Web Scrapping on
IPL 2008-2020

By:

K.Kishta Reddy

ABOUT US

Name:K.Kishta Reddy

Qualification:B.sc

Experience:Fresher

Batch:111

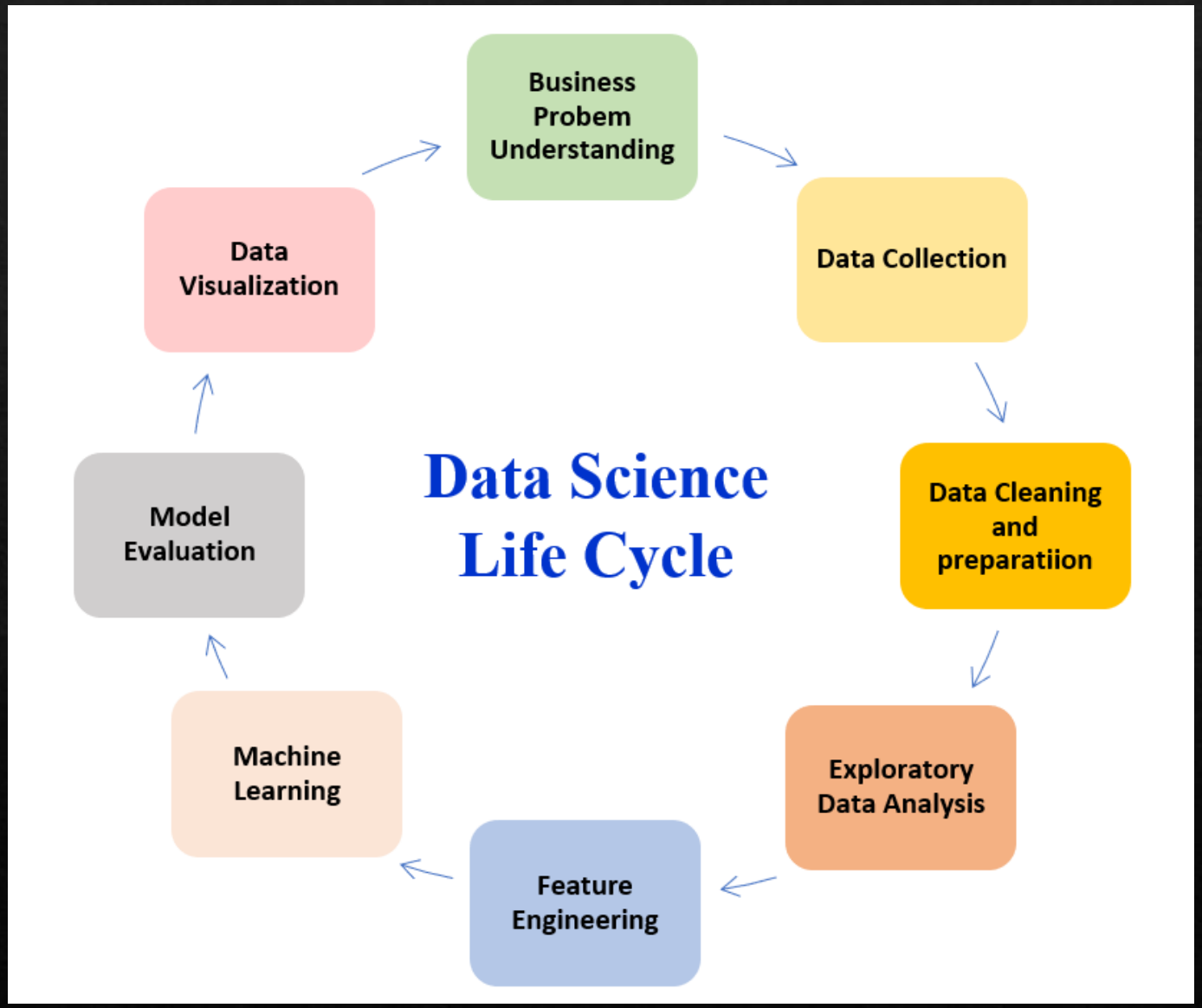
TABLE OF CONTENT

What is Web Scrapping & Why?

It is an automated method used to extract large amounts of data from websites. The data on the websites are unstructured. Web scrapping helps collect these unstructured data and store it in a structured form. There are different ways to scrape websites such as online services, we'll see how to implement web scrapping with python.

It is used to collect large information from websites.

DATA SCIENCE LIFECYCLE



LIBRARIES USED:

Warnings

Requests

Beautiful Soup

Regex (re)

Pandas

NumPy

Matplotlib

Seaborn

BUSINESS STATEMENT:

The following analysis let the player who scored more runs , fiftys , hundreds etc..

Website: IPLT20.COM

How did You scrap?

Step 1: Find the URL that you want to scrape
step 2: Inspecting the page
step 3: Find the data you want to extract
step 4: Write the code
step 5: Run the code & Extract the data
step 6: Store the data in a required format

CLEANING OF EXTRACTED DATA

Extracted data is in the form of unstructured format. So to convert it into structured format python has provided Pandas library. With the help of Pandas Library

we can convert unstructured data into structured data (Tabular Format).

The unstructured is always in the string format which contains some special characters and unwanted stuffs which we don't want. So to clean those unwanted things python has provided re Library. With the help of re Library we can clean the unwanted things and get the desired things which we want.

So after cleaning the unwanted things now we can convert the whole data into Data Frame by using Pandas Library.

DATA FRAME BEFORE CLEANING

	names	matches	innings	notouts	runs	balls	high_score	avg	strike_rate	fours	sixes	fiftys	hundreds	season
0	\n\n Shaun\n ...	\n 11\n	\n 11\n	\n 2\n	\n 616\n	\n 441\n	\n 115\n	\n 68.44\n	\n 139.68\n	\n 59\n	\n 26\n	\n 5\n	\n 1\n	2008
1	\n\n Gautam\n ...	\n 14\n	\n 14\n	\n 1\n	\n 534\n	\n 379\n	\n 86\n	\n 41.07\n	\n 140.89\n	\n 68\n	\n 8\n	\n 5\n	\n 0\n	2008
2	\n\n Sanath\n ...	\n 14\n	\n 14\n	\n 2\n	\n 518\n	\n 309\n	\n 114*\n	\n 43.16\n	\n 167.63\n	\n 58\n	\n 31\n	\n 2\n	\n 1\n	2008
3	\n\n Shane\n ...	\n 15\n	\n 15\n	\n 5\n	\n 472\n	\n 311\n	\n 76*\n	\n 47.20\n	\n 151.76\n	\n 47\n	\n 19\n	\n 4\n	\n 0\n	2008
4	\n\n Graeme\n ...	\n 11\n	\n 11\n	\n 2\n	\n 441\n	\n 362\n	\n 91\n	\n 49.00\n	\n 121.82\n	\n 54\n	\n 8\n	\n 3\n	\n 0\n	2008
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1832	\n\n Khaleel\n ...	\n 7\n	\n 1\n	\n 0\n	\n 0\n	\n 2\n	\n 0*\n	\n 0.00\n	\n 0.00\n	\n 0\n	\n 0\n	\n 0\n	\n 0\n	2020
1833	\n\n Arshdeep\n ...	\n 8\n	\n 1\n	\n 0\n	\n 0\n	\n 3\n	\n 0*\n	\n 0.00\n	\n 0.00\n	\n 0\n	\n 0\n	\n 0\n	\n 0\n	2020
1834	\n\n Daniel\n ...	\n 3\n	\n 1\n	\n 0\n	\n 0\n	\n 2\n	\n 0*\n	\n 0.00\n	\n 0.00\n	\n 0\n	\n 0\n	\n 0\n	\n 0\n	2020
1835	\n\n Shreevats\n ...	\n 2\n	\n 2\n	\n 0\n	\n 0\n	\n 4\n	\n 0*\n	\n 0.00\n	\n 0.00\n	\n 0\n	\n 0\n	\n 0\n	\n 0\n	2020
1836	\n\n Trent\n ...	\n 15\n	\n 1\n	\n 0\n	\n 0\n	\n 1\n	\n 0*\n	\n 0.00\n	\n 0.00\n	\n 0\n	\n 0\n	\n 0\n	\n 0\n	2020

1837 rows × 14 columns

In this Data Frame 1837 rows x 14 columns are present.

Before Cleaning DataFrame

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1837 entries, 0 to 1836
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   names        1837 non-null   object
 1   matches      1837 non-null   object
 2   innings      1837 non-null   object
 3   notouts      1837 non-null   object
 4   runs         1837 non-null   object
 5   balls        1837 non-null   object
 6   high_score   1837 non-null   object
 7   avg          1837 non-null   object
 8   strike_rate  1837 non-null   object
 9   fours        1837 non-null   object
 10  sixes        1837 non-null   object
 11  fiftys       1837 non-null   object
 12  hundreds     1837 non-null   object
 13  season       1837 non-null   int64 
dtypes: int64(1), object(13)
memory usage: 201.0+ KB
```

DATA FRAME AFTER CLEANING

	names	matches	innings	notouts	runs	balls	high_score	avg	strike_rate	fours	sixes	fiftys	hundreds	season
0	Shaun Marsh	11	11	2	616	441	115	68.44	139.68	59	26	5	1	2008
1	Gautam Gambhir	14	14	1	534	379	86	41.07	140.89	68	8	5	0	2008
2	Sanath Jayasuriya	14	14	2	518	309	114	43.16	167.63	58	31	2	1	2008
3	Shane Watson	15	15	5	472	311	76	47.20	151.76	47	19	4	0	2008
4	Graeme Smith	11	11	2	441	362	91	49.00	121.82	54	8	3	0	2008
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1832	Khaleel Ahmed	7	1	0	0	2	0	0.00	0.00	0	0	0	0	2020
1833	Arshdeep Singh	8	1	0	0	3	0	0.00	0.00	0	0	0	0	2020
1834	Daniel Sams	3	1	0	0	2	0	0.00	0.00	0	0	0	0	2020
1835	Shreevats Goswami	2	2	0	0	4	0	0.00	0.00	0	0	0	0	2020
1836	Trent Boult	15	1	0	0	1	0	0.00	0.00	0	0	0	0	2020

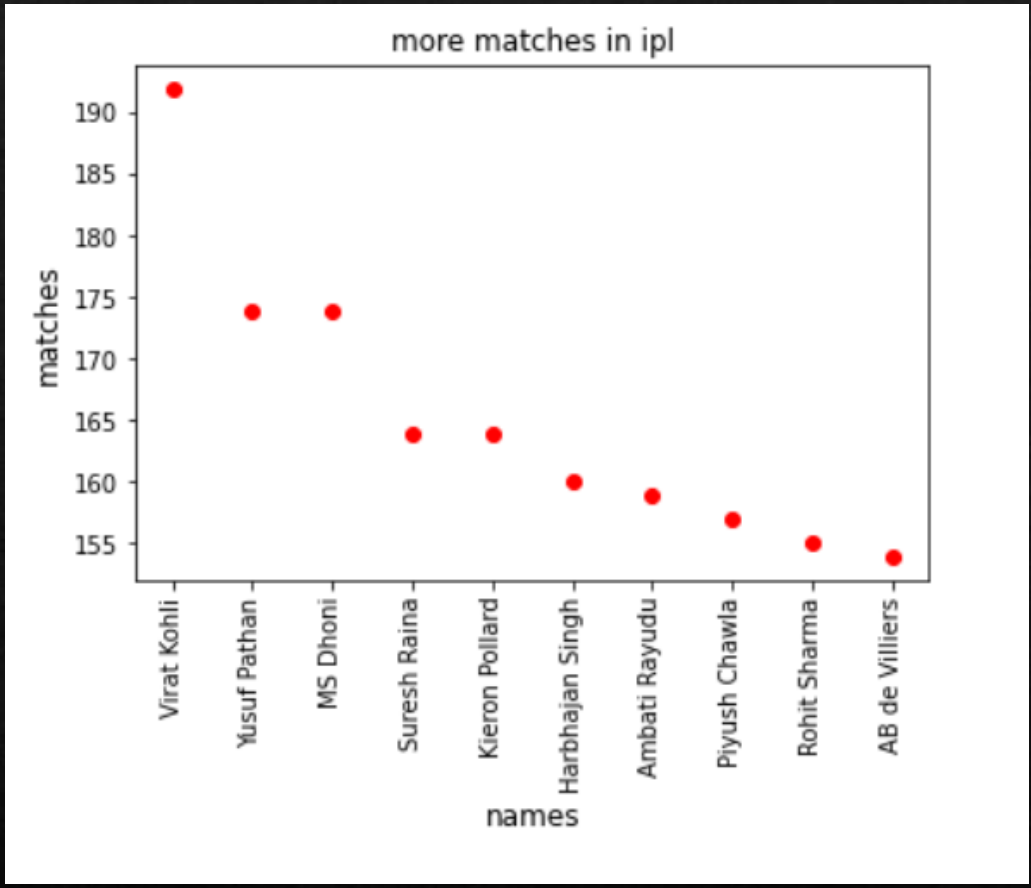
In this Data Frame 1837 rows x 14 columns are present.

After cleaning DataFrame

```
df_ipl.info()
```

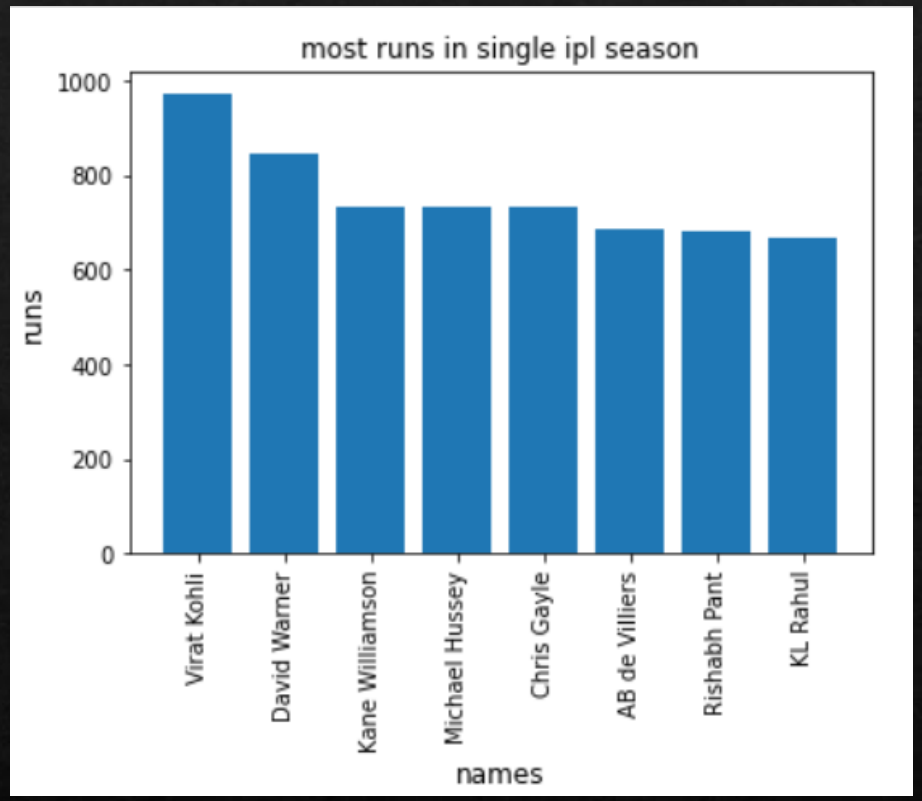
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1837 entries, 0 to 1836
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   names        1837 non-null   object 
 1   matches      1837 non-null   int64  
 2   innings      1837 non-null   int64  
 3   notouts      1837 non-null   int64  
 4   runs         1837 non-null   int64  
 5   balls        1837 non-null   int64  
 6   high_score   1837 non-null   int64  
 7   avg          1837 non-null   float64
 8   strike_rate  1837 non-null   float64
 9   fours        1837 non-null   int64  
 10  sixes        1837 non-null   int64  
 11  fiftys       1837 non-null   int64  
 12  hundreds     1837 non-null   int64  
 13  season       1837 non-null   int64  
dtypes: float64(2), int64(11), object(1)
memory usage: 201.0+ KB
```

WHO PLAYED MORE MATCHES IN IPL?



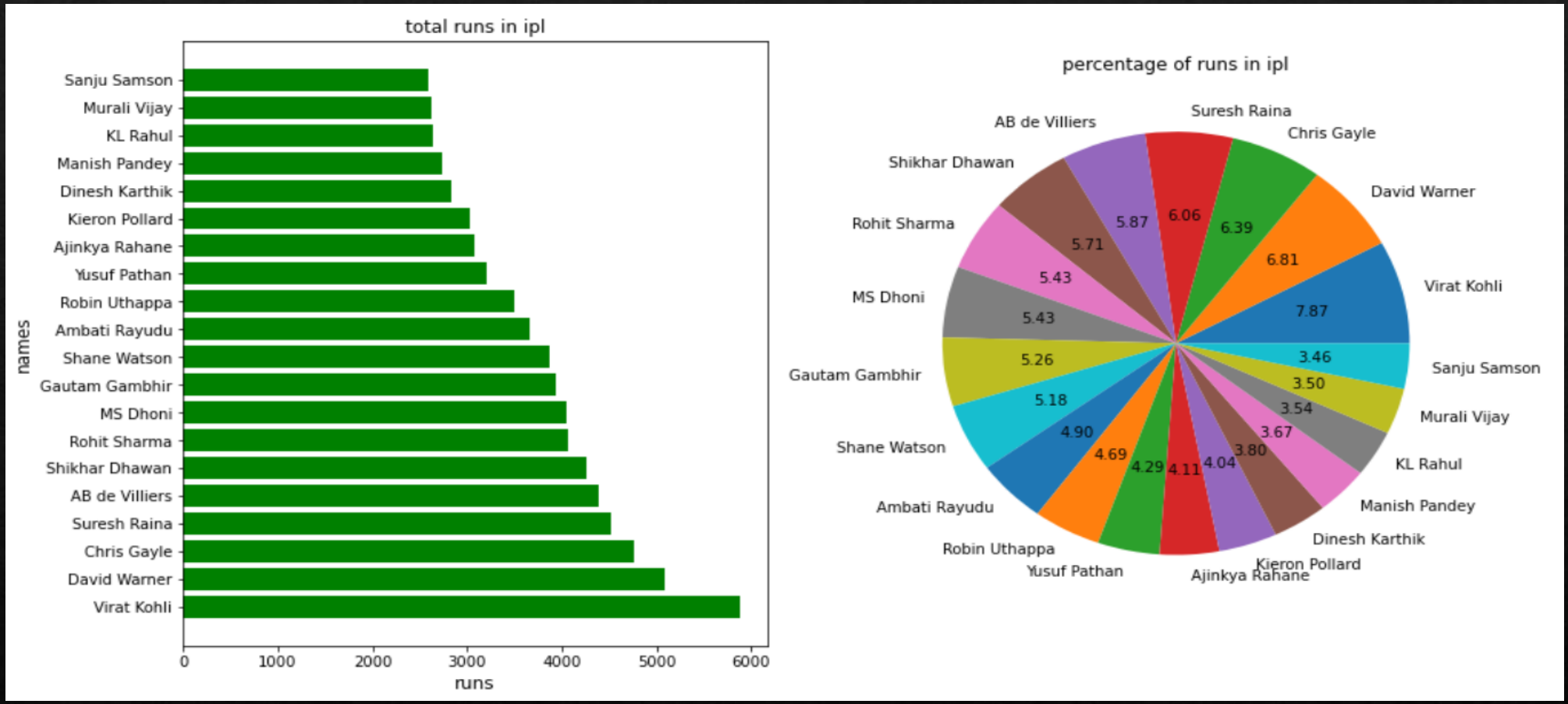
From above Scatter Plot **VIRAT KOHLI** played more matches throughout IPL

WHO SCORED MORE RUNS IN IPL SINGLE SEASON?



In single IPL season **VIRAT KOHLI** scored more runs

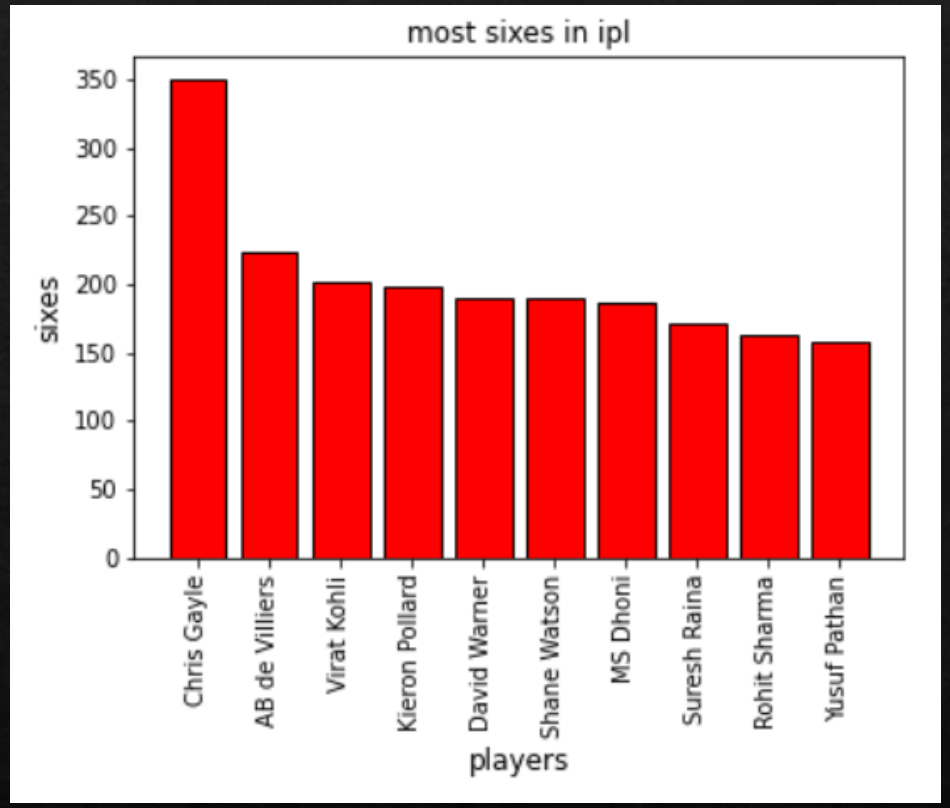
WHO SCORED MORE RUNS IN IPL?



Throughout all IPL seasons **VIRAT KOHLI** scored more runs &

The Percentage of runs is 7.87

WHO HIT MORE SIXES IN IPL?

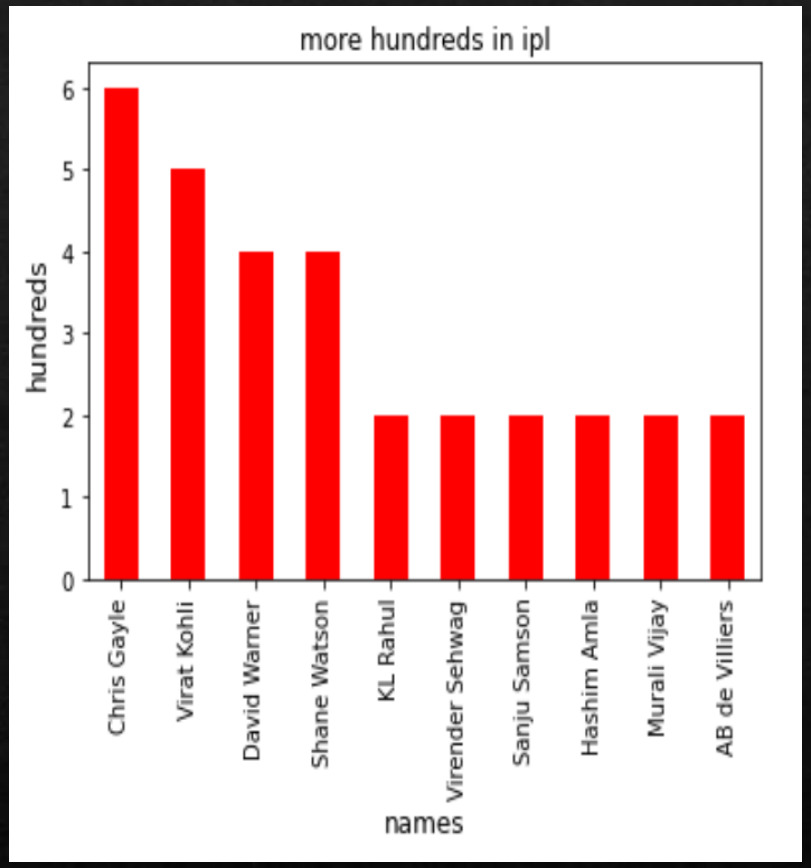


Chris Gayle hits more sixes in all IPL seasons and the difference between the first and second is Morethan 100sixes

WHO HIT MORE FIFTYS IN IPL?

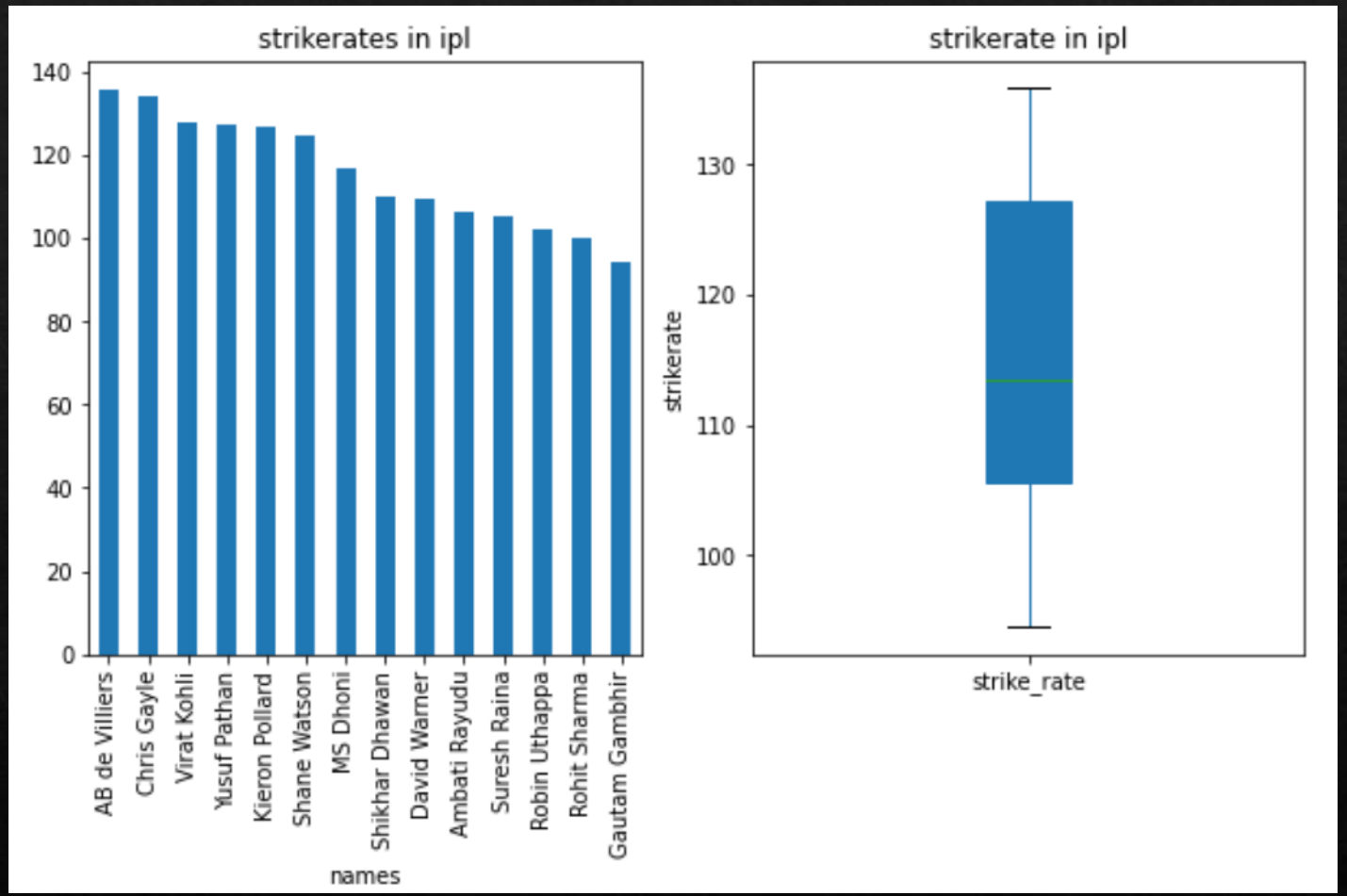
David Warner has more number of fiftys in ipl all seasons and second most is virat kohli

WHO HIT MORE NUMBER OF HUNDREDS IN IPL?



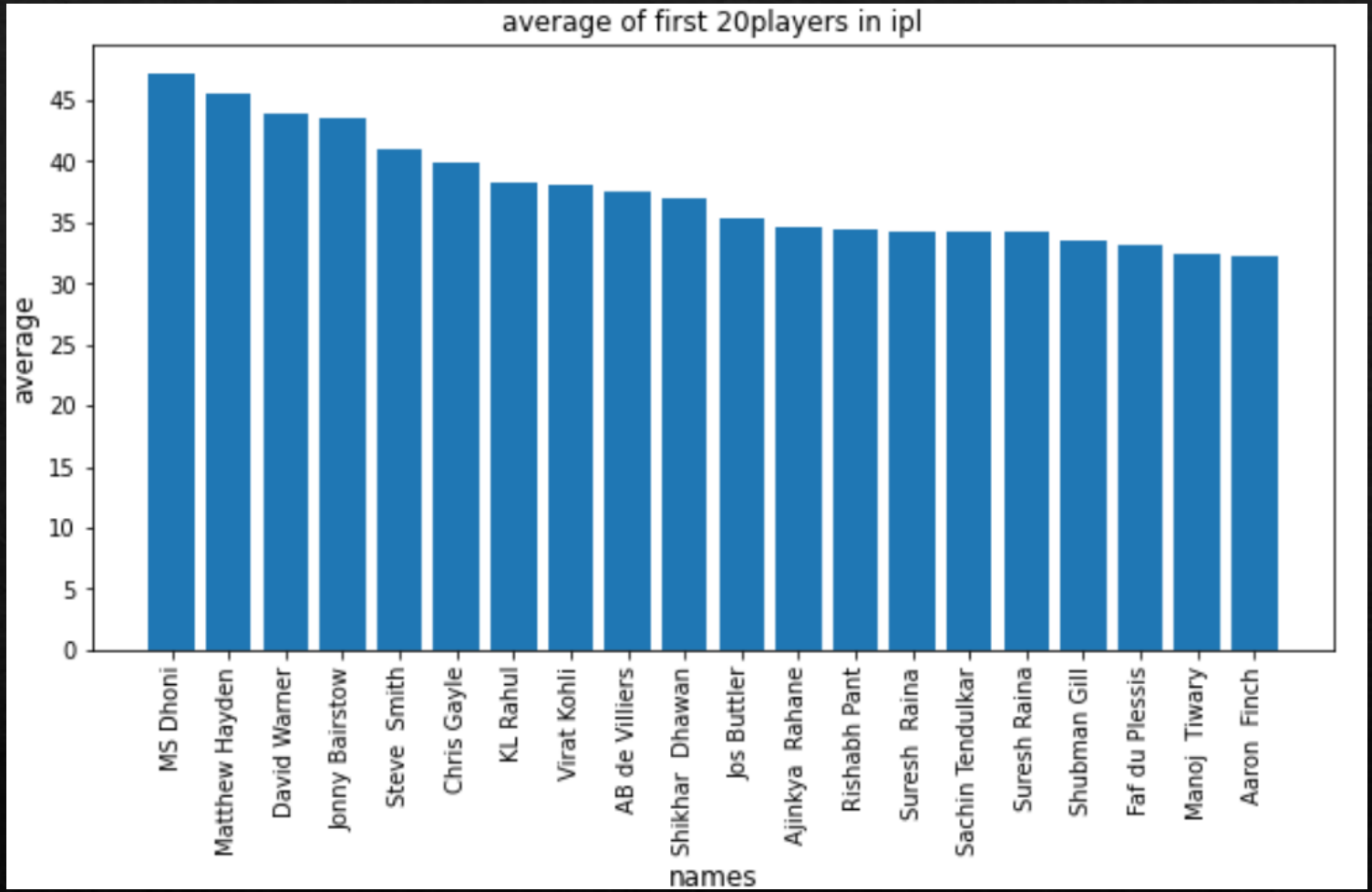
Chris Gayle hit more number of hundreds in ipl he has 6 ipl hundreds and next is Virat Kohli he has 5 hundreds in ipl..

BATSMEN'S STRIKERATE WHO SCORED MORETHAN 3000RUNS AND PLAYED MORETHAN 120INNINGS?



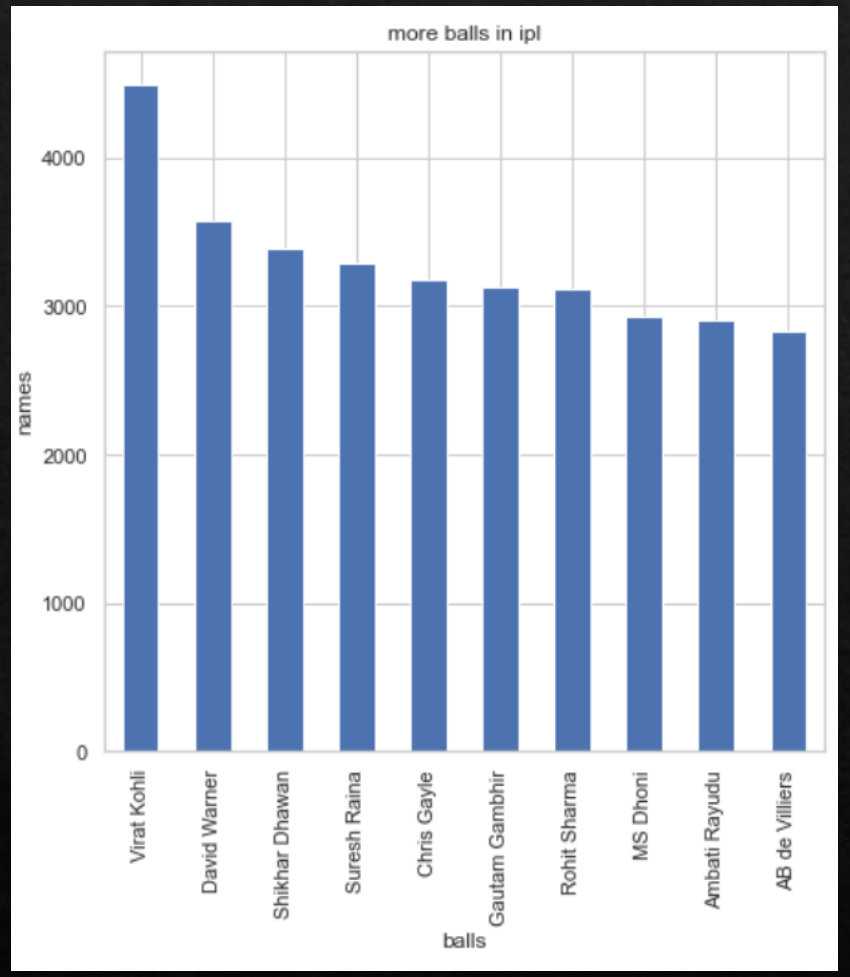
We can see that strikerates of different players who scored morethan 3000 runs and played morethan 120 innings in this AB Devilliers has more strikerate than all players in ipl and in another graph we can see that the median is between110 to 120 that is the average strikerate of all players in ipl who scored morethan 3000 runs and played morethan 120 innings

AVERAGE OF BATSMEN WHO SCORED MORETHAN 300RUNS IN SINGLE SEASON?



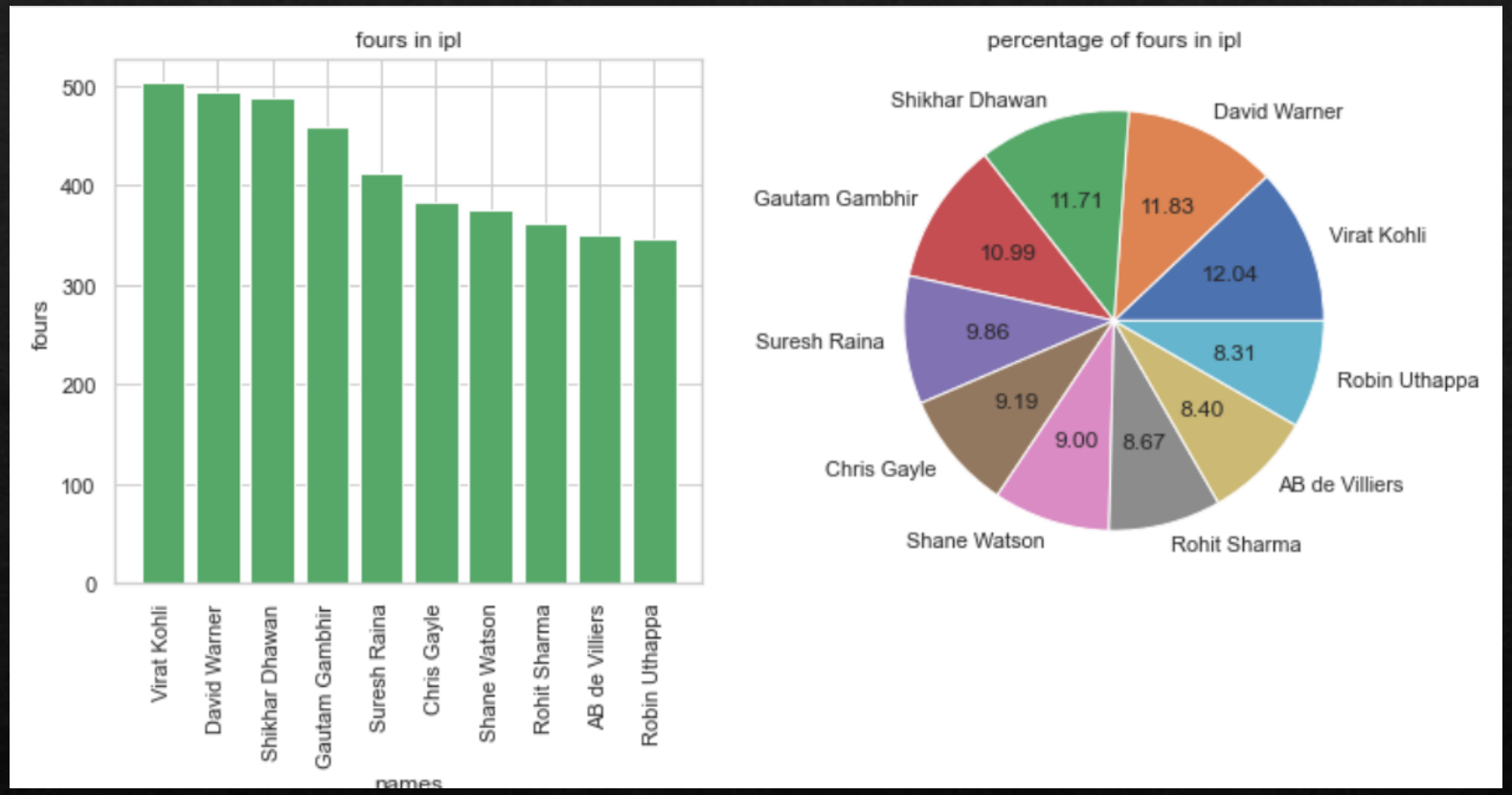
Averages of players who scored morethan 300 runs in single season in this MS Dhoni has the best average among all players

WHO FACED MORE BALLS IN ALL IPL SEASONS?



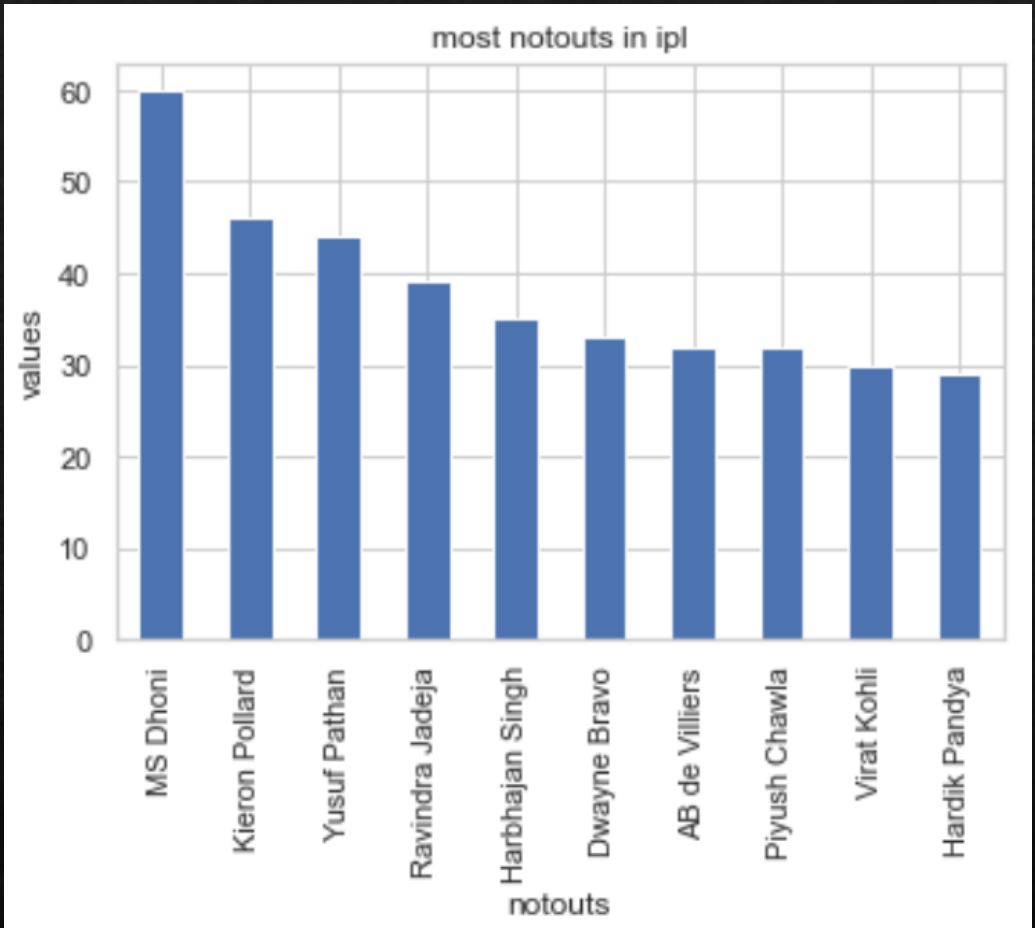
In this we can see that Virat Kohli played moreballs in all ipl seasons he played almost morethan 1000 balls than second player in the list that is David Warner

WHO HIT MORE FOURS IN IPL AND THE PERCENTAGE OF FOURS?



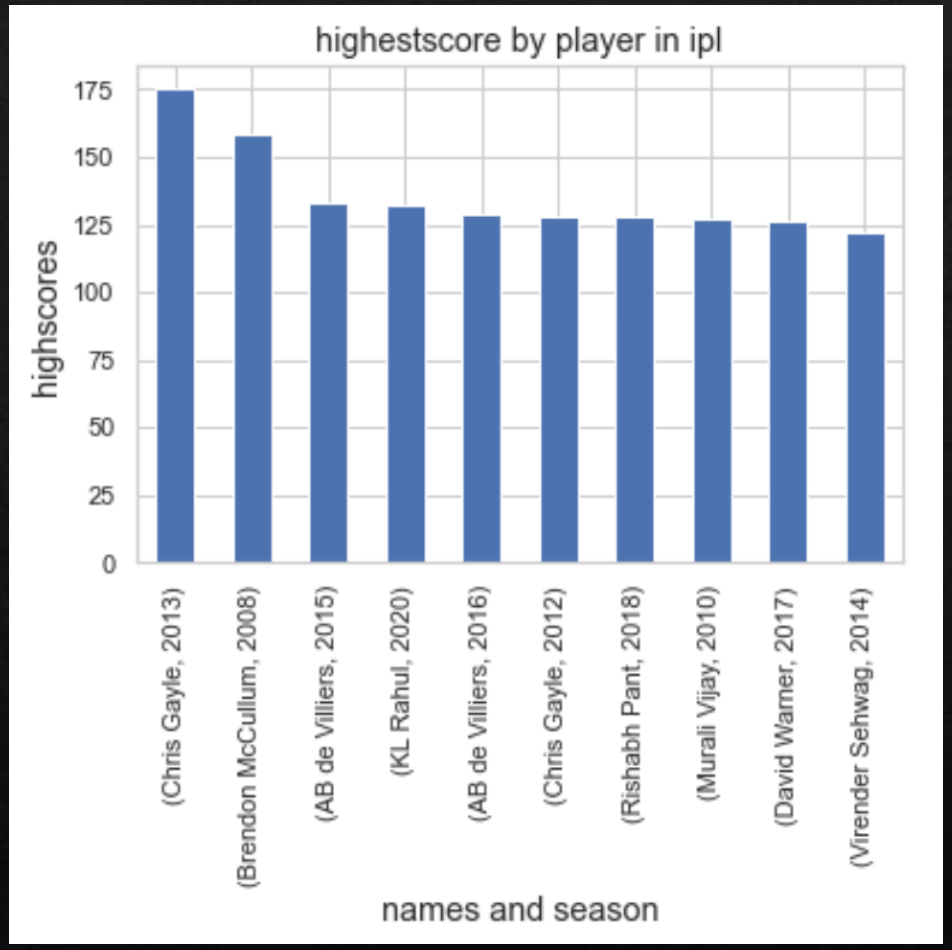
We can see that asusually Virat Kohli scored most number of fours in ipl all seasons the percentage of fours is 12.04 and next is David Warner with 11.83 percentage

WHICH PLAYER HAS THE MOST NUMBER OF NOTOUTS IN ALL IPL SEASONS?



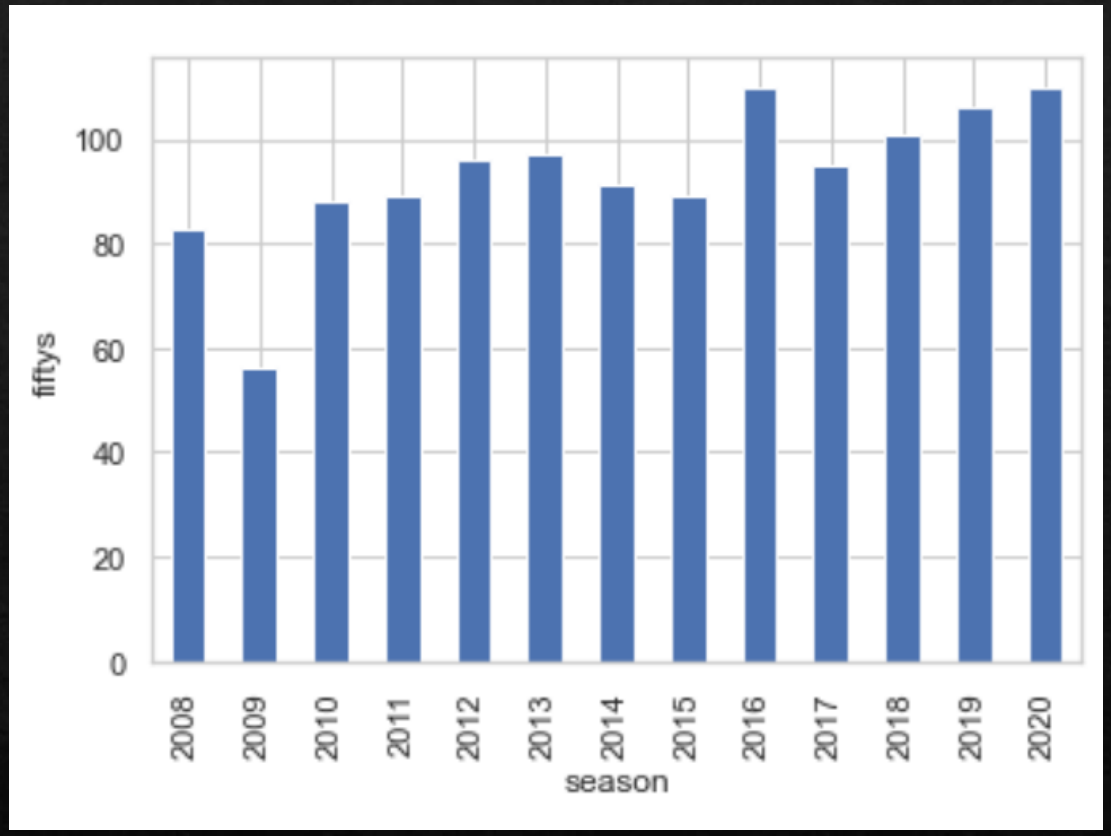
In this we can see that MS Dhoni has most number of notouts in ipl this is the reason why he has best average among all players in single season

HIGHESTSCORE OF PLAYER IN ALL IPL SEASON AND IN WHICH SEASON HE GOT?



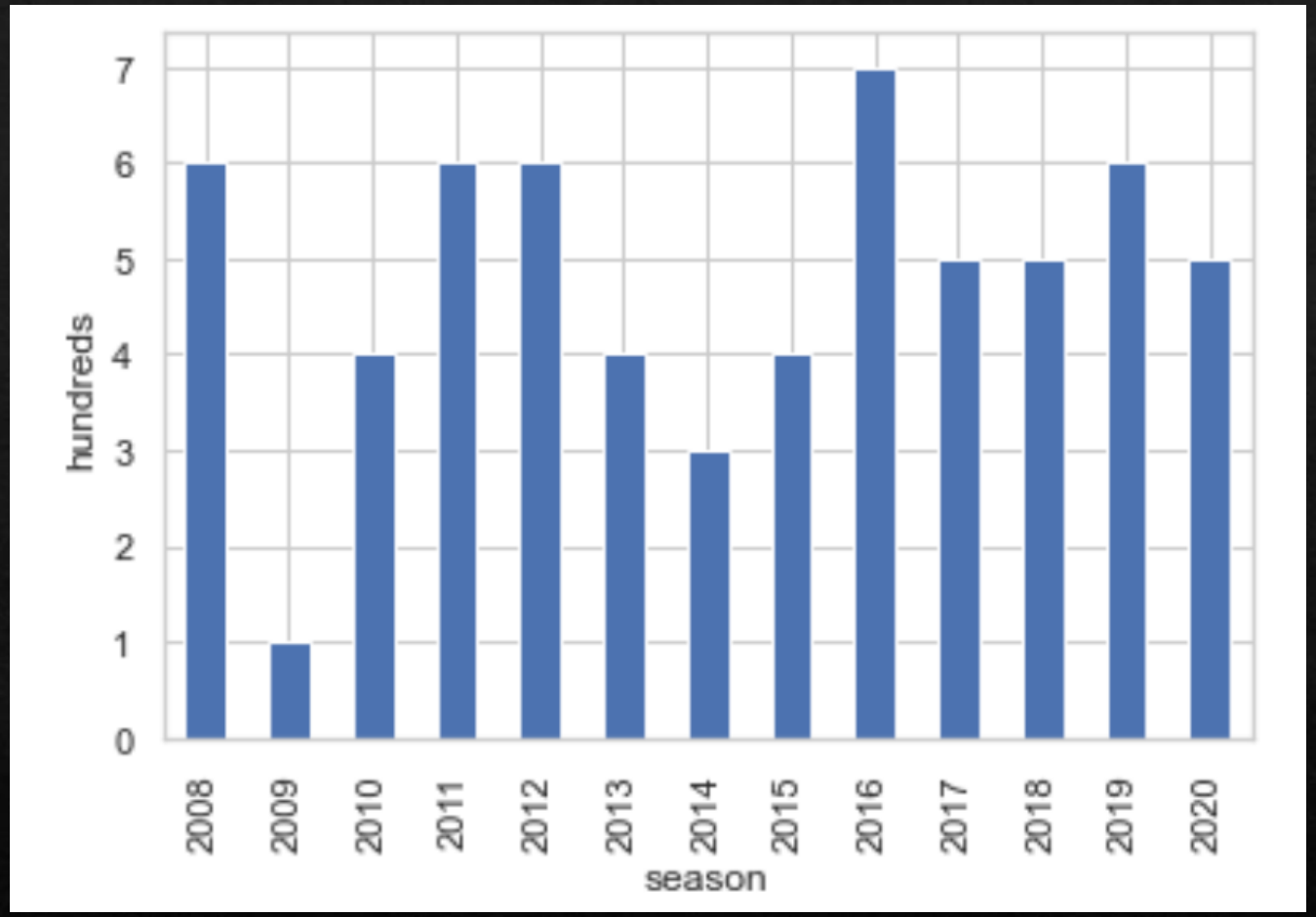
In this we can see that Chris Gayle scored highest score in 2013 which is 175 and second highest score is 158 scored by Brendon Mccullum in first season 2008

IN WHICH SEASON BATSMEN SCORED MORE NUMBER OF FIFTYS?



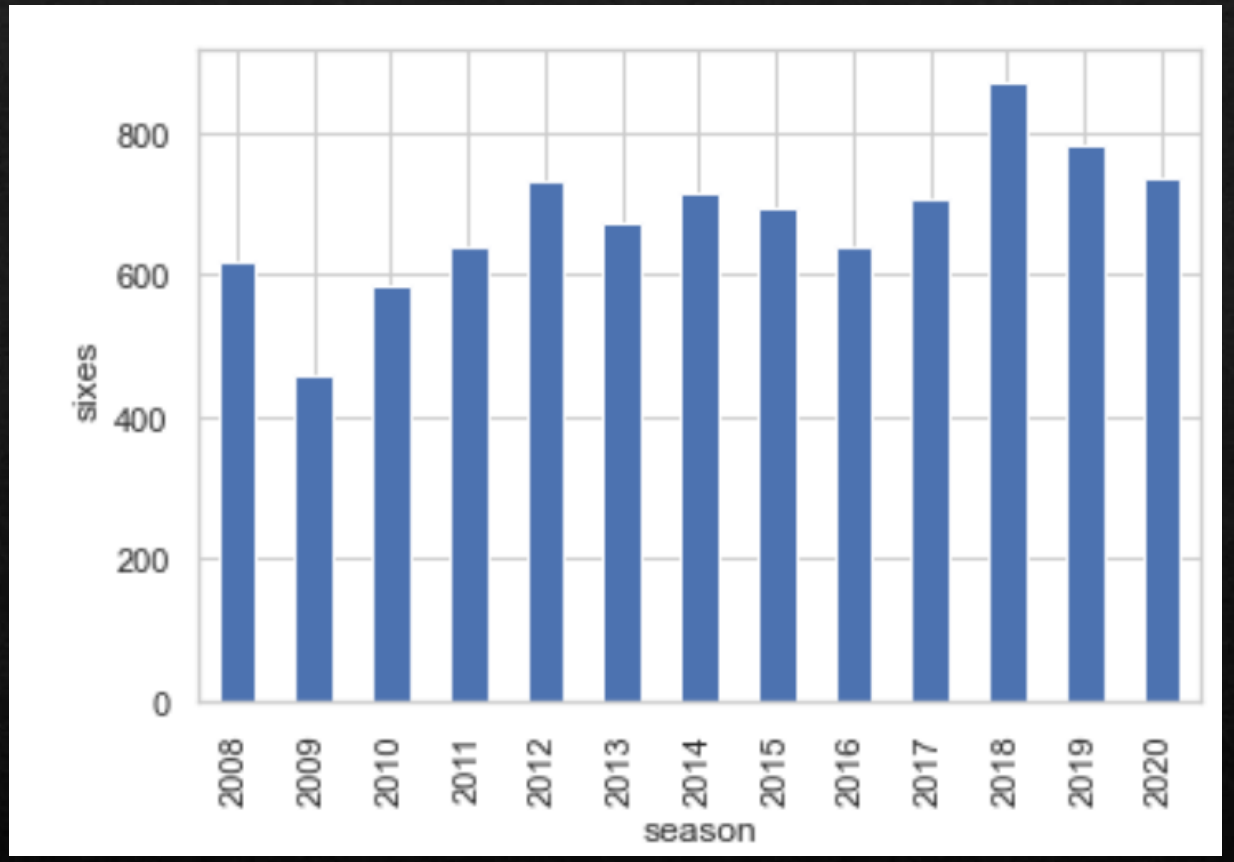
In the 2016 season batsmen got more number of fiftys and in 2020 also almost got same number of fiftys compared to 2016

IN WHICH SEASON BATSMEN SCORED MORE NUMBER OF HUNDREDS?



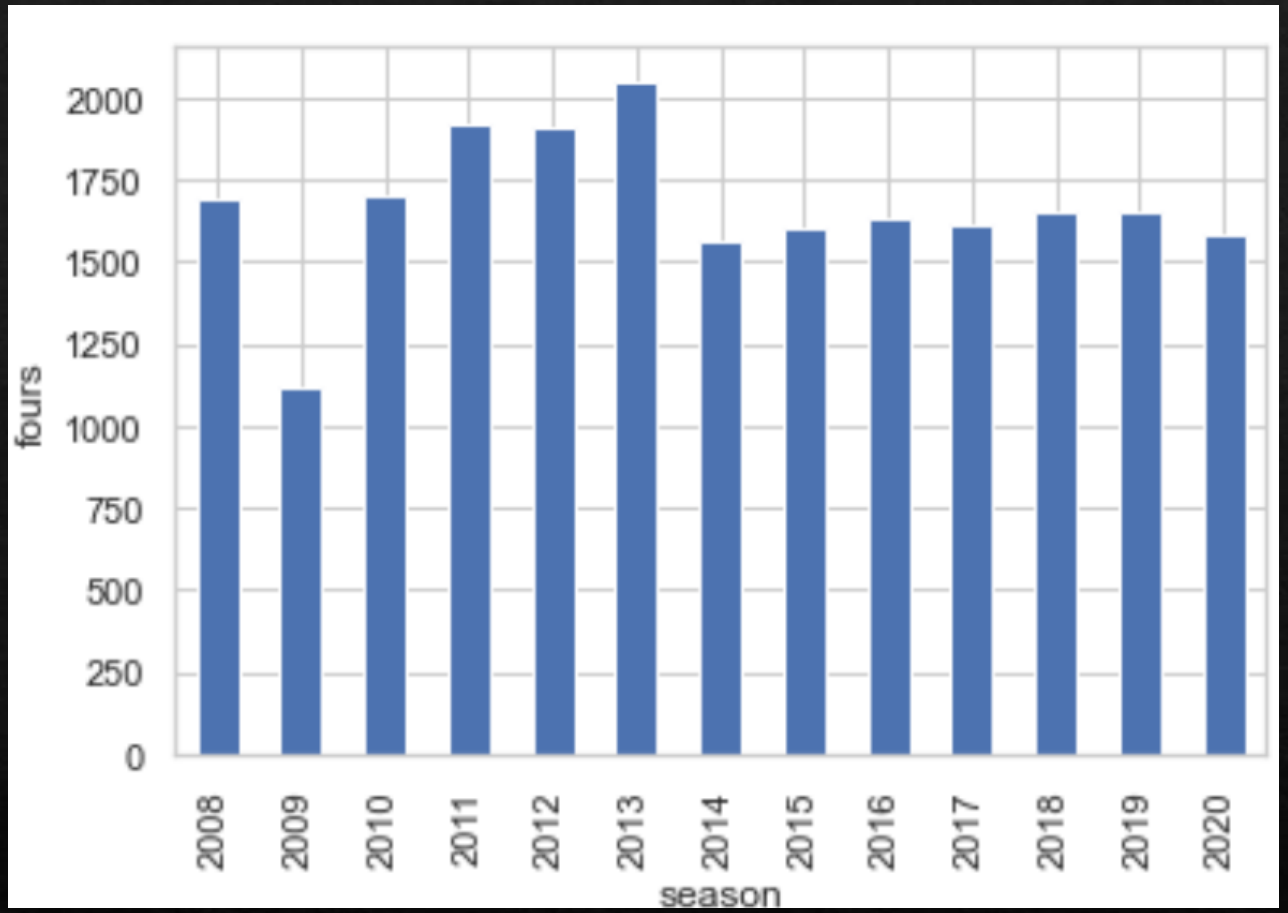
In 2016 season only batsmen scored more number of hundreds which is 7hundreds

IN WHICH SEASON BATSMEN SCORED MORE NUMBER OF SIXES?



In 2018 season batsmen get more number of sixes which is morethan 800 sixes

IN WHICH SEASON BATSMEN SCORED MORE NUMBER OF FOURS?



In 2013 season more number of fours are scored which is morethan 2000fours

CONCLUSION:

If any franchise wants to buy a batsmen I can suggest from above statistics Virat Kohli as best batsmen

In this Some of the records are notbroken which is like

dhoni 60notouts

Gayle 175 high score

Kohli highest runs in single season 973runs


THANK YOU