Forbers Highest 1990-2020

Introduce

Hi, my name is Dhanachote W. I practice writing queries, coding, and analyzing data to ensure I don't forget what I have learned.

By the way, I have many projects. Although they may not be impactful projects because I lack experience in data analysis, I pursue them independently and formulate questions based on the dataset.

This particular project is a mini project for practice, and I obtained the data from [Kaggle.com]

Also, I work in `VSCODE` using `Python` language for analysis and **Jupiter notebook** to create this project as a report.

Lastly, I am aware that my project is somewhat basic and that anyone can do it, and I believe someone could do it better than me. However, I am open to receiving recommendations to improve my skills in visualization, analysis, and coding. I welcome the opportunity to learn from your suggestions.

Data Name

Forbers Highest Paid Athletes 1990 - 2020

About Dataset

Context

Here is a completel list of the world's highest-paid athletes since the first list published by Forbes in 1990. In 2002,

they changed the reporting period from the full calendar year to June-to-June, and consequently, there are no records for 2001.

Content

The data is available from 1990 to 2019.

Acknowledgments

The data has been extracted from topendsports.com website

Data library

- S.NO: Serial number
- Name: Name of Forbes
- Nationality: Nationality of Forbes
- Current Rank: Current ranking based on earnings each year
- Previous Year Rank: Previous year's rank (may be NULL for some entries)
- Sport: Sport played by Forbes athlete
- Year: Year of earnings
- Earnings ($million): Earnings from sports

Prepare the dataset

Import library using for analyze the data

I use 3 library for analyze the dataset such as `pandas`, `numpy` and`*matplotlib` for create a chart.

Firstly, I import the file data set using by

`pd.read_csv` and `.info` to check the columns name and `NULL` in the dataset.

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

## load dataset from csv
data = pd.read_csv("Forbes_Richest_Athletes_1990-2020.csv")
data = pd.DataFrame(data)
data = data.reset_index()
print(data.describe())
print(data.info())
```

```
             index        S.NO  Current Rank         Year  earnings ($ million)
count   301.000000  301.000000    301.000000   301.000000            301.000000
mean    150.000000  151.000000      5.448505  2005.122924             45.516279
std      87.035433   87.035433      2.850995     9.063563             33.525337
min       0.000000    1.000000      1.000000  1990.000000              8.100000
25%      75.000000   76.000000      3.000000  1997.000000             24.000000
50%     150.000000  151.000000      5.000000  2005.000000             39.000000
75%     225.000000  226.000000      8.000000  2013.000000             59.400000
max     300.000000  301.000000     10.000000  2020.000000            300.000000
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 301 entries, 0 to 300
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   index                 301 non-null    int64
 1   S.NO                  301 non-null    int64
 2   Name                  301 non-null    object
 3   Nationality           301 non-null    object
 4   Current Rank          301 non-null    int64
 5   Previous Year Rank    277 non-null    object
 6   Sport                 301 non-null    object
 7   Year                  301 non-null    int64
 8   earnings ($ million)  301 non-null    float64
dtypes: float64(1), int64(4), object(4)
memory usage: 21.3+ KB
None
```

Rename Column

I found it difficult to type the column name `earnings ($million)` when coding for analysis, so I decided to rename it to `earnings_million` .

```
data = data.rename(columns={"earnings ($ million)" : "earning

data.head()
```

	index	S.NO	Name	Nationality	Current Rank	Previous Year Rank	Sport	Year	earnings_million
0	0	1	Mike Tyson	USA	1	NaN	boxing	1990	28.6
1	1	2	Buster Douglas	USA	2	NaN	boxing	1990	26.0
2	2	3	Sugar Ray Leonard	USA	3	NaN	boxing	1990	13.0
3	3	4	Ayrton Senna	Brazil	4	NaN	auto racing	1990	10.0
4	4	5	Alain Prost	France	5	NaN	auto racing	1990	9.0

Replace name of sport in Sport column

I found many types of sports that have the same meaning but different words. Therefore, I used `.replace` to standardize them.

```
data["Sport"].replace({"tennis": "Tennis",
                       "boxing":"Boxing",
                       "auto racing": "Auto Racing",
                       "auto Racing": "Auto Racing",
                       "Auto racing": "Auto Racing",
                       "golf": "Golf",
                       "basketball": "Basketball",
                       "NBA" : "Basketball",
                       "ice hockey": "Ice Hockey",
                       "Hockey": "Ice Hockey",
                       "baseball": "Baseball",
                       "Auto Racing (Nascar)": " NASCAR",
                       "F1 racing": "F1 Motorsports"
                       }, inplace = True)
data = data
data.head()
```

	index	S.NO	Name	Nationality	Current Rank	Previous Year Rank	Sport	Year	earnings ($ million)
0	0	1	Mike Tyson	USA	1	NaN	Boxing	1990	28.6
1	1	2	Buster Douglas	USA	2	NaN	Boxing	1990	26.0
2	2	3	Sugar Ray Leonard	USA	3	NaN	Boxing	1990	13.0
3	3	4	Ayrton Senna	Brazil	4	NaN	Auto Racing	1990	10.0
4	4	5	Alain Prost	France	5	NaN	Auto Racing	1990	9.0

Questions

Top 10 earning all the time

In this topic, I want to analyze the highest earners of all time in sports. I aim to extract data containing only the Name, Sport, and earnings_million to answer this question.

Firstly, I use `.groupby` on the columns Name and Sport and calculate the total earnings from 1990 to 2020 using the aggregation function `.agg(sum)`.

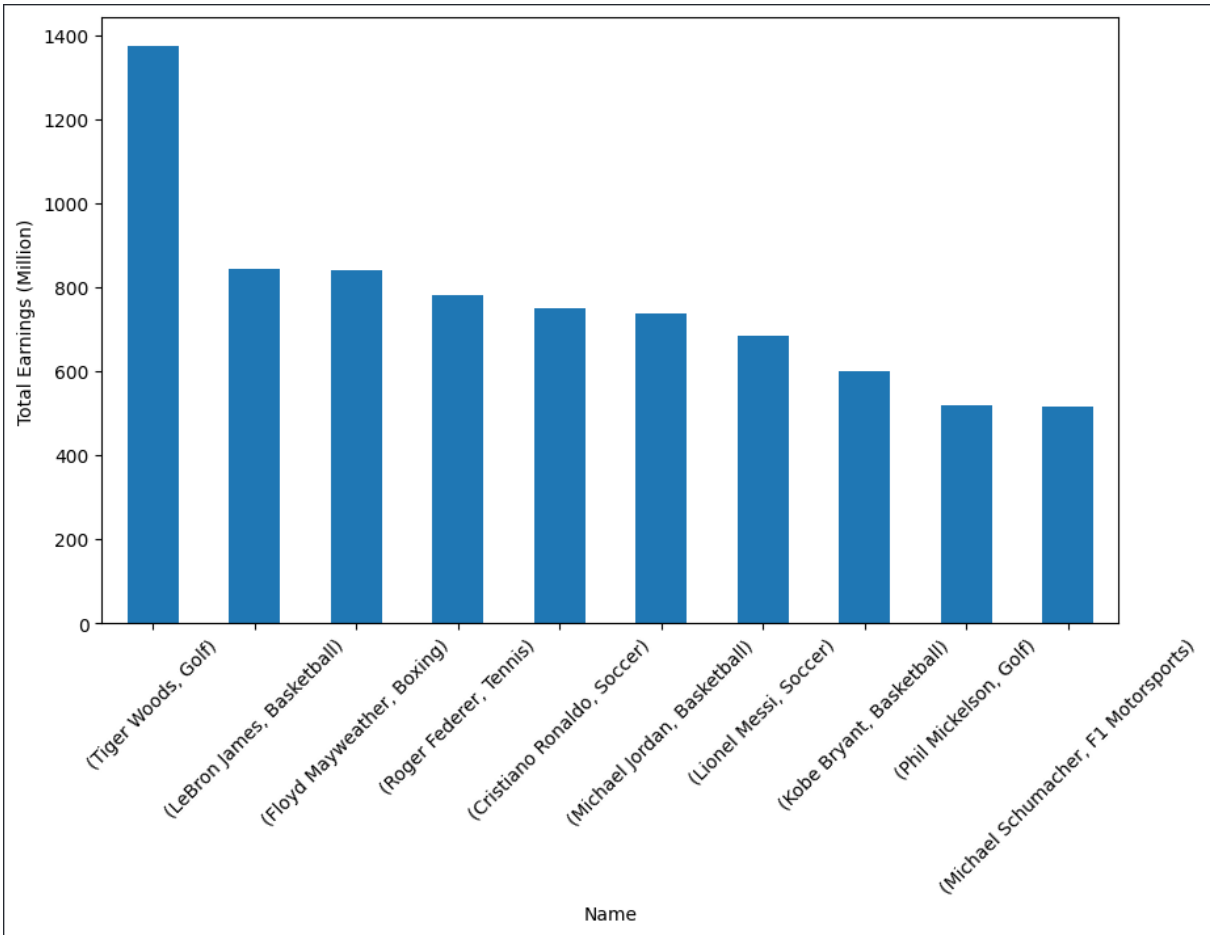
Then, I found that the highest earner is **Tiger Woods**, with earnings of $1,373.8 million.

Lastly, I created a bar chart to clarify and explain the question about the top 10 highest Forbers earners.

```
## Top earning from 1990-2020

forbes_hight = data.groupby(["Name","Sport"])["earnings_milli
forbes_hight_sort = forbes_hight.sort_values(ascending = Fals
print(pd.DataFrame(forbes_hight_sort).head(10))
forbes_hight_10 = forbes_hight_sort.head(10)

## create a bar chart

plt.figure(figsize=(10,6))
forbes_hight_10.plot(kind="bar")
```

```
plt.xlabel("Name")
plt.ylabel("Total Earnings (Million)")
plt.xticks(rotation=45)
plt.show()
```

		earnings_million
Name	Sport	
Tiger Woods	Golf	1373.8
LeBron James	Basketball	844.8
Floyd Mayweather	Boxing	840.0
Roger Federer	Tennis	781.1
Cristiano Ronaldo	Soccer	749.1
Michael Jordan	Basketball	738.8
Lionel Messi	Soccer	683.2
Kobe Bryant	Basketball	601.1
Phil Mickelson	Golf	519.9
Michael Schumacher	F1 Motorsports	516.0



Sport earning since 1990 and 2020

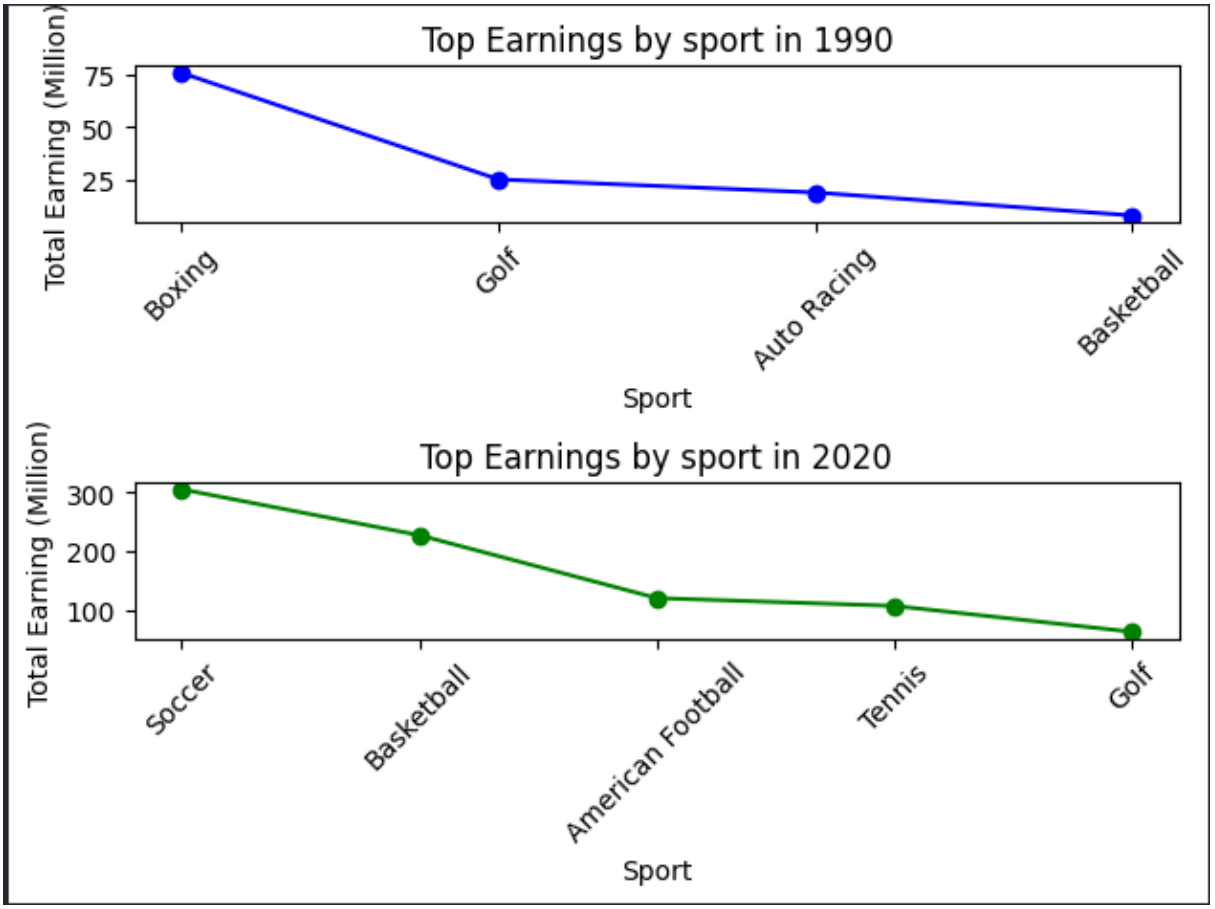
In this section, I aim to determine the earnings from 1990 to 2020 for each sport.

As depicted in the chart, various sports are represented, with Boxing emerging as the highest-earning sport in 1990. This suggests that Boxing was the most popular sport during that year.

By contrast, Soccer dominates in 2020, being the highest-earning sport. This indicates Soccer's widespread popularity in that year.

However, the most notable finding is the consistent increase in earnings for Basketball throughout the years 1990 to 2020.

```
## top earning since 1990 | 2020
earn1990 = data[data["Year"]== 1990][["Name", "Sport", "earni

earn1990 = earn1990.sort_values("earnings_million", ascending

earn2020 = data[data["Year"]==2020][["Name", "Sport", "earnin

earn2020 = earn2020.sort_values("earnings_million", ascending

fig, ax=plt.subplots(2,1)

## 1990
ax[0].plot(earn1990.index, earn1990["earnings_million"], colo
ax[0].set_title("Top Earnings by sport in 1990")
ax[0].set_xlabel("Sport")
ax[0].set_ylabel("Total Earning (Million)")


## 2020
ax[1].plot(earn2020.index, earn2020["earnings_million"], colo
ax[1].set_title("Top Earnings by sport in 2020")
ax[1].set_xlabel("Sport")
ax[1].set_ylabel("Total Earning (Million)")

for ax in ax:
    ax.tick_params(axis="x", rotation=45)

plt.tight_layout()
plt.show()
```

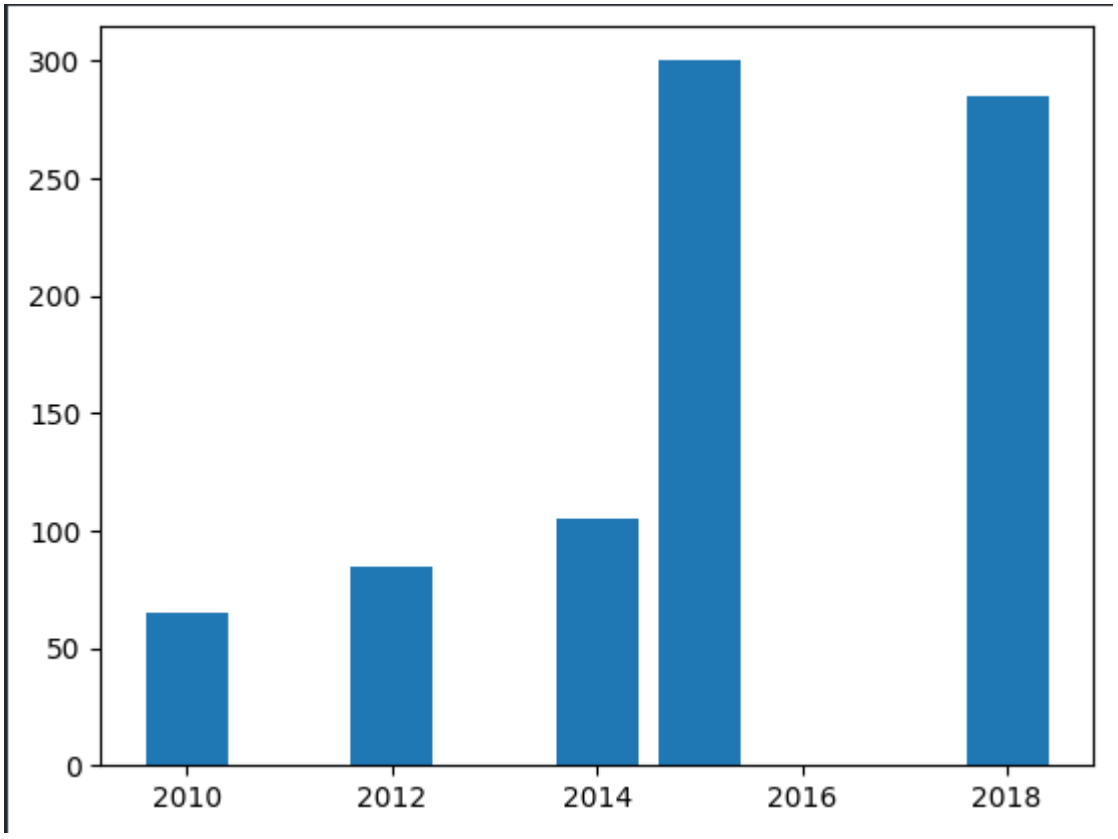


The earning chart of Floyd Mayweather

For example, I bring only Floyd Mayweather to see how much he earning in each year and his sport.
Therefore, he can make money from his sport $300 million in 2015 and decrease $15million for the next two years.

```
import matplotlib.pyplot as plt

fm = data[data["Name"]== "Floyd Mayweather"]
                    .sort_values("earnings_million", ascendin
print(fm[["Name", "Nationality", "Year", "earnings_million"]]

fig, ax=plt.subplots()
ax.bar(fm["Year"], fm["earnings_million"])
plt.show()
```

	Name	Nationality	Year	earnings_million
241	Floyd Mayweather	USA	2015	300.0
271	Floyd Mayweather	USA	2018	285.0
231	Floyd Mayweather	USA	2014	105.0
211	Floyd Mayweather	USA	2012	85.0
192	Floyd Mayweather	USA	2010	65.0



Sport Category

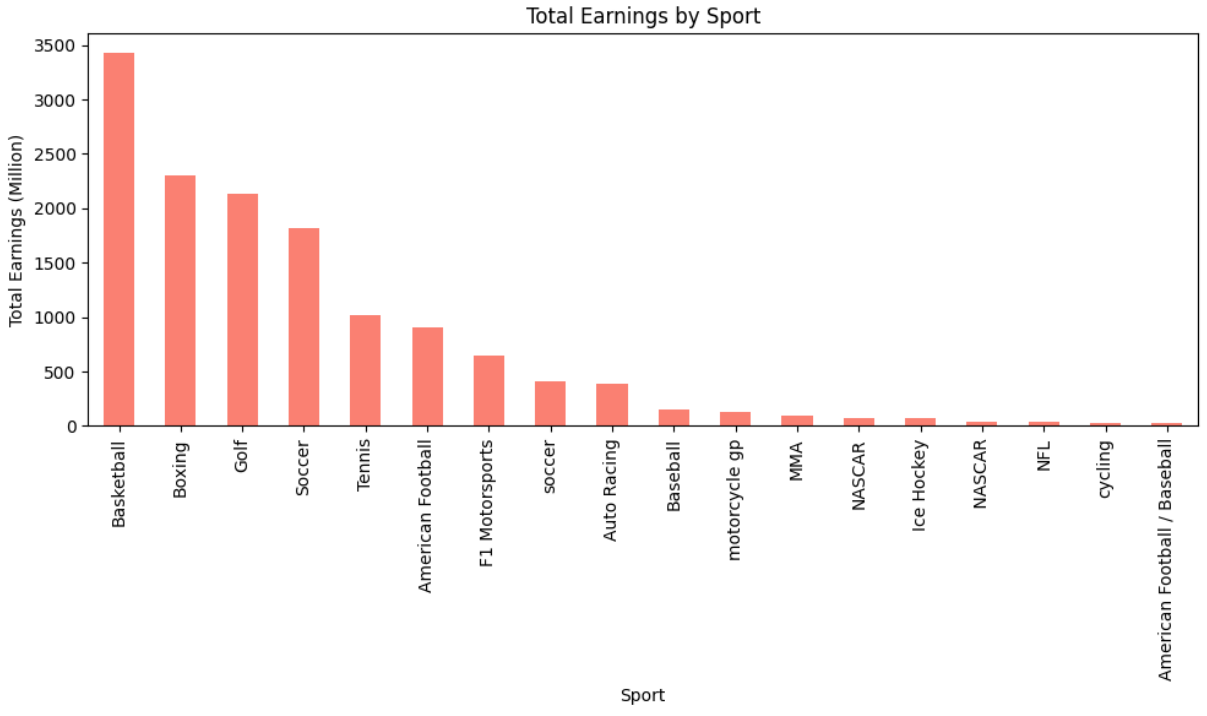
Even though Tiger Woods is the highest-earning Forbes athlete in the sport of Golf, the total earnings for Golf from 1990 to 2020 do not rank it as the highest-earning sport.

If you observe the bar chart, you'll notice that Basketball holds the highest earnings among all sports, with Golf ranking third.

Therefore, What does it mean, it means that Basketball

consistently shows an increase in earnings over time, surpassing both Golf and Boxing (which ranks second).

```
# Grouping data by sport and summing earnings
sport_earning = data.groupby("Sport")["earnings_million"].agg

# Sorting values in ascending order
sport_earning_sort = sport_earning.sort_values(ascending=Fals

# create the bar chart
plt.figure(figsize=(10,6))
sport_earning_sort.plot(kind="bar",
                        color="salmon")
plt.xlabel("Sport")
plt.ylabel("Total Earnings (Million)")
plt.title("Total Earnings by Sport")
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```



For the Last part

I am coding data within this scope to identify who earned the most between the years 2018 and 2020. After analyzing the data for 2018, it was found that the highest earner was **Floyd Mayweather**, in the sport of boxing.

However, upon directly examining the data, it becomes apparent that the total earnings from soccer surpass those from boxing because, in that year, Forbes reported higher earnings in soccer. Moreover, soccer sports continue to outperform even in 2020, with total earnings matching those of 2018.

In conclusion,
soccer emerges as the most lucrative sport from 2018 to 2022 going with basketball, based on the total earnings reported by Forbes, despite not holding the highest individual earnings.

```
## Total earning sport in 2018 : 2020
    ## 2018
earn2018 = data[data["Year"]== 2018][["Name", "Sport",  "earn

earn2018 = earn2018.sort_values("earnings_million", ascending=
print(earn2018.head(10))
count_2018 = earn2018.groupby(["Sport"]).size().sort_values(a
count_2018 = pd.DataFrame(count_2018)









    ## 2020
earn2020 = data[data["Year"]== 2020][["Name", "Sport", "earni

earn2020 = earn2020.sort_values("earnings_million", ascending=
print(earn2020.head(10))
count_2020 = earn2020.groupby(["Sport"]).size().sort_values(a
count_2020 = pd.DataFrame(count_2020)
```

Name	Sport	earnings_million
Floyd Mayweather	Boxing	285.0
Lionel Messi	Soccer	111.0
Cristiano Ronaldo	Soccer	108.0
Conor McGregor	MMA	99.0
Neymar	Soccer	90.0
LeBron James	Basketball	85.5
Roger Federer	Tennis	77.2
Stephen Curry	Basketball	76.9
Matt Ryan	American Football	67.3
Matthew Stafford	American Football	59.5

Name	Sport	earnings_million
Roger Federer	Tennis	106.3
Cristiano Ronaldo	Soccer	105.0
Lionel Messi	Soccer	104.0
Neymar	Soccer	95.5
LeBron James	Basketball	88.2
Stephen Curry	Basketball	74.4
Kevin Durant	Basketball	63.9
Tiger Woods	Golf	62.3
Kirk Cousins	American Football	60.5
Carson Wentz	American Football	59.1

```
##
count_2018, count_2020
```

```
(                  0
 Sport
 Soccer            3
 American Football 2
 Basketball        2
 Boxing            1
 MMA               1
 Tennis            1,
                   0
 Sport
 Basketball        3
 Soccer            3
 American Football 2
 Golf              1
 Tennis            1)
```

```
sns.countplot(data["Sport"].sort_values(ascending=False))
```