In [1]:

```
import warnings
warnings.filterwarnings('ignore')

import pandas as pd
import numpy as np

import matplotlib.pyplot as plt

#to display all rows columns
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
```

In [2]:

```
df = pd.read_csv('ODI_data.csv')
```

In [3]:

```
df.head(2)
```

Out[3]:

| | Innings Player | Innings Runs Scored | Innings Runs Scored Num | Innings Minutes Batted | Innings Batted Flag | Innings Not Out Flag | Innings Balls Faced | Innings Boundary Fours | Innings Boundary Sixes | Innings Batting Strike Rate |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | E Lewis | 65 | 65 | 128 | 1.0 | 0.0 | 80 | 8 | 1 | 81.25 |
| **1** | N Pooran | 42 | 42 | 69 | 1.0 | 0.0 | 52 | 4 | 1 | 80.76 |

In [4]:

```
len(df), len(df.columns)
```

Out[4]:

```
(171968, 28)
```

In [5]:

```
# Runs per innings
# SR
# 100's
# 50's
# Team contribution
```

In [6]:

```
df['Innings Runs Scored Num'].unique()
```

Out[6]:

```
array(['65', '42', '18', '17', '13', '11', '5', '0', '120', '71', '20',
       '16', '3', '2', '1', '-', nan, '40', '6', '4', '87', '54', '46',
       '30', '12', '69', '39', '14', '10', '9', '8', '7', '82', '52',
       '41', '15', '98', '43', '19', '111', '48', '36', '25', '67', '60',
       '84', '59', '55', '47', '85', '49', '45', '34', '29', '22', '74',
       '28', '77', '50', '32', '23', '35', '122', '100', '95', '103',
       '113', '53', '96', '27', '64', '58', '33', '31', '73', '56', '86',
       '62', '106', '24', '57', '104', '26', '66', '51', '118', '105',
       '101', '21', '79', '44', '102', '88', '80', '72', '97', '68', '89',
       '38', '83', '63', '148', '166', '90', '76', '37', '70', '124',
       '94', '140', '153', '107', '117', '121', '92', '78', '75', '114',
       '115', '130', '128', '151', '110', '138', '135', '109', '61',
       '179', '170', '112', '116', '91', '143', '93', '123', '145', '81',
       '150', '162', '108', '131', '133', '137', '146', '139', '125',
       '129', '157', '152', '144', '99', '127', '210', '147', '126',
       '181', '160', '180', '208', '176', '168', '141', '132', '119',
       '154', '185', '134', '156', '164', '173', '178', '171', '149',
       '237', '159', '161', '215', '264', '136', '169', '209', '174',
       '189', '183', '163', '219', '158', '175', '177', '200', '194',
       '142', '172', '186', '188', '167'], dtype=object)
```

In [7]:

```
df = df[df['Innings Runs Scored Num'] != '-']
```

In [8]:

```
df['Innings Runs Scored Num'].unique()
```

Out[8]:

```
array(['65', '42', '18', '17', '13', '11', '5', '0', '120', '71', '20',
       '16', '3', '2', '1', nan, '40', '6', '4', '87', '54', '46', '30',
       '12', '69', '39', '14', '10', '9', '8', '7', '82', '52', '41',
       '15', '98', '43', '19', '111', '48', '36', '25', '67', '60', '84',
       '59', '55', '47', '85', '49', '45', '34', '29', '22', '74', '28',
       '77', '50', '32', '23', '35', '122', '100', '95', '103', '113',
       '53', '96', '27', '64', '58', '33', '31', '73', '56', '86', '62',
       '106', '24', '57', '104', '26', '66', '51', '118', '105', '101',
       '21', '79', '44', '102', '88', '80', '72', '97', '68', '89', '38',
       '83', '63', '148', '166', '90', '76', '37', '70', '124', '94',
       '140', '153', '107', '117', '121', '92', '78', '75', '114', '115',
       '130', '128', '151', '110', '138', '135', '109', '61', '179',
       '170', '112', '116', '91', '143', '93', '123', '145', '81', '150',
       '162', '108', '131', '133', '137', '146', '139', '125', '129',
       '157', '152', '144', '99', '127', '210', '147', '126', '181',
       '160', '180', '208', '176', '168', '141', '132', '119', '154',
       '185', '134', '156', '164', '173', '178', '171', '149', '237',
       '159', '161', '215', '264', '136', '169', '209', '174', '189',
       '183', '163', '219', '158', '175', '177', '200', '194', '142',
       '172', '186', '188', '167'], dtype=object)
```

In [9]:

```
df = df.dropna(subset = ['Innings Runs Scored Num'])
```

In [10]:

```
df['Innings Runs Scored Num'].unique()
```

Out[10]:

```
array(['65', '42', '18', '17', '13', '11', '5', '0', '120', '71', '20',
       '16', '3', '2', '1', '40', '6', '4', '87', '54', '46', '30', '12',
       '69', '39', '14', '10', '9', '8', '7', '82', '52', '41', '15',
       '98', '43', '19', '111', '48', '36', '25', '67', '60', '84', '59',
       '55', '47', '85', '49', '45', '34', '29', '22', '74', '28', '77',
       '50', '32', '23', '35', '122', '100', '95', '103', '113', '53',
       '96', '27', '64', '58', '33', '31', '73', '56', '86', '62', '106',
       '24', '57', '104', '26', '66', '51', '118', '105', '101', '21',
       '79', '44', '102', '88', '80', '72', '97', '68', '89', '38', '83',
       '63', '148', '166', '90', '76', '37', '70', '124', '94', '140',
       '153', '107', '117', '121', '92', '78', '75', '114', '115', '130',
       '128', '151', '110', '138', '135', '109', '61', '179', '170',
       '112', '116', '91', '143', '93', '123', '145', '81', '150', '162',
       '108', '131', '133', '137', '146', '139', '125', '129', '157',
       '152', '144', '99', '127', '210', '147', '126', '181', '160',
       '180', '208', '176', '168', '141', '132', '119', '154', '185',
       '134', '156', '164', '173', '178', '171', '149', '237', '159',
       '161', '215', '264', '136', '169', '209', '174', '189', '183',
       '163', '219', '158', '175', '177', '200', '194', '142', '172',
       '186', '188', '167'], dtype=object)
```

In [11]:

```
df.head(1)
```

Out[11]:

| | Innings Player | Innings Runs Scored | Innings Runs Scored Num | Innings Minutes Batted | Innings Batted Flag | Innings Not Out Flag | Innings Balls Faced | Innings Boundary Fours | Innings Boundary Sixes | Innings Batting Strike Rate |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | E Lewis | 65 | 65 | 128 | 1.0 | 0.0 | 80 | 8 | 1 | 81.25 |

In [12]:

```
# convert to datetime
df['Innings Date'] = pd.to_datetime(df['Innings Date'])
```

In [13]:

```
df['year'] = df['Innings Date'].dt.year
```

In [14]:

```
df.tail(1)
```

Out[14]:

| | Innings Player | Innings Runs Scored | Innings Runs Scored Num | Innings Minutes Batted | Innings Batted Flag | Innings Not Out Flag | Innings Balls Faced | Innings Boundary Fours | Innings Boundary Sixes | In Ba |
|---|---|---|---|---|---|---|---|---|---|---|
| **171941** | RW Marsh | 10* | 10 | 24 | 1.0 | 1.0 | 18 | 2 | 0 | |

In [15]:

```
df['Innings Runs Scored Num'] = df['Innings Runs Scored Num'].astype('int')
```

In [16]:

```
df['Innings Balls Faced'] = df['Innings Balls Faced'].astype('int')
```

In [17]:

```
df['Innings Not Out Flag'] = df['Innings Not Out Flag'].astype('int')
```

In [ ]:

In [ ]:

In [18]:

```
# Sachin 1994 - 2004
# Virat 2009 - 2019
```

In [19]:

```
sachin_df = df[(df.year >= 1994) & (df.year <= 2004)]
```

In [20]:

```
kohli_df = df[(df.year >= 2009) & (df.year <= 2019)]
```

In [21]:

```
sachin_df.head(2)
```

Out[21]:

| | Innings Player | Innings Runs Scored | Innings Runs Scored Num | Innings Minutes Batted | Innings Batted Flag | Innings Not Out Flag | Innings Balls Faced | Innings Boundary Fours | Innings Boundary Sixes | Inn Bat S I |
|---|---|---|---|---|---|---|---|---|---|---|
| **77610** | V Sehwag | 70 | 70 | 85 | 1.0 | 0 | 52 | 9 | 2 | 13 |
| **77611** | Yuvraj Singh | 69 | 69 | 34 | 1.0 | 0 | 32 | 8 | 3 | 21 |

In [22]:

```
kohli_df.head(2)
```

Out[22]:

| | Innings Player | Innings Runs Scored | Innings Runs Scored Num | Innings Minutes Batted | Innings Batted Flag | Innings Not Out Flag | Innings Balls Faced | Innings Boundary Fours | Innings Boundary Sixes | Innings Batting Strike Rate |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | E Lewis | 65 | 65 | 128 | 1.0 | 0 | 80 | 8 | 1 | 81.25 |
| **1** | N Pooran | 42 | 42 | 69 | 1.0 | 0 | 52 | 4 | 1 | 80.76 |

In [23]:

```
# Runs per innings = Total Runs/Total Innings
# SR = 100*(Total Runs/Total Balls)
# 100's = sum(100's)
# 50's = sum(50's)
# Team contribution = Player Runs/Team Runs (ex: Virat 50/ Team Ind 150 => 50/150 : 33%)
```

In [24]:

```
# df.dtypes
```

In [25]:

```
# sachin_df.to_csv('sachin_data.csv')
```

In [26]:

```
# what is the total runs scored by sachin in these time frames?
```

In [27]:

```
# sachin_df.head(20)
```

In [28]:

```
# SR Tendulkar
sdf = sachin_df[sachin_df['Innings Player'] == 'SR Tendulkar']
```

In [29]:

```
sdf.head()
```

Out[29]:

| | Innings Player | Innings Runs Scored | Innings Runs Scored Num | Innings Minutes Batted | Innings Batted Flag | Innings Not Out Flag | Innings Balls Faced | Innings Boundary Fours | Innings Boundary Sixes | In B |
|---|---|---|---|---|---|---|---|---|---|---|
| **77614** | SR Tendulkar | 47 | 47 | 60 | 1.0 | 0 | 42 | 9 | 0 | |
| **77747** | SR Tendulkar | 19 | 19 | 46 | 1.0 | 0 | 32 | 3 | 0 | |
| **78054** | SR Tendulkar | 16 | 16 | 30 | 1.0 | 0 | 17 | 2 | 0 | |
| **79590** | SR Tendulkar | 74 | 74 | 170 | 1.0 | 0 | 100 | 7 | 1 | |
| **79681** | SR Tendulkar | 18 | 18 | 30 | 1.0 | 0 | 21 | 3 | 0 | |

In [30]:

```
sum(sdf['Innings Runs Scored Num'])
```

Out[30]:

```
11818
```

In [31]:

```
kdf = kohli_df[kohli_df['Innings Player'] == 'V Kohli']
```

In [32]:

```
# kohli_df['Innings Player'].unique()
```

In [33]:

```
kdf.head()
```

Out[33]:

| | Innings Player | Innings Runs Scored | Innings Runs Scored Num | Innings Minutes Batted | Innings Batted Flag | Innings Not Out Flag | Innings Balls Faced | Innings Boundary Fours | Innings Boundary Sixes | Inning Battir Strik Ra |
|---|---|---|---|---|---|---|---|---|---|---|
| **11** | V Kohli | 120 | 120 | 179 | 1.0 | 0 | 125 | 14 | 1 | 9 |
| **327** | V Kohli | 1 | 1 | 8 | 1.0 | 0 | 6 | 0 | 0 | 16.6 |
| **420** | V Kohli | 34* | 34 | 61 | 1.0 | 1 | 41 | 3 | 0 | 82.9 |
| **664** | V Kohli | 26 | 26 | 45 | 1.0 | 0 | 27 | 3 | 0 | 96.2 |
| **804** | V Kohli | 66 | 66 | 103 | 1.0 | 0 | 76 | 7 | 0 | 86.8 |

In [34]:

```
sum(kdf['Innings Runs Scored Num'])
```

Out[34]:

```
11247
```

In [35]:

```
len(kdf), len(sdf)
```

Out[35]:

```
(224, 271)
```

In [36]:

```
# RPI - Sachin, Virat
sum(kdf['Innings Runs Scored Num'])/len(kdf), sum(sdf['Innings Runs Scored Num'])/len(sdf)
```

Out[36]:

```
(50.20982142857143, 43.608856088560884)
```

In [37]:

```
# SR
100*sum(kdf['Innings Runs Scored Num'])/sum(kdf['Innings Balls Faced']), 100*sum(sdf['Innin
```

Out[37]:

```
(93.56126778138258, 88.21377920429947)
```

In [38]:

```
# 100's
sum(kdf["100's"]), sum(sdf["100's"])
```

Out[38]:

```
(42.0, 37.0)
```

In [39]:

```
# 50's
sum(kdf["50's"]), sum(sdf["50's"])
```

Out[39]:

```
(53.0, 57.0)
```

In [40]:

```
# Team Contribution - Runs score by each player, Runs by team
sum(kdf['Innings Runs Scored Num']), sum(sdf['Innings Runs Scored Num'])
```

Out[40]:

```
(11247, 11818)
```

In [41]:

```
# 1994 - 2004 = All players
sum(sachin_df[sachin_df.Country == 'India']['Innings Runs Scored Num'])
```

Out[41]:

```
69715
```

In [42]:

```
# 2009 - 2019 = All players
sum(kohli_df[kohli_df.Country == 'India']['Innings Runs Scored Num'])
```

Out[42]:

```
63867
```

In [43]:

```
100*sum(kdf['Innings Runs Scored Num'])/sum(kohli_df[kohli_df.Country == 'India']['Innings
```

Out[43]:

```
17.610033350556627
```

In [44]:

```
100*sum(sdf['Innings Runs Scored Num'])/sum(sachin_df[sachin_df.Country == 'India']['Inning
```

Out[44]:

```
16.951875493078965
```

## Visualizations:

In [45]:

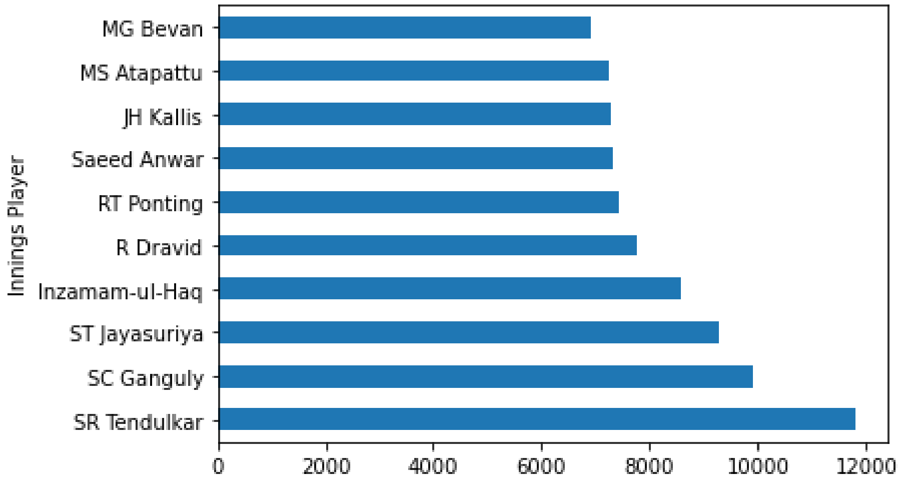
```
sachin_df.groupby(['Innings Player'])['Innings Runs Scored Num'].sum().sort_values(ascendin
```

Out[45]:

```
Innings Player
SR Tendulkar      11818
SC Ganguly        9911
ST Jayasuriya     9297
Inzamam-ul-Haq    8561
R Dravid          7751
RT Ponting        7422
Saeed Anwar       7320
JH Kallis         7267
MS Atapattu       7253
MG Bevan          6912
Name: Innings Runs Scored Num, dtype: int32
```

In [46]:

```
sachin_df.groupby(['Innings Player'])['Innings Runs Scored Num'].sum().sort_values(ascendin
plt.show()
```



In [47]:

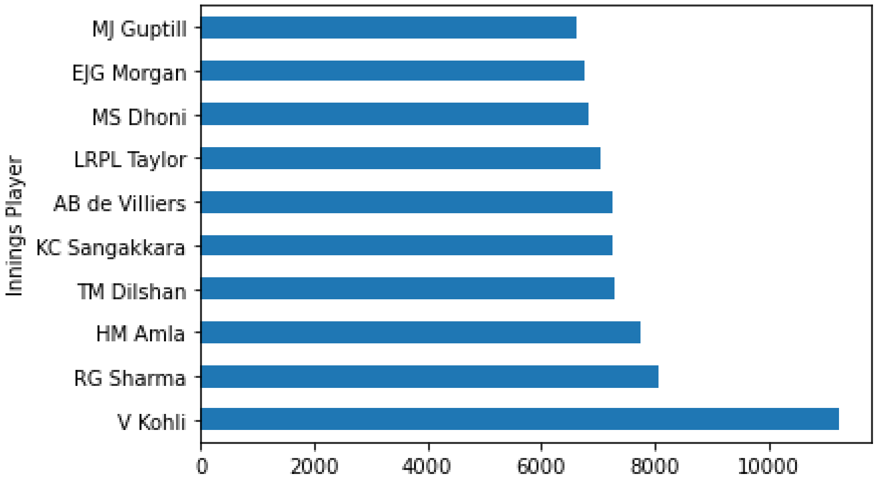
```
kohli_df.groupby(['Innings Player'])['Innings Runs Scored Num'].sum().sort_values(ascending
```

Out[47]:

```
Innings Player
V Kohli            11247
RG Sharma          8083
HM Amla            7745
TM Dilshan         7296
KC Sangakkara      7275
AB de Villiers     7247
LRPL Taylor        7059
MS Dhoni           6838
EJG Morgan         6748
MJ Guptill         6626
Name: Innings Runs Scored Num, dtype: int32
```

In [48]:

```
kohli_df.groupby(['Innings Player'])['Innings Runs Scored Num'].sum().sort_values(ascending
plt.show()
```



In [49]:

```
sdf.head(1)
```

Out[49]:

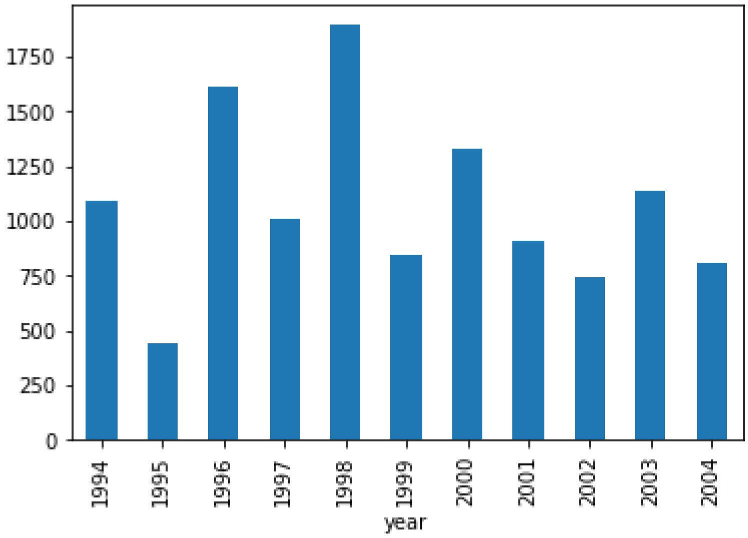
| | Innings Player | Innings Runs Scored | Innings Runs Scored Num | Innings Minutes Batted | Innings Batted Flag | Innings Not Out Flag | Innings Balls Faced | Innings Boundary Fours | Innings Boundary Sixes | In B |
|---|---|---|---|---|---|---|---|---|---|---|
| **77614** | SR Tendulkar | 47 | 47 | 60 | 1.0 | 0 | 42 | 9 | 0 | |

In [50]:

```
sdf.groupby(['year'])['Innings Runs Scored Num'].sum().plot(kind = 'bar')
```

Out[50]:

```
<AxesSubplot:xlabel='year'>
```

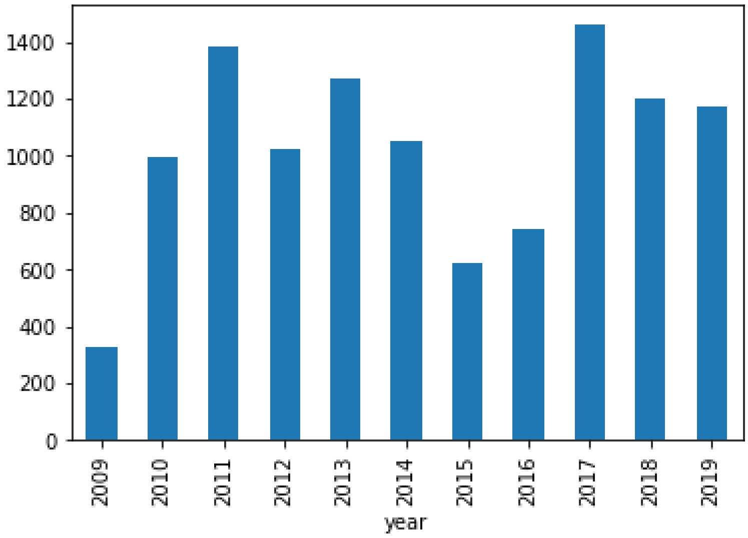


In [51]:

```
kdf.groupby(['year'])['Innings Runs Scored Num'].sum().plot(kind = 'bar')
```

Out[51]:

```
<AxesSubplot:xlabel='year'>
```



## Normalization:

In [52]:

```
# RPI - Sachin, Virat
sum(kdf['Innings Runs Scored Num'])/len(kdf), sum(sdf['Innings Runs Scored Num'])/len(sdf)
```

Out[52]:

```
(50.20982142857143, 43.608856088560884)
```

In [53]:

```
# Kohli_df = player runs with Kohli
# player runs excluding Kohli => not_kohli = kohli_df[kohli_df.player_name != 'V Kohli']
```

In [54]:

```
# RPI - Sachin, Virat
sum(kohli_df['Innings Runs Scored Num'])/len(kohli_df)
```

Out[54]:

```
24.99673202614379
```

In [55]:

```
kohli_df.head(1)
```

Out[55]:

| | Innings Player | Innings Runs Scored | Innings Runs Scored Num | Innings Minutes Batted | Innings Batted Flag | Innings Not Out Flag | Innings Balls Faced | Innings Boundary Fours | Innings Boundary Sixes | Innings Batting Strike Rate |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | E Lewis | 65 | 65 | 128 | 1.0 | 0 | 80 | 8 | 1 | 81.25 |

In [56]:

```
non_kohli_df = kohli_df[kohli_df['Innings Player'] != 'V Kohli']
```

In [57]:

```
non_sachin_df = sachin_df[sachin_df['Innings Player'] != 'SR Tendulkar']
```

In [58]:

```
# Avg = 25 runs
# Kohli = 50
```

In [59]:

```
(sum(kdf['Innings Runs Scored Num'])/len(kdf))/(sum(non_kohli_df['Innings Runs Scored Num']
```

Out[59]:

```
2.029683688052565
```

In [60]:

```
(sum(sdf['Innings Runs Scored Num'])/len(sdf))/(sum(non_sachin_df['Innings Runs Scored Num'
```

Out[60]:

```
1.9437755485945407
```

In [61]:

```
# kohli => other
# SR = V = 93, Other = 80, V/other, S/others
# 100s - Number of matches to score a 100
# 50s - Number of matches to score a 50
# Team contribution - V_cont/O_cont
```

In [62]:

```
200/40, 200/37
```

Out[62]:

```
(5.0, 5.405405405405405)
```

## Strike Rate:

In [63]:

```
# sr of sachin
sum(sdf['Innings Runs Scored Num'])/sum(sdf['Innings Balls Faced'])
```

Out[63]:

```
0.8821377920429947
```

In [64]:

```
# sr of sachin's peers
sum(non_sachin_df['Innings Runs Scored Num'])/sum(non_sachin_df['Innings Balls Faced'])
```

Out[64]:

```
0.7233808936558636
```

In [65]:

```
# sr of kohli
sum(kdf['Innings Runs Scored Num'])/sum(kdf['Innings Balls Faced'])
```

Out[65]:

```
0.9356126778138258
```

In [66]:

```
# sr of kohli's peers
sum(non_kohli_df['Innings Runs Scored Num'])/sum(non_kohli_df['Innings Balls Faced'])
```

Out[66]:

```
0.8342743413330611
```

In [67]:

```
# normalized sachin's value
sachin_sr = sum(sdf['Innings Runs Scored Num'])/sum(sdf['Innings Balls Faced'])
sachin_peer_sr = sum(non_sachin_df['Innings Runs Scored Num'])/sum(non_sachin_df['Innings B
sachin_sr/sachin_peer_sr
```

Out[67]:

```
1.2194651528391862
```

In [68]:

```
# normalized kohli's value
kohli_sr = sum(kdf['Innings Runs Scored Num'])/sum(kdf['Innings Balls Faced'])
kohli_peer_sr = sum(non_kohli_df['Innings Runs Scored Num'])/sum(non_kohli_df['Innings Ball
kohli_sr/kohli_peer_sr
```

Out[68]:

```
1.121468839996732
```

## 100's: Number of matches to score a 100

In [69]:

```
# sachin matches per 100
len(sdf)/sum(sdf["100's"])
```

Out[69]:

```
7.324324324324325
```

In [70]:

```
# sachin peers - matches per 100
len(non_sachin_df)/sum(non_sachin_df["100's"])
```

Out[70]:

```
47.377969762419006
```

In [71]:

```
# kohli matches per 100
len(kdf)/sum(kdf["100's"])
```

Out[71]:

```
5.333333333333333
```

In [72]:

```
# kohli peers - matches per 100
len(non_kohli_df)/sum(non_kohli_df["100's"])
```

Out[72]:

```
29.311827956989248
```

In [73]:

```
# normalized sachin value
sachin_mper_100 = len(sdf)/sum(sdf["100's"])
sachin_peers_mper_100 = len(non_sachin_df)/sum(non_sachin_df["100's"])
sachin_mper_100/sachin_peers_mper_100
```

Out[73]:

```
0.15459346107595562
```

In [74]:

```
# normalized virat value
kohli_mper_100 = len(kdf)/sum(kdf["100's"])
kohli_peers_mper_100 = len(non_kohli_df)/sum(non_kohli_df["100's"])
kohli_mper_100/kohli_peers_mper_100
```

Out[74]:

```
0.18195157740278795
```

## 50's: Number of matches to score a 50

In [75]:

```
# sachin matches per 100
len(sdf)/sum(sdf["50's"])
```

Out[75]:

```
4.754385964912281
```

In [76]:

```
# sachin peers - matches per 100
len(non_sachin_df)/sum(non_sachin_df["50's"])
```

Out[76]:

```
8.33751425313569
```

In [77]:

```python
# kohli matches per 100
len(kdf)/sum(kdf["50's"])
```

Out[77]:

```
4.226415094339623
```

In [78]:

```python
# kohli peers - matches per 100
len(non_kohli_df)/sum(non_kohli_df["50's"])
```

Out[78]:

```
7.673469387755102
```

In [79]:

```python
# normalized sachin value
sachin_mper_50 = len(sdf)/sum(sdf["50's"])
sachin_peers_mper_50 = len(non_sachin_df)/sum(non_sachin_df["50's"])
sachin_mper_50/sachin_peers_mper_50
```

Out[79]:

```
0.5702402203539483
```

In [80]:

```python
# normalized virat value
kohli_mper_50 = len(kdf)/sum(kdf["50's"])
kohli_peers_mper_50 = len(non_kohli_df)/sum(non_kohli_df["50's"])
kohli_mper_50/kohli_peers_mper_50
```

Out[80]:

```
0.5507828181453231
```

## Team Contribution: Here we are already comparing with peers, hence no need of a normalization

In [81]:

```python
# % of team runs by sachin
100*sum(sdf['Innings Runs Scored Num'])/(sum(non_sachin_df[non_sachin_df.Country == 'India'
```

Out[81]:

```
16.951875493078965
```

In [82]:

```
# % of team runs by kohli
100*sum(kdf['Innings Runs Scored Num'])/(sum(non_kohli_df[non_kohli_df.Country == 'India'][
```

Out[82]:

```
17.610033350556627
```