In [1]:
```
import pandas as pd
import pickle
import numpy as np
```

In [2]:
```
df = pd.read_csv("t20_wc.csv")
df.head()
```

Out[2]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	venue
0	2	Australia	Sri Lanka	0.1	0	0	NaN	Melbourne Cricket Ground
1	2	Australia	Sri Lanka	0.2	0	0	NaN	Melbourne Cricket Ground
2	2	Australia	Sri Lanka	0.3	1	0	NaN	Melbourne Cricket Ground
3	2	Australia	Sri Lanka	0.4	2	0	NaN	Melbourne Cricket Ground
4	2	Australia	Sri Lanka	0.5	0	0	NaN	Melbourne Cricket Ground

In [3]:
```
df.shape
```

Out[3]:
```
(63888, 8)
```

In [4]:
```
df.isnull().sum()
```

Out[4]:
```
match_id               0
batting_team           0
bowling_team           0
ball                   0
runs                   0
player_dismissed       0
city                8548
venue                  0
dtype: int64
```

In [5]:
```
df["venue"].mask
```

Out[5]:
```
<bound method Series.mask of 0        Melbourne Cricket Ground
1        Melbourne Cricket Ground
2        Melbourne Cricket Ground
3        Melbourne Cricket Ground
4        Melbourne Cricket Ground
                   ...
63883         R Premadasa Stadium
63884         R Premadasa Stadium
63885         R Premadasa Stadium
63886         R Premadasa Stadium
63887         R Premadasa Stadium
Name: venue, Length: 63888, dtype: object>
```

In [6]:
```
df[df['city'].isnull()]["venue"].value_counts()
```

Out[6]:
```
Dubai International Cricket Stadium        2969
Pallekele International Cricket Stadium    2066
Melbourne Cricket Ground                   1453
Sydney Cricket Ground                       749
Adelaide Oval                               498
Harare Sports Club                          372
Sharjah Cricket Stadium                     249
Sylhet International Cricket Stadium        128
Carrara Oval                                 64
Name: venue, dtype: int64
```

In [7]:
```
df["venue"].str.split().apply(lambda x: x[0])
```

Out[7]:
```
0        Melbourne
1        Melbourne
2        Melbourne
3        Melbourne
4        Melbourne
           ...
63883            R
63884            R
63885            R
63886            R
63887            R
Name: venue, Length: 63888, dtype: object
```

In [8]:
```
df["city"] = df["city"].mask(df['city'].isnull(),df['venue'].str.split().str
```

In [9]:
```
df[df["city"].isnull()]["venue"].value_counts()
```

Out[9]:
```
Series([], Name: venue, dtype: int64)
```

In [10]:
```
df.isnull().sum()
```

Out[10]:
```
match_id            0
batting_team        0
bowling_team        0
ball                0
runs                0
player_dismissed    0
city                0
venue               0
dtype: int64
```

In [11]:
```
df.drop(columns = ["venue"] , inplace = True)
```

In [12]:
```
df
```

Out[12]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne
...	...	...	...	...	...	...	...
63883	964	Sri Lanka	Australia	19.3	1	0	Colombo
63884	964	Sri Lanka	Australia	19.4	0	0	Colombo
63885	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo
63886	964	Sri Lanka	Australia	19.6	2	0	Colombo
63887	964	Sri Lanka	Australia	19.7	1	0	Colombo

63888 rows × 7 columns

Data Filtering

5 innings played

In [13]:
```
(6*20)*5
```

Out[13]:
```
600
```

In [14]:
```
eligible_cities = df["city"].value_counts()
eligible_cities
```

Out[14]:
```
Colombo          4086
Mirpur           3420
Johannesburg     3331
Dubai            2969
Auckland         2532
                 ... 
Nairobi           123
Potchefstroom     122
Dharamsala        122
Ahmedabad         121
Carrara            64
Name: city, Length: 86, dtype: int64
```

In [15]:
```
eligible_cities = eligible_cities[eligible_cities >600].index.tolist()
```

In [16]:

```
eligible_cities
```

Out[16]:

```
['Colombo',
 'Mirpur',
 'Johannesburg',
 'Dubai',
 'Auckland',
 'Cape Town',
 'London',
 'Pallekele',
 'Barbados',
 'Sydney',
 'Melbourne',
 'Durban',
 'St Lucia',
 'Wellington',
 'Lauderhill',
 'Hamilton',
 'Centurion',
 'Manchester',
 'Abu Dhabi',
 'Mumbai',
 'Nottingham',
 'Southampton',
 'Mount Maunganui',
 'Chittagong',
 'Kolkata',
 'Lahore',
 'Delhi',
 'Nagpur',
 'Chandigarh',
 'Adelaide',
 'Bangalore',
 'St Kitts',
 'Cardiff',
 'Christchurch',
 'Trinidad']
```

In [17]:

```
df = df[df["city"].isin(eligible_cities)]
df
```

Out[17]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne
...	...	...	...	...	...	...	...
63883	964	Sri Lanka	Australia	19.3	1	0	Colombo
63884	964	Sri Lanka	Australia	19.4	0	0	Colombo
63885	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo
63886	964	Sri Lanka	Australia	19.6	2	0	Colombo
63887	964	Sri Lanka	Australia	19.7	1	0	Colombo

50501 rows × 7 columns

In [18]:

```
df.groupby("match_id")["runs"].cumsum().iloc[115:150]
```

Out[18]:

```
115    153
116    153
117    154
118    158
119    158
120    160
121    161
122    162
123    164
124    168
248      0
249      0
250      0
251      0
252      1
253      1
254      2
255      8
256      9
257      9
258     13
259     14
260     15
261     15
262     19
263     20
264     21
265     21
266     25
267     26
268     27
269     31
270     31
271     31
272     35
Name: runs, dtype: int64
```

In [19]:

```
df ["current_score"] = df.groupby("match_id")["runs"].cumsum()
```

```
C:\Users\azfer\AppData\Local\Temp\ipykernel_8084\2622827675.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df ["current_score"] = df.groupby("match_id")["runs"].cumsum()
```

In [20]:

```
df
```

Out[20]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	curre
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
63883	964	Sri Lanka	Australia	19.3	1	0	Colombo	
63884	964	Sri Lanka	Australia	19.4	0	0	Colombo	
63885	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo	
63886	964	Sri Lanka	Australia	19.6	2	0	Colombo	
63887	964	Sri Lanka	Australia	19.7	1	0	Colombo	

50501 rows × 8 columns

In [21]:

```
df["over"] = df['ball'].astype(int)
```

```
C:\Users\azfer\AppData\Local\Temp\ipykernel_8084\3842803776.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["over"] = df['ball'].astype(int)
```

In [22]:
```
df
```

Out[22]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	curre
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
63883	964	Sri Lanka	Australia	19.3	1	0	Colombo	
63884	964	Sri Lanka	Australia	19.4	0	0	Colombo	
63885	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo	
63886	964	Sri Lanka	Australia	19.6	2	0	Colombo	
63887	964	Sri Lanka	Australia	19.7	1	0	Colombo	

50501 rows × 9 columns

In [23]:
```
df["ball_no"] = df["ball"].astype(str).str.extract("\d.(\d)").astype(int)
```

```
C:\Users\azfer\AppData\Local\Temp\ipykernel_8084\536658437.py:1: SettingWi
thCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["ball_no"] = df["ball"].astype(str).str.extract("\d.(\d)").astype(in
t)
```

In [24]:
```
df
```

Out[24]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	curre
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
63883	964	Sri Lanka	Australia	19.3	1	0	Colombo	
63884	964	Sri Lanka	Australia	19.4	0	0	Colombo	
63885	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo	
63886	964	Sri Lanka	Australia	19.6	2	0	Colombo	
63887	964	Sri Lanka	Australia	19.7	1	0	Colombo	

50501 rows × 10 columns

In [25]:
```
df["total_deliveries"] = (df["over"]*6)+df['ball_no']
```

```
C:\Users\azfer\AppData\Local\Temp\ipykernel_8084\4049483542.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["total_deliveries"] = (df["over"]*6)+df['ball_no']
```

In [26]:

```
df
```

Out[26]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	curre
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
63883	964	Sri Lanka	Australia	19.3	1	0	Colombo	
63884	964	Sri Lanka	Australia	19.4	0	0	Colombo	
63885	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo	
63886	964	Sri Lanka	Australia	19.6	2	0	Colombo	
63887	964	Sri Lanka	Australia	19.7	1	0	Colombo	

50501 rows × 11 columns

In [27]:

```
df["balls_left"] = 120 - df["total_deliveries"]
```

```
C:\Users\azfer\AppData\Local\Temp\ipykernel_8084\4070253561.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["balls_left"] = 120 - df["total_deliveries"]
```

In [28]:
```
df
```

Out[28]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	curre
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
63883	964	Sri Lanka	Australia	19.3	1	0	Colombo	
63884	964	Sri Lanka	Australia	19.4	0	0	Colombo	
63885	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo	
63886	964	Sri Lanka	Australia	19.6	2	0	Colombo	
63887	964	Sri Lanka	Australia	19.7	1	0	Colombo	

50501 rows × 12 columns

In [29]:
```
df["balls_left"].mask(df["balls_left"] <0 , 0 , inplace = True)
```

```
C:\Users\azfer\AppData\Local\Temp\ipykernel_8084\1644300400.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["balls_left"].mask(df["balls_left"] <0 , 0 , inplace = True)
```

In [30]:
```
df
```

Out[30]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	curre
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
63883	964	Sri Lanka	Australia	19.3	1	0	Colombo	
63884	964	Sri Lanka	Australia	19.4	0	0	Colombo	
63885	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo	
63886	964	Sri Lanka	Australia	19.6	2	0	Colombo	
63887	964	Sri Lanka	Australia	19.7	1	0	Colombo	

50501 rows × 12 columns

In [31]:
```
df["player_dismissed"].unique()
```

Out[31]:
```
array(['0', 'M Klinger', 'AJ Finch', 'MC Henriques', 'TM Head',
       'AJ Turner', 'TD Paine', 'BR Dunk', 'JP Faulkner', 'L Ronchi',
       'KS Williamson', 'CJ Anderson', 'C Munro', 'C de Grandhomme',
       'JDS Neesham', 'MJ Santner', 'TC Bruce', 'Q de Kock',
       'F du Plessis', 'HM Amla', 'AB de Villiers', 'F Behardien',
       'JP Duminy', 'JT Smuts', 'RR Hendricks', 'DA Miller', 'JJ Roy',
       'AD Hales', 'DJ Malan', 'SW Billings', 'LS Livingstone',
       'LE Plunkett', 'JC Buttler', 'DJ Willey', 'V Kohli', 'SK Raina',
       'Yuvraj Singh', 'KL Rahul', 'MK Pandey', 'HH Pandya', 'A Mishra',
       'MS Dhoni', 'J Charles', 'AD Russell', 'E Lewis', 'CR Brathwait
e',
       'KA Pollard', 'LMP Simmons', 'MN Samuels', 'ADS Fletcher',
       'DJ Bravo', 'S Badree', 'N Pooran', 'SP Narine', 'JE Taylor',
       'Sharjeel Khan', 'Babar Azam', 'Khalid Latif', 'Shoaib Malik',
       'CAK Walton', 'MJ Guptill', 'TA Blundell', 'LRPL Taylor',
       'TG Southee', 'IS Sodhi', 'JM Vince', 'MS Chapman', 'TL Seifert',
       'LA Dawson', 'R Powell', 'JO Holder', 'Kamran Akmal',
       'Ahmed Shehzad', 'Fakhar Zaman', 'Sarfraz Ahmed', 'Sohail Tanvi
r',
```

In [32]:
```
df["player_dismissed"] = df["player_dismissed"].mask(df["player_dismissed"]
```

```
C:\Users\azfer\AppData\Local\Temp\ipykernel_8084\2179371311.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["player_dismissed"] = df["player_dismissed"].mask(df["player_dismisse
d"] != "0",1).astype(int)
```

In [33]:
```
df
```

Out[33]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	currei
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
63883	964	Sri Lanka	Australia	19.3	1	0	Colombo	
63884	964	Sri Lanka	Australia	19.4	0	0	Colombo	
63885	964	Sri Lanka	Australia	19.5	0	1	Colombo	
63886	964	Sri Lanka	Australia	19.6	2	0	Colombo	
63887	964	Sri Lanka	Australia	19.7	1	0	Colombo	

50501 rows × 12 columns

In [34]:
```
df["player_dismissed"] = df.groupby("match_id")["player_dismissed"].cumsum()
```

```
C:\Users\azfer\AppData\Local\Temp\ipykernel_8084\3595609500.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["player_dismissed"] = df.groupby("match_id")["player_dismissed"].cums
um()
```

In [35]:
```
df
```

Out[35]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	curre
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
63883	964	Sri Lanka	Australia	19.3	1	8	Colombo	
63884	964	Sri Lanka	Australia	19.4	0	8	Colombo	
63885	964	Sri Lanka	Australia	19.5	0	9	Colombo	
63886	964	Sri Lanka	Australia	19.6	2	9	Colombo	
63887	964	Sri Lanka	Australia	19.7	1	9	Colombo	

50501 rows × 12 columns

In [36]:
```
df["wickets_left"] = 10 - df["player_dismissed"]
```

```
C:\Users\azfer\AppData\Local\Temp\ipykernel_8084\2835093938.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["wickets_left"] = 10 - df["player_dismissed"]
```

In [37]:
```
df
```

Out[37]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	curre
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
63883	964	Sri Lanka	Australia	19.3	1	8	Colombo	
63884	964	Sri Lanka	Australia	19.4	0	8	Colombo	
63885	964	Sri Lanka	Australia	19.5	0	9	Colombo	
63886	964	Sri Lanka	Australia	19.6	2	9	Colombo	
63887	964	Sri Lanka	Australia	19.7	1	9	Colombo	

50501 rows × 13 columns

In [38]:
```
df["crr"] = (df["current_score"]*6)/df["total_deliveries"]
```

```
C:\Users\azfer\AppData\Local\Temp\ipykernel_8084\2462965211.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  df["crr"] = (df["current_score"]*6)/df["total_deliveries"]
```

In [39]:
```
df
```

Out[39]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	curre
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
63883	964	Sri Lanka	Australia	19.3	1	8	Colombo	
63884	964	Sri Lanka	Australia	19.4	0	8	Colombo	
63885	964	Sri Lanka	Australia	19.5	0	9	Colombo	
63886	964	Sri Lanka	Australia	19.6	2	9	Colombo	
63887	964	Sri Lanka	Australia	19.7	1	9	Colombo	

50501 rows × 14 columns

In [40]:
```
total_run_table = df.groupby(["match_id"])["runs"].sum().reset_index()
total_run_table
```

Out[40]:

	match_id	runs
0	2	168
1	4	187
2	10	195
3	11	194
4	12	185
...	...	...
411	958	129
412	960	150
413	961	120
414	963	263
415	964	128

416 rows × 2 columns

In [41]:
```
df
```

Out[41]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	currei
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
63883	964	Sri Lanka	Australia	19.3	1	8	Colombo	
63884	964	Sri Lanka	Australia	19.4	0	8	Colombo	
63885	964	Sri Lanka	Australia	19.5	0	9	Colombo	
63886	964	Sri Lanka	Australia	19.6	2	9	Colombo	
63887	964	Sri Lanka	Australia	19.7	1	9	Colombo	

50501 rows × 14 columns

In [42]:
```
df = df.merge(total_run_table , on = "match_id")
df
```

Out[42]:

	match_id	batting_team	bowling_team	ball	runs_x	player_dismissed	city	cur
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
50496	964	Sri Lanka	Australia	19.3	1	8	Colombo	
50497	964	Sri Lanka	Australia	19.4	0	8	Colombo	
50498	964	Sri Lanka	Australia	19.5	0	9	Colombo	
50499	964	Sri Lanka	Australia	19.6	2	9	Colombo	
50500	964	Sri Lanka	Australia	19.7	1	9	Colombo	

50501 rows × 15 columns

In [43]:
```
df["last_five"] = df.groupby("match_id")["runs_x"].rolling(window = 30).sum(
```

In [44]:
```
df
```

Out[44]:

	match_id	batting_team	bowling_team	ball	runs_x	player_dismissed	city	cur
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	
...	...	...	...	...	...	...	...	
50496	964	Sri Lanka	Australia	19.3	1	8	Colombo	
50497	964	Sri Lanka	Australia	19.4	0	8	Colombo	
50498	964	Sri Lanka	Australia	19.5	0	9	Colombo	
50499	964	Sri Lanka	Australia	19.6	2	9	Colombo	
50500	964	Sri Lanka	Australia	19.7	1	9	Colombo	

50501 rows × 16 columns

In [45]:
```
df.columns
```

Out[45]:
```
Index(['match_id', 'batting_team', 'bowling_team', 'ball', 'runs_x',
       'player_dismissed', 'city', 'current_score', 'over', 'ball_no',
       'total_deliveries', 'balls_left', 'wickets_left', 'crr', 'runs_y',
       'last_five'],
      dtype='object')
```

In [46]:
```
final_df = df[['batting_team', 'bowling_team','current_score','balls_left',
final_df
```

Out[46]:

	batting_team	bowling_team	current_score	balls_left	wickets_left	crr	last_five
0	Australia	Sri Lanka	0	119	10	0.000000	NaN
1	Australia	Sri Lanka	0	118	10	0.000000	NaN
2	Australia	Sri Lanka	1	117	10	2.000000	NaN
3	Australia	Sri Lanka	3	116	10	4.500000	NaN
4	Australia	Sri Lanka	3	115	10	3.600000	NaN
...	...	...	...	...	...	...	...
50496	Sri Lanka	Australia	125	3	2	6.410256	32.0
50497	Sri Lanka	Australia	125	2	2	6.355932	32.0
50498	Sri Lanka	Australia	125	1	1	6.302521	32.0
50499	Sri Lanka	Australia	127	0	1	6.350000	33.0
50500	Sri Lanka	Australia	128	0	1	6.347107	32.0

50501 rows × 8 columns

In [47]:
```
final_df.isnull().sum()
```

Out[47]:
```
batting_team        0
bowling_team        0
current_score       0
balls_left          0
wickets_left        0
crr                 0
last_five       12024
runs_y              0
dtype: int64
```

In [48]:
```
final_df.dropna(inplace = True)
```

```
C:\Users\azfer\AppData\Local\Temp\ipykernel_8084\1587496580.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  final_df.dropna(inplace = True)
```

In [49]:
```
final_df
```

Out[49]:

	batting_team	bowling_team	current_score	balls_left	wickets_left	crr	last_five
29	Australia	Sri Lanka	43	90	10	8.600000	43.0
30	Australia	Sri Lanka	44	89	10	8.516129	44.0
31	Australia	Sri Lanka	45	88	10	8.437500	45.0
32	Australia	Sri Lanka	45	87	10	8.181818	44.0
33	Australia	Sri Lanka	45	86	10	7.941176	42.0
...	...	...	...	...	...	...	...
50496	Sri Lanka	Australia	125	3	2	6.410256	32.0
50497	Sri Lanka	Australia	125	2	2	6.355932	32.0
50498	Sri Lanka	Australia	125	1	1	6.302521	32.0
50499	Sri Lanka	Australia	127	0	1	6.350000	33.0
50500	Sri Lanka	Australia	128	0	1	6.347107	32.0

38477 rows × 8 columns

In [50]:
```
final_df.sample(10)
```

Out[50]:

	batting_team	bowling_team	current_score	balls_left	wickets_left	crr	last_five
43136	South Africa	Australia	75	69	10	8.823529	53.0
44416	Australia	South Africa	115	26	3	7.340426	24.0
6042	India	Bangladesh	68	64	10	7.285714	39.0
24281	South Africa	Afghanistan	84	48	7	7.000000	28.0
20935	South Africa	West Indies	55	82	9	8.684211	53.0
26898	Pakistan	New Zealand	131	21	5	7.939394	38.0
14032	South Africa	England	151	16	7	8.711538	66.0
36741	West Indies	Bangladesh	69	68	8	7.961538	43.0
36577	Sri Lanka	Australia	111	23	6	6.865979	46.0
27582	Australia	England	41	83	10	6.648649	37.0

In [51]:
```
final_df.sample(final_df.shape[0])
```

Out[51]:

	batting_team	bowling_team	current_score	balls_left	wickets_left	crr	last_five
22863	South Africa	England	239	1	4	12.050420	45.0
5210	India	South Africa	152	15	6	8.685714	47.0
16528	South Africa	England	151	3	3	7.743590	58.0
16651	Pakistan	Sri Lanka	179	4	4	9.258621	50.0
21671	Pakistan	South Africa	51	77	8	7.116279	23.0
...	...	...	...	...	...	...	...
34190	Pakistan	Bangladesh	61	68	9	7.038462	27.0
38762	New Zealand	Bangladesh	78	69	9	9.176471	40.0
46548	South Africa	Sri Lanka	63	66	6	7.000000	43.0
16506	South Africa	England	105	23	4	6.494845	24.0
36231	Pakistan	Australia	150	1	6	7.563025	47.0

38477 rows × 8 columns

In [52]:
```
final_df
```

Out[52]:

	batting_team	bowling_team	current_score	balls_left	wickets_left	crr	last_five
29	Australia	Sri Lanka	43	90	10	8.600000	43.0
30	Australia	Sri Lanka	44	89	10	8.516129	44.0
31	Australia	Sri Lanka	45	88	10	8.437500	45.0
32	Australia	Sri Lanka	45	87	10	8.181818	44.0
33	Australia	Sri Lanka	45	86	10	7.941176	42.0
...	...	...	...	...	...	...	...
50496	Sri Lanka	Australia	125	3	2	6.410256	32.0
50497	Sri Lanka	Australia	125	2	2	6.355932	32.0
50498	Sri Lanka	Australia	125	1	1	6.302521	32.0
50499	Sri Lanka	Australia	127	0	1	6.350000	33.0
50500	Sri Lanka	Australia	128	0	1	6.347107	32.0

38477 rows × 8 columns

In [53]:
```
X = final_df.drop(columns = ['runs_y'])
y = final_df['runs_y']
```

In [54]:
```
from sklearn.model_selection import train_test_split
```

In [55]:
```
X_train , X_test , y_train , y_test = train_test_split(X,y,test_size = 0.2,r
```

In [56]:
```
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.metrics import r2_score,mean_absolute_error
```

In [57]:
```
trf = ColumnTransformer([
    ('trf' , OneHotEncoder(sparse = False),['batting_team' , 'bowling_team']
], remainder = 'passthrough')
```

In [58]:
```
pipe = Pipeline(steps = [
    ('step1' , trf),
    ('step2' , StandardScaler()),
    ('step3' , XGBRegressor(n_estimators = 1000 , learning_rate = 0.2 , max_
])
```

In [59]:

```
pipe.fit(X_train , y_train)
y_pred = pipe.predict(X_test)
print(r2_score(y_test,y_pred))
print(mean_absolute_error(y_test,y_pred))
```

```
C:\Users\azfer\anaconda3\Lib\site-packages\sklearn\preprocessing\_encoder
s.py:972: FutureWarning: `sparse` was renamed to `sparse_output` in versio
n 1.2 and will be removed in 1.4. `sparse_output` is ignored unless you le
ave `sparse` to its default value.
  warnings.warn(

0.9456152685147481
3.852135229011583
```

In [60]:

```
pickle.dump(pipe, open ('pipe.pkl' , 'wb'))
```

In [66]:

```
final_df['batting_team'].unique()
```

Out[66]:

```
array(['Australia', 'New Zealand', 'South Africa', 'England', 'India',
       'West Indies', 'Pakistan', 'Bangladesh', 'Afghanistan',
       'Sri Lanka'], dtype=object)
```

In [68]:

```
df['city'].unique().tolist()
```

Out[68]:

```
['Melbourne',
 'Adelaide',
 'Mount Maunganui',
 'Auckland',
 'Southampton',
 'Cardiff',
 'Nagpur',
 'Bangalore',
 'Lauderhill',
 'Dubai',
 'Abu Dhabi',
 'Sydney',
 'Wellington',
 'Hamilton',
 'Barbados',
 'Trinidad',
 'Colombo',
 'St Kitts',
 'Manchester',
 'Delhi',
 'Lahore',
 'Johannesburg',
 'Centurion',
 'Cape Town',
 'Mumbai',
 'Kolkata',
 'Durban',
 'Chandigarh',
 'Christchurch',
 'London',
 'Nottingham',
 'St Lucia',
 'Pallekele',
 'Mirpur',
 'Chittagong']
```

```
In [ ]:
```