INNOMATICS
RESEARCH LABS

WEB SCRAPING ON IPL 2015 - 2021

By :
S.Sai Surya Sanjeev

TABLE OF CONTENTS

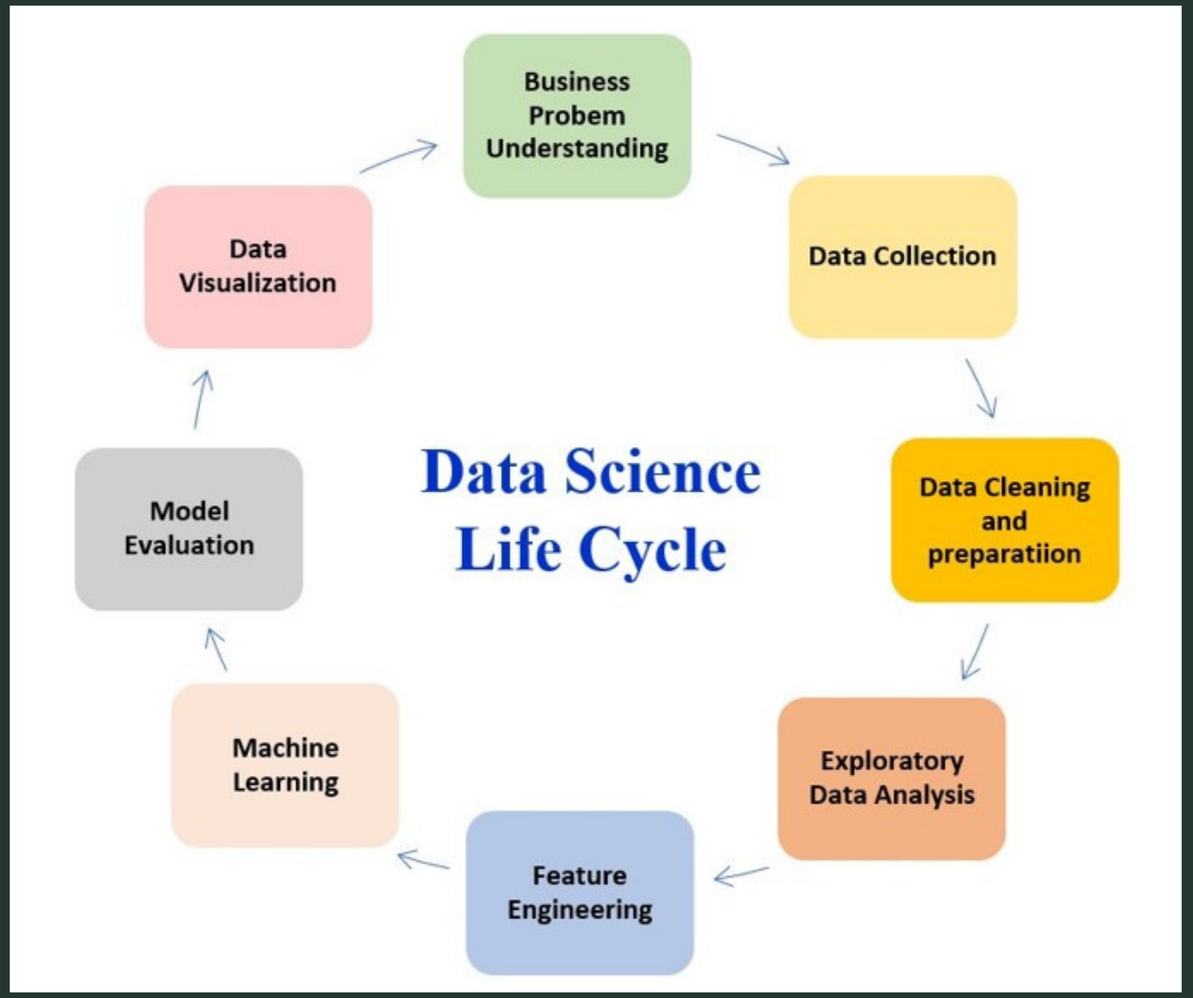
WHAT IS WEB SCRAPING & WHY?

- It is an automated method used to extract large amounts of data from websites. The data on the websites are unstructured from.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.

INNOMATICS
RESEARCH LABS

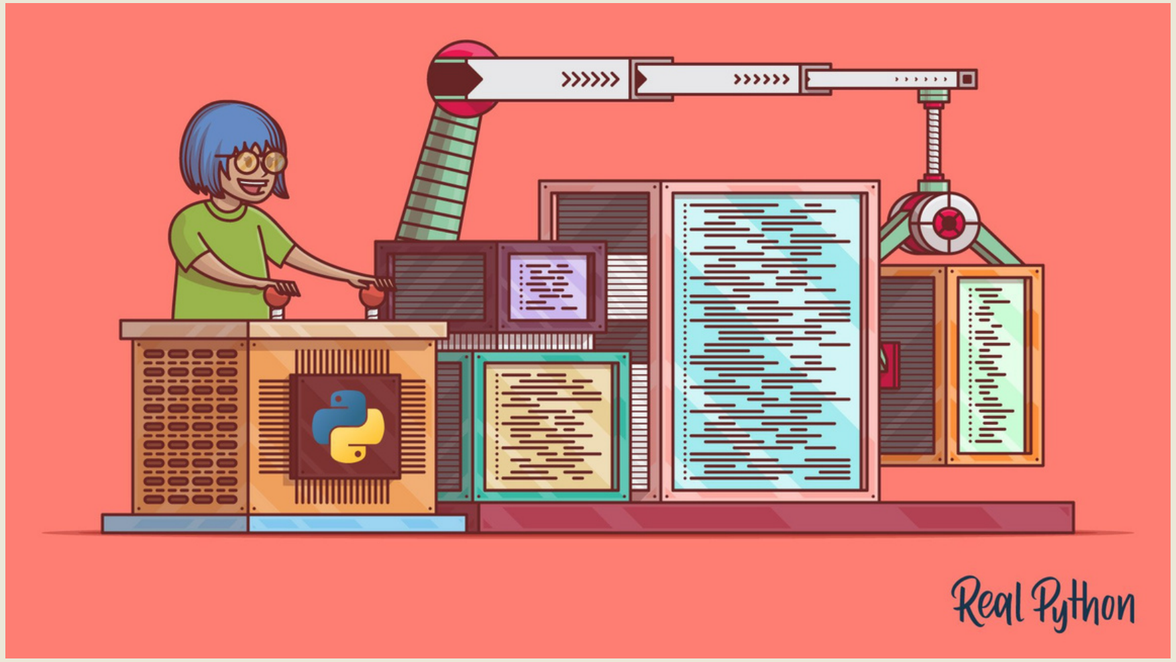
DATA SCIENCE LIFECYCLE



INNOMATICS RESEARCH LABS

LIBRARIES USED:

- Requests
- Beautiful Soup
- Regex (re)
- Pandas
- NumPy
- Matplotlib
- Seaborn



INNOMATICS
RESEARCH LABS

BUSINESS STATEMENTS:

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

Website : iplt20.com

INNOMATICS
RESEARCH LABS

Article

BCCI ANNOUNCES THE SUCCESSFUL BIDDERS FOR TWO NEW INDIAN PREMIER LEAGUE FRANCHISES

25 Oct 2021

Standings | Playoffs

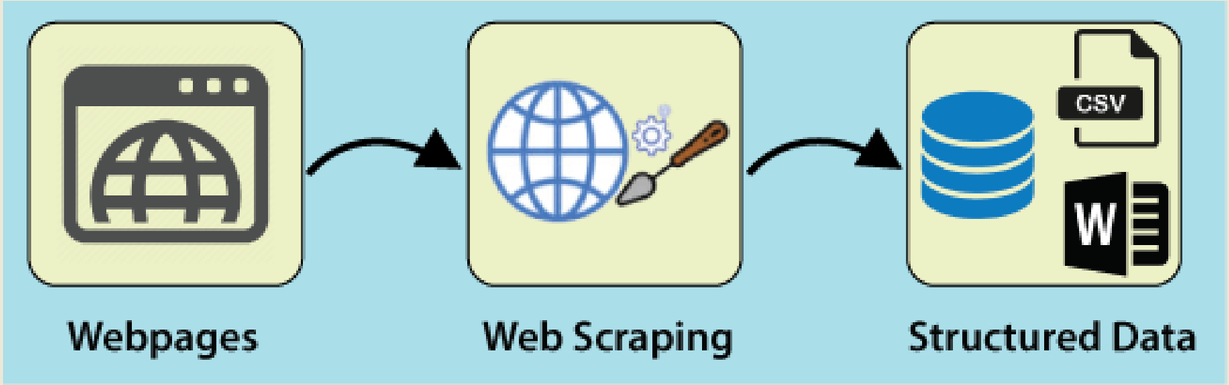
TEAM	PLD	NET RR	PTS	FORM
DC	14	+0.481	20	L W W
CSK	14	+0.455	18	L L L
RCB	14	-0.140	18	W L W
KKR	14	+0.587	14	W W L
MI	14	+0.116	14	W W L
PBKS	14	-0.001	12	W L W
RR	14	-0.993	10	L L W
SRH	14	-0.545	6	L W L

View Full Table | Tweet

HOW DID YOU SCRAPE?

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,Which 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. Which 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.

INNOMATICS
RESEARCH LABS

DATA FRAME BEFORE CLEANING

	index	Unnamed: 0	names	matches	innings	notouts	runs	balls	high_score	avg	strike_rate	fours	sixes	fiftys	hundreds	season
0	0	0	\n\n\n\nDavid Warner\n\n\n\n	14	14	1	562	359	91	43.23	156.54	65	21	0	7	2015
1	1	1	\n\n\n\nLendl Simmons\n\n\n	13	13	1	540	441	71	45.00	122.44	56	21	0	6	2015
2	2	2	\n\n\n\nAjinkya Rahane\n\n\n	14	13	2	540	413	91	49.09	130.75	53	13	0	4	2015
3	3	3	\n\n\n\nAB de Villiers\n\n\n	16	14	3	513	293	133	46.63	175.08	60	22	1	2	2015
4	4	4	\n\n\n\nVirat Kohli\n\n\n	16	16	5	505	386	82	45.90	130.82	35	23	0	3	2015
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
967	144	144	\n\n\n\nKuldip Yadav\n\n\n	1	1	1	0	4	0	0.00	0.00	0	0	0	0	2021
968	145	145	\n\n\n\nSarfaraz Khan\n\n\n	2	2	0	0	4	0	0.00	0.00	0	0	0	0	2021
969	146	146	\n\n\n\nAnuj Rawat\n\n\n	2	1	0	0	1	0	0.00	0.00	0	0	0	0	2021
970	147	147	\n\n\n\nJimmy Neesham\n\n\n	3	2	0	0	2	0	0.00	0.00	0	0	0	0	2021
971	148	148	\n\n\n\nPiyush Chawla\n\n\n	1	1	0	0	2	0	0.00	0.00	0	0	0	0	2021

972 rows × 16 columns

In this Data Frame 972 rows x 16 columns are present.

BEFORE CLEANING DATA FRAME

```
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	David Warner	14	14	1	562	359	91	43.23	156.54	65	21	0	7	2015
1	Lendl Simmons	13	13	1	540	441	71	45.00	122.44	56	21	0	6	2015
2	Ajinkya Rahane	14	13	2	540	413	91	49.09	130.75	53	13	0	4	2015
3	AB de Villiers	16	14	3	513	293	133	46.63	175.08	60	22	1	2	2015
4	Virat Kohli	16	16	5	505	386	82	45.90	130.82	35	23	0	3	2015
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
967	Kuldip Yadav	1	1	1	0	4	0	0.00	0.00	0	0	0	0	2021
968	Sarfaraz Khan	2	2	0	0	4	0	0.00	0.00	0	0	0	0	2021
969	Anuj Rawat	2	1	0	0	1	0	0.00	0.00	0	0	0	0	2021
970	Jimmy Neesham	3	2	0	0	2	0	0.00	0.00	0	0	0	0	2021
971	Piyush Chawla	1	1	0	0	2	0	0.00	0.00	0	0	0	0	2021

972 rows × 14 columns

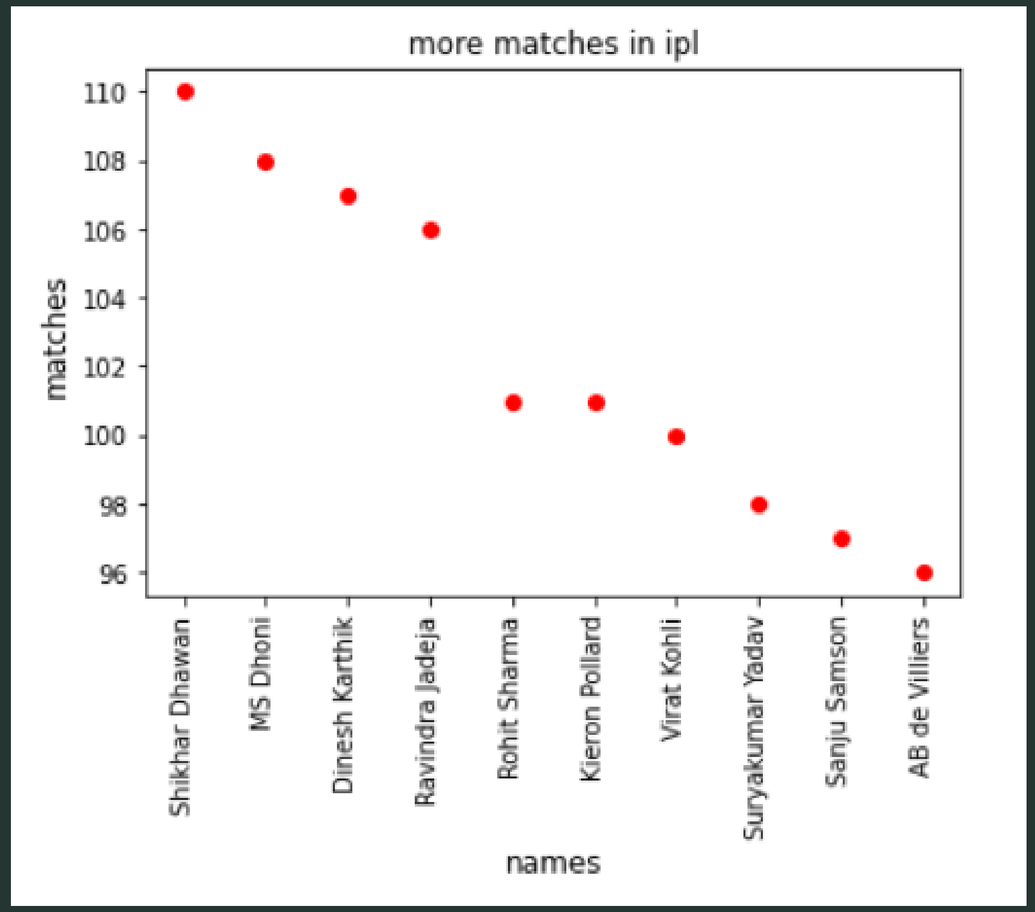
In this Data Frame 972 rows x 14 columns are present.

AFTER CLEANING DATA FRAME

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 972 entries, 0 to 971
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   names        972 non-null    object 
 1   matches      972 non-null    int64  
 2   innings      972 non-null    int64  
 3   notouts      972 non-null    int64  
 4   runs         972 non-null    int64  
 5   balls        972 non-null    int64  
 6   high_score   972 non-null    int64  
 7   avg          972 non-null    float64
 8   strike_rate  972 non-null    float64
 9   fours        972 non-null    int64  
 10  sixes        972 non-null    int64  
 11  fiftys       972 non-null    int64  
 12  hundreds     972 non-null    int64  
 13  season       972 non-null    int64  
dtypes: float64(2), int64(11), object(1)
memory usage: 106.4+ KB
```

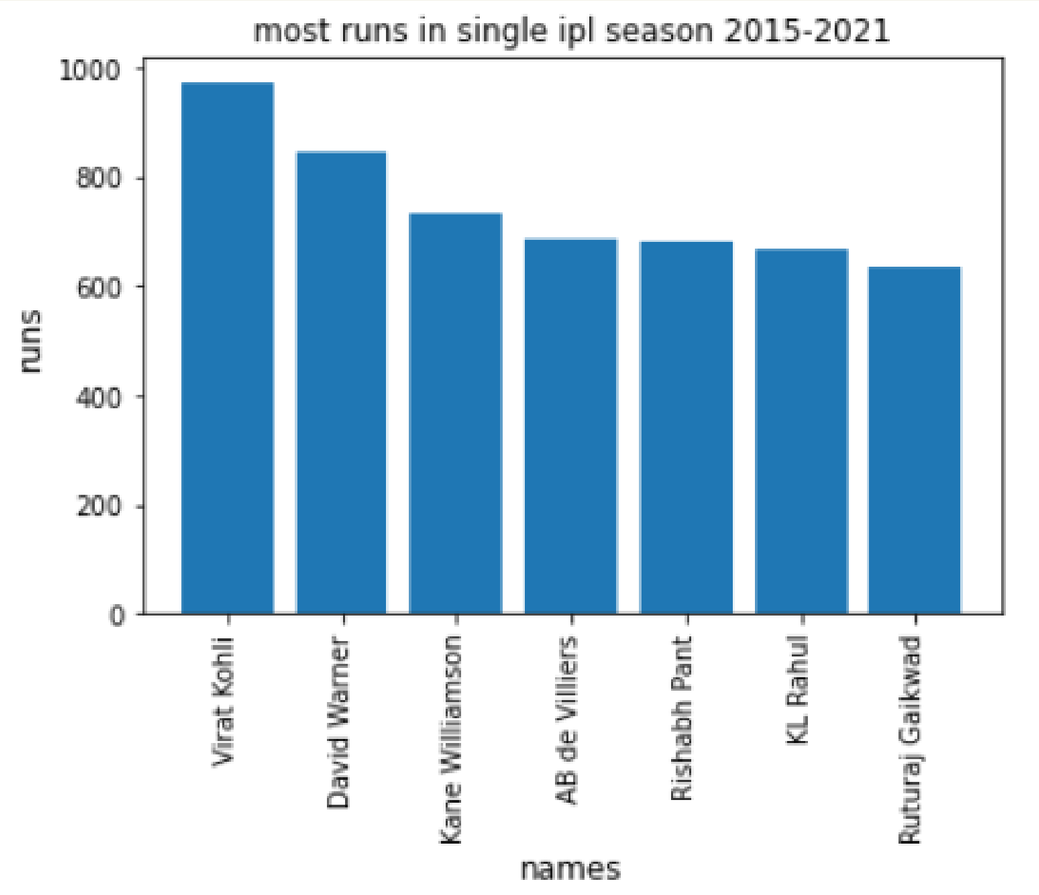
WHO PLAYED MORE MATCHES IN IPL FROM 2015 - 2021

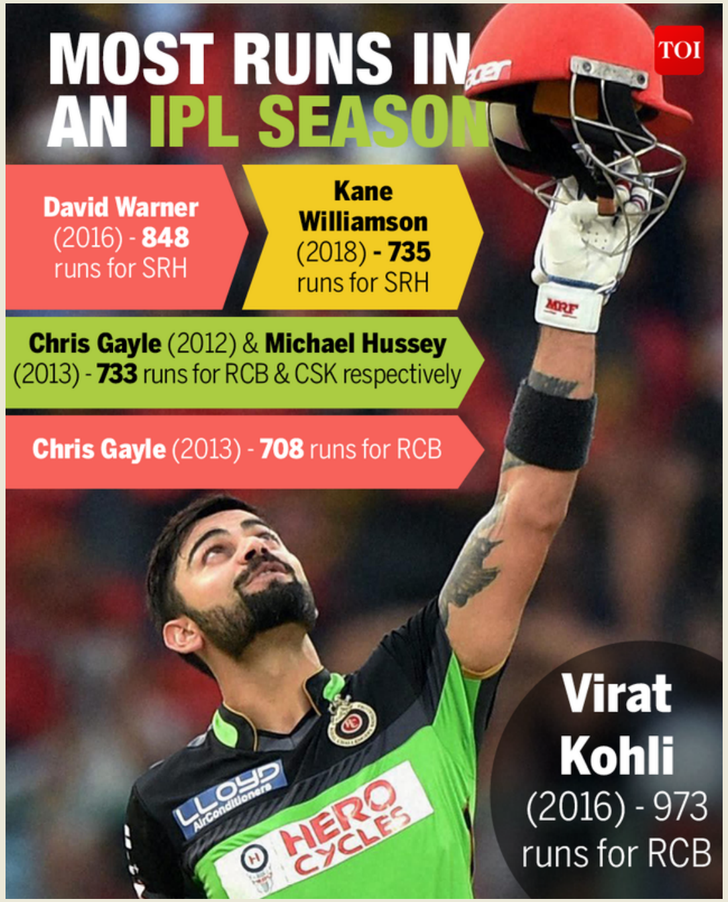


From above Scatter Plot SHIKHAR DHAWAN played more matches throughout IPL

INNOMATICS
RESEARCH LABS

WHO SCORED MORE RUNS IN IPL SINGLE SEASON FROM 2015-2021

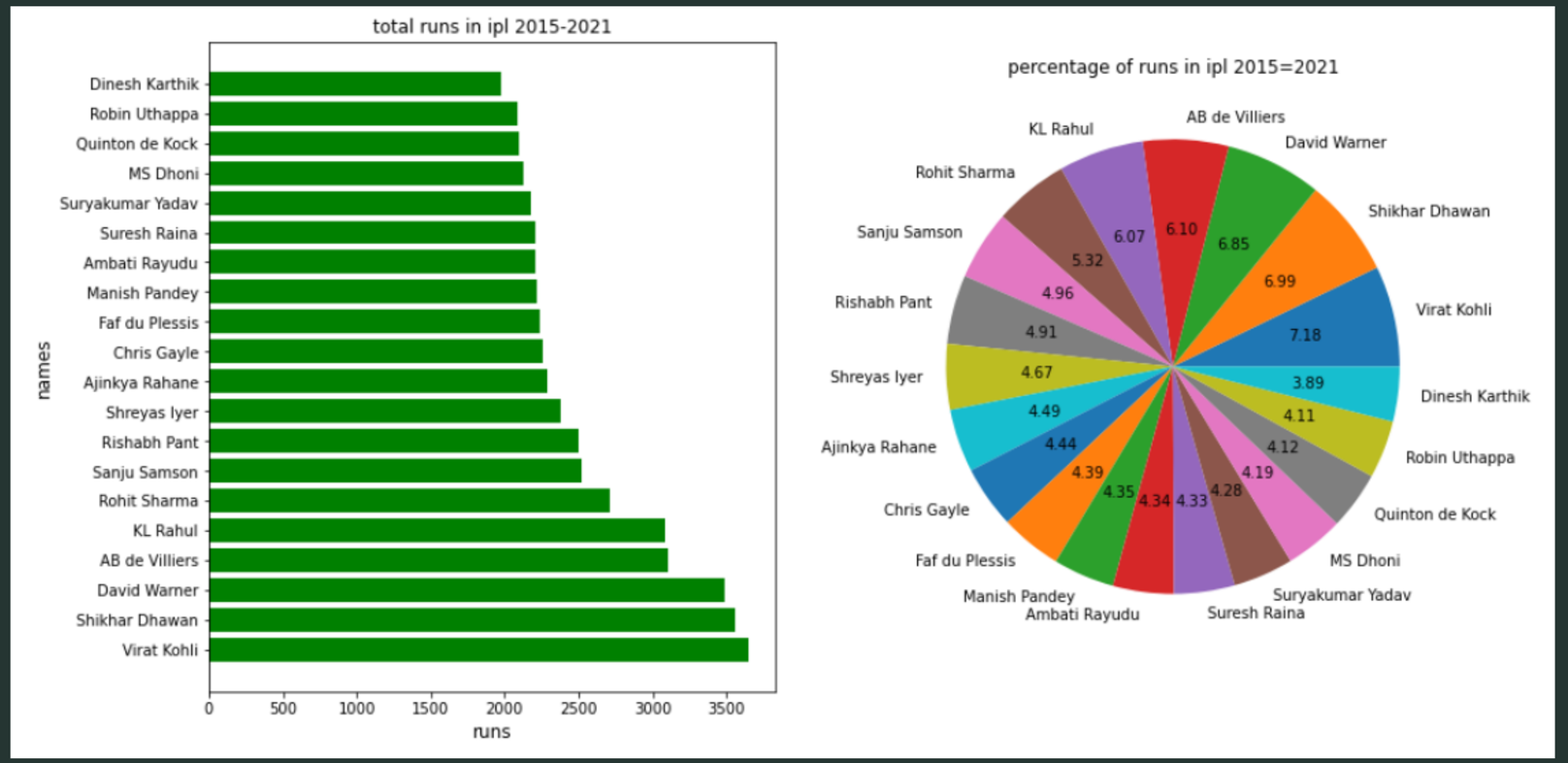




In single IPL season VIRAT KOHLI scored more runs 973 -(2016)

INNOMATICS RESEARCH LABS

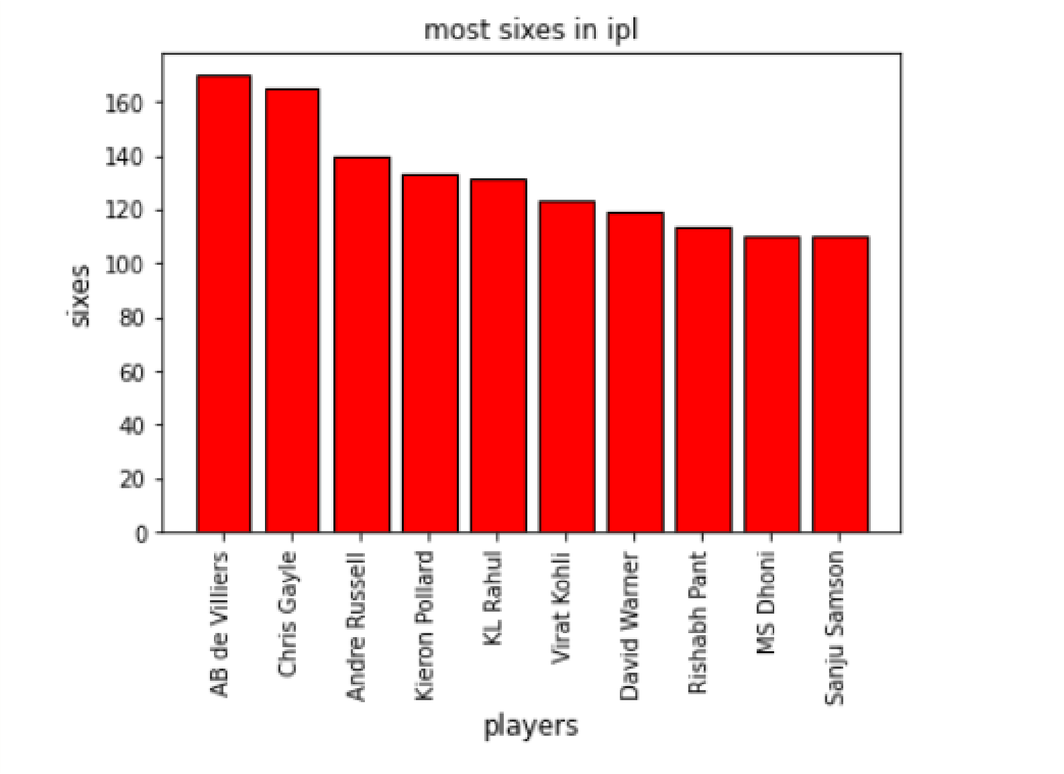
WHO SCORED MORE RUNS IN IPL FROM 2015-2021 ?



From 2015-2021 IPL seasons VIRAT KOHLLI scored more runs & The Percentage of runs is 7.87

INNOMATICS RESEARCH LABS

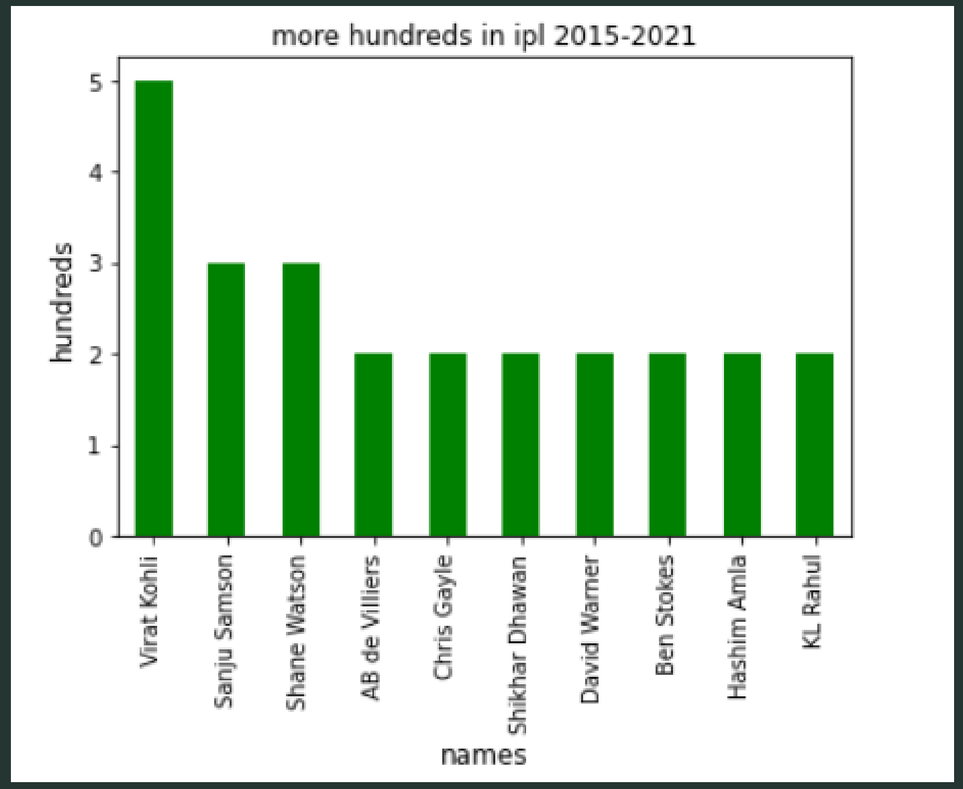
WHO HIT MORE SIXES IN IPL 2015-2021



AB de Villiers hits more sixes in ipl 2015-2021 seasons

INNOMATICS
RESEARCH LABS

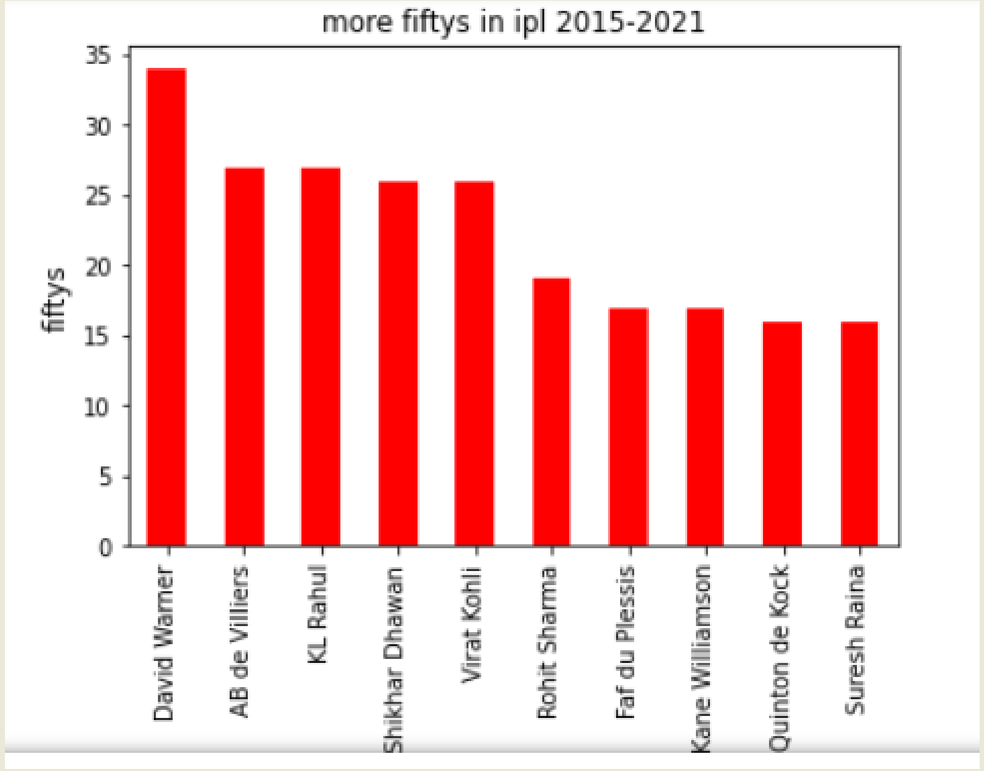
WHO HIT MORE HUNDRED'S IN IPL 2015-2021



Virat Kohil has more number of Hundred's in ipl 2015-2021 and second most is Sanju Samson

INNOMATICS RESEARCH LABS

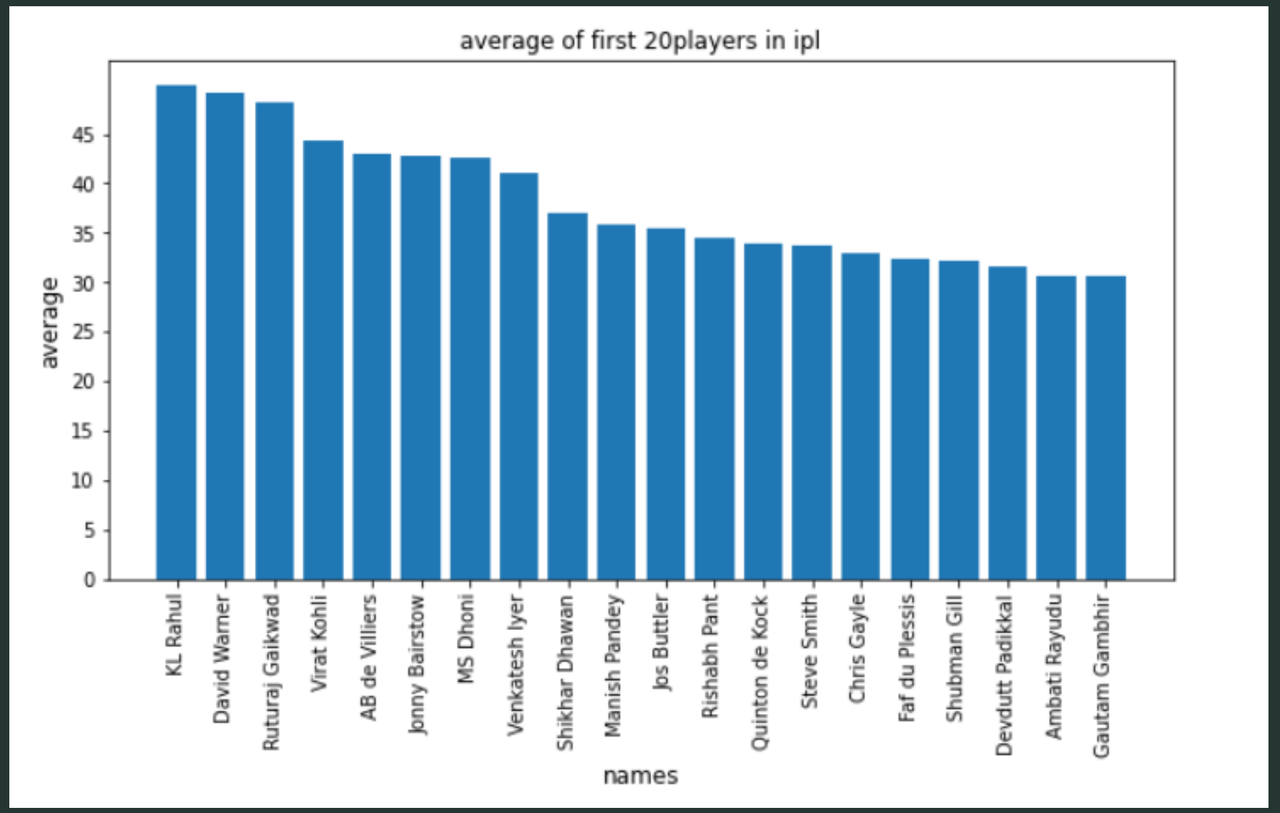
who hit more fifty's in ipl 2015-2021



David Warner has more number of fiftys in ipl 2015-2021 and second most is AB de Villiers

INNOMATICS RESEARCH LABS

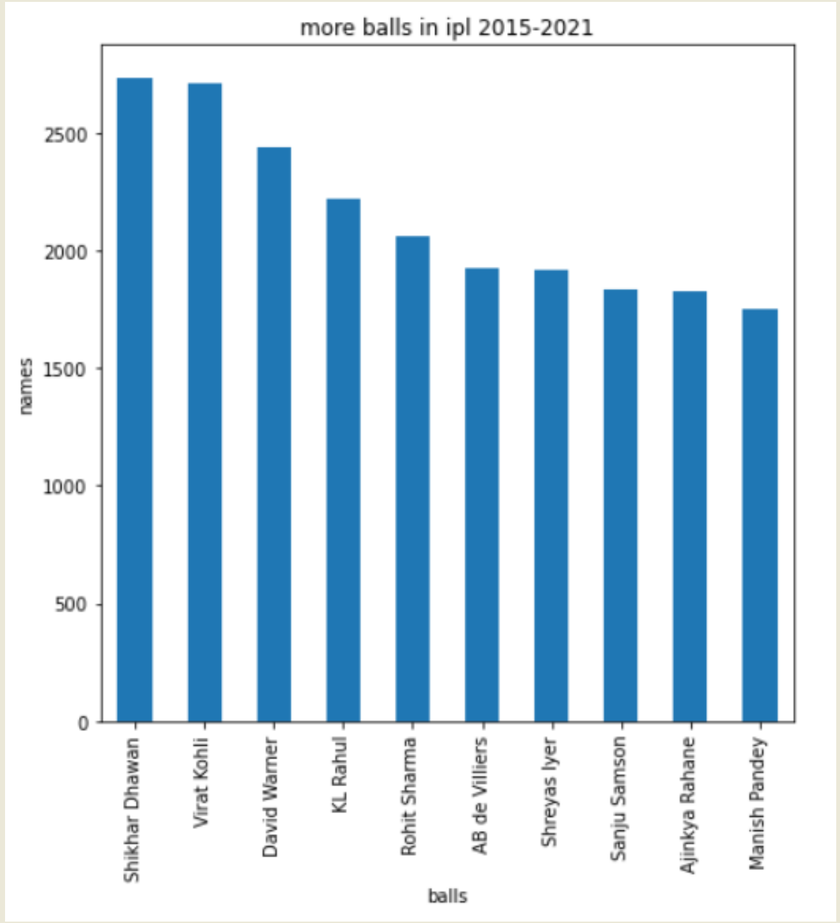
AVERAGE OF BATSMEN WHO SCORED MORETHAN 300RUNS IN SINGLE SEASON ?



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

INNOMATICS RESEARCH LABS

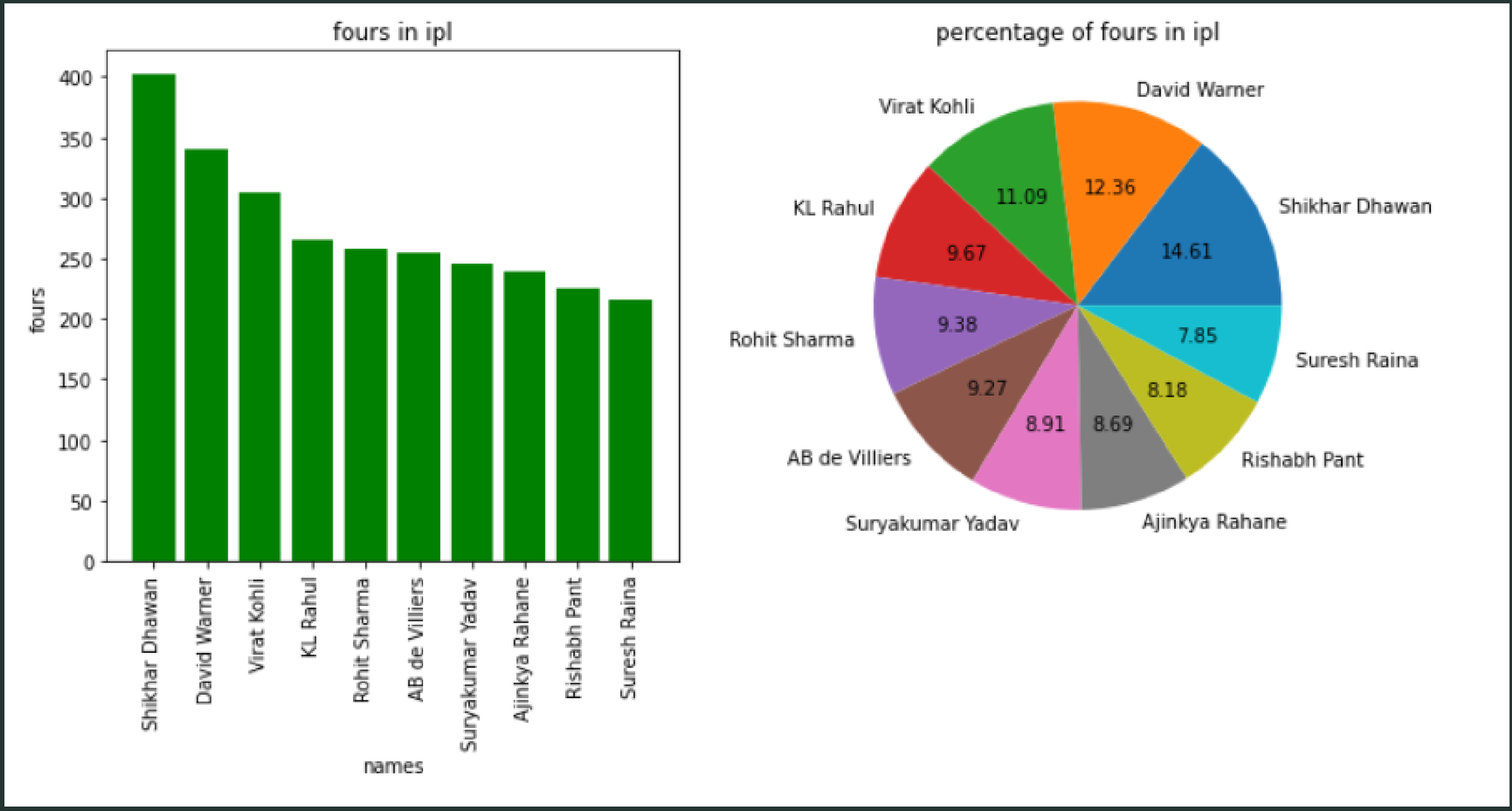
WHO FACED MORE BALLS IN 2015-2021 IPL SEASONS ?



In this we can see that Shikhar Dhawan played moreballs in 2015 - 2021 ipl seasons he played almost more balls than second player in the list that is Virat Kohli

INNOMATICS
RESEARCH LABS

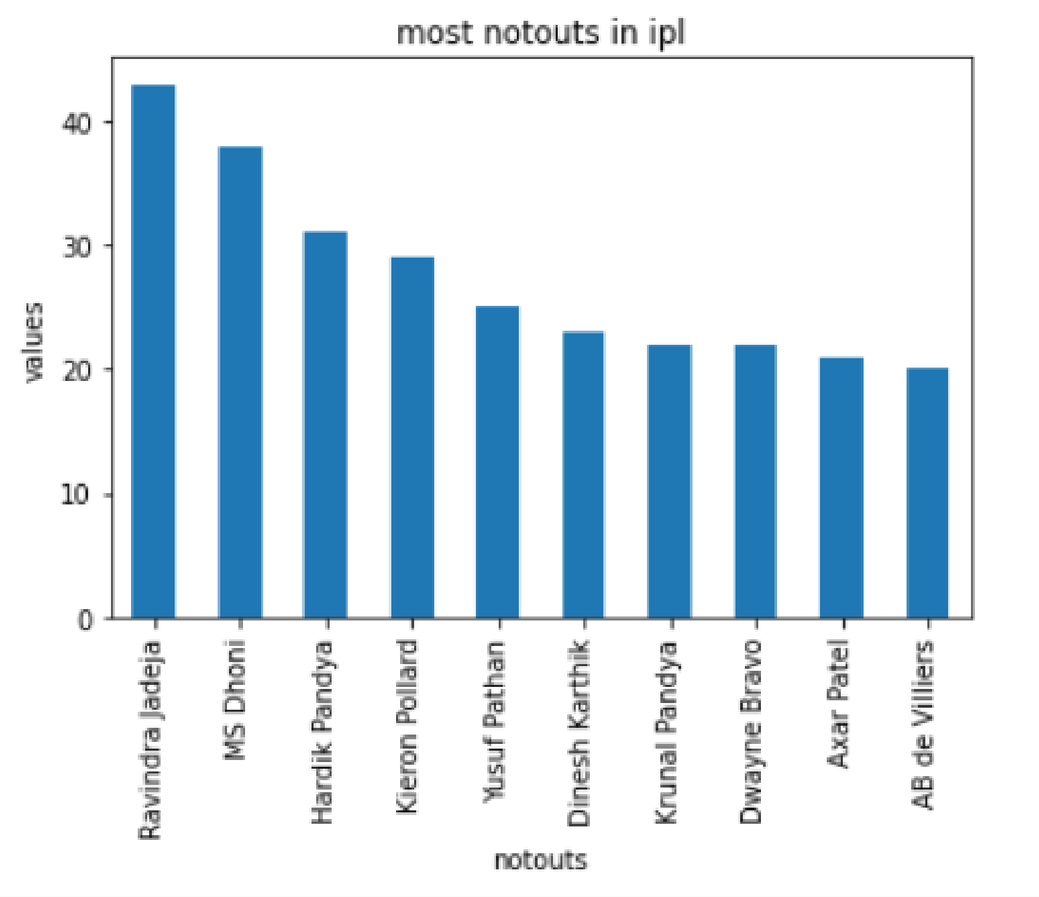
WHO HIT MORE FOURS IN IPL AND THE PERCENTAGE OF FOURS ?



We can see that asusually Shikhar Dhawan scored most number of fours in ipl 2015-2021 seasons the precentage of fours is 14.61

INNOMATICS RESEARCH LABS

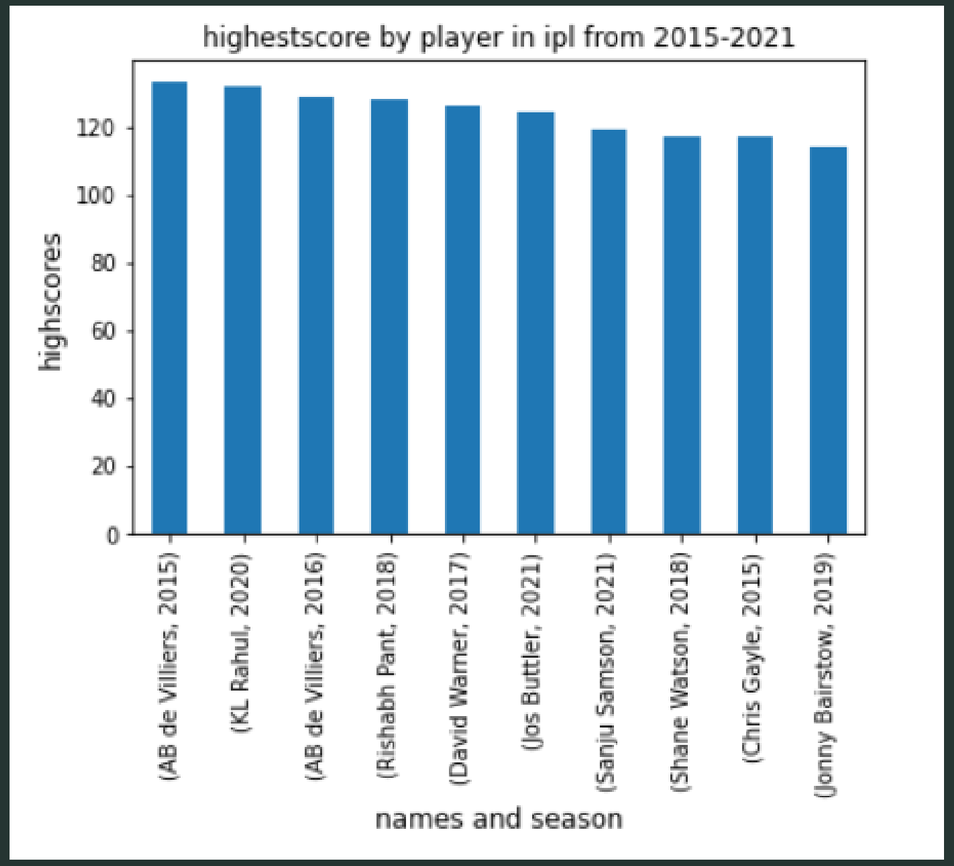
WHICH PLAYER HAS THE MOST NUMBER OF NOTOUTS IN 2015-2021 IPL



In this we can see that Ravindra Jadeja has most number of notouts in 2015-2021 ipl

INNOMATICS
RESEARCH LABS

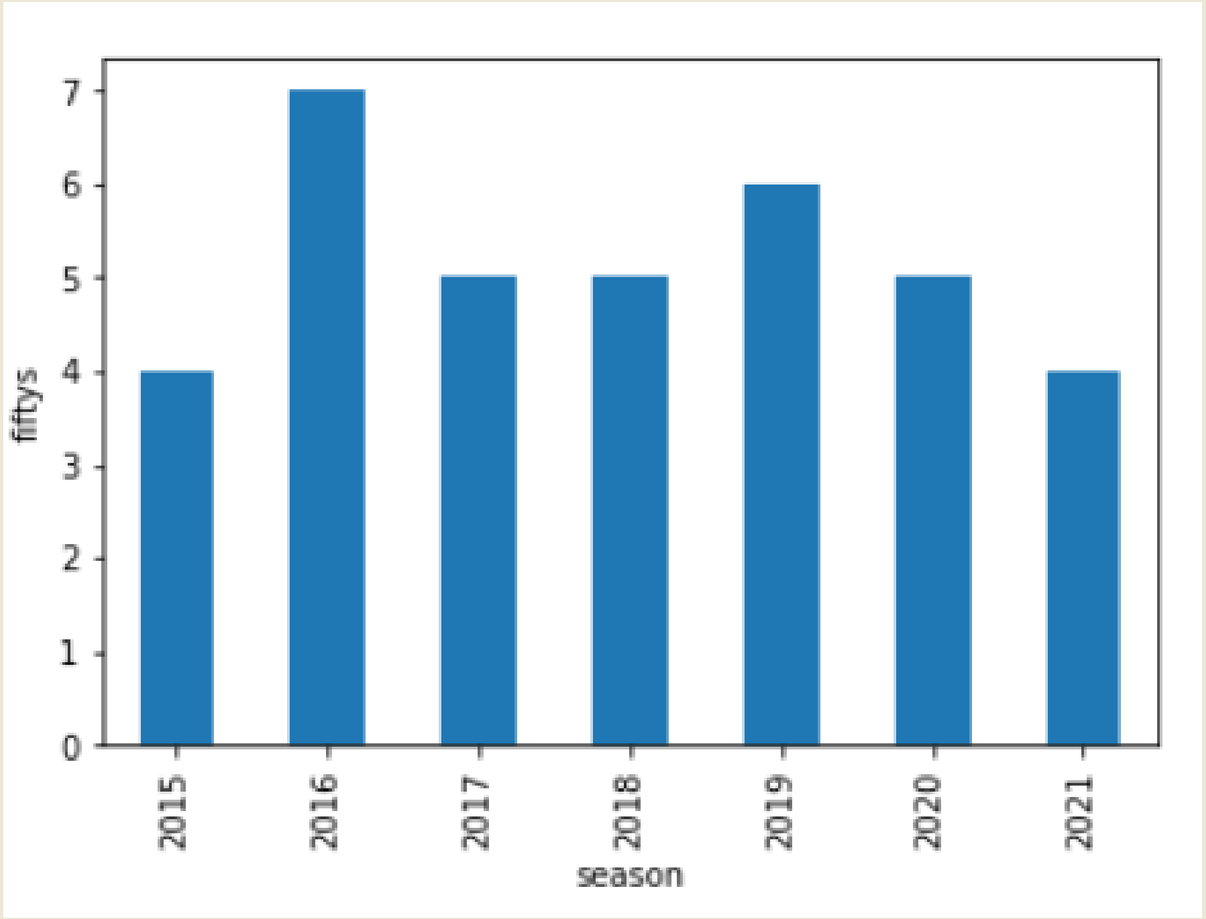
HIGHESTSCORE OF PLAYER IN 2015-2021 IPL SEASON AND IN WHICH SEASON HE GOT



In this we can see that AB de Villiers scored highest score in 2015 which is 133 and second highest score is 132 scored by KL Rahul in 2020 ipl season.

INNOMATICS
RESEARCH LABS

IN WHICH SEASON BATSMEN SCORED MORE NUMBER OF FIFTYS



In the 2016 season batsmen got more number of fiftys

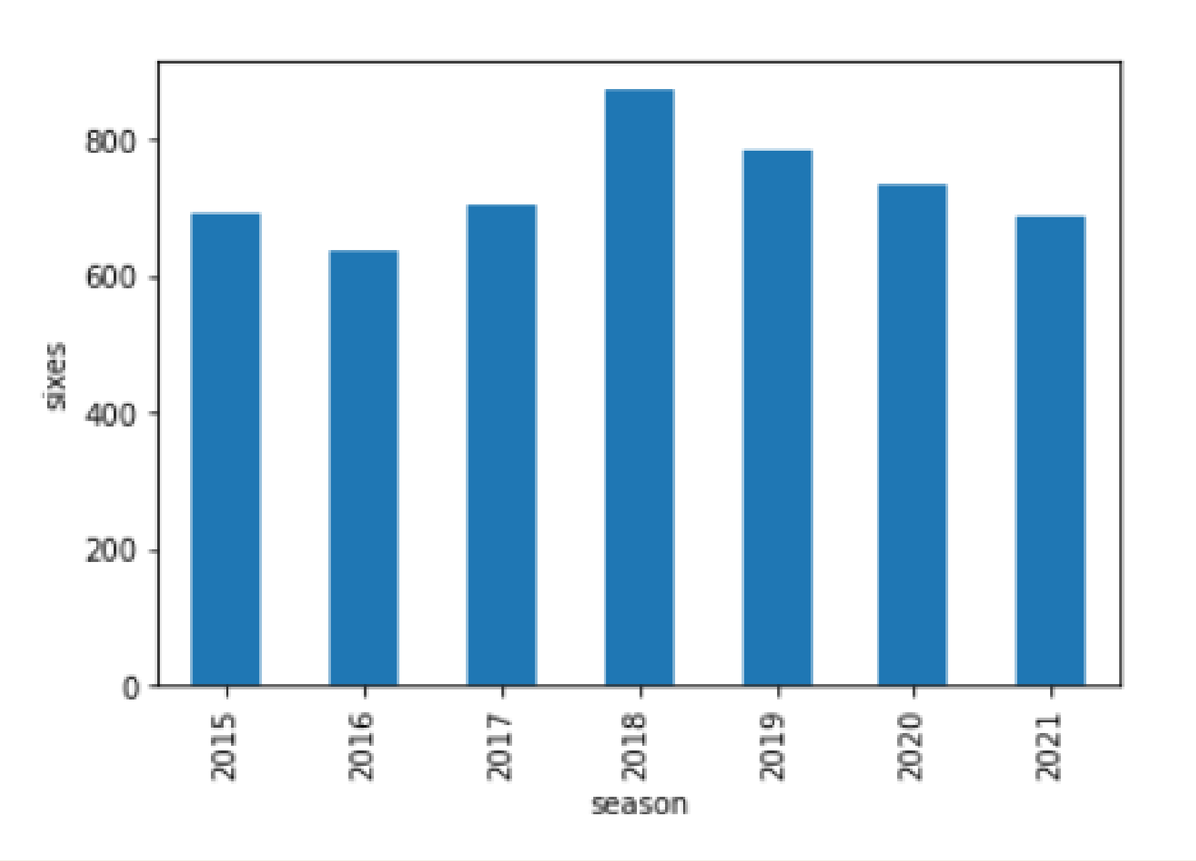
INNOMATICS
RESEARCH LABS

IN WHICH SEASON BATSMEN SCORED MORE NUMBER OF HUNDERDS

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

INNOMATICS
RESEARCH LABS

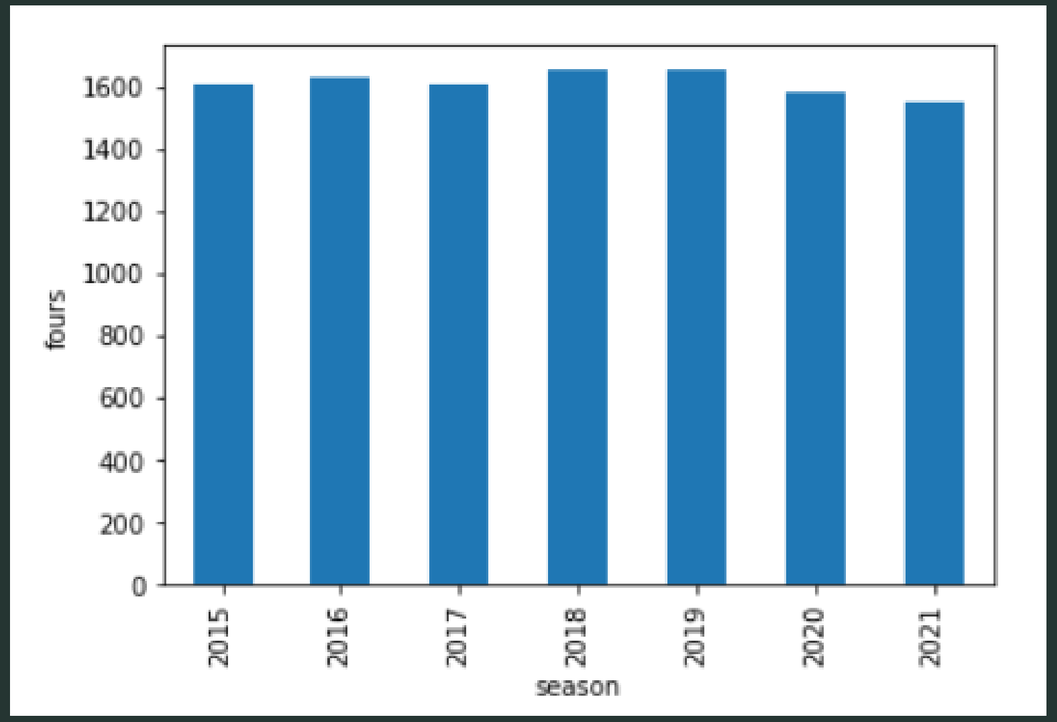
IN WHICH SEASON BATSMEN SCORED MORE NUMBER OF SIXES



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

INNOMATICS
RESEARCH LABS

IN WHICH SEASON BATSMENS SCORED MORE NUMBER OF FOURS



In 2018 & 2019 season more number of fours are scored which is morethan 1500 fours.

INNOMATICS RESEARCH LABS

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
- Ravindra Jadeja 45 notouts
- AB de Villiers 133 high score
- Kohli highest runs in single season 973 runs

INNOMATICS RESEARCH LABS

THANK YOU

INNOMATICS
RESEARCH LABS