In [1]:

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

In [2]:

```
df = pd.read_csv('Forbes Richest Atheletes (Forbes Richest Athletes 1990-2020).csv')
df
```

Out[2]:

	S.NO	Name	Nationality	Current Rank	Previous Year Rank	Sport	Year	earnings ($ million)
0	1	Mike Tyson	USA	1	NaN	boxing	1990	28.6
1	2	Buster Douglas	USA	2	NaN	boxing	1990	26.0
2	3	Sugar Ray Leonard	USA	3	NaN	boxing	1990	13.0
3	4	Ayrton Senna	Brazil	4	NaN	auto racing	1990	10.0
4	5	Alain Prost	France	5	NaN	auto racing	1990	9.0
...	...	...	...	...	...	...	...	...
296	297	Stephen Curry	USA	6	9	Basketball	2020	74.4
297	298	Kevin Durant	USA	7	10	Basketball	2020	63.9
298	299	Tiger Woods	USA	8	11	Golf	2020	62.3
299	300	Kirk Cousins	USA	9	>100	American Football	2020	60.5
300	301	Carson Wentz	USA	10	>100	American Football	2020	59.1

301 rows × 8 columns

In [3]:

```
display(df.dtypes)
```

```
S.NO                        int64
Name                       object
Nationality                object
Current Rank                int64
Previous Year Rank         object
Sport                      object
Year                        int64
earnings ($ million)      float64
dtype: object
```

In [4]:

```
df.isna().sum()
```

Out[4]:

```
S.NO                     0
Name                     0
Nationality              0
Current Rank             0
Previous Year Rank      24
Sport                    0
Year                     0
earnings ($ million)     0
dtype: int64
```

In [5]:

```
dfnull = df[df['Previous Year Rank'].isna()]
dfnull
```

Out[5]:

	S.NO	Name	Nationality	Current Rank	Previous Year Rank	Sport	Year	earnings ($ million)
0	1	Mike Tyson	USA	1	NaN	boxing	1990	28.6
1	2	Buster Douglas	USA	2	NaN	boxing	1990	26.0
2	3	Sugar Ray Leonard	USA	3	NaN	boxing	1990	13.0
3	4	Ayrton Senna	Brazil	4	NaN	auto racing	1990	10.0
4	5	Alain Prost	France	5	NaN	auto racing	1990	9.0
5	6	Jack Nicklaus	USA	6	NaN	golf	1990	8.6
6	7	Greg Norman	Australia	7	NaN	golf	1990	8.5
7	8	Michael Jordan	USA	8	NaN	basketball	1990	8.1
8	9	Arnold Palmer	USA	8	NaN	golf	1990	8.1
9	10	Evander Holyfield	USA	8	NaN	boxing	1990	8.1

80	81	Michael Jordan	USA	1	NaN	Basketball	1998	69.0
81	82	Michael Schumacher	Germany	2	NaN	F1 Motorsports	1998	38.0
82	83	Sergei Federov	Russia	3	NaN	Ice Hockey	1998	29.8
83	84	Tiger Woods	USA	4	NaN	Golf	1998	26.8
84	85	Dale Earnhardt	USA	5	NaN	NASCAR	1998	24.1
85	86	Grant Hill	USA	6	NaN	Basketball	1998	21.6
86	87	Oscar De La Hoya	USA	7	NaN	Boxing	1998	18.5
87	88	Patrick Ewing	USA	8	NaN	Basketball	1998	18.3
88	89	Arnold Palmer	USA	9	NaN	Golf	1998	18.1
89	90	Gary Sheffield	USA	10	NaN	Baseball	1998	17.2
266	267	Andrew Luck	USA	6	NaN	American Football	2017	50.0
268	269	Stephen Curry	USA	8	NaN	Basketball	2017	47.3
269	270	James Harden	USA	9	NaN	Basketball	2017	46.6
270	271	Lewis Hamilton	UK	10	NaN	auto racing	2017	46.0

Filling null value

In [6]:

```
mode=df["Previous Year Rank"].mode()
mode
```

Out[6]:

```
0    >10
dtype: object
```

In [7]:

```
df["Previous Year Rank"].fillna(mode,inplace=True)
```

In [8]:

```
df.isnull().sum()
```

Out[8]:

```
S.NO                     0
Name                     0
Nationality              0
Current Rank             0
Previous Year Rank      23
Sport                    0
Year                     0
earnings ($ million)     0
dtype: int64
```

In [9]:

```
#Checking for duplicate data
df.duplicated().sum()
```

Out[9]:

```
0
```

In [10]:

```
df.describe(include="all")
```

Out[10]:

	S.NO	Name	Nationality	Current Rank	Previous Year Rank	Sport	Year	earnings ($ million)
count	301.000000	301	301	301.000000	278	301	301.000000	301.000000
unique	NaN	82	22	NaN	36	29	NaN	NaN
top	NaN	Tiger Woods	USA	NaN	>10	Basketball	NaN	NaN
freq	NaN	19	206	NaN	37	54	NaN	NaN
mean	151.000000	NaN	NaN	5.448505	NaN	NaN	2005.122924	45.516279
std	87.035433	NaN	NaN	2.850995	NaN	NaN	9.063563	33.525337
min	1.000000	NaN	NaN	1.000000	NaN	NaN	1990.000000	8.100000
25%	76.000000	NaN	NaN	3.000000	NaN	NaN	1997.000000	24.000000
50%	151.000000	NaN	NaN	5.000000	NaN	NaN	2005.000000	39.000000
75%	226.000000	NaN	NaN	8.000000	NaN	NaN	2013.000000	59.400000

max	301.000000	NaN	NaN	10.000000	NaN	NaN	2020.000000	300.000000

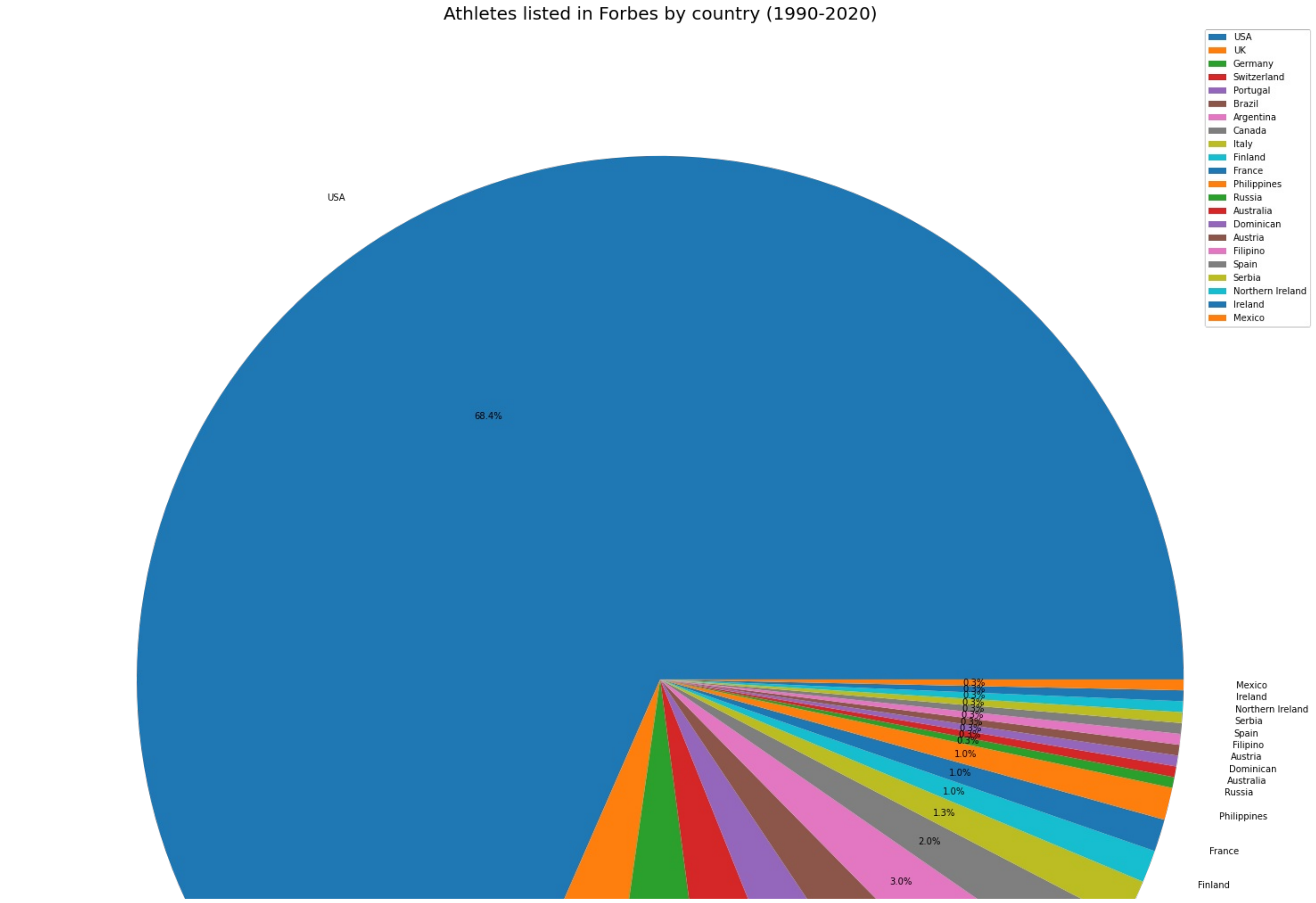
Athletes listed in Forbes by country (1990-2020

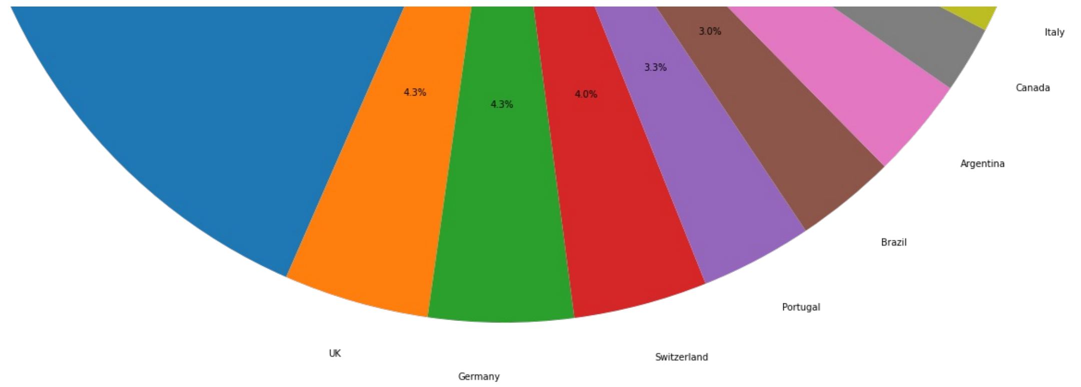
In [11]:
```
count_by_nationality=df.Nationality.value_counts()
```

In [12]:
```
count_by_nationality
```

Out[12]:
```
USA                  206
UK                    13
Germany               13
Switzerland           12
Portugal              10
Brazil                 9
Argentina              9
Canada                 6
Italy                  4
Finland                3
France                 3
Philippines            3
Russia                 1
Australia              1
Dominican              1
Austria                1
Filipino               1
Spain                  1
Serbia                 1
Northern Ireland       1
Ireland                1
Mexico                 1
Name: Nationality, dtype: int64
```

In [13]:
```
plt.figure(figsize=(30,27))
plt.pie(df.Nationality.value_counts().to_frame().values.flatten(),
        labels=df.Nationality.value_counts().to_frame().index.tolist(),
        autopct='%.1f%%')
plt.title('Athletes listed in Forbes by country (1990-2020)',fontsize=20)
plt.legend(df.Nationality.value_counts().to_frame().index)
plt.show()
```





Which sport has maximum number of athletes in Forbes, listed till 2020?

In [14]:

```
no_of_athletes=df.Sport.str.lower().value_counts()
no_of_athletes
```

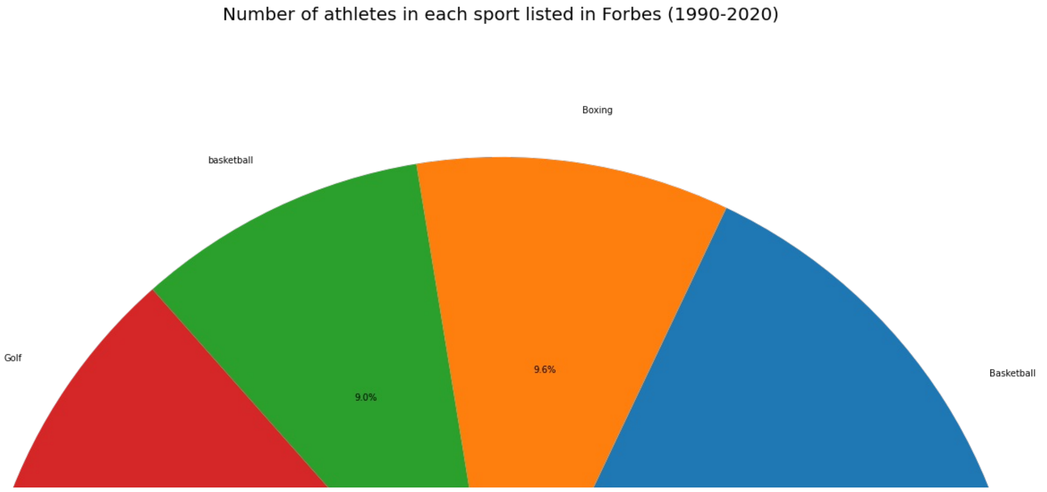
Out[14]:

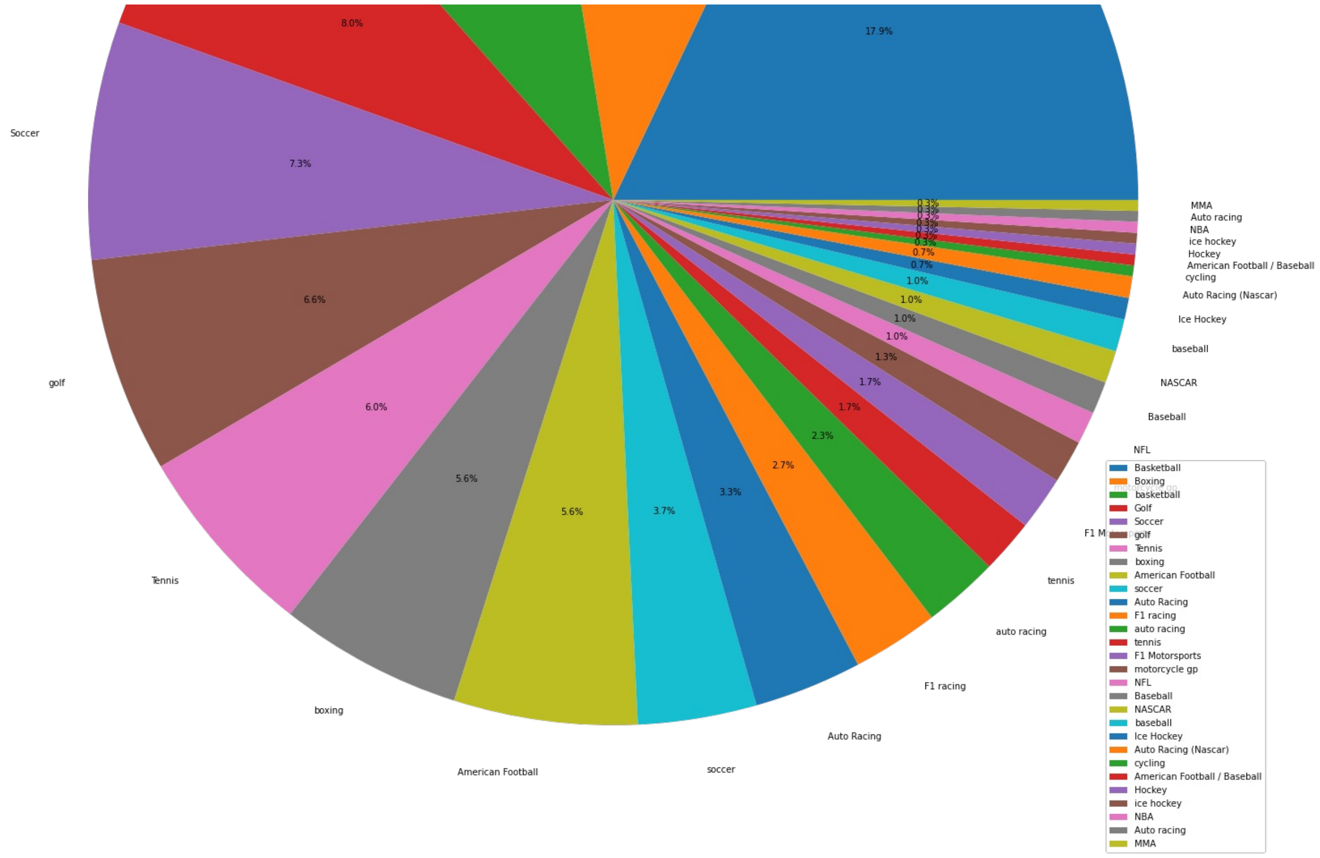
```
basketball                    81
boxing                        46
golf                          44
soccer                        33
tennis                        23
auto racing                   18
american football             17
f1 racing                      8
baseball                       6
f1 motorsports                 5
motorcycle gp                  4
nascar                         3
ice hockey                     3
nfl                            3
auto racing (nascar)           2
american football / baseball   1
hockey                         1
nba                            1
cycling                        1
mma                            1
Name: Sport, dtype: int64
```

Number of athletes in each sport listed in Forbes (1990-2020)

In [15]:

```
plt.figure(figsize=(30,27))
plt.pie(df.Sport.value_counts().to_frame().values.flatten(),
        labels=df.Sport.value_counts().to_frame().index.tolist(),
        autopct='%.1f%%')
plt.title('Number of athletes in each sport listed in Forbes (1990-2020)',fontsize=20)
plt.legend(df.Sport.value_counts().to_frame().index)
plt.show()
```





Top 10 highest paid Athletes (1990-2020)

In [16]:

```
top_ten=df.sort_values('earnings ($ million)',ascending=False).head(10)
```

In [17]:

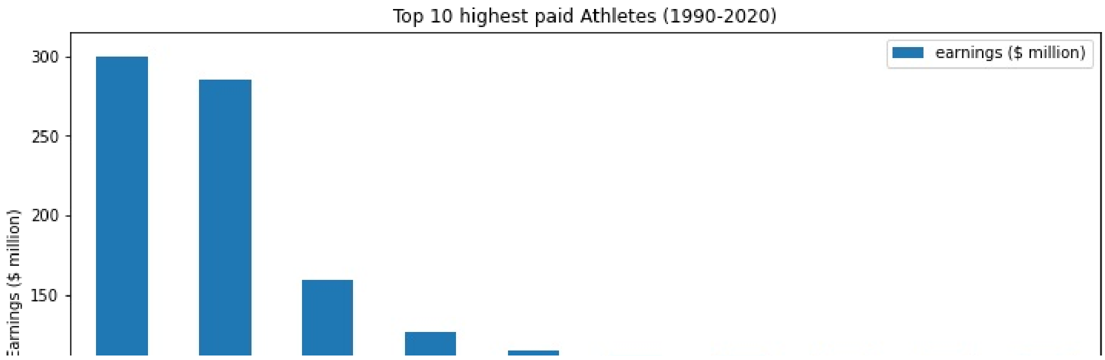
```
top_ten
```

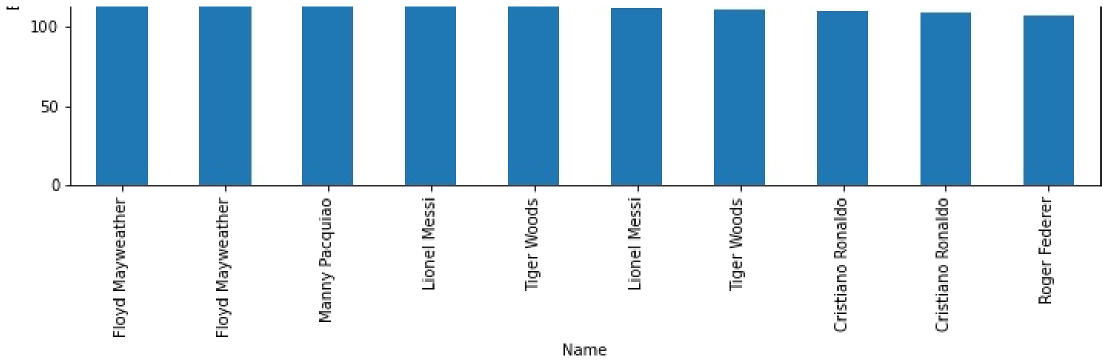
Out[17]:

	S.NO	Name	Nationality	Current Rank	Previous Year Rank	Sport	Year	earnings ($ million)
241	242	Floyd Mayweather	USA	1	1	Boxing	2015	300.0
271	272	Floyd Mayweather	USA	1	>100	Boxing	2018	285.0
242	243	Manny Pacquiao	Philippines	2	11	Boxing	2015	160.0
281	282	Lionel Messi	Argentina	1	2	Soccer	2019	127.0
171	172	Tiger Woods	USA	1	1	golf	2008	115.0
272	273	Lionel Messi	Argentina	2	3	Soccer	2018	111.0
181	182	Tiger Woods	USA	1	1	golf	2009	110.0
282	283	Cristiano Ronaldo	Portugal	2	3	Soccer	2019	109.0
273	274	Cristiano Ronaldo	Portugal	3	1	Soccer	2018	108.0
291	292	Roger Federer	Switzerland	1	5	Tennis	2020	106.3

In [18]:

```
plot=top_ten.plot.bar(x='Name',y='earnings ($ million)',figsize=(12,6));
plt.title('Top 10 highest paid Athletes (1990-2020)');
plt.ylabel('Earnings ($ million)');
```

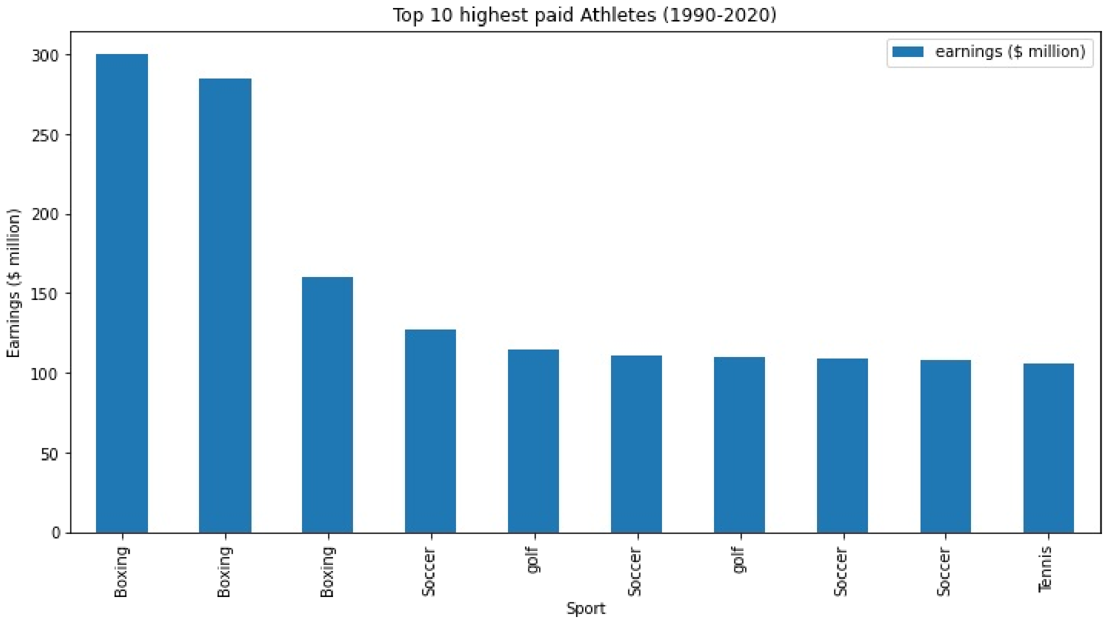




Sports in which top 10 athelets are

In [19]:

```
plot=top_ten.plot.bar(x='Sport',y='earnings ($ million)',figsize=(12,6));
plt.title('Top 10 highest paid Athletes (1990-2020)');
plt.ylabel('Earnings ($ million)');
```



Who is the most listed athlete in 'Forbes highest Paid Athletes' history, also include sport, nationality and year in which they listed in Forbes?

In [20]:

```
new_df=df.copy()#copy of original dataframe
```

In [21]:

```
new_df.set_index('Name',inplace=True)#make name column index
```

In [22]:

```
max_listed_athletes=df.Name.mode().tolist()
```

In [23]:

```
new_df.loc[max_listed_athletes][['Nationality','Sport','Year']]
```

Out[23]:

	Nationality	Sport	Year
Name			
Michael Jordan	USA	basketball	1990
Michael Jordan	USA	basketball	1991
Michael Jordan	USA	Basketball	1992
Michael Jordan	USA	Basketball	1993
Michael Jordan	USA	Basketball	1994

Michael Jordan	USA	basketball	1995
Michael Jordan	USA	Basketball	1996
Michael Jordan	USA	Basketball	1997
Michael Jordan	USA	Basketball	1998
Michael Jordan	USA	Basketball	1999
Michael Jordan	USA	Basketball	2000
Michael Jordan	USA	Basketball	2002
Michael Jordan	USA	Basketball	2003
Michael Jordan	USA	basketball	2004
Michael Jordan	USA	basketball	2005
Michael Jordan	USA	basketball	2006
Michael Jordan	USA	basketball	2007
Michael Jordan	USA	basketball	2008
Michael Jordan	USA	basketball	2009
Tiger Woods	USA	Golf	1997
Tiger Woods	USA	Golf	1998
Tiger Woods	USA	Golf	1999
Tiger Woods	USA	Golf	2000
Tiger Woods	USA	Golf	2002
Tiger Woods	USA	Golf	2003
Tiger Woods	USA	golf	2004
Tiger Woods	USA	golf	2005
Tiger Woods	USA	golf	2006
Tiger Woods	USA	golf	2007
Tiger Woods	USA	golf	2008
Tiger Woods	USA	golf	2009
Tiger Woods	USA	golf	2010
Tiger Woods	USA	golf	2011
Tiger Woods	USA	Golf	2012
Tiger Woods	USA	Golf	2013
Tiger Woods	USA	Golf	2014
Tiger Woods	USA	Golf	2015
Tiger Woods	USA	Golf	2020

average income of athletes

In [24]:

```
sport_df=df.groupby(df.Sport)[['earnings ($ million)']].mean()
```

In [25]:

```
sport_df=sport_df.sort_values('earnings ($ million)')
```

In [26]:

```
sport_df
```

Out[26]:

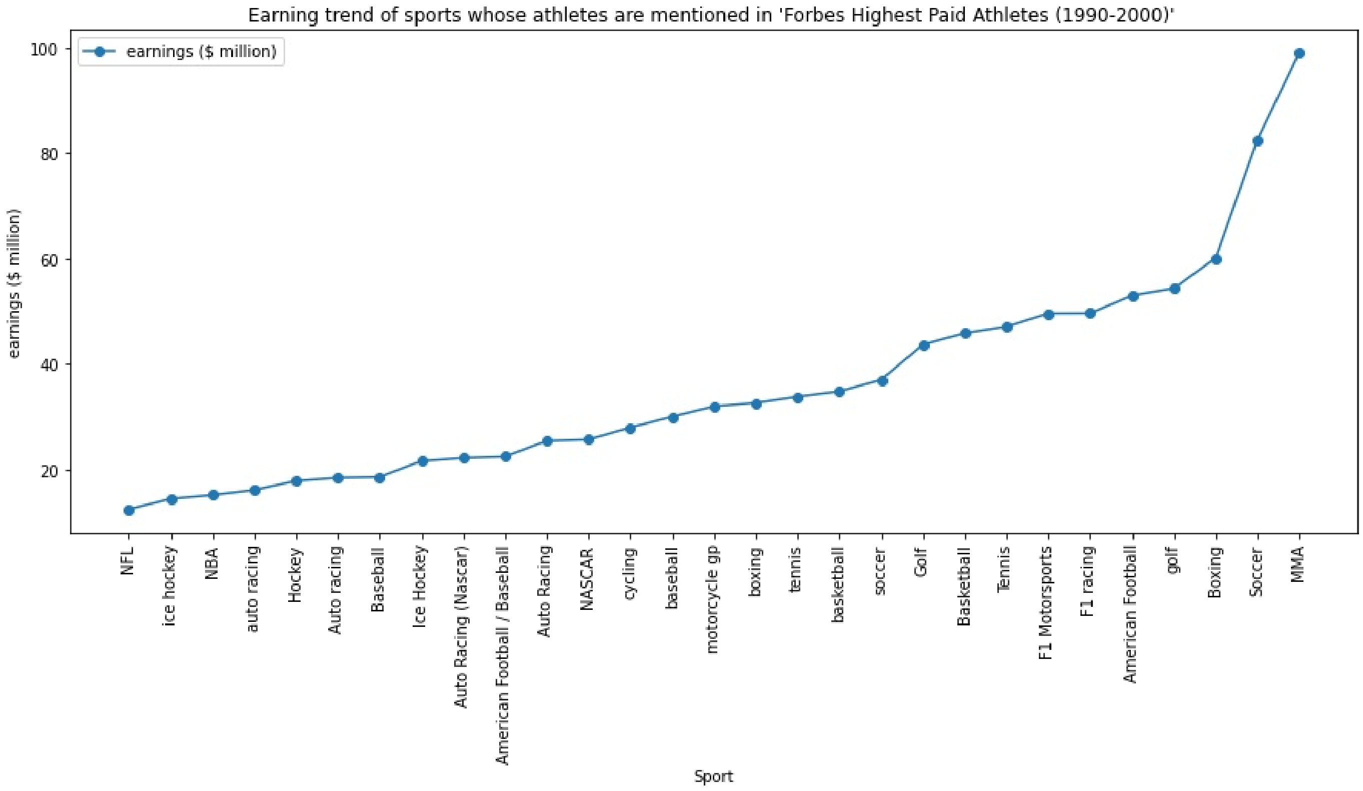
	earnings ($ million)
Sport	
NFL	12.500000
ice hockey	14.500000
NBA	15.200000
auto racing	16.142857
Hockey	17.900000
Auto racing	18.500000
Baseball	18.633333
Ice Hockey	21.650000
Auto Racing (Nascar)	22.250000

American Football / Baseball	22.500000
Auto Racing	25.480000
NASCAR	25.733333
cycling	28.000000
baseball	30.066667
motorcycle gp	32.000000
boxing	32.682353
tennis	33.840000
basketball	34.837037
soccer	37.045455
Golf	43.733333
Basketball	45.879630
Tennis	47.116667
F1 Motorsports	49.600000
F1 racing	49.625000
American Football	53.011765
golf	54.345000
Boxing	60.110345
Soccer	82.545455
MMA	99.000000

In [27]:

```
plt.figure(figsize=(15,6))

plt.plot(sport_df[['earnings ($ million)']].index[:].tolist(),sport_df[['earnings ($ million)']],marker='o')
plt.xticks(rotation=90)
plt.xlabel('Sport')
plt.ylabel('earnings ($ million)')
plt.title("Earning trend of sports whose athletes are mentioned in 'Forbes Highest Paid Athletes (1990-2000)' ")
plt.legend(sport_df[['earnings ($ million)']]);
```

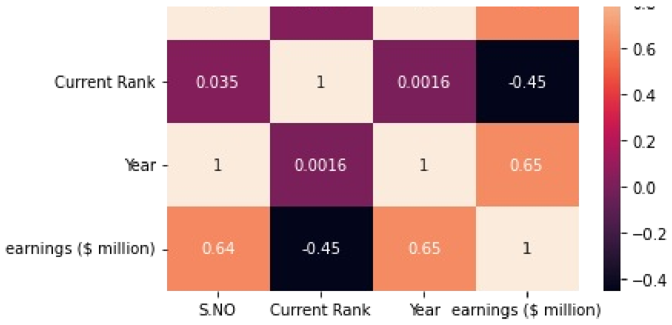


In [28]:

```
sns.heatmap(df.corr(),annot=True)
```

Out[28]:

```
<AxesSubplot:>
```



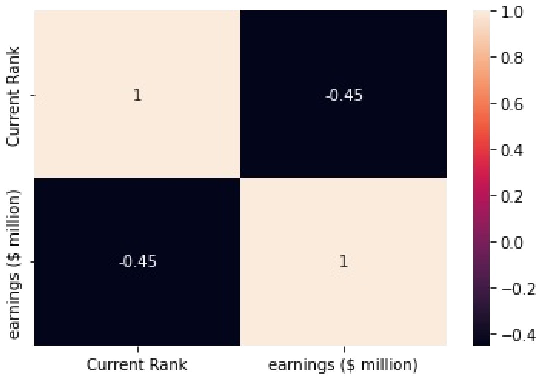
There is a strong correlation between earnings and year of earning. As well as, there is a good correlation between serial number and earnings which doesn't really make any sense since its just a serial number.

In [29]:

```
#correlation between rank and earning.
data=df[['Current Rank','earnings ($ million)']]
corr=data.corr()
sns.heatmap(corr, annot=True)
```

Out[29]:

```
<AxesSubplot:>
```



according to the above correlation grid, there is a weak correlation between the athletes current rank and the earnings.

which sport pays the most to its athletes?

In [30]:

```
df1=df[['Sport','earnings ($ million)']]
df1.head()
```

Out[30]:

	Sport	earnings ($ million)
0	boxing	28.6
1	boxing	26.0
2	boxing	13.0
3	auto racing	10.0
4	auto racing	9.0

In [31]:

```
df1.max()
```

Out[31]:

```
Sport                   tennis
earnings ($ million)     300.0
dtype: object
```

In [32]:

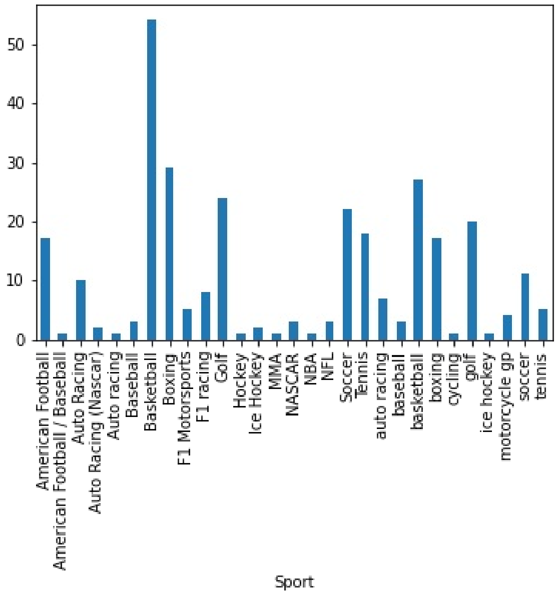
```
#question: who is the highest paid athlete?
df2=df[['Name','Sport','earnings ($ million)']]
df2.max()
```

Out[32]:

```
Name                    Wayne Gretzky
Sport                          tennis
earnings ($ million)            300.0
dtype: object
```

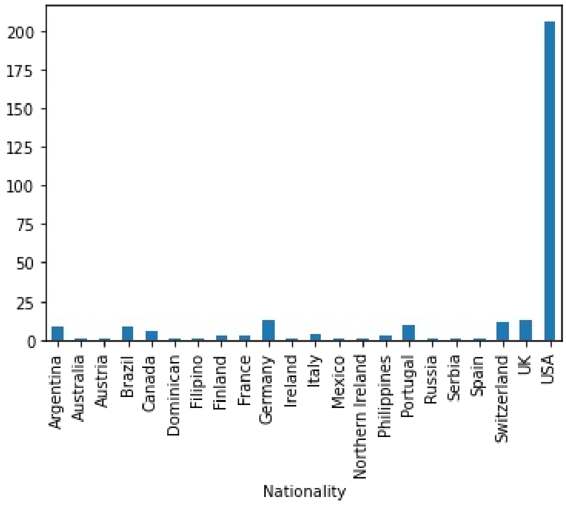
In [33]:

```
sport=df.groupby(['Sport'])
earning=df.groupby(['earnings ($ million)'])
plt.clf()
df.groupby('Sport').size().plot(kind='bar')
plt.show()
```



In [34]:

```
Nationality=df.groupby(['Nationality'])
Nationality
plt.clf()
df.groupby(['Nationality']).size().plot(kind='bar')
plt.show()
```



Top Paid Athlete for Each Year

In [61]:

```
Top_paid_each_year = df[df['Current Rank'] == 1].sort_values(by='Year',ascending=False)

z = Top_paid_each_year[['Name','Sport','Nationality','earnings ($ million)']]
```

```
z.style.background_gradient(cmap='Reds')
```

Out[61]:

	Name	Sport	Nationality	earnings ($ million)
291	Roger Federer	Tennis	Switzerland	106.300000
281	Lionel Messi	Soccer	Argentina	127.000000
271	Floyd Mayweather	Boxing	USA	285.000000
261	Cristiano Ronaldo	Soccer	Portugal	93.000000
251	Cristiano Ronaldo	Soccer	Portugal	88.000000
241	Floyd Mayweather	Boxing	USA	300.000000
231	Floyd Mayweather	Boxing	USA	105.000000
221	Tiger Woods	Golf	USA	78.100000
211	Floyd Mayweather	Boxing	USA	85.000000
201	Tiger Woods	golf	USA	75.000000
191	Tiger Woods	golf	USA	105.000000
181	Tiger Woods	golf	USA	110.000000
171	Tiger Woods	golf	USA	115.000000
161	Tiger Woods	golf	USA	100.000000
151	Tiger Woods	golf	USA	90.000000
141	Tiger Woods	golf	USA	87.000000
131	Tiger Woods	golf	USA	80.300000
121	Tiger Woods	Golf	USA	78.000000
110	Tiger Woods	Golf	USA	69.000000
100	Michael Schumacher	Auto Racing	Germany	59.000000
90	Michael Schumacher	Auto Racing	Germany	49.000000
80	Michael Jordan	Basketball	USA	69.000000
70	Michael Jordan	Basketball	USA	78.300000
60	Mike Tyson	Boxing	USA	75.000000
50	Michael Jordan	basketball	USA	43.900000
40	Michael Jordan	Basketball	USA	30.000000
30	Michael Jordan	Basketball	USA	36.000000
20	Michael Jordan	Basketball	USA	35.900000
10	Evander Holyfield	boxing	USA	60.500000
0	Mike Tyson	boxing	USA	28.600000

2020=Roger Federer,2019=Lionel Messi,2018=Floyd Mayweather highest earning per year

In [54]:

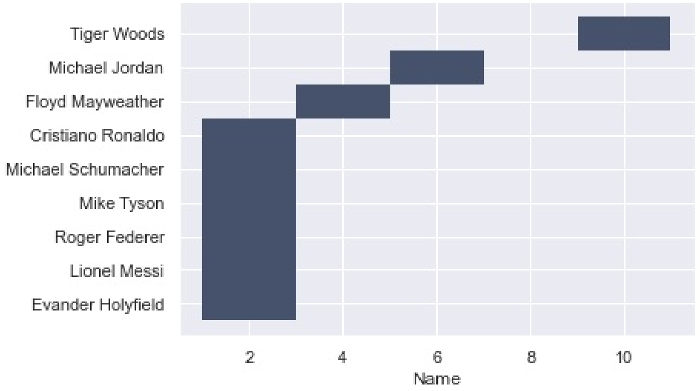
```
counts_top = Top_paid_each_year['Name'].value_counts().to_frame()



sns.histplot(
                    y = counts_top.index,
                    x = counts_top['Name'] ,


                    )

iplot(fig)
```



Athletes appearing maximum time on the list

In [68]:

```
s = df['Name'].value_counts().to_frame()[:5]
s
```

Out[68]:

	Name
Tiger Woods	19
Michael Jordan	19
Kobe Bryant	14
LeBron James	13
Michael Schumacher	13

People who have appeared once on the list.

In [76]:

```
names = df['Name'].value_counts().to_frame()
names[names['Name']==1].index
```

Out[76]:

```
Index(['Matthew Stafford', 'Aaron Rodgers', 'Rafael Nadal', 'Kirk Cousins',
       'Aaron Rogers', 'Novak Djokovic', 'Jordan Spieth', 'Cam Newton',
       'Canelo Alvarez', 'Andrew Luck', 'Rory McIlroy', 'Drew Brees',
       'James Harden', 'Lewis Hamilton', 'Russell Wilson', 'Conor McGregor',
       'Deion Sanders', 'Donovan "Razor" Ruddock', 'Terrell Suggs',
       'Eli Manning', 'Emmit Smith', 'Dennis Rodman', 'Gerhard Berger',
       'Joe Sakic', 'Cecil Fielder', 'Sergei Federov', 'Gary Sheffield',
       'Jeff Gordon', 'Buster Douglas', 'Monica Seles', 'Michael Vick',
       'Lance Armstrong', 'Muhammad Ali', 'Tom Brady', 'Michael Moorer',
       'Dale Earnhardt Jr.', 'Greg Norman', 'Carson Wentz'],
      dtype='object')
```

Only women souce google

In [71]:

In [71]:

```
monica = df[df['Name'] == 'Monica Seles']
monica
```

Out[71]:

	S.NO	Name	Nationality	Current Rank	Previous Year Rank	Sport	Year	earnings ($ million)
29	30	Monica Seles	USA	10	12	Tennis	1992	8.5

Top 5 earners of all time

In [77]:

```
top_earners_alltime = pd.pivot_table(df, index='Name',values="earnings ($ million)", aggfunc='sum')
top5_earners_all = top_earners_alltime.sort_values(by="earnings ($ million)",ascending=False)[:5]
```

In [78]:

```
top3_earners_all.style.background_gradient(cmap='Reds')
```

Out[78]:

	earnings ($ million)
Name	
Tiger Woods	1373.800000
LeBron James	844.800000
Floyd Mayweather	840.000000
Cristiano Ronaldo	787.100000
Roger Federer	781.100000

In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js