PROJECT : DESIGN A WEB SCRAPE DATA FROM A WEBSITE. ANALYZE THE DATA AND MAKE A REPORT ON THE ANALYSIS.

(https://github.com/HritikaGupta22/WEB_SCRAPPINGPROJECT_045020/blob/main/WEB_SCRAPPING(045020).ipynb)

REPORT :

OBJECTIVES :

1. To collect men's T20 cricket statistics from a ESPNCRICINFO(Men’s t20 data) source.
2. Perform some statistical operation on whole data
3. To identify the top 8 countries based on the total runs scored by players from each country.
4. To compare players from the top 8 countries based on specific performance metrics.
5. To analyze Indian players' statistics and identify key insights.
6. To calculate correlations and p-values to understand relationships within the data of Indian players.

GENERAL DESCRIPTIONS OF DATA

Data Sources :

I scrap the data from Website : “EspnCricinfo” url : https://www.espncricinfo.com/records/most-runs-in-career-282827.

The data consist of T20 batting records for T20I matches which data of many players from different countries.

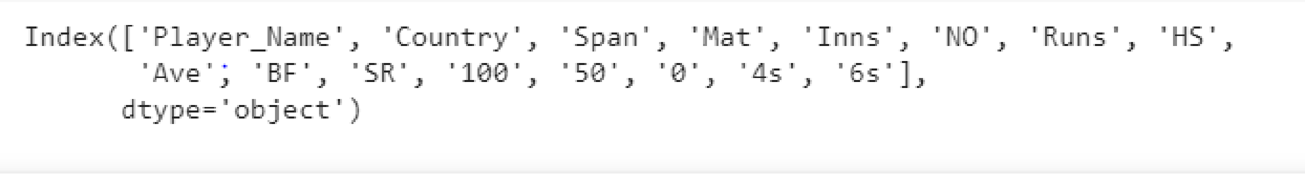
Libraries which are used by this projects are :

```
from bs4 import BeautifulSoup
import requests
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import pearsonr
```

Data Preprocessing steps

1. seperating name and country
2. Replacing - with 0 where - representing 0
3. Removing *
4. Modifying the required columns data type
5. Extract the country code from the ‘Country’ column

Columns in the table



```
Index(['Player_Name', 'Country', 'Span', 'Mat', 'Inns', 'NO', 'Runs', 'HS',
       'Ave', 'BF', 'SR', '100', '50', '0', '4s', '6s'],
      dtype='object')
```

Player_Name – This column consist of Players Name

Country- This column consist of country for which player play

Span- This column consist of time period a player play. For Example : 2011-2023

Mat – No. of matches a player play

Inns – No. of innings a player get opportunity to bat

NO – no. of times a player remains not out.

Runs – Total no. of Runs scored by a player

HS- highest runs scored by a player

Ave- average of a player

BF-total no. of ball faced

SR – the player play with what strike rate

100 – total no. of 100 scored by a player

50 – total no. of 50 scored by a player

0-total number of dismissals on zero

4s- the total number of sixes hit by each player in T20 matches.

6s - the total number of sixes hit by each player in T20 matches.

Analysis:

Basic Statistics of the data:

	Mat	**Inns**	**Runs**	**Ave**	**SR**	**100**	**50**	**0**	**4s**	**6s**
count	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000
mean	67.093333	62.393333	1487.406667	28.439000	129.506000	0.493333	8.426667	3.986667	133.280000	55.946667
Std	24.593704	22.270725	626.703146	6.305679	12.162105	0.817154	5.634680	2.702191	64.934879	29.980931
Min	26.000000	26.000000	861.000000	17.560000	101.360000	0.000000	1.000000	0.000000	56.000000	10.000000
25%	49.000000	46.250000	1038.250000	23.532500	120.555000	0.000000	5.000000	2.000000	90.250000	33.250000
50%	62.500000	57.000000	1271.500000	27.970000	128.955000	0.000000	7.000000	3.000000	114.000000	50.500000
75%	79.750000	73.750000	1683.750000	31.797500	137.302500	1.000000	10.000000	6.000000	157.750000	69.000000
Max	148.000000	140.000000	4008.000000	52.730000	172.700000	4.000000	37.000000	13.000000	394.000000	182.000000

This gives us the mean, stand deviation,min,25%,50%,75%,max of matches,innings,total run scored, 50s, 100s, 4s, 6s, ball faced , highest score of players.

Count represent total no. of rows

Mean represent the average. It is a measure of central tendency. It represents the "typical" or "central" value in a dataset.

To calculate the mean, we add up all the values in a dataset and then divide by the total number of values.

The standard deviation is a measure of the dispersion or spread of data points around the mean.

For Example : The mean of 28.439 indicates that, on average, players approximately 28.439 runs per inning. The standard deviation of 6.305679 quantifies the spread or variability in runs scored by players in your dataset. In the context of runs, a standard deviation of 6.305679 means that the individual run scores of players tend to deviate from the mean (average) by approximately 6.305679 runs on average.

It quantifies how much individual data points differ from the mean on average. A higher standard deviation indicates greater variability in the data.

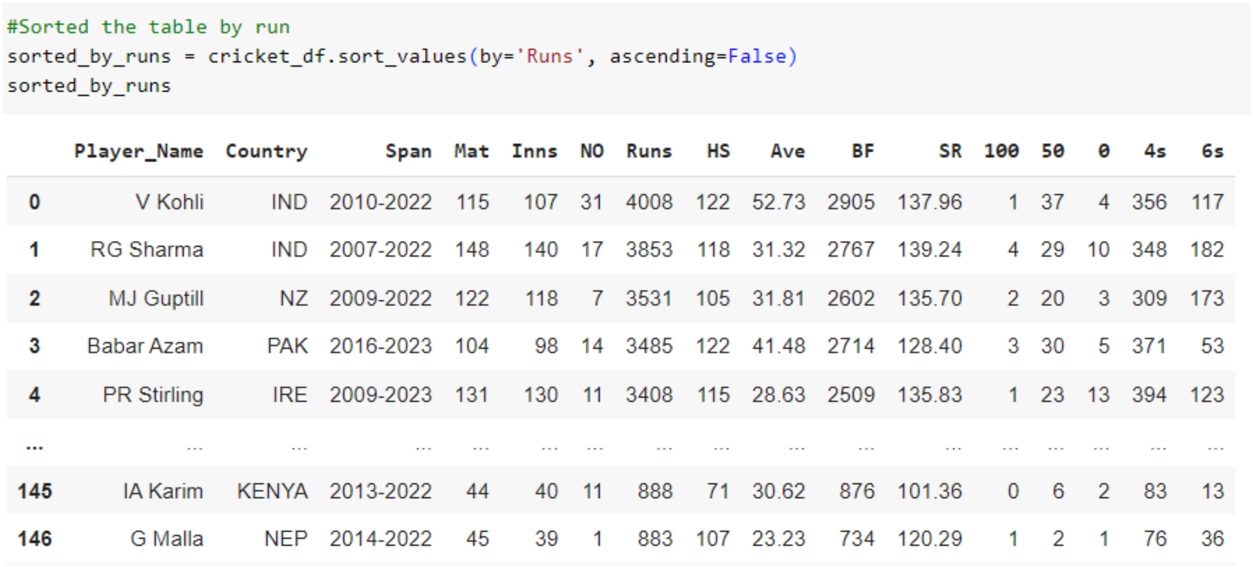
In statistics, the 25th percentile,50th percentile and 75th percentile are values that divide a dataset into four equal parts, with each part representing 25% of the data. These percentiles are also known as quartiles. The 25th percentile, denoted as Q1, represents the value below which 25% of the data falls. The 50th percentile, denoted as Q2 or simply the median, represents the middle value in the dataset when it's sorted in ascending order. The 75th percentile, denoted as Q3, represents the value below which 75% of the data falls. Percentiles are particularly useful for understanding the spread and distribution of data, especially in situations where data may contain outliers or be skewed. They provide a way to describe where data points are concentrated within a dataset and help identify potential data points that are higher or lower than the majority of the observations. the distribution of data and identify key summary statistics.

Min and Max will represent minimum and maximum value of a particular column.

17.56 is the minimum average

52.73 is the maximum average

When Data is Sorted by run



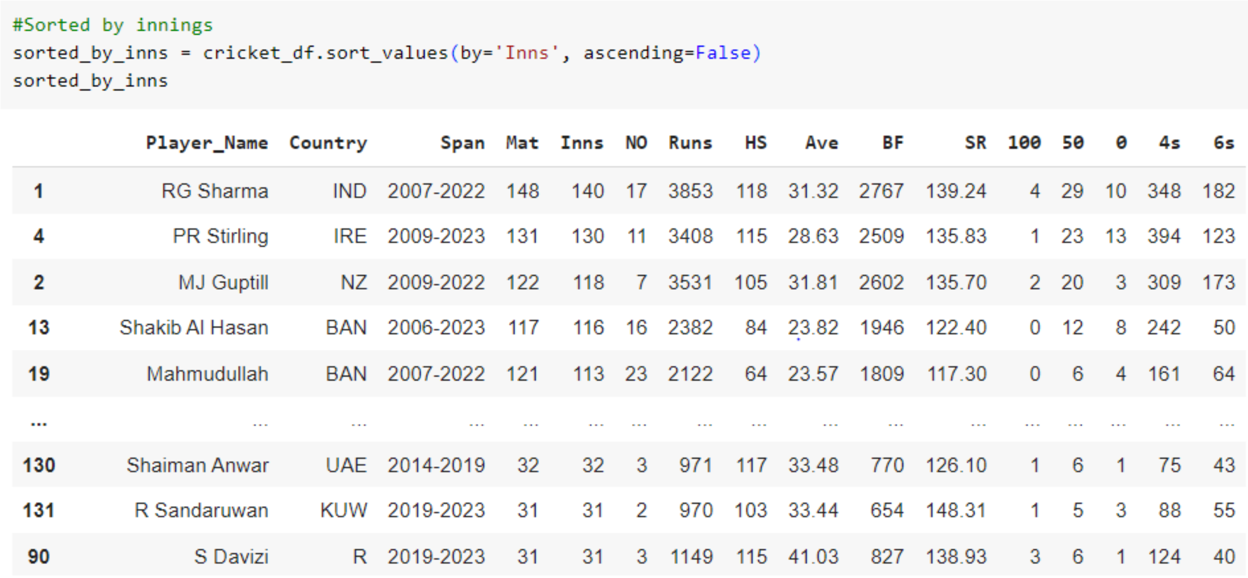
```
#Sorted the table by run
sorted_by_runs = cricket_df.sort_values(by='Runs', ascending=False)
sorted_by_runs
```

	Player_Name	Country	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s
0	V Kohli	IND	2010-2022	115	107	31	4008	122	52.73	2905	137.96	1	37	4	356	117
1	RG Sharma	IND	2007-2022	148	140	17	3853	118	31.32	2767	139.24	4	29	10	348	182
2	MJ Guptill	NZ	2009-2022	122	118	7	3531	105	31.81	2602	135.70	2	20	3	309	173
3	Babar Azam	PAK	2016-2023	104	98	14	3485	122	41.48	2714	128.40	3	30	5	371	53
4	PR Stirling	IRE	2009-2023	131	130	11	3408	115	28.63	2509	135.83	1	23	13	394	123
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
145	IA Karim	KENYA	2013-2022	44	40	11	888	71	30.62	876	101.36	0	6	2	83	13
146	G Malla	NEP	2014-2022	45	39	1	883	107	23.23	734	120.29	1	2	1	76	36

- ➔ Max runs are scored by Virat Kohli

- ➔ He is one of the important player for his team performance.

When data is Sorted by Innings



```
#Sorted by innings
sorted_by_inns = cricket_df.sort_values(by='Inns', ascending=False)
sorted_by_inns
```

	Player_Name	Country	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s
1	RG Sharma	IND	2007-2022	148	140	17	3853	118	31.32	2767	139.24	4	29	10	348	182
4	PR Stirling	IRE	2009-2023	131	130	11	3408	115	28.63	2509	135.83	1	23	13	394	123
2	MJ Guptill	NZ	2009-2022	122	118	7	3531	105	31.81	2602	135.70	2	20	3	309	173
13	Shakib Al Hasan	BAN	2006-2023	117	116	16	2382	84	23.82	1946	122.40	0	12	8	242	50
19	Mahmudullah	BAN	2007-2022	121	113	23	2122	64	23.57	1809	117.30	0	6	4	161	64
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
130	Shaiman Anwar	UAE	2014-2019	32	32	3	971	117	33.48	770	126.10	1	6	1	75	43
131	R Sandaruwan	KUW	2019-2023	31	31	2	970	103	33.44	654	148.31	1	5	3	88	55
90	S Davizi	R	2019-2023	31	31	3	1149	115	41.03	827	138.93	3	6	1	124	40

- ➔ Max innings are played by Rohit Sharma
- ➔ He is playing for a long time and also an opener for his team

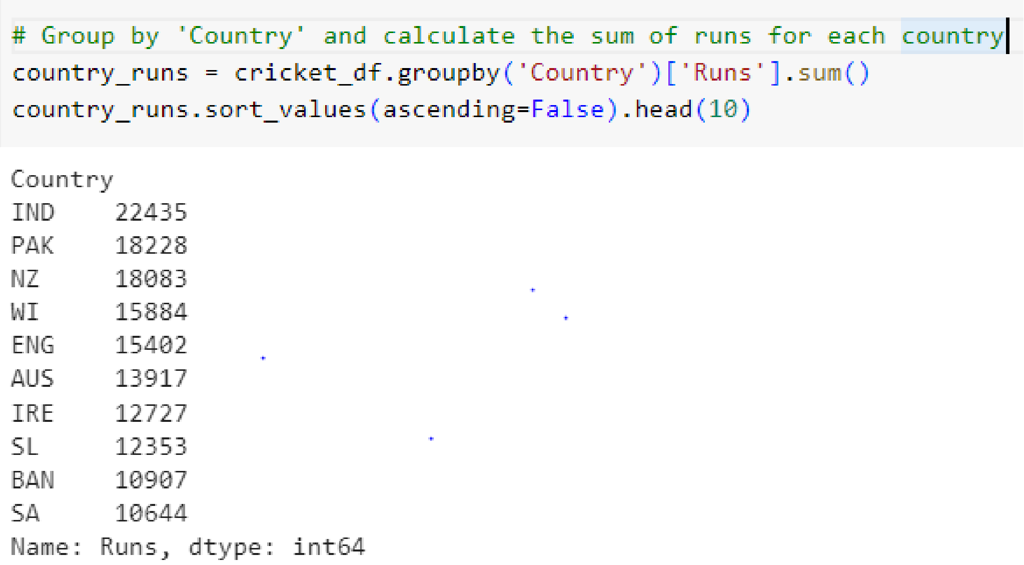
Extracting the data of top 8 countries Players

The criteria we use to extract top 8 countries is total runs scored by playears from each countries.

We choose this criteria because:

Total runs scored is a fundamental performance metric in cricket. Countries with higher total runs often have a strong and consistent batting lineup. It allows for a straightforward comparison between countries.

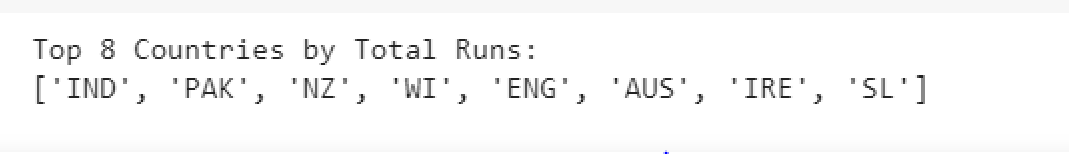
Count we have countries from each country :



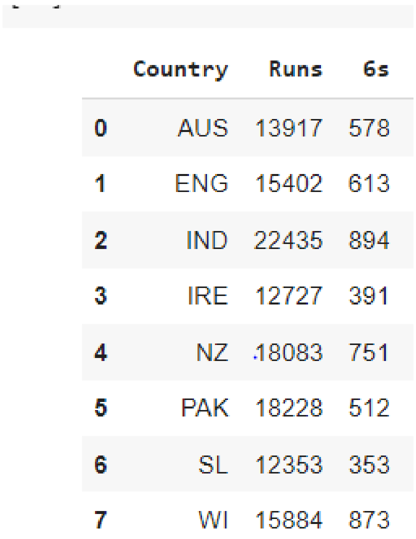
```
# Group by 'Country' and calculate the sum of runs for each country
country_runs = cricket_df.groupby('Country')['Runs'].sum()
country_runs.sort_values(ascending=False).head(10)
```

```
Country
IND    22435
PAK    18228
NZ     18083
WI     15884
ENG    15402
AUS    13917
IRE    12727
SL     12353
BAN    10907
SA     10644
Name: Runs, dtype: int64
```

The Top 8 Countries are by Total Runs are:



```
Top 8 Countries by Total Runs:
['IND', 'PAK', 'NZ', 'WI', 'ENG', 'AUS', 'IRE', 'SL']
```

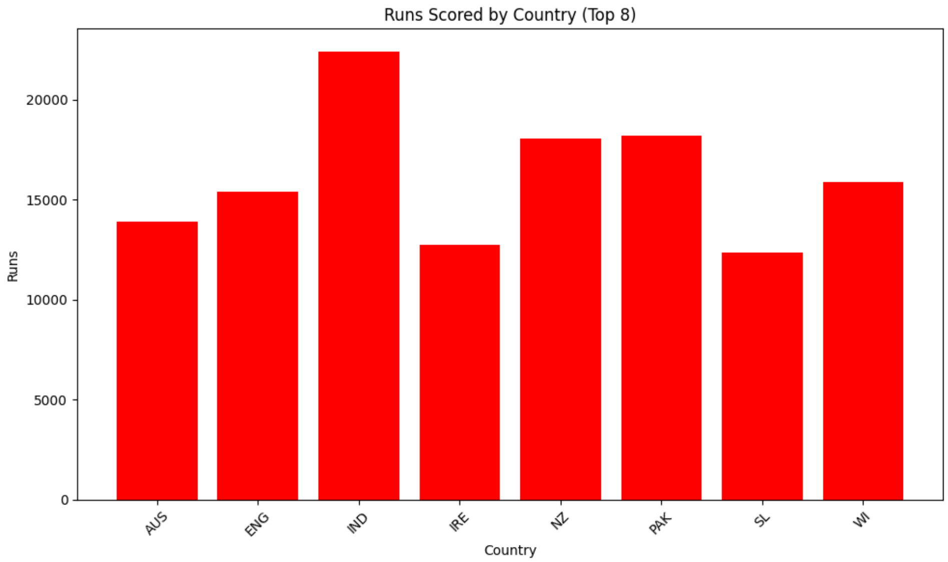


	Country	Runs	6s
0	AUS	13917	578
1	ENG	15402	613
2	IND	22435	894
3	IRE	12727	391
4	NZ	18083	751
5	PAK	18228	512
6	SL	12353	353
7	WI	15884	873

FINDINGS AND INFERENCES

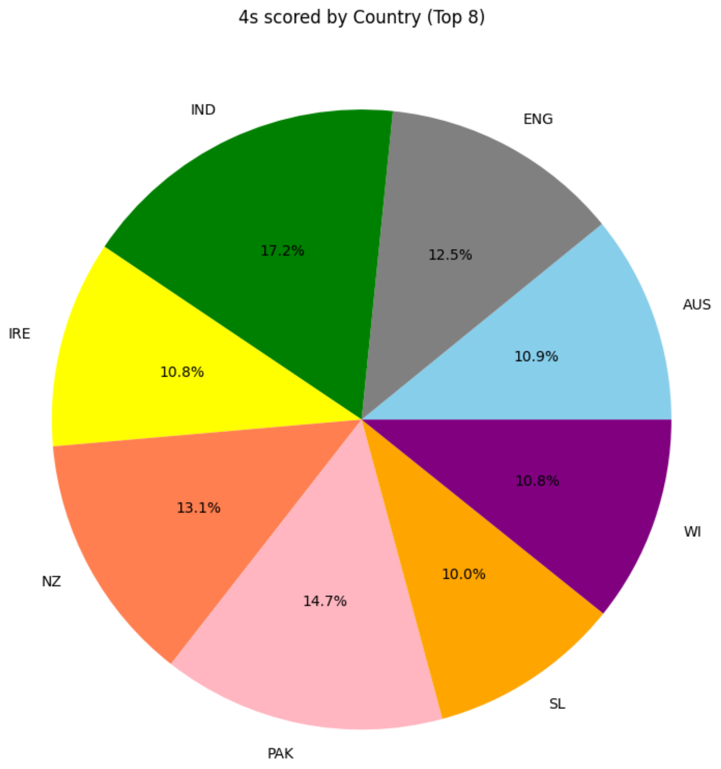
Comparisions between top 8 countries on the basis of :

i) Runs



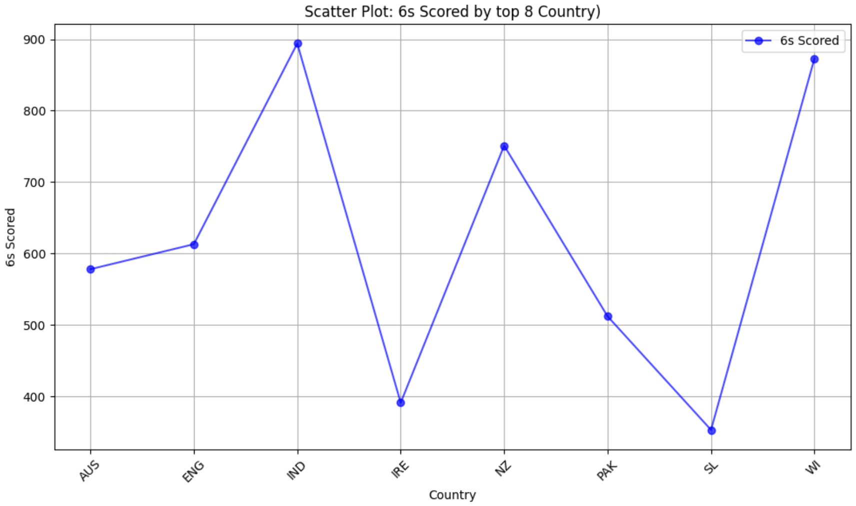
On the basis of the bar chart, we can say maximum total runs are scored by India players

ii) 4s



By this pie chart, we can say that max 4s are hit by India players

iii) 6s

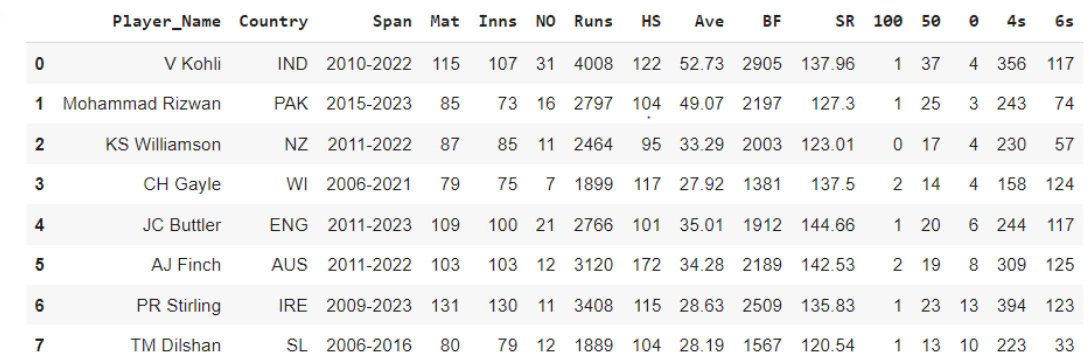


By this line graph, we can say that max 6s are hit by combining Indian players

There could be three reasons for India topped in each factors are :

1. The conditions of the pitches where India played their matches could favor batting and aggressive strokeplay.
2. India have strong and agressive batters that consistently scores a high number of runs.
3. India may have batters who are exceptionally skilled at finding the gaps and hitting boundaries. This could result in a higher number of fours.

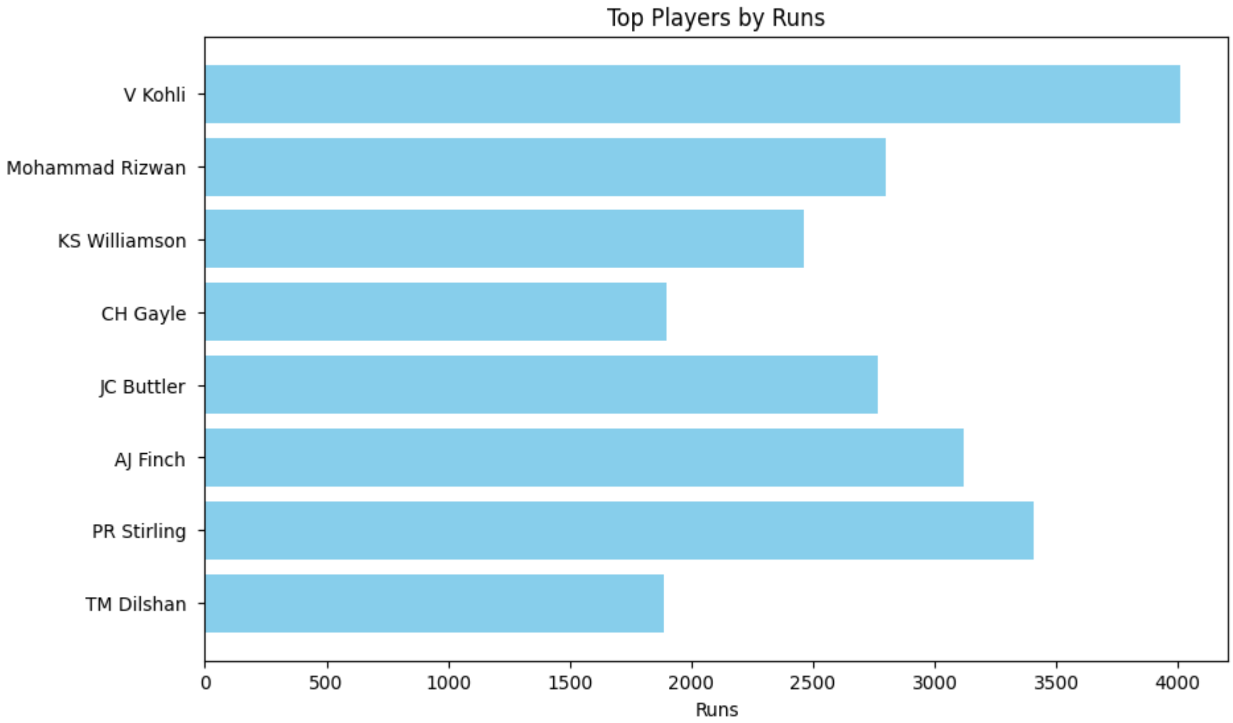
<u>Extracting the top player from each country on the basis that he has highest averge if a player has scored more than 1500 runs and played more than 70 innings</u>



	Player_Name	Country	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s
0	V Kohli	IND	2010-2022	115	107	31	4008	122	52.73	2905	137.96	1	37	4	356	117
1	Mohammad Rizwan	PAK	2015-2023	85	73	16	2797	104	49.07	2197	127.3	1	25	3	243	74
2	KS Williamson	NZ	2011-2022	87	85	11	2464	95	33.29	2003	123.01	0	17	4	230	57
3	CH Gayle	WI	2006-2021	79	75	7	1899	117	27.92	1381	137.5	2	14	4	158	124
4	JC Buttler	ENG	2011-2023	109	100	21	2766	101	35.01	1912	144.66	1	20	6	244	117
5	AJ Finch	AUS	2011-2022	103	103	12	3120	172	34.28	2189	142.53	2	19	8	309	125
6	PR Stirling	IRE	2009-2023	131	130	11	3408	115	28.63	2509	135.83	1	23	13	394	123
7	TM Dilshan	SL	2006-2016	80	79	12	1889	104	28.19	1567	120.54	1	13	10	223	33

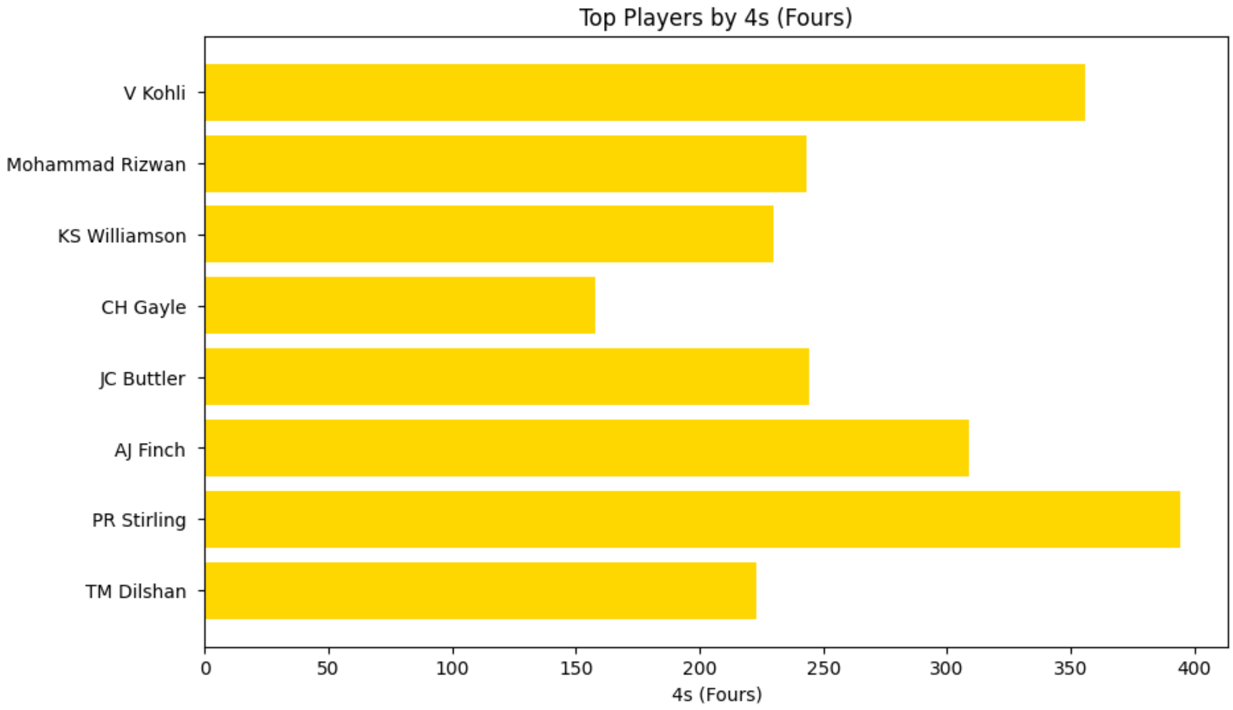
Comparing the top player from each country by :

i) Runs



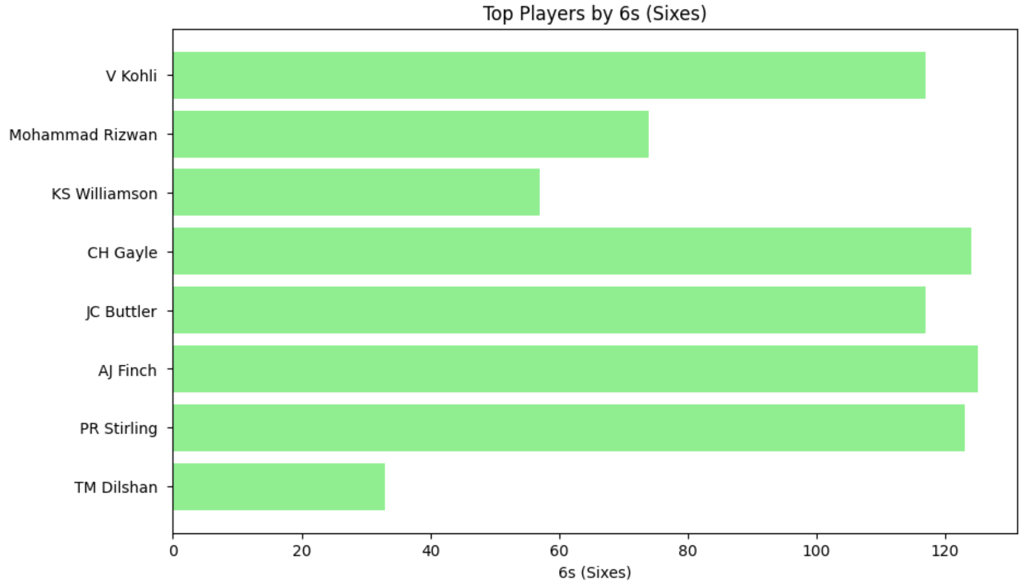
The most runs is scored by Virat Kohli(IND) And the second most is scored by PR stirling (IRE) among the top player from each country

ii) 4s



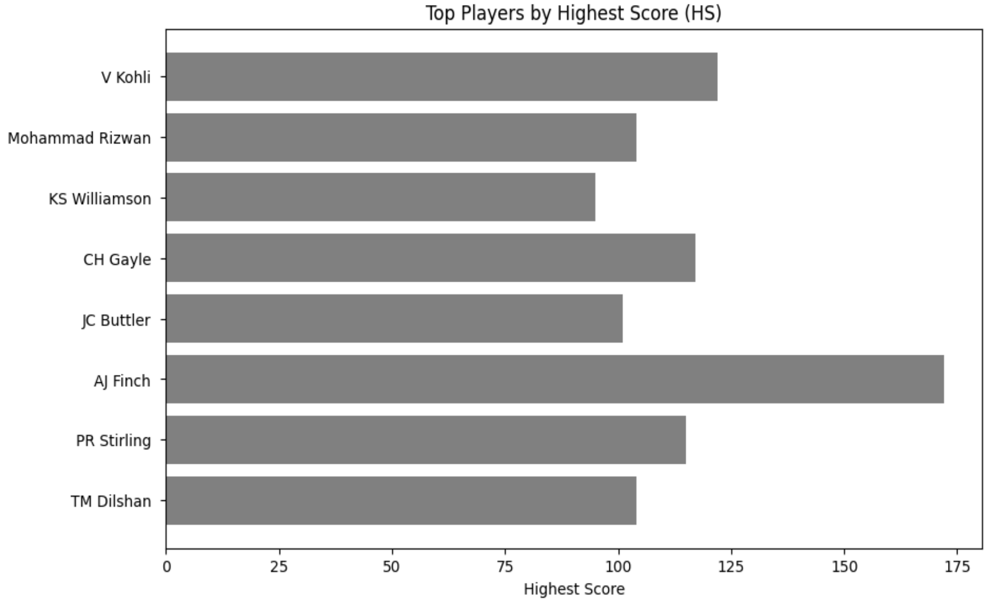
Most 4s are scored by PR STIRLING And second most by Virat Kohli among the top player from each country

iii) 6s

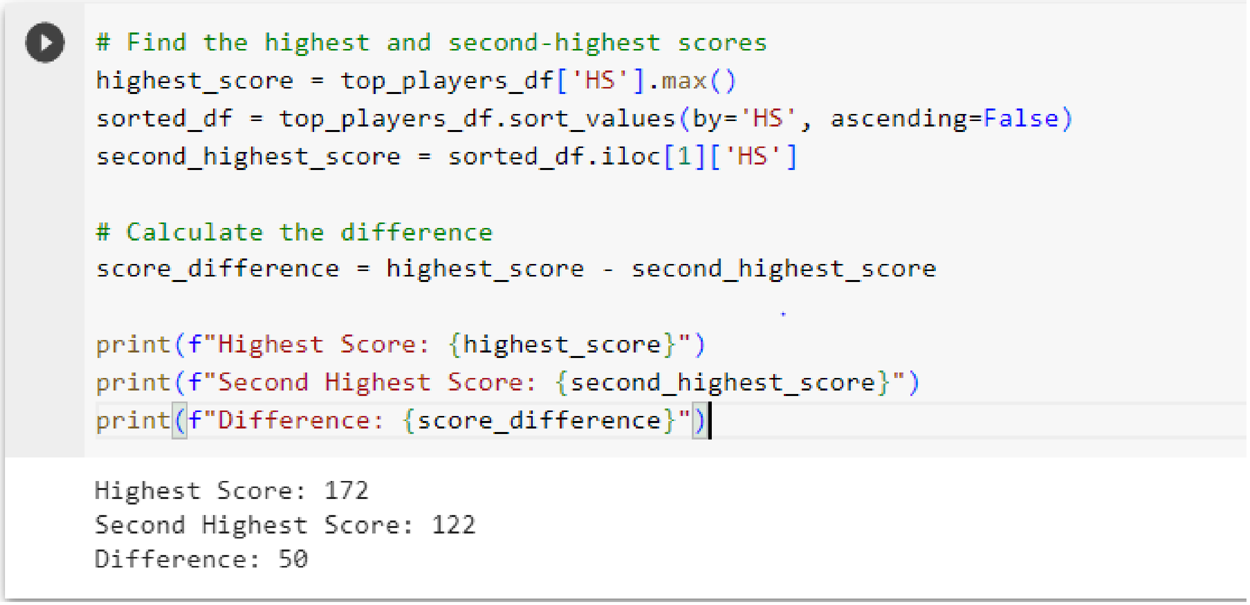


Most no. of 6s are hit by AJ Finch and 2nd most 6s hit by CH Gayle among the top player from each country

iv) Highest Score



The highest score is made by AJ Finch which is 172



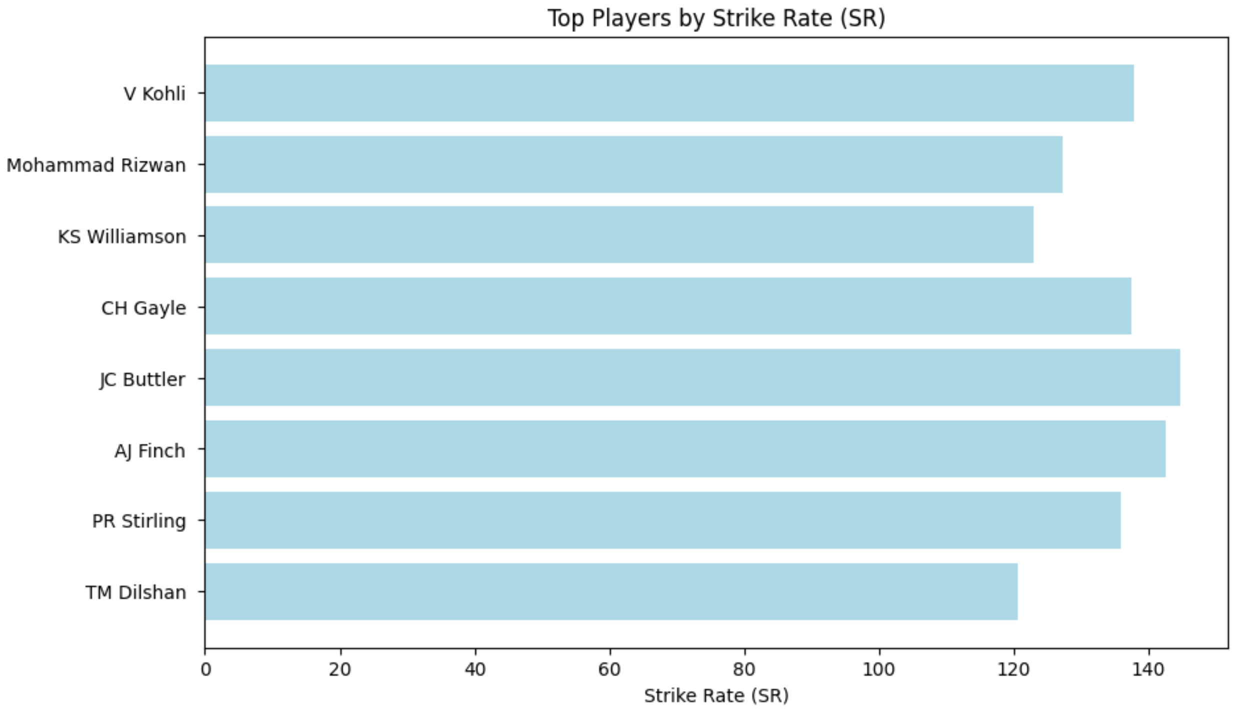
```
# Find the highest and second-highest scores
highest_score = top_players_df['HS'].max()
sorted_df = top_players_df.sort_values(by='HS', ascending=False)
second_highest_score = sorted_df.iloc[1]['HS']

# Calculate the difference
score_difference = highest_score - second_highest_score

print(f"Highest Score: {highest_score}")
print(f"Second Highest Score: {second_highest_score}")
print(f"Difference: {score_difference}")
```

```
Highest Score: 172
Second Highest Score: 122
Difference: 50
```

vi) Strike Rate

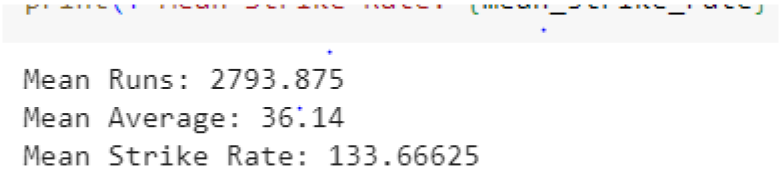


Players with Higher Strike :JC BUTTLER

Player with 2ND Highest strike rates : AJ FINCH

Player with Lower Strike rate : TM Dilshan among the top player from each country. This represent JC Buttler play with intent most of time.

MEAN OF RUNS, AVERAGE AND, STRIKE RATE SCORED BY TOP PLAYER OF TOP 8 COUNTRIES

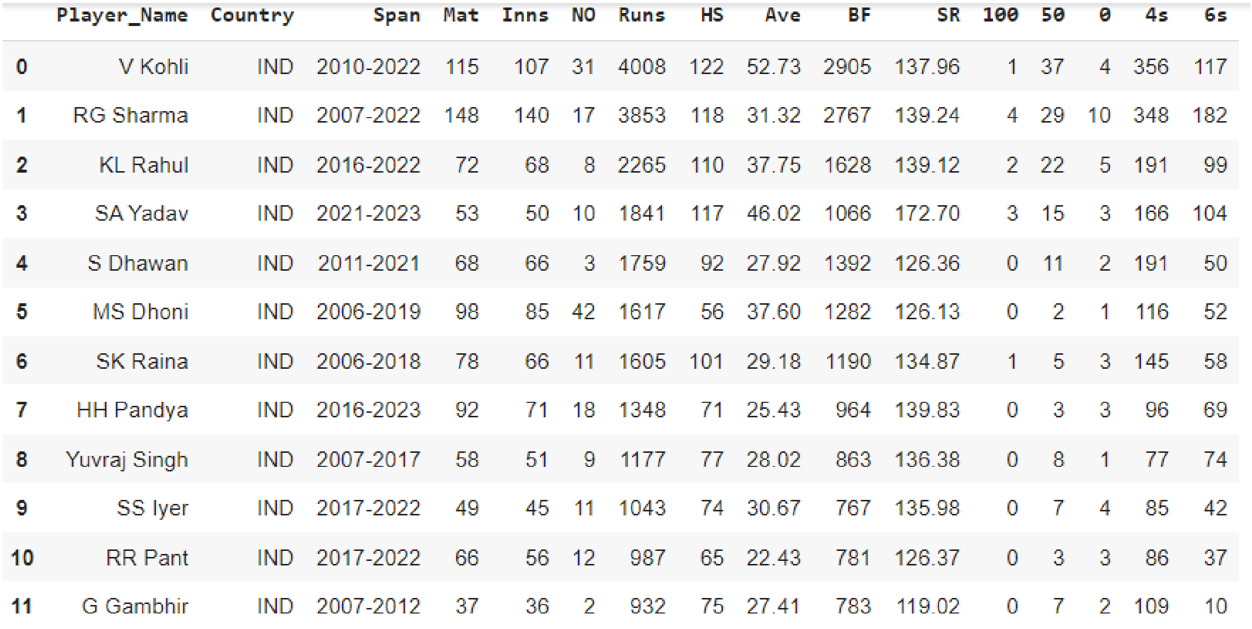


```
print(f"Mean Strike Rate: {mean_strike_rate}")

Mean Runs: 2793.875
Mean Average: 36.14
Mean Strike Rate: 133.66625
```

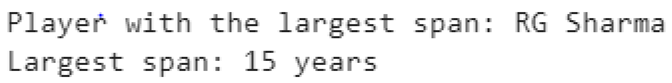
WE CAN SAY VIRAT KOHLI IS THE BEST PLAYER BECAUSE IN MAXIMUM ASPECT, HE IS IN TOP 1 OR TOP 2.

Extracting data of Indian Players out of the total players



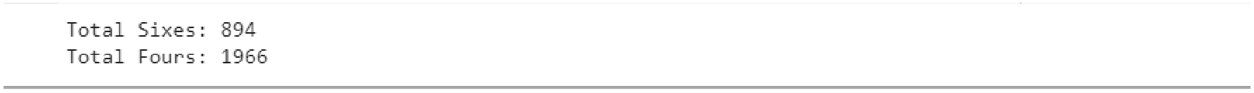
	Player_Name	Country	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s
0	V Kohli	IND	2010-2022	115	107	31	4008	122	52.73	2905	137.96	1	37	4	356	117
1	RG Sharma	IND	2007-2022	148	140	17	3853	118	31.32	2767	139.24	4	29	10	348	182
2	KL Rahul	IND	2016-2022	72	68	8	2265	110	37.75	1628	139.12	2	22	5	191	99
3	SA Yadav	IND	2021-2023	53	50	10	1841	117	46.02	1066	172.70	3	15	3	166	104
4	S Dhawan	IND	2011-2021	68	66	3	1759	92	27.92	1392	126.36	0	11	2	191	50
5	MS Dhoni	IND	2006-2019	98	85	42	1617	56	37.60	1282	126.13	0	2	1	116	52
6	SK Raina	IND	2006-2018	78	66	11	1605	101	29.18	1190	134.87	1	5	3	145	58
7	HH Pandya	IND	2016-2023	92	71	18	1348	71	25.43	964	139.83	0	3	3	96	69
8	Yuvraj Singh	IND	2007-2017	58	51	9	1177	77	28.02	863	136.38	0	8	1	77	74
9	SS Iyer	IND	2017-2022	49	45	11	1043	74	30.67	767	135.98	0	7	4	85	42
10	RR Pant	IND	2017-2022	66	56	12	987	65	22.43	781	126.37	0	3	3	86	37
11	G Gambhir	IND	2007-2012	37	36	2	932	75	27.41	783	119.02	0	7	2	109	10

PLAYER WITH LARGEST SPAN



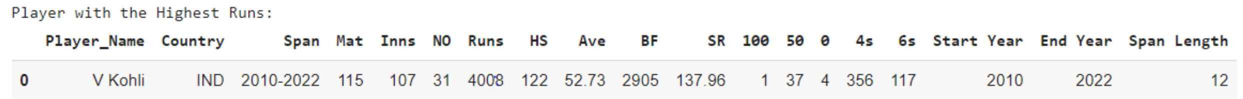
```
Player with the largest span: RG Sharma
Largest span: 15 years
```

Total 4s and 6s by Indian players:



```
    Total Sixes: 894
    Total Fours: 1966
```

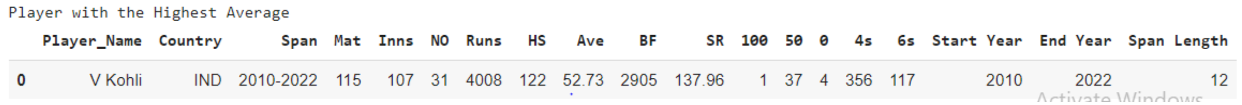
Player with Highest Run getter :



Player with the Highest Runs:

	Player_Name	Country	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s	Start Year	End Year	Span Length
0	V Kohli	IND	2010-2022	115	107	31	4008	122	52.73	2905	137.96	1	37	4	356	117	2010	2022	12

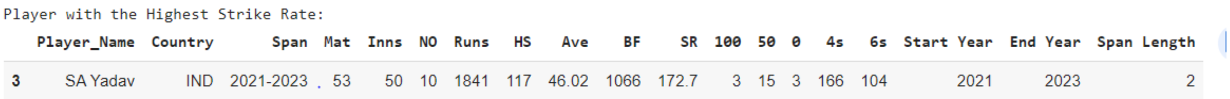
Player with Highest Average :



Player with the Highest Average

	Player_Name	Country	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s	Start Year	End Year	Span Length
0	V Kohli	IND	2010-2022	115	107	31	4008	122	52.73	2905	137.96	1	37	4	356	117	2010	2022	12

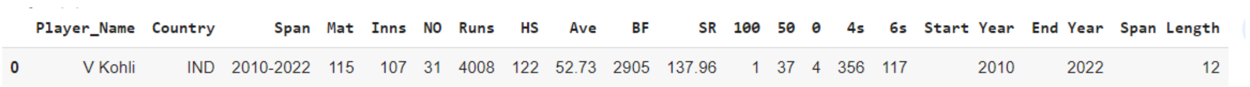
Player with Highest Strike Rate :



Player with the Highest Strike Rate:

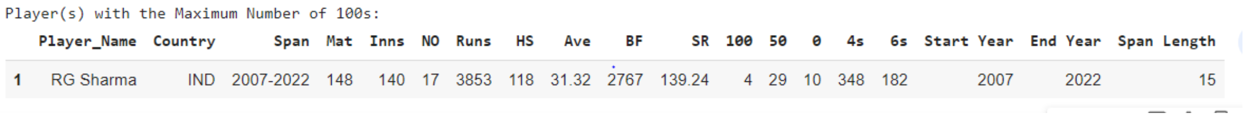
	Player_Name	Country	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s	Start Year	End Year	Span Length
3	SA Yadav	IND	2021-2023	53	50	10	1841	117	46.02	1066	172.7	3	15	3	166	104	2021	2023	2

Player with Maximum 50 :



	Player_Name	Country	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s	Start Year	End Year	Span Length
0	V Kohli	IND	2010-2022	115	107	31	4008	122	52.73	2905	137.96	1	37	4	356	117	2010	2022	12

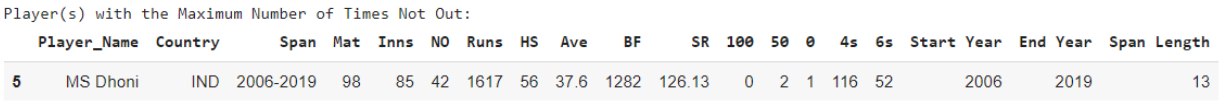
Player with maximum 100 :



Player(s) with the Maximum Number of 100s:

	Player_Name	Country	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s	Start Year	End Year	Span Length
1	RG Sharma	IND	2007-2022	148	140	17	3853	118	31.32	2767	139.24	4	29	10	348	182	2007	2022	15

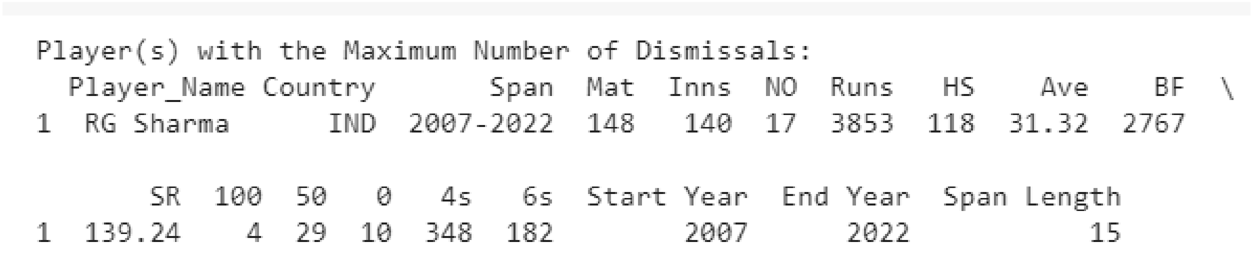
Player with Maximum no. of Not Outs



Player(s) with the Maximum Number of Times Not Out:

	Player_Name	Country	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s	Start Year	End Year	Span Length
5	MS Dhoni	IND	2006-2019	98	85	42	1617	56	37.6	1282	126.13	0	2	1	116	52	2006	2019	13

Player with max no. of Dismissals :



```
Player(s) with the Maximum Number of Dismissals:
  Player_Name Country       Span  Mat  Inns  NO  Runs   HS    Ave    BF  \
1   RG Sharma     IND  2007-2022  148   140  17  3853  118  31.32  2767   

       SR  100  50   0   4s   6s  Start Year  End Year  Span Length  
1  139.24    4  29  10  348  182        2007      2022           15  
```

If we consider only INDIAN players Different player are good in different aspect :

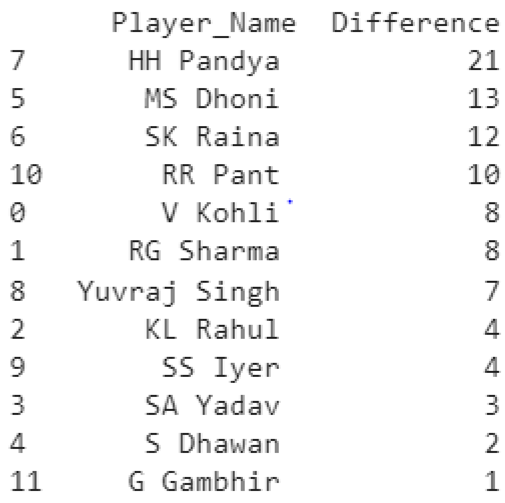
For Example :

VIRAT KOHLI is the highest run getter and has maximum 50 and highest average which shows he is the consistent player.

SA YADAV is the player with highest Strike Rate which shows he is attacking player, He play with intent.

ROHIT SHARMA is the player with maximum no. of runs and most no. of dismissals, which shows when he plays good he goes on scoring very high, otherwise zero. We don’t have enough data to know how he got out and more. So we can say if he get the flow, he is a very good batsman, otherwise not.

MS Dhoni is the player who has maximum no. of not outs, with the help of difference between matches and innings(given below)->

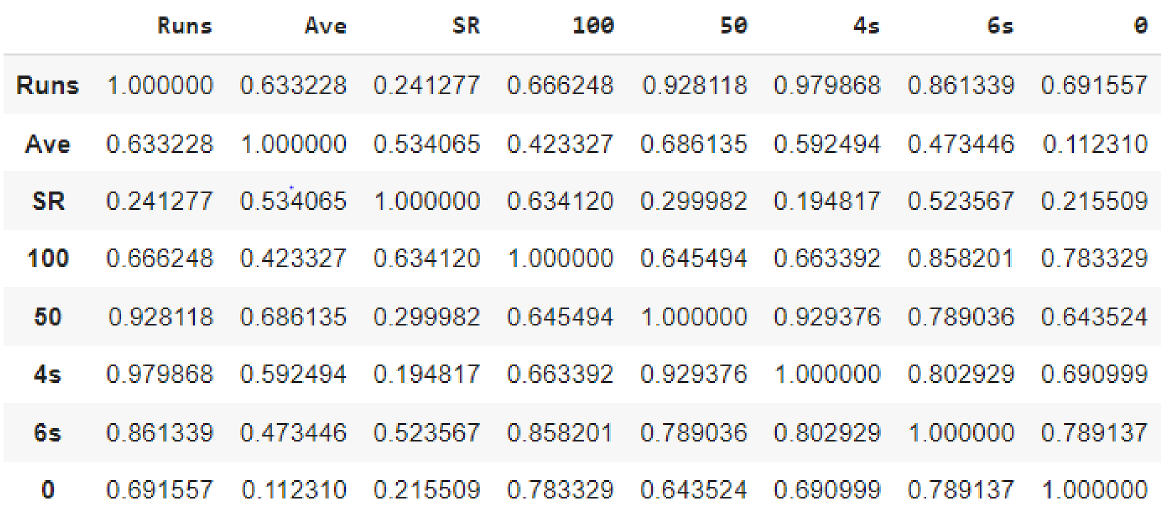


```
      Player_Name  Difference
7       HH Pandya          21
5        MS Dhoni          13
6        SK Raina          12
10        RR Pant          10
0         V Kohli           8
1       RG Sharma           8
8    Yuvraj Singh           7
2        KL Rahul           4
9         SS Iyer           4
3        SA Yadav           3
4       S Dhawan            2
11      G Gambhir           1
```

We can say say, He did not have much score compared to other player who have played approximately equal match. And Out of the 98 matches he play only 85 innings, which is the 2nd largest difference. So, he did not get batting in many matches as well because top order playes all the overs. So, we can say he come down the order and finishes the game.

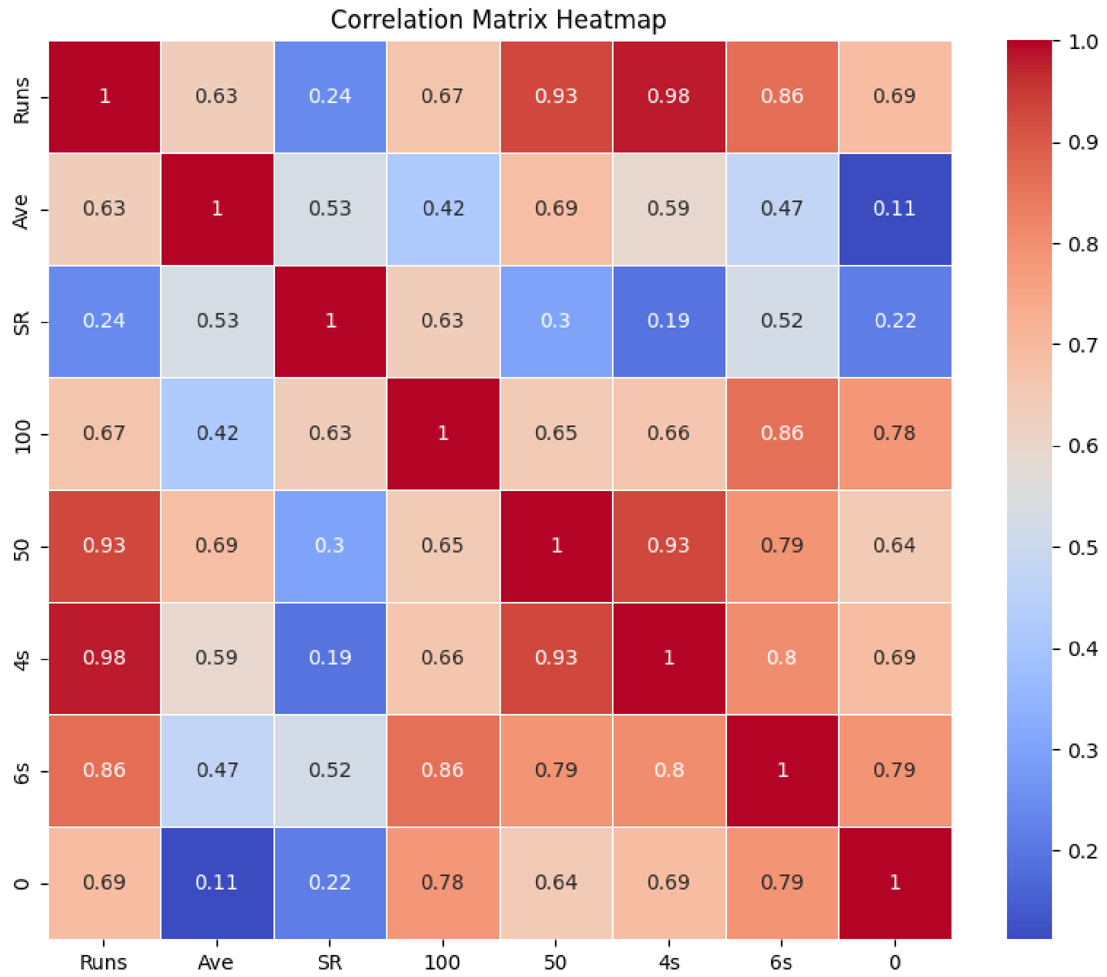
Correlation table

Correlation Matrix:

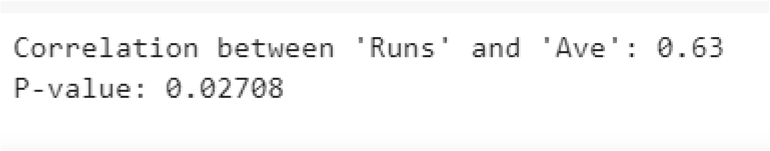


	Runs	Ave	SR	100	50	4s	6s	0
Runs	1.000000	0.633228	0.241277	0.666248	0.928118	0.979868	0.861339	0.691557
Ave	0.633228	1.000000	0.534065	0.423327	0.686135	0.592494	0.473446	0.112310
SR	0.241277	0.534065	1.000000	0.634120	0.299982	0.194817	0.523567	0.215509
100	0.666248	0.423327	0.634120	1.000000	0.645494	0.663392	0.858201	0.783329
50	0.928118	0.686135	0.299982	0.645494	1.000000	0.929376	0.789036	0.643524
4s	0.979868	0.592494	0.194817	0.663392	0.929376	1.000000	0.802929	0.690999
6s	0.861339	0.473446	0.523567	0.858201	0.789036	0.802929	1.000000	0.789137
0	0.691557	0.112310	0.215509	0.783329	0.643524	0.690999	0.789137	1.000000

Correlaion Matrix Heatmap :



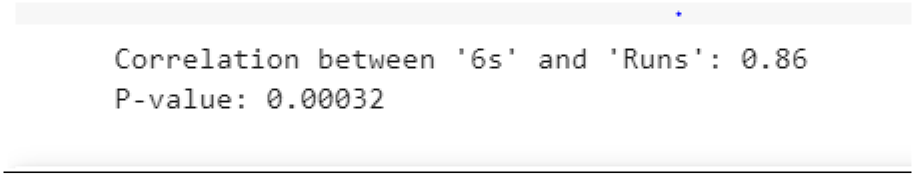
Calculate the correlation and p-value for ‘Runs’ and ‘Ave’ columns.



```
Correlation between 'Runs' and 'Ave': 0.63
P-value: 0.02708
```

This tells : The correlation coefficient(r-value) between 'Runs' and 'Ave' is 0.63. This positive correlation suggests that as a player's batting average ('Ave') increases, their total runs ('Runs') tend to increase as well. The p-value is 0.02708, which is less than the typical significance level of 0.05. This indicates that the correlation is statistically significant, suggesting a meaningful relationship between batting average and total runs.

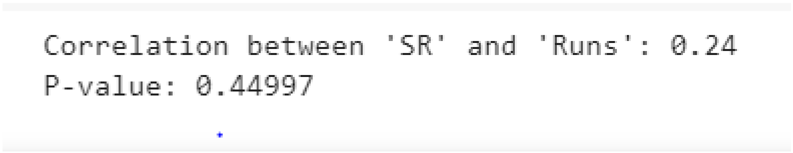
Calculate the correlation and p-value for ‘6s’ and ‘Runs’ columns:



```
Correlation between '6s' and 'Runs': 0.86
P-value: 0.00032
```

This tells :he correlation coefficient (r-value) between '6s' (sixes hit) and 'Runs' is 0.86. This strong positive correlation suggests that as a player scores more runs, they tend to hit more sixes. The p-value is 0.00032, which is much less than 0.05. This indicates that the correlation is highly statistically significant, confirming a strong relationship between the number of sixes and total runs.

Calculate the correlation and p-value for ‘SR’ and ‘Runs’ columns:



```
Correlation between 'SR' and 'Runs': 0.24
P-value: 0.44997
```

This tells : The correlation coefficient (r-value) between 'SR' (strike rate) and 'Runs' is 0.24. This positive correlation nearby 0 suggests that there is a weak relationship between a player's strike rate and their total runs. The p-value is 0.44997, which is greater than 0.05. This indicates that the correlation is not statistically significant, suggesting that the relationship between strike rate and total runs may not be meaningful.

By this we can conclude that if you hit more 4s and more 6s , your runs increases. But if a player play with good strike chances of making more run is low. The reason could be more risk is assosciated with in playing with good strike rate

MANAGERIAL INSIGHTS

- Identify players with a high number of 50s (half-centuries) as they consistently contribute with substantial runs. These players can be reliable middle-order batsmen.
- Focus on players who have hit a significant number of 4s and 6s, as they can provide quick runs and keep the scoreboard ticking as there is a positive correlations between runs and 4s and runs and 6s. These players are valuable in T20 cricket.
- According to p-value and r-value Runs and strike rate are weakly related. But In T20 cricket Strike Rate and Runs both are very important. So, choose a player who is a need to focus on both factors simultaneously.
- Identify players who have strike rate between, because some of the stars players of different countries have strike rate above 140 like JC Buttler and A Finch
- Identify Players who has average above 36.14, and strike rate above 136.66 as this is the mean of top players of top 8 countries.
- A correlation coefficient of 0.78 between no. of dismissals and maximum no. of 100 shows a relatively strong positive relationship so if an individual have max dismissals on 0 .then also he can win matches when he is in form.

- Players with a high number of not outs may have the ability to finish innings effectively and ensure that their team reaches a competitive total.