**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

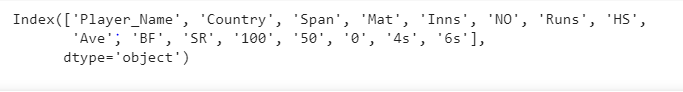
import matplotlib.pyplot as plt

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**



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.

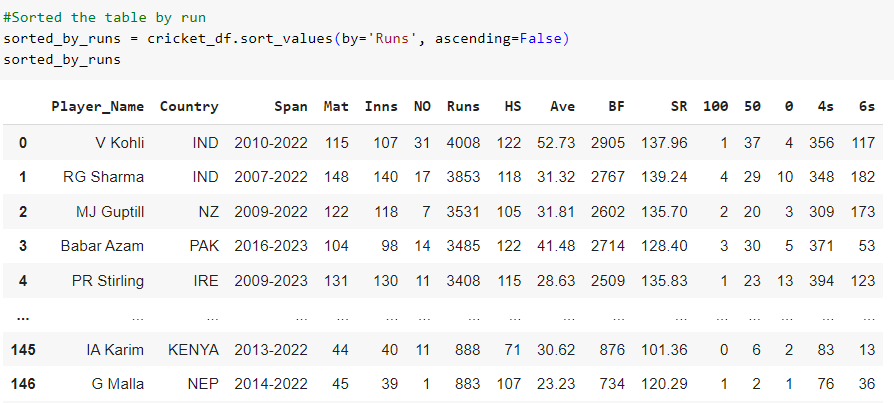
**Analysis:**

**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 |

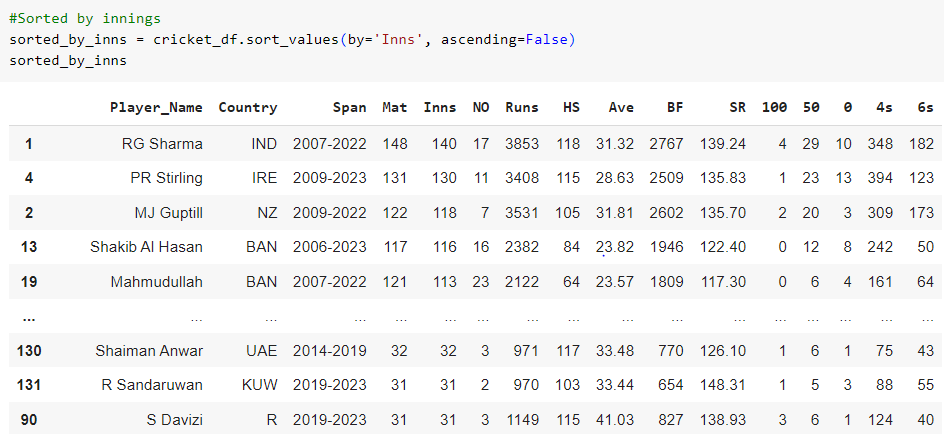
This gives us the mean, stand deviation,min,25%,50%,75%,max of matches,innings,total run scored, 50s, 100s, 4s, 6s, ball faced , highest score of players.

**When Data is Sorted by run**

****

* Max runs are scored by Virat Kohli
* He is one of the important player for his team performance.

**When data is Sorted by Innings**



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

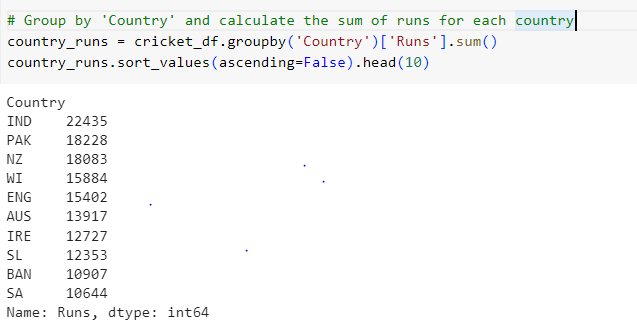
**Extracting the data of top 8 countries Players**

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

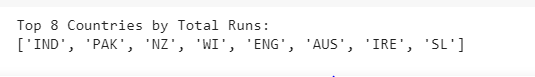
**We choose this criteria because:**

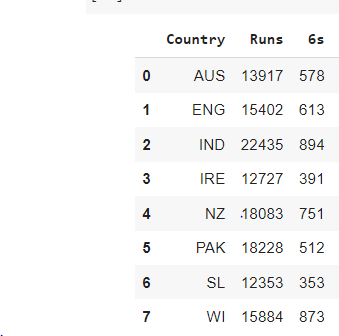
**Total runs scored is a fundamental performance metric in cricket. Countries with higher total runs often have a strong and consistent batting lineup. It allows for a straightforward comparison between countries.**

Count we have countries from each country :



The Top 8 Countries are by Total Runs are:

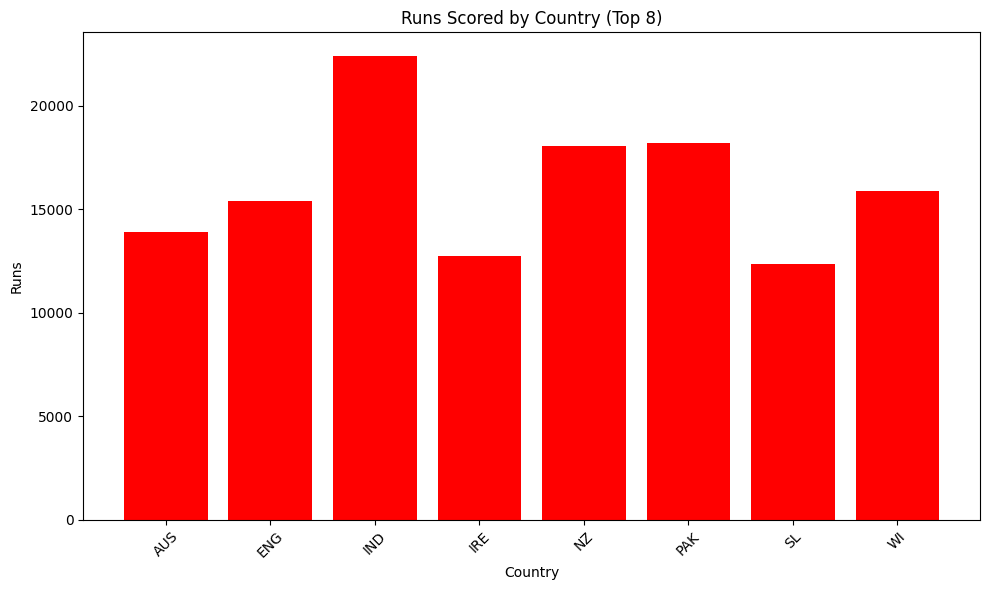
****



**FINDINGS AND INFERENCES**

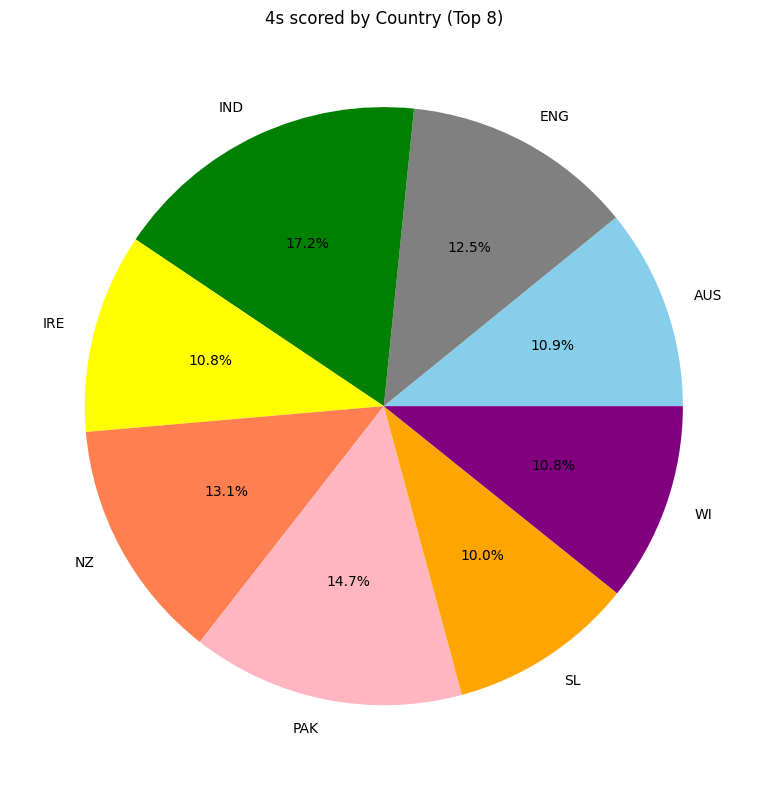
**Comparisions between top 8 countries on the basis of :**

1. **Runs**

****

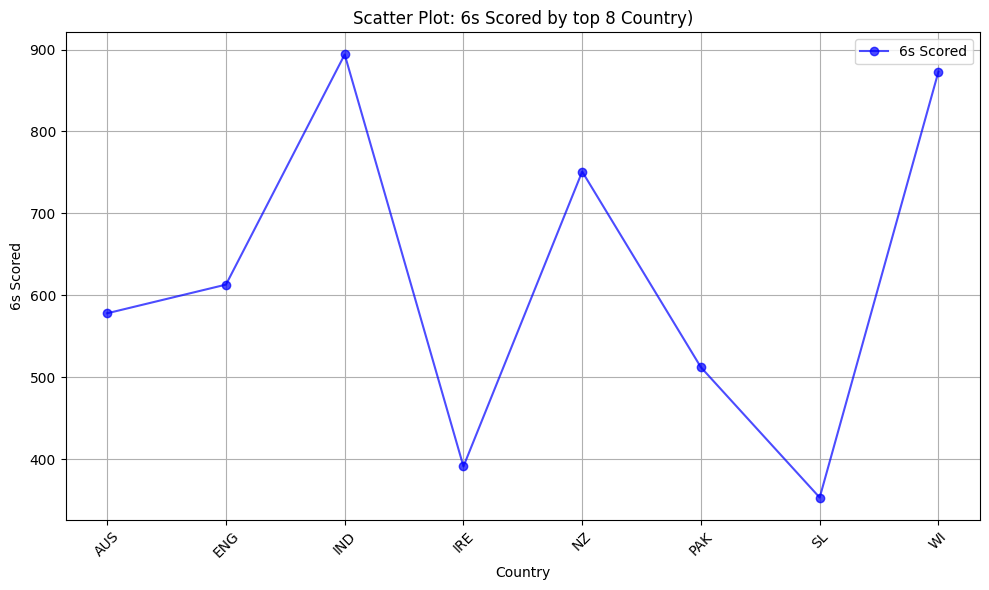
**On the basis of the bar chart, we can say maximum total runs are scored by India players**

1. **4s**



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

1. **6s**

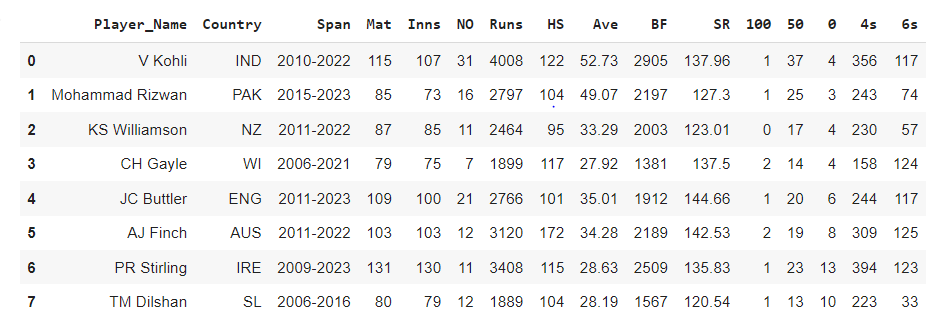


**By this line graph, we can say that max 6s are hit by combining Indian players**

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

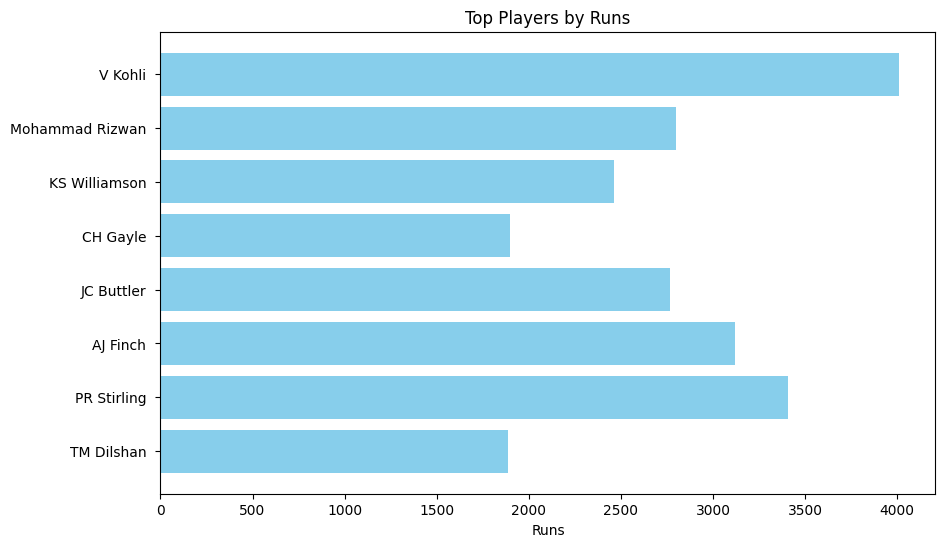
1. The conditions of the pitches where India played their matches could favor batting and aggressive strokeplay.
2. India have strong and agressive batters that consistently scores a high number of runs.
3. India may have batters who are exceptionally skilled at finding the gaps and hitting boundaries.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

**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**



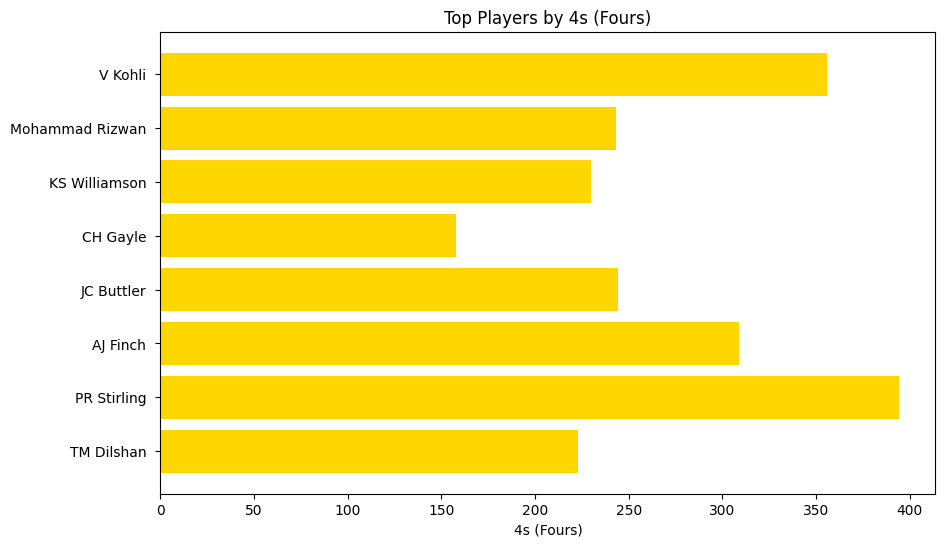
**Comparing the top player from each country by :**

1. **Runs**

****

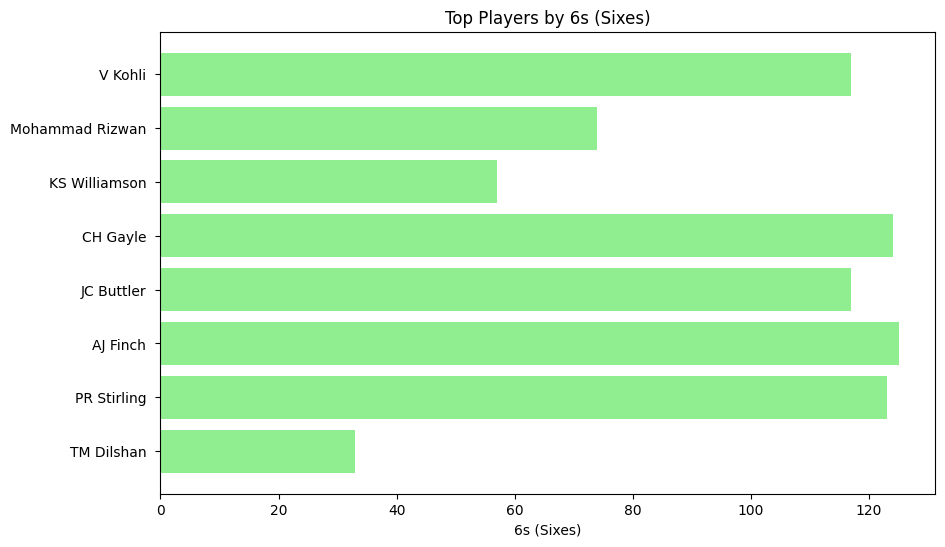
**The most runs is scored by Virat Kohli(IND) And the second most is scored by PR stirling (IRE) among the top player from each country**

1. **4s**

****

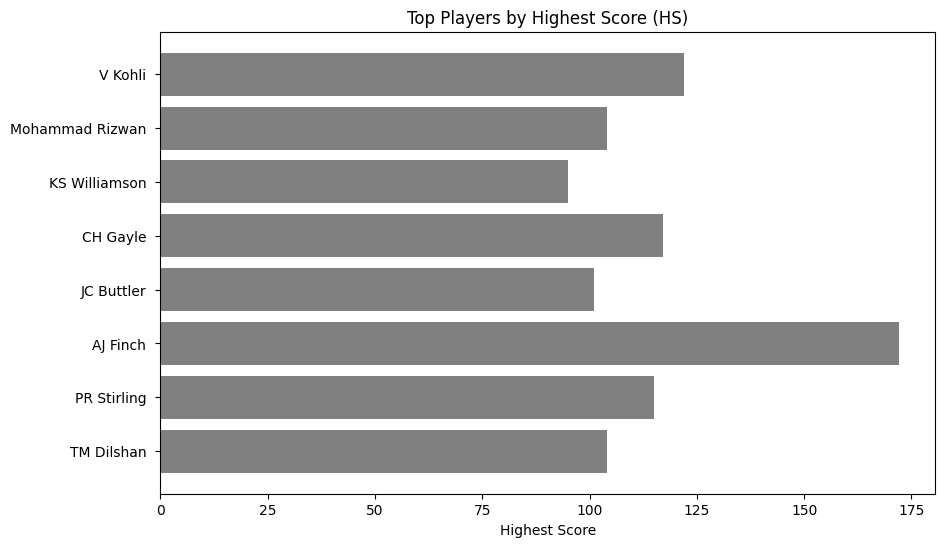
**Most 4s are scored by PR STIRLING And second most by Virat Kohli among the top player from each country**

1. **6s**

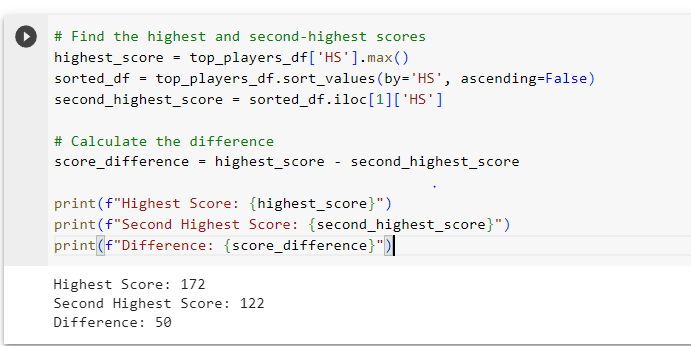
****

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

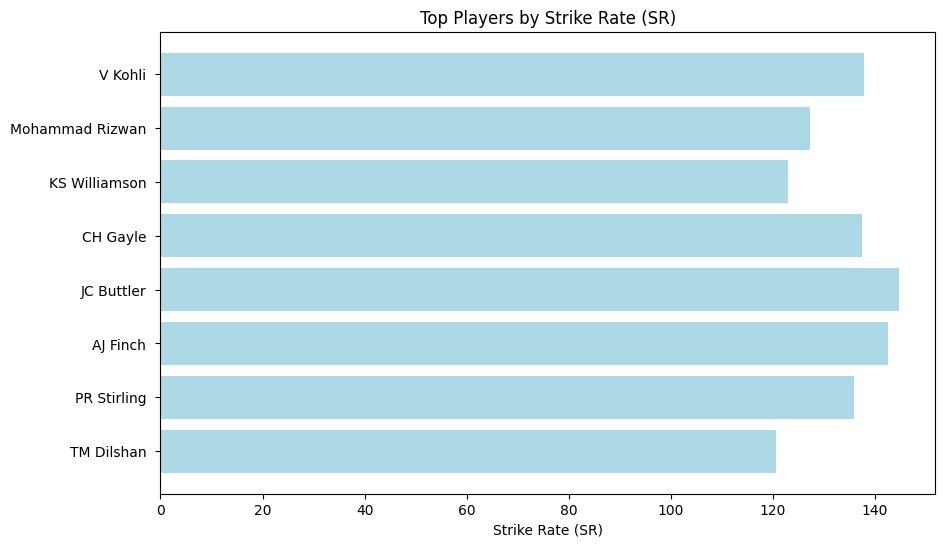
1. **Highest Score**

****

**The highest score is made by AJ Finch which is 172**

****

**vi) Strike Rate**

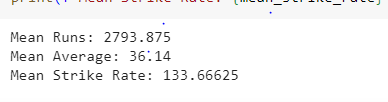


**Players with Higher Strike :JC BUTTLER**

**Player with 2ND Highest strike rates : AJ FINCH**

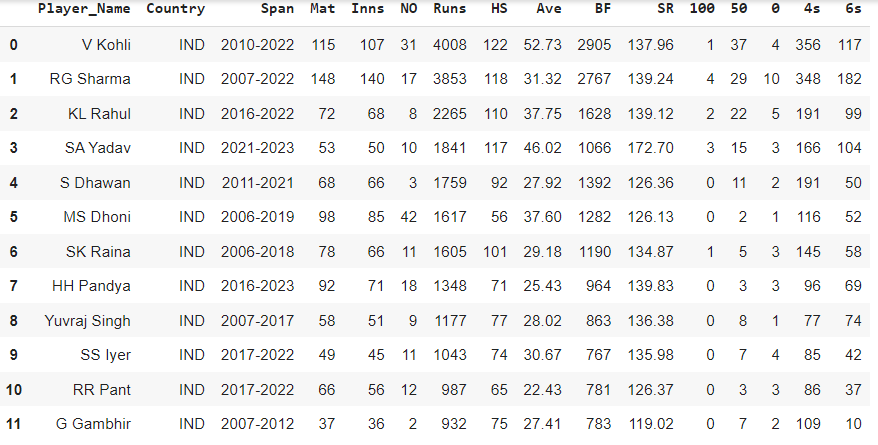
**Player with Lower Strike rate : TM Dilshan among the top player from each country. This represent JC Buttler play with intent most of time.**

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



**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**



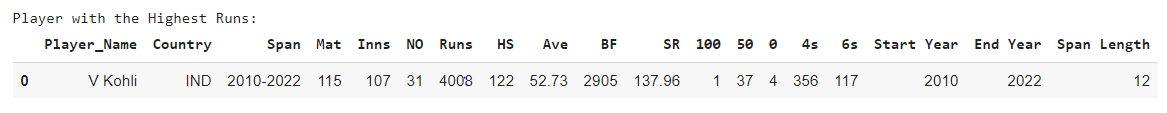
**PLAYER WITH LARGEST SPAN**



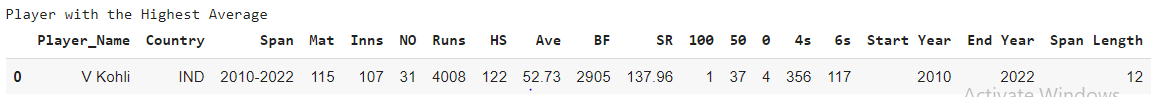
Total 4s and 6s by Indian players:



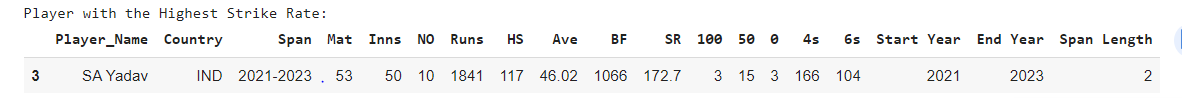
**Player with Highest Run getter :**



**Player with Highest Average :**



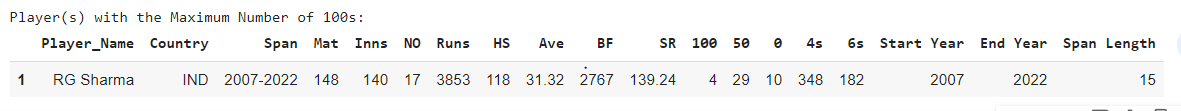
**Player with Highest Strike Rate :**



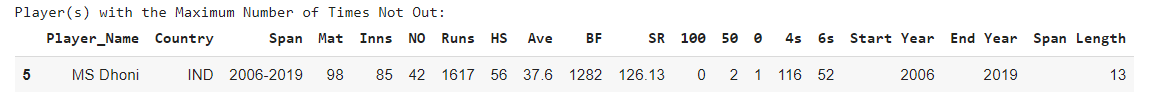
**Player with Maximum 50 :**



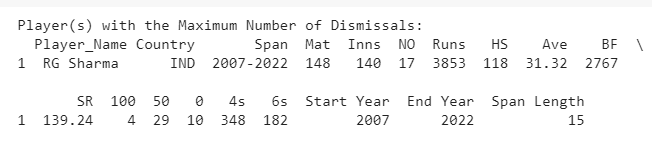
**Player with maximum 100 :**



**Player with Maximum no. of Not Outs**



**Player with max no. of Dismissals :**



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

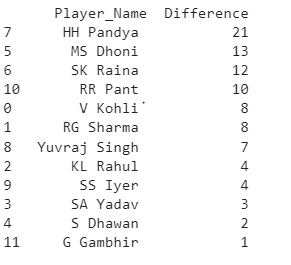
**For Example :**

**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.**

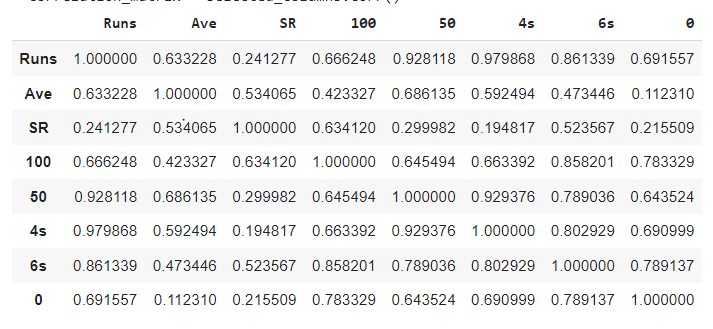
**MS Dhoni is the player who has maximum no. of not outs, with the help of difference between matches and innings(given below)->**



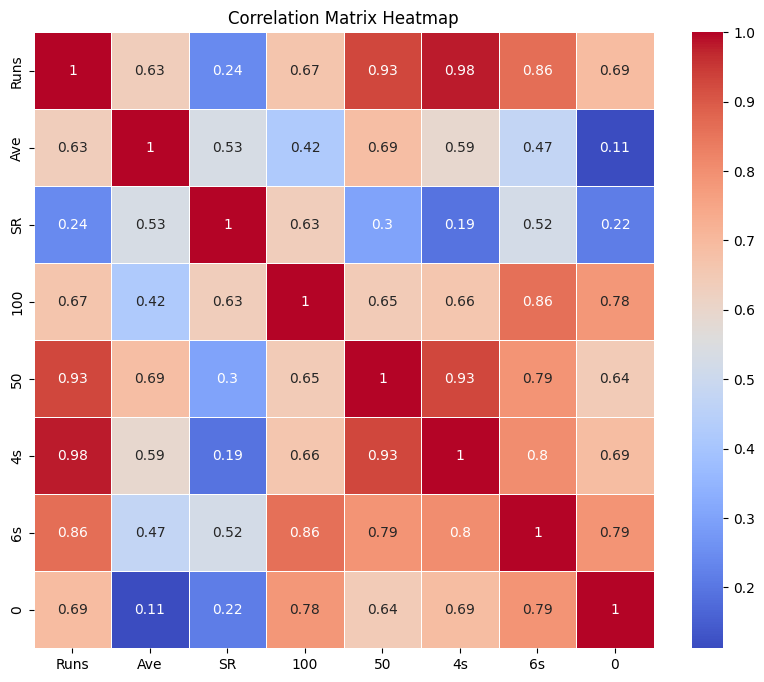
**We can say say, He did not have much score compared to other player who have played approximately equal match. And Out of the 98 matches he play only 85 innings, which is the 2nd largest difference. So, he did not get batting in many matches as well because top order playes all the overs. So, we can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

**Correlation table**

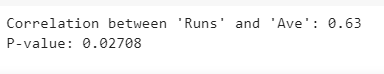
**Correlation Matrix:**



**Correlaion Matrix Heatmap :**

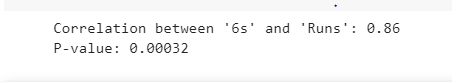
****

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



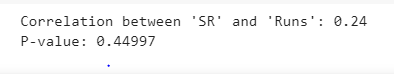
This tells : The correlation coefficient(r-value) between 'Runs' and 'Ave' is 0.63. This positive correlation suggests that as a player's batting average ('Ave') increases, their total runs ('Runs') tend to increase as well. The p-value is 0.02708, which is less than the typical significance level of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



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:**



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.