","TaxiOut"]]

sns.set(rc={"figure.figsize":(15, 15)})
sns.scatterplot(data=airlines_df, x="ActualElapsedTime", y="CRSElapsedTime", color = 'green
                marker = '+', hue = "UniqueCarrier", alpha=.5, s= 150)

plt05.xlabel("Actual Elapsed Time")
plt05.ylabel("CRS Elapsed Time")
plt05.legend(loc='best')
plt05.tight_layout()
plt05.show()
```

In [484]:

```
#Mètodes de mostreig: 2.mostra sistemàtica
```

In [485]:

```
#És una variant del mostreig aleatori simple (MAS) en què després de definir la població d'
#i es té la mostra ordenada en una llista s'inicia la presa del primer element de la mostra
#Després cal calcular una constant, que es denomina coeficient d'elevació k = N / n;
#on N és la mida de la població estudiada i n la mida de la mostra.
#Després cal triar a l'atzar un nombre entre 1 i k, i prendre els elements de k en k al lla
#Ocasionalment, és convenient tenir en compte la periodicitat del fenomen.

#S'utilitza quan l'univers o població és de gran mida, o l'estudi s'ha d'estendre en el tem
#Primer cal identificar les unitats i relacionar-les amb el calendari (quan escaigui).
#Això vol dir que si tenim un determinat nombre de persones que és la població i es vol esc
#d'aquesta població un nombre més petit que és la mostra, es divideix el nombre de la pobla
#pel nombre de la mostra que es vol prendre i el resultat d'aquesta operació serà l'interva
#llavors s'escull un nombre a l'atzar des d'un fins al número de l'interval, i a partir d'a
#escollim els altres seguint l'ordre de l'interval.
```

In [486]:

```
#Àrea d'estudi: partits de World Cup
#Grandària: 852 partits
mostra_s = wcupmatch_df.sample(n=40).shape
mostra_s[0] # files 40
mostra_s[1] # columnes 20
```

Out[486]:

20

In [487]:

```
#Mostra de dades de 10 partits de World Cup
wcupmatch_df.sample(n=10, random_state=0)[['Year','Datetime','Stadium','City','Home Team Na
```

Out[487]:

	Year	Datetime	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name
31	1934	03 Jun 1934 - 16:30	San Siro	Milan	Italy	1	0	Austria
150	1962	03 Jun 1962 - 15:00	Estadio El Teniente-Codelco	Rancagua	Hungary	6	1	Bulgaria
334	1982	22 Jun 1982 - 21:00	La Rosaleda	Malaga	Soviet Union	2	2	Scotland
527	1998	14 Jun 1998 - 17:30	Stade Geoffroy Guichard	Saint-Etienne	Yugoslavia	1	0	Iran
596	2002	05 Jun 2002 - 20:30	Kashima Stadium	Ibaraki	Germany	1	1	rn">Republic of Ireland
666	2006	16 Jun 2006 - 21:00	FIFA World Cup Stadium, Hanover	Hanover	Mexico	0	0	Angola
680	2006	21 Jun 2006 - 16:00	Zentralstadion	Leipzig	Iran	1	1	Angola
686	2006	22 Jun 2006 - 21:00	FIFA World Cup Stadium, Dortmund	Dortmund	Japan	1	4	Brazil
735	2010	20 Jun 2010 - 16:00	Mbombela Stadium	Nelspruit	Italy	1	1	New Zealand
784	2014	16 Jun 2014 - 16:00	Arena da Baixada	Curitiba	IR Iran	0	0	Nigeria

In [661]:

```
# Calcul del coeficient d'elevació k = N / n
# N és la mida total estudiada.
# n la mida de la mostra.
import random
import pandas as pd
import numpy as np
import sklearn

# Exemple:
# Suposem una mostra de 40 partits
# Si partim d'una mostra total de 852 partits - [wcupmatch_df.count]

N = 852
n = 40
k = int(N/n)
print('Mostra World Cup Matches. Coeficient elevació k:', k)

sample_df = pd.DataFrame() #dataframe contenidor de les instàncies de la mostra sistemàtica

pos=0
while pos < k+1:
    sample_df = sample_df.append(wcupmatch_df.iloc[k*pos], ignore_index=False, sort=True) #
    pos+=1

print('Visualització de', k, 'instàncies del dataframe World Cup Matches utilitzant mostra
sample_df
```

```
Mostra World Cup Matches. Coeficient elevació k: 21
Visualització de 21 instàncies del dataframe World Cup Matches utilitzant mo
stra sistemàtica.
```

Out[661]:

	Assistant 1	Assistant 2	Attendance	Away Team Goals	Away Team Initials	Away Team Name	
0	CRISTOPHE Henry (BEL)	REGO Gilberto (BRA)	4444.0	1.0	MEX	Mexico	Monte
21	CARRARO Albino (ITA)	TURBIANI Giuseppe (ITA)	14000.0	2.0	ARG	Argentina	Bo
42	CAPDEVILLE Pierre (FRA)	MARENCO Paul (FRA)	8000.0	1.0	ROU	Romania	Tou
63	BERANEK Alois (AUT)	DA COSTA VIEIRA Jose (POR)	142429.0	0.0	YUG	Yugoslavia	R Ja
84	DOERFLINGER Ernst (SUI)	GULDE Josef (SUI)	26000.0	0.0	TCH	Czechoslovakia	Ż
105	GRIFFITHS Benjamin (WAL)	BROZZI Juan (ARG)	16518.0	3.0	PAR	Paraguay	Norrkï¿½

	Assistant 1	Assistant 2	Attendance	Away Team Goals	Away Team Initials	Away Team Name	
126	FERNANDES CAMPOS Joaquim (POR)	AHLNER Sten (SWE)	6196.0	1.0	TCH	Czechoslovakia	Mali
147	GOLDSTEIN Leo (USA)	BUERGO Fernando (MEX)	66057.0	0.0	ITA	Italy	Santiag
168	BAKHRAMOV Tofik (URS)	RUMENTCHEV Dimitar (BUL)	87148.0	0.0	URU	Uruguay	Lo
189	GARDEAZABAL Juan (ESP)	CODESAL Jose Maria (URU)	24129.0	1.0	BUL	Bulgaria	Manch
210	EMSBERGER Gyula (HUN)	LORAUX Vital (BEL)	56818.0	1.0	TCH	Czechoslovakia	Guada
231	SCHEURER Ruedi (SUI)	COEREZZA Norberto Angel (ARG)	107412.0	1.0	ITA	Italy	Mexic
252	BIWERSI Ferdinand (GER)	ESCHWEILER Walter (GER)	53300.0	4.0	NED	Netherlands	Dort
273	NDIAYE Youssou (SEN)	PARTRIDGE Pat (ENG)	71615.0	1.0	HUN	Hungary	Buenos
294	GONZALEZ ARCHUNDIA Alfonso (MEX)	COMESANA Miguel (ARG)	67547.0	0.0	ITA	Italy	Buenos
315	CASTRO Gaston (CHI)	COELHO Arnaldo (BRA)	44172.0	1.0	FRA	France	E
336	LAMO CASTILLO Augusto (ESP)	LACARNE Belaid (ALG)	32500.0	0.0	SLV	El Salvador	Ali
357	GALLER Bruno (SUI)	VALENTINE Robert (SCO)	70000.0	3.0	FRA	France	S
378	AGNOLIN Luigi (ITA)	NEMETH Lajos (HUN)	28000.0	2.0	ESP	Spain	Guada
399	MARQUEZ RAMIREZ Antonio (MEX)	SNODDY Alan (NIR)	45000.0	0.0	POL	Poland	Guada
420	FREDRIKSSON Erik (SWE)	ROETHLISBERGER Kurt (SUI)	35238.0	1.0	IRL	rn">Republic of Ireland	Ca
441	LORENC Richard (AUS)	PETROVIC Zoran (SRB)	34857.0	1.0	USA	USA	Flo

In [489]:

```
#Nivell 2
```

In [490]:

```
#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).
```

In [491]:

```
#Mètodes de mostreig: 3.mostra estratificada
```

In [492]:

```
#És una altra variant del Mètode Aleatori Simple. En aquest cas cal definir la mida de la p
#i els estrats en què es classificaran cadascun dels individus de la mostra (p. ex.: barri
#És d'interès quan es poden presentar variacions de les variables d'estudi segons l'estrat
#de la mostra. En estudis sobre poblacions humanes resulta d'interès, ja que permet creuar
#amb dades recollides en censos de població. No obstant afegeix una nova variable a recolli
```

In [493]:

```
#Àrea d'estudi: Celebracions de World Cup
#Grandària: 20 world cups
mostra_wc = wcup_df.sample(n=10).shape
mostra_wc[0] # files 10
mostra_wc[1] # columnes 20
```

Out[493]:

```
10
```

In [592]:

```
wcup_df.head(11)
```

Out[592]:

	Year	Country	Winner	Runners-Up	Third	Fourth	GoalsScored	QualifiedTea
0	1930	Uruguay	Uruguay	Argentina	USA	Yugoslavia	70	
1	1934	Italy	Italy	Czechoslovakia	Germany	Austria	70	
2	1938	France	Italy	Hungary	Brazil	Sweden	84	
3	1950	Brazil	Uruguay	Brazil	Sweden	Spain	88	
4	1954	Switzerland	Germany FR	Hungary	Austria	Uruguay	140	
5	1958	Sweden	Brazil	Sweden	France	Germany FR	126	
6	1962	Chile	Brazil	Czechoslovakia	Chile	Yugoslavia	89	
7	1966	England	England	Germany FR	Portugal	Soviet Union	89	
8	1970	Mexico	Brazil	Italy	Germany FR	Uruguay	95	
9	1974	Germany	Germany FR	Netherlands	Poland	Brazil	97	
10	1978	Argentina	Argentina	Netherlands	Brazil	Italy	102	

In [ ]:

```
#World Cup by decades
```

In [ ]:

```
#The 30's
wcup_df.loc[wcup_df['Year'].isin([1930,1934,1938])]
```

In [ ]:

```
d Cup
] = wcup_df['Attendance'].astype(str).str.replace('.','').astype(int) # Transformem el valor
d','QualifiedTeams','MatchesPlayed','Attendance']].loc[wcup_df['Year'].isin([1930,1934,1938]
```

In [ ]:

```
#The 50's
wcup_df.loc[wcup_df['Year'].isin([1950,1954,1958])]
```

In [ ]:

```
#Summatory data World Cup
wcup_df[['GoalsScored','QualifiedTeams','MatchesPlayed','Attendance']].loc[wcup_df['Year'].
```

In [ ]:

```
#The 60's
wcup_df.loc[wcup_df['Year'].isin([1962,1966])]
```

In [ ]:

```
#Summatory data World Cup
wcup_df[['GoalsScored','QualifiedTeams','MatchesPlayed','Attendance']].loc[wcup_df['Year'].
```

In [ ]:

```
#The 70's
wcup_df.loc[wcup_df['Year'].isin([1970,1974,1978])]
```

In [ ]:

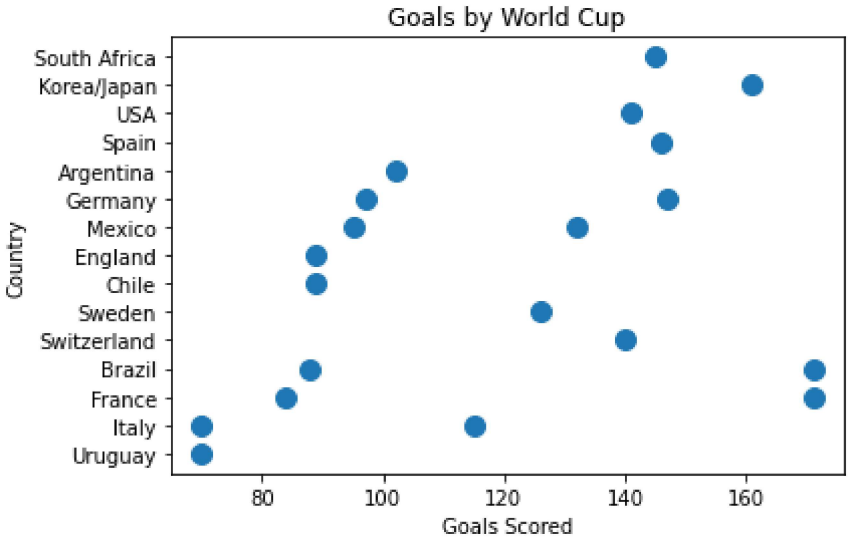
```
#Summatory data World Cup
wcup_df[['GoalsScored','QualifiedTeams','MatchesPlayed','Attendance']].loc[wcup_df['Year'].
```

In [654]:

```
#Graphic World Cup: Goals Scored by World Cup
import matplotlib.pyplot as plt

x = wcup_df['GoalsScored']
y = wcup_df['Country']

plt.scatter(x, y, s=100)
plt.title('Goals by World Cup')
plt.xlabel('Goals Scored')
plt.ylabel('Country')
plt.show()
```

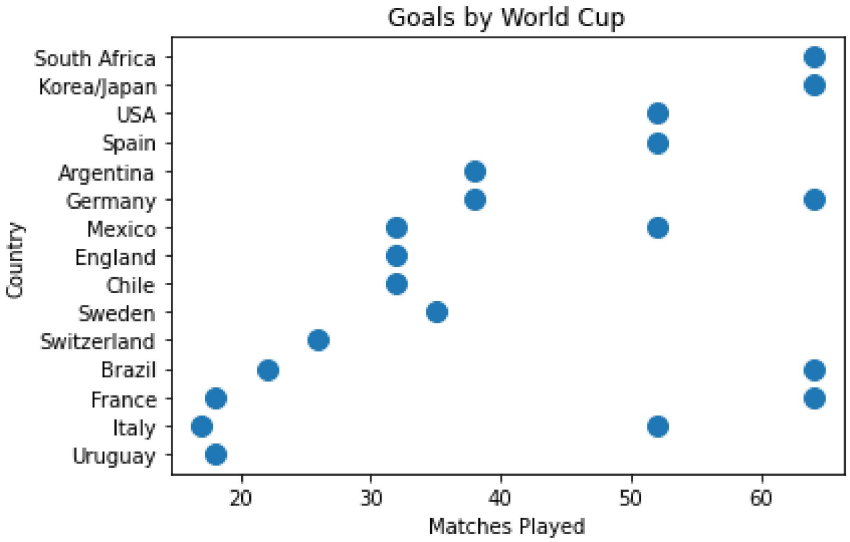


In [660]:

```
#Graphic World Cup: Matches Played by World Cup
import matplotlib.pyplot as plt

x = wcup_df['MatchesPlayed']
y = wcup_df['Country']

plt.scatter(x, y, s=100)
plt.title('Goals by World Cup')
plt.xlabel('Matches Played')
plt.ylabel('Country')
plt.show()
```



In [ ]:

```
#Mètodes de mostreig: 3.mostra estratificada
```

In [678]:

```
# Transformem el valor de l'atribut Attendance a número.
wcup_df['Attendance'] = wcup_df['Attendance'].astype(str).str.replace('.','').astype(int)

data_x=wcup_df
data_y=wcup_df['Attendance']

x_train, x_test, y_train, y_test = train_test_split(data_x,data_y,test_size=0.33, random_st

print(x_train,'\n')
print(x_test,'\n')
print(y_train,'\n')
print(y_test)
```

```
    Year       Country      Winner      Runners-Up        Third  \
3   1950        Brazil     Uruguay          Brazil       Sweden
18  2010  South Africa       Spain     Netherlands      Germany
16  2002   Korea/Japan      Brazil         Germany       Turkey
13  1990         Italy  Germany FR       Argentina        Italy
2   1938        France       Italy         Hungary       Brazil
9   1974       Germany  Germany FR     Netherlands       Poland
19  2014        Brazil     Germany       Argentina  Netherlands
4   1954   Switzerland  Germany FR         Hungary      Austria
12  1986        Mexico   Argentina      Germany FR       France
7   1966       England     England      Germany FR     Portugal
10  1978     Argentina   Argentina     Netherlands       Brazil
14  1994           USA      Brazil           Italy       Sweden
6   1962         Chile      Brazil  Czechoslovakia        Chile

            Fourth  GoalsScored  QualifiedTeams  MatchesPlayed  Attendance
3            Spain           88              13             22     1045246
18         Uruguay          145              32             64     3178856
16  Korea Republic          161              32             64     2705197
13         England          115              24             52     2516215
2           Sweden           84              15             18      375700
9           Brazil           97              16             38     1865753
19          Brazil          171              32             64     3386810
4          Uruguay          140              16             26      768607
12         Belgium          132              24             52     2394031
7     Soviet Union           89              16             32     1563135
10           Italy          102              16             38     1545791
14        Bulgaria          141              24             52     3587538
6       Yugoslavia           89              16             32      893172

    Year  Country   Winner      Runners-Up       Third       Fourth  \
0   1930  Uruguay  Uruguay       Argentina         USA   Yugoslavia
17  2006  Germany    Italy          France     Germany     Portugal
15  1998   France   France          Brazil     Croatia  Netherlands
1   1934    Italy    Italy  Czechoslovakia     Germany      Austria
8   1970   Mexico   Brazil           Italy  Germany FR      Uruguay
5   1958   Sweden   Brazil          Sweden      France   Germany FR
11  1982    Spain    Italy      Germany FR      Poland       France

    GoalsScored  QualifiedTeams  MatchesPlayed  Attendance
0            70              13             18      590549
17          147              32             64     3359439
15          171              32             64     2785100
1            70              16             17      363000
8            95              16             32     1603975
5           126              16             35      819810
```

```
11          146              24              52      2109723

3     1045246
18    3178856
16    2705197
13    2516215
2      375700
9     1865753
19    3386810
4      768607
12    2394031
7     1563135
10    1545791
14    3587538
6      893172
Name: Attendance, dtype: int32

0      590549
17    3359439
15    2785100
1      363000
8     1603975
5      819810
11    2109723
Name: Attendance, dtype: int32
```

In [ ]:

```
#Mètodes de mostreig: 4.mostra utilitzant SMOTE (Synthetic Minority Oversampling Technique)
```

In [ ]:

```
#SMOTE  (Synthetic Minority Oversampling Technique) ens permet generar mostres sintetiques
#Hi ha una diferencia entre un valor i els valors propers i es multiplica per un valor entr
```

In [ ]:

```
#SMOTE (Synthetic Minority Oversampling Technique) funciona seleccionant exemples propers a
#característiques, dibuixant una línia entre els exemples de l'espai de característiques i
#dibuixant una nova mostra en un punt al llarg d'aquesta línia. Concretament, primer es tri
#exemple aleatori de la classe minoritària i llavors es troben k dels veïns més propers.
#Aleshores es tria un veí seleccionat aleatòriament i es crea un exemple sintètic en un pun
#seleccionat aleatòriament entre els dos exemples de l'espai de característiques.
```

In [563]:

```
#Llibreria per Synthetic Minority Oversampling Technique
from imblearn.over_sampling import SMOTE
#Llibreria per evitar warnings
from warnings import simplefilter

#Desactivació de Future warnings
simplefilter(action='ignore', category=FutureWarning)

#Implementing sample of World Cup Players
sample = SMOTE(wcupplay_df['Player Name'], sampling_strategy='auto', random_state=None, k_n
sample
```

Out[563]:

```
SMOTE(sampling_strategy=0             Alex THEPOT
1          Oscar BONFIGLIO
2         Marcel LANGILLER
3             Juan CARRENO
4          Ernest LIBERATI
               ...
37779             ALVAREZ
37780             KHEDIRA
37781              AGUERO
37782             MUSTAFI
37783             BASANTA
Name: Player Name, Length: 37784, dtype: object)
```

In [590]:

```
#Implementació de la mostra de World Cup Coachers
sample = SMOTE(wcupplay_df['Coach Name'].sample(n=6), sampling_strategy='minority', k_neigh
sample
```

Out[590]:

```
SMOTE(sampling_strategy=4153      NAUSCH Walter (AUT)
11199     ZAGALLO Mario (BRA)
8421      SCHOEN Helmut (FRG)
33000    WEISS Vladimir (SVK)
7026     RIERA Fernando (CHI)
22467       BROWN Craig (SCO)
Name: Coach Name, dtype: object)
```

In [569]:

```
#Library for Synthetic Minority Oversampling Technique
from imblearn.over_sampling import SMOTE
#Implementing sample of World Cup Coachers
sample = SMOTE(wcupplay_df['Coach Name'], k_neighbors=5)
sample
```

Out[569]:

```
SMOTE(sampling_strategy=0            CAUDRON Raoul (FRA)
1               LUQUE Juan (MEX)
2            CAUDRON Raoul (FRA)
3               LUQUE Juan (MEX)
4            CAUDRON Raoul (FRA)
                  ...
37779    SABELLA Alejandro (ARG)
37780         LOEW Joachim (GER)
37781    SABELLA Alejandro (ARG)
37782         LOEW Joachim (GER)
37783    SABELLA Alejandro (ARG)
Name: Coach Name, Length: 37784, dtype: object)
```

In [547]:

```
#Població d'estudi: jugadors de World Cup
#Grandària: 37784 individus - wcupplay_df['Player Name'].unique
mostra = wcupplay_df.sample(n=60).shape
mostra[0] # files 60
mostra[1] # columnes 9
```

Out[547]:

```
9
```

In [ ]:

```
#Nivell 3
```

In [ ]:

```
#Exercici 3
#Continua amb el conjunt de dades de tema esportiu.
#Genera una mostra utilitzant el mètode Reservoir sampling.
```

In [ ]:

```
#Reservoir sampling és una familia de algorismes aleatoris per escollir una mostra aleatori
#sense substitució, de k items d'una població d'un tamany n desconegut d'una vegada sobre e
#El tamany de la població n no és conegut per l'algorisme i sovint és massa extens pels n i
#per accedir a la memoria principal. La població és donada a l'algorisme tota l'estona, i l
#no pot trobar els items anteriors. En qualsevol moment, la situació actual de l'algorisme
#l'extracció d'una mostra aleatoria simple sense modificar el tamany k de la part de la pob
```

In [528]:

```
#Reservoir Sampling
#Població d'estudi: partits de World Cup
#Grandària: 852 partits

k=5 #nombre de files (items) que es retornaran.
sample = []

for i, row in enumerate(wcupmatch_df.values):
    if i+1<= k:
        sample.append(row)
    else:
        if random.random() < k/(i+1):
            sample[random.choice(range(0,k))] = row

print(sample)
```

```
[array([1950, '09 Jul 1950 - 15:00 ', 'Group 6', 'Pacaembu', 'Sao Paulo ',
       'Uruguay', 2, 2, 'Spain', ' ', 44802.0, 1, 2,
       'GRIFFITHS Benjamin (WAL)', 'DATTILO Generoso (ITA)',
       'ALVAREZ Alfredo (BOL)', 209, 1207, 'URU', 'ESP'], dtype=object), arr
ay([1986, '02 Jun 1986 - 12:00 ', 'Group C', 'Estadio Irapuato',
       'Irapuato ', 'Soviet Union', 6, 0, 'Hungary', ' ', 16500.0, 3, 0,
       'AGNOLIN Luigi (ITA)', 'COURTNEY George (ENG)',
       'BRUMMEIER Horst (AUT)', 308, 610, 'URS', 'HUN'], dtype=object), arra
y([1930, '16 Jul 1930 - 14:45 ', 'Group 1', 'Parque Central',
       'Montevideo ', 'Chile', 3, 0, 'Mexico', ' ', 9249.0, 1, 0,
       'CRISTOPHE Henry (BEL)', 'APHESTEGUY Martin (URU)',
       'LANGENUS Jean (BEL)', 201, 1095, 'CHI', 'MEX'], dtype=object), array
([1998, '25 Jun 1998 - 16:00 ', 'Group E', 'Stade Geoffroy Guichard',
       'Saint-Etienne ', 'Netherlands', 2, 2, 'Mexico', ' ', 30600.0, 2,
       0, 'ALZEID Abdulrahman (KSA)', 'TRESACO GRACIA Fernando (ESP)',
       'GHADANFARI Hussain (KUW)', 1014, 8766, 'NED', 'MEX'], dtype=object),
array([1950, '25 Jun 1950 - 15:00 ', 'Group 2', 'Durival de Brito',
       'Curitiba ', 'Spain', 3, 1, 'USA', ' ', 9511.0, 0, 1,
       'VIANA Mario (BRA)', 'DA COSTA VIEIRA Jose (POR)',
       'DE LA SALLE Charles (FRA)', 208, 1208, 'ESP', 'USA'], dtype=object)]
```

In [ ]: