web scraping basketball game

In [1]:
```
year=2018
str = "https://www.basketball-reference.com/leagues/NBA_{}_per_game.html"
url=str.format(year)
```

import pandas as pd

In [2]:
```
# Read html webpage using pandas, where df is Dataframe and pd= pandas

df=pd.read_html(url,header =0)
```

In [4]:
```
print(df)
```

```
[      Rk           Player Pos Age   Tm   G  GS    MP   FG  FGA  ...   FT%  ORB  \
0      1    Álex Abrines  SG  24  OKC  75   8  15.1  1.5  3.9  ...  .848  0.3
1      2      Quincy Acy  PF  27  BRK  70   8  19.4  1.9  5.2  ...  .817  0.6
2      3   Steven Adams   C  24  OKC  76  76  32.7  5.9  9.4  ...  .559  5.1
3      4    Bam Adebayo   C  20  MIA  69  19  19.8  2.5  4.9  ...  .721  1.7
4      5  Arron Afflalo  SG  32  ORL  53   3  12.9  1.2  3.1  ...  .846  0.1
..   ...            ...  ..  ..  ...  ..  ..   ...  ...  ...  ...   ...  ...
685  537   Tyler Zeller   C  28  BRK  42  33  16.7  3.0  5.5  ...  .667  1.5
686  537   Tyler Zeller   C  28  MIL  24   1  16.9  2.6  4.4  ...  .895  2.0
687  538    Paul Zipser  SF  23  CHI  54  12  15.3  1.5  4.3  ...  .760  0.2
688  539     Ante Žižić   C  21  CLE  32   2   6.7  1.5  2.1  ...  .724  0.8
689  540    Ivica Zubac   C  20  LAL  43   0   9.5  1.4  2.8  ...  .765  1.0

     DRB  TRB  AST  STL  BLK  TOV   PF   PTS
0    1.2  1.5  0.4  0.5  0.1  0.3  1.7   4.7
1    3.1  3.7  0.8  0.5  0.4  0.9  2.1   5.9
2    4.0  9.0  1.2  1.2  1.0  1.7  2.8  13.9
3    3.8  5.5  1.5  0.5  0.6  1.0  2.0   6.9
4    1.2  1.2  0.6  0.1  0.2  0.4  1.1   3.4
..   ...  ...  ...  ...  ...  ...  ...   ...
685  3.1  4.6  0.7  0.2  0.5  0.8  1.9   7.1
686  2.7  4.6  0.8  0.3  0.6  0.5  2.0   5.9
687  2.2  2.4  0.9  0.4  0.3  0.8  1.6   4.0
688  1.1  1.9  0.2  0.1  0.4  0.3  0.9   3.7
689  1.8  2.9  0.6  0.2  0.3  0.6  1.1   3.7

[690 rows x 30 columns]]
```

In [7]:
```
# to check the lenght of table
len(df)
```

Out[7]:
```
1
```

In [9]:
```
# to select the first table
df[0]
```

Out[9]:

	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
0	1	Álex Abrines	SG	24	OKC	75	8	15.1	1.5	3.9	...	.848	0.3	1.2	1.5	0.4	0.5	0.1	0.3	1.7	4.7
1	2	Quincy Acy	PF	27	BRK	70	8	19.4	1.9	5.2	...	.817	0.6	3.1	3.7	0.8	0.5	0.4	0.9	2.1	5.9
2	3	Steven Adams	C	24	OKC	76	76	32.7	5.9	9.4	...	.559	5.1	4.0	9.0	1.2	1.2	1.0	1.7	2.8	13.9
3	4	Bam Adebayo	C	20	MIA	69	19	19.8	2.5	4.9	...	.721	1.7	3.8	5.5	1.5	0.5	0.6	1.0	2.0	6.9
4	5	Arron Afflalo	SG	32	ORL	53	3	12.9	1.2	3.1	...	.846	0.1	1.2	1.2	0.6	0.1	0.2	0.4	1.1	3.4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
685	537	Tyler Zeller	C	28	BRK	42	33	16.7	3.0	5.5	...	.667	1.5	3.1	4.6	0.7	0.2	0.5	0.8	1.9	7.1
686	537	Tyler Zeller	C	28	MIL	24	1	16.9	2.6	4.4	...	.895	2.0	2.7	4.6	0.8	0.3	0.6	0.5	2.0	5.9
687	538	Paul Zipser	SF	23	CHI	54	12	15.3	1.5	4.3	...	.760	0.2	2.2	2.4	0.9	0.4	0.3	0.8	1.6	4.0
688	539	Ante Žižić	C	21	CLE	32	2	6.7	1.5	2.1	...	.724	0.8	1.1	1.9	0.2	0.1	0.4	0.3	0.9	3.7
689	540	Ivica Zubac	C	20	LAL	43	0	9.5	1.4	2.8	...	.765	1.0	1.8	2.9	0.6	0.2	0.3	0.6	1.1	3.7

In [15]:
```
# assigning df table into 2018 i.e
df2018 = df[0]
```

In [25]:
```
# for this table every 20  row there is another header so to remove the header to contain subsequent header

df2018[df2018.Age == "Age"]
```

Out[25]:

	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
20	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
47	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
73	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
98	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
127	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
564	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
587	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
612	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
640	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
663	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS

In [26]:
```
## the length of df2018
len(df2018[df2018.Age == "Age"])
```

Out[26]:
```
26
```

In [27]:
```
# to drop the header
df = df2018.drop(df2018[df2018.Age == "Age"].index)
```

In [28]:
```
# to check the rows and column of the dataframe after drop
df.shape
```

Out[28]:
```
(664, 30)
```

In [29]:
```
# before drop
df2018.shape
```

Out[29]:
```
(690, 30)
```

to implement Exploratory Data Analysis (EDA)

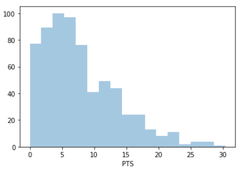
In [30]:
```
import seaborn as sns
```

making histogram

In [40]:
```
# kde is False because to retain original frequency, please note if the kds = True is means probability.
sns.distplot(df.PTS,kde=False)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
<AxesSubplot:xlabel='PTS'>
```

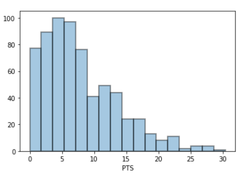
Out[40]:



change the color line

In [42]:
```
sns.distplot(df.PTS,
             kde=False,
             hist_kws=dict(edgecolor="black", linewidth=2))
```

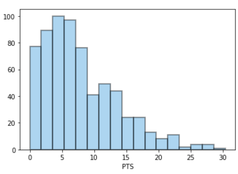
Out[42]:
```
<AxesSubplot:xlabel='PTS'>
```



change the bar fill color

In [43]:
```
sns.distplot(df.PTS,
             kde=False,
             hist_kws=dict(edgecolor="black",linewidth=2),
              color="#3498DB")
```

Out[43]:
```
<AxesSubplot:xlabel='PTS'>
```



In [ ]:
```
```