DDS_Project_Data_Extraction

November 26, 2023

[ ]:
```
import pandas as pd
file_path = '/content/all_matches.csv'
data = pd.read_csv(file_path)
```

[ ]:
```
data
```

[ ]:
```
       match_id   season  start_date                         venue  innings  \
0        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
1        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
2        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
3        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
4        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
…           …        …           …                            …      …
35256   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35257   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35258   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35259   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35260   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2

       ball batting_team bowling_team           striker     non_striker  …  \
0       0.1  Afghanistan   Bangladesh  Mohammad Shahzad  Najeeb Tarakai  …
1       0.2  Afghanistan   Bangladesh      Gulbadin Naib  Najeeb Tarakai  …
2       0.3  Afghanistan   Bangladesh      Gulbadin Naib  Najeeb Tarakai  …
3       0.4  Afghanistan   Bangladesh      Gulbadin Naib  Najeeb Tarakai  …
4       0.5  Afghanistan   Bangladesh      Gulbadin Naib  Najeeb Tarakai  …
…      …            …          …                …              …  …
35256  18.2      England     Pakistan            MM Ali        BA Stokes  …
35257  18.3      England     Pakistan     LS Livingstone       BA Stokes  …
35258  18.4      England     Pakistan         BA Stokes  LS Livingstone  …
35259  18.5      England     Pakistan         BA Stokes  LS Livingstone  …
35260  18.6      England     Pakistan         BA Stokes  LS Livingstone  …

      extras  wides  noballs  byes  legbyes  penalty  wicket_type  \
0          0    NaN      NaN   NaN      NaN      NaN       caught
1          1    1.0      NaN   NaN      NaN      NaN          NaN
2          0    NaN      NaN   NaN      NaN      NaN          NaN
3          0    NaN      NaN   NaN      NaN      NaN          NaN
```

```
4          0    NaN      NaN   NaN      NaN      NaN          NaN
…      …    …      …   …      …      …          …
35256      0    NaN      NaN   NaN      NaN      NaN       bowled
35257      0    NaN      NaN   NaN      NaN      NaN          NaN
35258      0    NaN      NaN   NaN      NaN      NaN          NaN
35259      0    NaN      NaN   NaN      NaN      NaN          NaN
35260      0    NaN      NaN   NaN      NaN      NaN          NaN

       player_dismissed other_wicket_type other_player_dismissed
0      Mohammad Shahzad               NaN                    NaN
1                   NaN               NaN                    NaN
2                   NaN               NaN                    NaN
3                   NaN               NaN                    NaN
4                   NaN               NaN                    NaN
…                 …               …                    …
35256            MM Ali               NaN                    NaN
35257               NaN               NaN                    NaN
35258               NaN               NaN                    NaN
35259               NaN               NaN                    NaN
35260               NaN               NaN                    NaN

[35261 rows x 22 columns]
```

[ ]:
```
data['venue'].unique()
data['venue_id'], unique_values = pd.factorize(data['venue'])
```

[ ]:
```
data['venue'], data['venue_id']
```

[ ]:
```
(0        Shere Bangla National Stadium
 1        Shere Bangla National Stadium
 2        Shere Bangla National Stadium
 3        Shere Bangla National Stadium
 4        Shere Bangla National Stadium
                       …
 35256         Melbourne Cricket Ground
 35257         Melbourne Cricket Ground
 35258         Melbourne Cricket Ground
 35259         Melbourne Cricket Ground
 35260         Melbourne Cricket Ground
 Name: venue, Length: 35261, dtype: object,
 0         0
 1         0
 2         0
 3         0
 4         0
          ..
 35256    18
 35257    18
 35258    18
 35259    18
 35260    18
 Name: venue_id, Length: 35261, dtype: int64)
```

[ ]:

```
batting_team_players = data.groupby('batting_team')['striker', 'non_striker'].
 ↪agg(lambda x: list(set(x))).reset_index()
bowling_team_players = data.groupby('bowling_team')['bowler'].agg(lambda x:␣
 ↪list(set(x))).reset_index()
country = {}

for index, row in batting_team_players.iterrows():
    if row['batting_team'] in country:
        country[row['batting_team']].append(row['striker'])
    else:
        country[row['batting_team']] = [row['striker']]

for index, row in bowling_team_players.iterrows():
    if row['bowling_team'] in country:
        country[row['bowling_team']].append(row['bowler'])
    else:
        country[row['bowling_team']] = [row['bowler']]
```

[ ]:

```
country
```

[ ]:

```
unique_values_dict = {}
for key, value in country.items():
  unique_values_dict[key] = list(set(value[0] + value[1]))
```

[ ]:

```
unique_values_dict
```

[ ]:

```
import csv
with open('players_list.csv', 'w', newline='') as f:
    writer = csv.writer(f)
    writer.writerow(['Key', 'Values'])
    for key, values in unique_values_dict.items():
        for value in values:
            writer.writerow([key, value])
```

[ ]:

```
data[['over_no','bowl_no']] = data['ball'].astype(str).str.split('.',␣
 ↪expand=True)
data['catcher_id'] = None
data['wicket'] = data['wicket_type'].notna()
print(data)
```

```
      match_id   season  start_date                          venue  innings  \
0        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
1        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
2        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
3        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
4        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
…         …      …         …                            …      …
35256   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35257   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35258   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35259   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35260   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2

       ball batting_team bowling_team           striker     non_striker  …  \
0       0.1  Afghanistan   Bangladesh  Mohammad Shahzad  Najeeb Tarakai  …
1       0.2  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
2       0.3  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
3       0.4  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
4       0.5  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
…     …          …           …               …             …  …
35256  18.2      England     Pakistan            MM Ali       BA Stokes  …
35257  18.3      England     Pakistan    LS Livingstone       BA Stokes  …
35258  18.4      England     Pakistan         BA Stokes  LS Livingstone  …
35259  18.5      England     Pakistan         BA Stokes  LS Livingstone  …
35260  18.6      England     Pakistan         BA Stokes  LS Livingstone  …

      legbyes  penalty  wicket_type  player_dismissed  other_wicket_type  \
0         NaN      NaN       caught  Mohammad Shahzad                NaN
1         NaN      NaN          NaN               NaN                NaN
2         NaN      NaN          NaN               NaN                NaN
3         NaN      NaN          NaN               NaN                NaN
4         NaN      NaN          NaN               NaN                NaN
…       …      …          …               …                …
35256     NaN      NaN       bowled            MM Ali                NaN
35257     NaN      NaN          NaN               NaN                NaN
35258     NaN      NaN          NaN               NaN                NaN
35259     NaN      NaN          NaN               NaN                NaN
35260     NaN      NaN          NaN               NaN                NaN

       other_player_dismissed  over_no  bowl_no catcher_id wicket
0                         NaN        0        1       None   True
1                         NaN        0        2       None  False
2                         NaN        0        3       None  False
3                         NaN        0        4       None  False
4                         NaN        0        5       None  False
…                       …      …      …        …    …
35256                     NaN       18        2       None   True
35257                     NaN       18        3       None  False
35258                     NaN       18        4       None  False
```

```
35259                     NaN       18        5       None  False
35260                     NaN       18        6       None  False

[35261 rows x 26 columns]
```

```
[ ]: data2 = pd.read_csv('/content/team.csv')
```

```
[ ]: data2
```

```
[ ]:     team_id             team_name               coach  home_ground  \
     0         1           Afghanistan         Graham Reid            1
     1         2             Australia     Andrew McDonald            2
     2         3            Bangladesh     Russell Domingo            3
     3         4               England        Matthew Mott            4
     4         5             Hong Kong     Simon Muggleton            5
     5         6                 India        Rahul Dravid            6
     6         7               Ireland         Graham Ford            7
     7         8               Namibia     Pierre de Bruyn            8
     8         9                 Nepal  Pubudu Dassanayake            9
     9        10           Netherlands       Ryan Campbell           10
     10       11           New Zealand          Gary Stead           11
     11       12                  Oman       Duleep Mendis           12
     12       13              Pakistan     Saqlain Mushtaq           13
     13       14      Papua New Guinea     Simon Lavington           14
     14       15              Scotland        Shane Burger           15
     15       16          South Africa        Mark Boucher           16
     16       17             Sri Lanka    Chris Silverwood           17
     17       18  United Arab Emirates         Robin Singh           18
     18       19           West Indies        Phil Simmons           19
     19       20              Zimbabwe        Dave Houghton           20

                                        owner
     0              Afghanistan Cricket Board
     1                      Cricket Australia
     2               Bangladesh Cricket Board
     3        England and Wales Cricket Board
     4          Hong Kong Cricket Association
     5  Board of Control for Cricket in India
     6                        Cricket Ireland
     7                        Cricket Namibia
     8           Cricket Association of Nepal
     9    Koninklijke Nederlandse Cricket Bond
     10                  New Zealand Cricket
     11                         Oman Cricket
     12                Pakistan Cricket Board
     13              Cricket Papua New Guinea
     14                      Cricket Scotland
```

```
15                    Cricket South Africa
16                       Sri Lanka Cricket
17                   Emirates Cricket Board
18                     Cricket West Indies
19                        Zimbabwe Cricket
```

[ ]:
```
data2 = data2[['team_id','team_name']]
```

[ ]:
```
data2 = data2.rename(columns={'team_id':'batting_team_id'})
```

[ ]:
```
data2
```

[ ]:
```
    batting_team_id             team_name
0                 1           Afghanistan
1                 2             Australia
2                 3            Bangladesh
3                 4               England
4                 5             Hong Kong
5                 6                 India
6                 7               Ireland
7                 8               Namibia
8                 9                 Nepal
9                10           Netherlands
10               11           New Zealand
11               12                  Oman
12               13              Pakistan
13               14      Papua New Guinea
14               15              Scotland
15               16          South Africa
16               17             Sri Lanka
17               18  United Arab Emirates
18               19           West Indies
19               20              Zimbabwe
```

[ ]:
```
data = data.reset_index().merge(data2, how='left', left_on='batting_team',␣
 ↪right_on='team_name').set_index('index')
print(data)
```

```
       match_id   season  start_date                          venue  innings  \
index
0        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
1        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
2        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
3        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
4        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
…           …       …          …                             …       …
35256   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
```

```
35257   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35258   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35259   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35260   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2

       ball batting_team bowling_team           striker     non_striker  …  \
index                                                                    …
0       0.1  Afghanistan   Bangladesh  Mohammad Shahzad  Najeeb Tarakai  …
1       0.2  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
2       0.3  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
3       0.4  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
4       0.5  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
…     …          …          …               …             …  …
35256  18.2      England     Pakistan            MM Ali       BA Stokes  …
35257  18.3      England     Pakistan    LS Livingstone       BA Stokes  …
35258  18.4      England     Pakistan         BA Stokes  LS Livingstone  …
35259  18.5      England     Pakistan         BA Stokes  LS Livingstone  …
35260  18.6      England     Pakistan         BA Stokes  LS Livingstone  …

      wicket_type  player_dismissed  other_wicket_type  \
index
0          caught  Mohammad Shahzad                NaN
1             NaN               NaN                NaN
2             NaN               NaN                NaN
3             NaN               NaN                NaN
4             NaN               NaN                NaN
…           …               …                 …
35256      bowled            MM Ali                NaN
35257         NaN               NaN                NaN
35258         NaN               NaN                NaN
35259         NaN               NaN                NaN
35260         NaN               NaN                NaN

       other_player_dismissed  over_no  bowl_no  catcher_id  wicket  \
index
0                        NaN        0        1        None    True
1                        NaN        0        2        None   False
2                        NaN        0        3        None   False
3                        NaN        0        4        None   False
4                        NaN        0        5        None   False
…                       …      …      …         …     …
35256                    NaN       18        2        None    True
35257                    NaN       18        3        None   False
35258                    NaN       18        4        None   False
35259                    NaN       18        5        None   False
35260                    NaN       18        6        None   False

      batting_team_id    team_name
```

```
index
0                    1  Afghanistan
1                    1  Afghanistan
2                    1  Afghanistan
3                    1  Afghanistan
4                    1  Afghanistan
…                 …          …
35256                4      England
35257                4      England
35258                4      England
35259                4      England
35260                4      England

[35261 rows x 28 columns]
```

[ ]:

```
data2 = data2.rename(columns={'batting_team_id':'bowling_team_id'})
```

[ ]:

```
data2
```

[ ]:

```
    bowling_team_id             team_name
0                 1           Afghanistan
1                 2             Australia
2                 3            Bangladesh
3                 4               England
4                 5             Hong Kong
5                 6                 India
6                 7               Ireland
7                 8               Namibia
8                 9                 Nepal
9                10           Netherlands
10               11           New Zealand
11               12                  Oman
12               13              Pakistan
13               14      Papua New Guinea
14               15              Scotland
15               16          South Africa
16               17             Sri Lanka
17               18  United Arab Emirates
18               19           West Indies
19               20              Zimbabwe
```

[ ]:

```
data = data.reset_index().merge(data2, how='left', left_on='bowling_team',␣
 ↪right_on='team_name').set_index('index')
print(data)
```

```
       match_id   season  start_date                        venue  innings  \
index
0        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
```

```
1        682897  2013/14  2014-03-16  Shere Bangla National Stadium       1
2        682897  2013/14  2014-03-16  Shere Bangla National Stadium       1
3        682897  2013/14  2014-03-16  Shere Bangla National Stadium       1
4        682897  2013/14  2014-03-16  Shere Bangla National Stadium       1
…         …      …         …                            …      …
35256   1298179  2022/23  2022-11-13       Melbourne Cricket Ground       2
35257   1298179  2022/23  2022-11-13       Melbourne Cricket Ground       2
35258   1298179  2022/23  2022-11-13       Melbourne Cricket Ground       2
35259   1298179  2022/23  2022-11-13       Melbourne Cricket Ground       2
35260   1298179  2022/23  2022-11-13       Melbourne Cricket Ground       2

       ball batting_team bowling_team           striker     non_striker  …  \
index                                                                    …
0       0.1  Afghanistan   Bangladesh  Mohammad Shahzad  Najeeb Tarakai  …
1       0.2  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
2       0.3  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
3       0.4  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
4       0.5  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
…     …          …          …               …             …  …
35256  18.2      England     Pakistan           MM Ali       BA Stokes  …
35257  18.3      England     Pakistan     LS Livingstone       BA Stokes  …
35258  18.4      England     Pakistan         BA Stokes  LS Livingstone  …
35259  18.5      England     Pakistan         BA Stokes  LS Livingstone  …
35260  18.6      England     Pakistan         BA Stokes  LS Livingstone  …

      other_wicket_type  other_player_dismissed  over_no  bowl_no  catcher_id  \
index
0                   NaN                     NaN        0        1        None
1                   NaN                     NaN        0        2        None
2                   NaN                     NaN        0        3        None
3                   NaN                     NaN        0        4        None
4                   NaN                     NaN        0        5        None
…                 …                     …      …      …         …
35256               NaN                     NaN       18        2        None
35257               NaN                     NaN       18        3        None
35258               NaN                     NaN       18        4        None
35259               NaN                     NaN       18        5        None
35260               NaN                     NaN       18        6        None

       wicket  batting_team_id  team_name_x bowling_team_id team_name_y
index
0        True                1  Afghanistan               3  Bangladesh
1       False                1  Afghanistan               3  Bangladesh
2       False                1  Afghanistan               3  Bangladesh
3       False                1  Afghanistan               3  Bangladesh
4       False                1  Afghanistan               3  Bangladesh
…       …              …          …             …         …
35256    True                4      England              13    Pakistan
```

```
35257   False              4      England              13     Pakistan
35258   False              4      England              13     Pakistan
35259   False              4      England              13     Pakistan
35260   False              4      England              13     Pakistan

[35261 rows x 30 columns]
```

[ ]:
```
data3 = pd.read_csv('/content/players.csv')
```

[ ]:
```
data3 = data3[['player_id','player_name','team_id']]
```

[ ]:
```
data3 = data3.rename(columns={'player_id':'batsman_id'})
```

[ ]:
```
data3
```

[ ]:
```
     batsman_id         player_name  team_id
0             1          Aftab Alam        1
1             2          Amir Hamza        1
2             3    Asghar Stanikzai        1
3             4  Azmatullah Omarzai        1
4             5      Darwish Rasooli       1
..          …                 …      …
505         506       T Panyangara       20
506         507         TL Chatara       20
507         508          V Sibanda       20
508         509        W Madhevere       20
509         510       WP Masakadza       20

[510 rows x 3 columns]
```

[ ]:
```
data = data.reset_index().merge(data3, how='left',␣
 ↪left_on=['striker','batting_team_id'], right_on=['player_name','team_id']).
 ↪set_index('index')
print(data)
```

```
       match_id   season  start_date                          venue  innings  \
index
0        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
1        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
2        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
3        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
4        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
…          …       …        …                             …      …
35256   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35257   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35258   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35259   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35260   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
```

```
       ball batting_team bowling_team           striker     non_striker  …  \
index                                                                    …
0       0.1 Afghanistan   Bangladesh  Mohammad Shahzad  Najeeb Tarakai  …
1       0.2 Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
2       0.3 Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
3       0.4 Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
4       0.5 Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
…     …          …          …               …             …  …
35256  18.2      England     Pakistan            MM Ali         BA Stokes  …
35257  18.3      England     Pakistan    LS Livingstone         BA Stokes  …
35258  18.4      England     Pakistan         BA Stokes  LS Livingstone  …
35259  18.5      England     Pakistan         BA Stokes  LS Livingstone  …
35260  18.6      England     Pakistan         BA Stokes  LS Livingstone  …

      bowl_no  catcher_id  wicket  batting_team_id  team_name_x  \
index
0            1        None    True                1  Afghanistan
1            2        None   False                1  Afghanistan
2            3        None   False                1  Afghanistan
3            4        None   False                1  Afghanistan
4            5        None   False                1  Afghanistan
…       …         …     …              …          …
35256        2        None    True                4      England
35257        3        None   False                4      England
35258        4        None   False                4      England
35259        5        None   False                4      England
35260        6        None   False                4      England

       bowling_team_id  team_name_y  batsman_id       player_name team_id
index
0                    3   Bangladesh         17  Mohammad Shahzad       1
1                    3   Bangladesh          9     Gulbadin Naib       1
2                    3   Bangladesh          9     Gulbadin Naib       1
3                    3   Bangladesh          9     Gulbadin Naib       1
4                    3   Bangladesh          9     Gulbadin Naib       1
…                …          …         …               …     …
35256               13     Pakistan        119            MM Ali       4
35257               13     Pakistan        116    LS Livingstone       4
35258               13     Pakistan        101         BA Stokes       4
35259               13     Pakistan        101         BA Stokes       4
35260               13     Pakistan        101         BA Stokes       4

[35261 rows x 33 columns]
```

```
[ ]: data3 = data3.rename(columns={'batsman_id':'non_striker_id'})
```

[ ]:
```
data3
```

[ ]:
```
     non_striker_id           player_name  team_id
0                 1           Aftab Alam        1
1                 2           Amir Hamza        1
2                 3     Asghar Stanikzai        1
3                 4   Azmatullah Omarzai        1
4                 5      Darwish Rasooli        1
..              …                  …        …
505             506         T Panyangara       20
506             507          TL Chatara       20
507             508            V Sibanda       20
508             509          W Madhevere       20
509             510         WP Masakadza       20

[510 rows x 3 columns]
```

[ ]:
```
data = data.reset_index().merge(data3, how='left',␣
 ↪left_on=['non_striker','batting_team_id'],␣
 ↪right_on=['player_name','team_id']).set_index('index')
print(data)
```

```
       match_id   season  start_date                          venue  innings  \
index
0        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
1        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
2        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
3        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
4        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
…           …      …         …                              …      …
35256   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35257   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35258   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35259   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35260   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2

       ball batting_team bowling_team          striker     non_striker  …  \
index                                                                   …
0       0.1  Afghanistan   Bangladesh  Mohammad Shahzad  Najeeb Tarakai  …
1       0.2  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
2       0.3  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
3       0.4  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
4       0.5  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
…       …          …          …              …            …  …
35256  18.2      England     Pakistan           MM Ali       BA Stokes  …
35257  18.3      England     Pakistan    LS Livingstone       BA Stokes  …
35258  18.4      England     Pakistan        BA Stokes  LS Livingstone  …
```

```
35259  18.5      England     Pakistan         BA Stokes  LS Livingstone  …
35260  18.6      England     Pakistan         BA Stokes  LS Livingstone  …

      batting_team_id  team_name_x  bowling_team_id  team_name_y  batsman_id  \
index
0                    1  Afghanistan                3   Bangladesh          17
1                    1  Afghanistan                3   Bangladesh           9
2                    1  Afghanistan                3   Bangladesh           9
3                    1  Afghanistan                3   Bangladesh           9
4                    1  Afghanistan                3   Bangladesh           9
…                …          …              …          …         …
35256               4      England               13     Pakistan         119
35257               4      England               13     Pakistan         116
35258               4      England               13     Pakistan         101
35259               4      England               13     Pakistan         101
35260               4      England               13     Pakistan         101

          player_name_x  team_id_x  non_striker_id   player_name_y team_id_y
index
0      Mohammad Shahzad          1             19  Najeeb Tarakai          1
1         Gulbadin Naib          1             19  Najeeb Tarakai          1
2         Gulbadin Naib          1             19  Najeeb Tarakai          1
3         Gulbadin Naib          1             19  Najeeb Tarakai          1
4         Gulbadin Naib          1             19  Najeeb Tarakai          1
…                  …        …             …              …       …
35256            MM Ali          4            101       BA Stokes          4
35257    LS Livingstone          4            101       BA Stokes          4
35258         BA Stokes          4            116  LS Livingstone          4
35259         BA Stokes          4            116  LS Livingstone          4
35260         BA Stokes          4            116  LS Livingstone          4

[35261 rows x 36 columns]
```

[ ]:

```
data3 = data3.rename(columns={'non_striker_id':'bowler_id'})
```

[ ]:

```
data = data.reset_index().merge(data3, how='left',␣
 ↪left_on=['bowler','bowling_team_id'], right_on=['player_name','team_id']).
 ↪set_index('index')
print(data)
```

```
       match_id   season  start_date                          venue  innings  \
index
0        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
1        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
2        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
3        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
4        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
…         …      …          …                             …      …
```

```
35256   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35257   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35258   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35259   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35260   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2

       ball batting_team bowling_team           striker     non_striker  …  \
index                                                                    …
0       0.1  Afghanistan   Bangladesh  Mohammad Shahzad  Najeeb Tarakai  …
1       0.2  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
2       0.3  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
3       0.4  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
4       0.5  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
…     …          …          …               …             …  …
35256  18.2      England     Pakistan           MM Ali       BA Stokes  …
35257  18.3      England     Pakistan    LS Livingstone       BA Stokes  …
35258  18.4      England     Pakistan         BA Stokes  LS Livingstone  …
35259  18.5      England     Pakistan         BA Stokes  LS Livingstone  …
35260  18.6      England     Pakistan         BA Stokes  LS Livingstone  …

      team_name_y  batsman_id     player_name_x  team_id_x  non_striker_id  \
index
0      Bangladesh          17  Mohammad Shahzad          1              19
1      Bangladesh           9     Gulbadin Naib          1              19
2      Bangladesh           9     Gulbadin Naib          1              19
3      Bangladesh           9     Gulbadin Naib          1              19
4      Bangladesh           9     Gulbadin Naib          1              19
…           …         …               …        …             …
35256    Pakistan         119            MM Ali          4             101
35257    Pakistan         116    LS Livingstone          4             101
35258    Pakistan         101         BA Stokes          4             116
35259    Pakistan         101         BA Stokes          4             116
35260    Pakistan         101         BA Stokes          4             116

        player_name_y  team_id_y  bowler_id      player_name team_id
index
0      Najeeb Tarakai          1         72  Mashrafe Mortaza       3
1      Najeeb Tarakai          1         72  Mashrafe Mortaza       3
2      Najeeb Tarakai          1         72  Mashrafe Mortaza       3
3      Najeeb Tarakai          1         72  Mashrafe Mortaza       3
4      Najeeb Tarakai          1         72  Mashrafe Mortaza       3
…               …        …        …               …     …
35256       BA Stokes          4        311    Mohammad Wasim      13
35257       BA Stokes          4        311    Mohammad Wasim      13
35258  LS Livingstone          4        311    Mohammad Wasim      13
35259  LS Livingstone          4        311    Mohammad Wasim      13
35260  LS Livingstone          4        311    Mohammad Wasim      13
```

```
[35261 rows x 39 columns]
```

[ ]:

```
data3 = data3.rename(columns={'bowler_id':'player_out_id'})
```

[ ]:

```
data = data.reset_index().merge(data3, how='left',␣
 ↪left_on=['player_dismissed','batting_team_id'],␣
 ↪right_on=['player_name','team_id']).set_index('index')
print(data)
```

```
       match_id   season  start_date                          venue  innings  \
index
0        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
1        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
2        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
3        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
4        682897  2013/14  2014-03-16  Shere Bangla National Stadium        1
…         …      …         …                            …      …
35256   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35257   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35258   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35259   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2
35260   1298179  2022/23  2022-11-13       Melbourne Cricket Ground        2

       ball batting_team bowling_team           striker     non_striker  …  \
index                                                                    …
0       0.1  Afghanistan   Bangladesh  Mohammad Shahzad  Najeeb Tarakai  …
1       0.2  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
2       0.3  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
3       0.4  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
4       0.5  Afghanistan   Bangladesh     Gulbadin Naib  Najeeb Tarakai  …
…     …          …          …               …             …  …
35256  18.2      England     Pakistan            MM Ali       BA Stokes  …
35257  18.3      England     Pakistan    LS Livingstone       BA Stokes  …
35258  18.4      England     Pakistan         BA Stokes  LS Livingstone  …
35259  18.5      England     Pakistan         BA Stokes  LS Livingstone  …
35260  18.6      England     Pakistan         BA Stokes  LS Livingstone  …

       team_id_x  non_striker_id   player_name_y team_id_y  bowler_id   \
index
0              1              19  Najeeb Tarakai         1         72
1              1              19  Najeeb Tarakai         1         72
2              1              19  Najeeb Tarakai         1         72
3              1              19  Najeeb Tarakai         1         72
4              1              19  Najeeb Tarakai         1         72
…          …             …             …       …        …
35256          4             101       BA Stokes         4        311
35257          4             101       BA Stokes         4        311
35258          4             116  LS Livingstone         4        311
```

```
35259          4             116  LS Livingstone        4        311
35260          4             116  LS Livingstone        4        311

          player_name_x  team_id_x  player_out_id     player_name_y team_id_y
index
0      Mashrafe Mortaza          3           17.0  Mohammad Shahzad       1.0
1      Mashrafe Mortaza          3            NaN               NaN       NaN
2      Mashrafe Mortaza          3            NaN               NaN       NaN
3      Mashrafe Mortaza          3            NaN               NaN       NaN
4      Mashrafe Mortaza          3            NaN               NaN       NaN
…                   …         …             …                 …         …
35256    Mohammad Wasim         13          119.0            MM Ali       4.0
35257    Mohammad Wasim         13            NaN               NaN       NaN
35258    Mohammad Wasim         13            NaN               NaN       NaN
35259    Mohammad Wasim         13            NaN               NaN       NaN
35260    Mohammad Wasim         13            NaN               NaN       NaN

[35261 rows x 42 columns]
```

```
<ipython-input-50-8d60fc0a8f09>:1: FutureWarning: Passing 'suffixes' which cause
duplicate columns {'team_id_x', 'player_name_x'} in the result is deprecated and
will raise a MergeError in a future version.
  data = data.reset_index().merge(data3, how='left',
left_on=['player_dismissed','batting_team_id'],
right_on=['player_name','team_id']).set_index('index')
```

```
[ ]: data = data.rename(columns={'innings': 'inning_id', 'runs_off_bat':␣
      ↪'runs_by_batter', 'extras': 'runs_by_extras'})
     data = data[['match_id', 'inning_id', 'bowl_no', 'over_no', 'wicket',␣
      ↪'runs_by_batter', 'runs_by_extras', 'batsman_id', 'bowler_id',␣
      ↪'non_striker_id', 'wicket_type', 'catcher_id', 'player_out_id']]
```

```
[ ]: data.to_csv('new_file.csv', index=False)
```

```
[ ]: len(data['match_id'].unique())
```

```
[ ]: 154
```

```
[ ]: unique_batsman_ids = data.groupby('match_id')['batsman_id'].unique()
```

```
[ ]: unique_bowler_ids = data.groupby('match_id')['bowler_id'].unique()
```

```
[ ]: unique_non_batsman_ids = data.groupby('match_id')['non_striker_id'].unique()
```

```
[ ]: combined_ids = unique_batsman_ids.combine(unique_bowler_ids, lambda x1, x2:␣
      ↪list(set(list(x1) + list(x2))))
     combined_ids2 = combined_ids.combine(unique_non_batsman_ids, lambda x1, x2:␣
      ↪list(set(list(x1) + list(x2))))
```

```
print(combined_ids2)
```

```
match_id
682897      [1, 6, 9, 14, 16, 17, 19, 22, 26, 27, 28, 60, …
682899      [128, 130, 132, 133, 134, 136, 138, 139, 140, …
682901      [172, 173, 174, 179, 181, 182, 186, 188, 192, …
682903      [430, 431, 432, 437, 438, 440, 442, 444, 445, …
682905      [128, 130, 3, 132, 133, 134, 2, 136, 137, 9, 6…
                                 …
1298175     [293, 297, 299, 306, 308, 309, 311, 312, 315, …
1298176     [147, 148, 149, 152, 156, 158, 160, 163, 164, …
1298177     [257, 258, 260, 262, 266, 268, 269, 293, 297, …
1298178     [147, 148, 149, 152, 156, 158, 160, 163, 164, …
1298179     [293, 297, 299, 306, 308, 309, 311, 312, 315, …
Length: 154, dtype: object
```

```
[ ]: df = combined_ids2.apply(pd.Series).stack().reset_index(level=1, drop=True).
       ↪reset_index()
     df.columns = ['Key', 'Value']
     df.to_csv('output2.csv', index=False)
```