t20-data-preprocessing

July 18, 2023

T20 World Cup Cricket Data Pre Processing

[85]:
```
#import necessary libraries
import pandas as pd
import json
```

(1) Process Match Results

[86]:
```
with open('t20_json_files/t20_wc_match_results.json') as f:
    data = json.load(f)

df_match = pd.DataFrame(data[0]['matchSummary'])
df_match.head()
```

[86]:
```
         team1        team2       winner     margin   ground     matchDate  \
0      Namibia    Sri Lanka      Namibia    55 runs  Geelong  Oct 16, 2022
1  Netherlands       U.A.E.  Netherlands  3 wickets  Geelong  Oct 16, 2022
2     Scotland  West Indies     Scotland    42 runs   Hobart  Oct 17, 2022
3      Ireland     Zimbabwe     Zimbabwe    31 runs   Hobart  Oct 17, 2022
4      Namibia  Netherlands  Netherlands  5 wickets  Geelong  Oct 18, 2022

     scorecard
0  T20I # 1823
1  T20I # 1825
2  T20I # 1826
3  T20I # 1828
4  T20I # 1830
```

[87]:
```
df_match.shape
```

[87]:
```
(45, 7)
```

Use scorecard as a match id to link with other tables

[88]:
```
df_match.rename({'scorecard': 'match_id'}, axis = 1, inplace = True)
df_match.head()
```

[88]:
```
         team1        team2       winner     margin   ground     matchDate  \
0      Namibia    Sri Lanka      Namibia    55 runs  Geelong  Oct 16, 2022
```

```
1  Netherlands       U.A.E.  Netherlands  3 wickets  Geelong  Oct 16, 2022
2     Scotland  West Indies     Scotland    42 runs   Hobart  Oct 17, 2022
3      Ireland     Zimbabwe     Zimbabwe    31 runs   Hobart  Oct 17, 2022
4      Namibia  Netherlands  Netherlands  5 wickets  Geelong  Oct 18, 2022

      match_id
0  T20I # 1823
1  T20I # 1825
2  T20I # 1826
3  T20I # 1828
4  T20I # 1830
```

Create a match ids dictionary that maps team names to a unique match id. This will be useful later on to link with other tables

[89]:

```
match_ids_dict = {}

for index, row in df_match.iterrows():
    key1 = row['team1'] + ' Vs ' + row['team2']
    key2 = row['team2'] + ' Vs ' + row['team1']
    match_ids_dict[key1] = row['match_id']
    match_ids_dict[key2] = row['match_id']
```

[90]:

```
df_match.to_csv('t20_csv_files/dim_match_summary.csv', index = False)
```

(2) Process Batting Summary

[91]:

```
with open('t20_json_files/t20_wc_batting_summary.json') as f:
    data = json.load(f)
    all_records = []
    for rec in data:
        all_records.extend(rec['battingSummary'])

df_batting = pd.DataFrame(all_records)
df_batting.head(11)
```

[91]:

```
                    match teamInnings  battingPos            batsmanName  \
0   Namibia Vs Sri Lanka     Namibia           1      Michael van Lingen
1   Namibia Vs Sri Lanka     Namibia           2           Divan la Cock
2   Namibia Vs Sri Lanka     Namibia           3  Jan Nicol Loftie-Eaton
3   Namibia Vs Sri Lanka     Namibia           4           Stephan Baard
4   Namibia Vs Sri Lanka     Namibia           5      Gerhard Erasmus(c)
5   Namibia Vs Sri Lanka     Namibia           6            Jan Frylinck
6   Namibia Vs Sri Lanka     Namibia           7             David Wiese
7   Namibia Vs Sri Lanka     Namibia           8                 JJ Smit
8   Namibia Vs Sri Lanka   Sri Lanka           1          Pathum Nissanka
9   Namibia Vs Sri Lanka   Sri Lanka           2          Kusal Mendisâ€
10  Namibia Vs Sri Lanka   Sri Lanka           3      Dhananjaya de Silva
```

```
                         dismissal runs balls 4s 6s      SR
0      c Pramod Madushan b Chameera    3      6  0  0   50.00
1       c Shanaka b Pramod Madushan    9      9  1  0  100.00
2         c â€ Mendis b Karunaratne   20     12  1  2  166.66
3   c DM de Silva b Pramod Madushan   26     24  2  0  108.33
4      c Gunathilaka b PWH de Silva   20     24  0  0   83.33
5   run out (Gunathilaka/â€ Mendis)   44     28  4  0  157.14
6          c â€ Mendis b Theekshana    0      1  0  0    0.00
7                                     31     16  2  2  193.75
8                 c Smit b Shikongo    9     10  1  0   90.00
9                c â€ Green b Wiese    6      6  0  0  100.00
10            c Shikongo b Frylinck   12     11  1  0  109.09
```

[92]:
```
df_batting['out/not_out'] = df_batting.dismissal.apply(lambda x: "out" if␣
 ↪len(x)>0 else "not_out")
df_batting.head(11)
```

[92]:
```
                   match teamInnings  battingPos            batsmanName  \
0   Namibia Vs Sri Lanka     Namibia           1     Michael van Lingen
1   Namibia Vs Sri Lanka     Namibia           2          Divan la Cock
2   Namibia Vs Sri Lanka     Namibia           3 Jan Nicol Loftie-Eaton
3   Namibia Vs Sri Lanka     Namibia           4          Stephan Baard
4   Namibia Vs Sri Lanka     Namibia           5     Gerhard Erasmus(c)
5   Namibia Vs Sri Lanka     Namibia           6           Jan Frylinck
6   Namibia Vs Sri Lanka     Namibia           7            David Wiese
7   Namibia Vs Sri Lanka     Namibia           8                JJ Smit
8   Namibia Vs Sri Lanka   Sri Lanka           1        Pathum Nissanka
9   Namibia Vs Sri Lanka   Sri Lanka           2        Kusal Mendisâ€
10  Namibia Vs Sri Lanka   Sri Lanka           3    Dhananjaya de Silva

                         dismissal runs balls 4s 6s      SR out/not_out
0      c Pramod Madushan b Chameera    3      6  0  0   50.00         out
1       c Shanaka b Pramod Madushan    9      9  1  0  100.00         out
2         c â€ Mendis b Karunaratne   20     12  1  2  166.66         out
3   c DM de Silva b Pramod Madushan   26     24  2  0  108.33         out
4      c Gunathilaka b PWH de Silva   20     24  0  0   83.33         out
5   run out (Gunathilaka/â€ Mendis)   44     28  4  0  157.14         out
6          c â€ Mendis b Theekshana    0      1  0  0    0.00         out
7                                     31     16  2  2  193.75     not_out
8                 c Smit b Shikongo    9     10  1  0   90.00         out
9                c â€ Green b Wiese    6      6  0  0  100.00         out
10            c Shikongo b Frylinck   12     11  1  0  109.09         out
```

[93]:
```
df_batting['match_id'] = df_batting['match'].map(match_ids_dict)
df_batting.head()
```

[93]:

```
                  match teamInnings  battingPos             batsmanName  \
0  Namibia Vs Sri Lanka     Namibia           1      Michael van Lingen
1  Namibia Vs Sri Lanka     Namibia           2           Divan la Cock
2  Namibia Vs Sri Lanka     Namibia           3  Jan Nicol Loftie-Eaton
3  Namibia Vs Sri Lanka     Namibia           4           Stephan Baard
4  Namibia Vs Sri Lanka     Namibia           5      Gerhard Erasmus(c)

                         dismissal runs balls 4s 6s      SR out/not_out  \
0     c Pramod Madushan b Chameera    3     6  0  0   50.00         out
1      c Shanaka b Pramod Madushan    9     9  1  0  100.00         out
2        c â€ Mendis b Karunaratne   20    12  1  2  166.66         out
3  c DM de Silva b Pramod Madushan   26    24  2  0  108.33         out
4     c Gunathilaka b PWH de Silva   20    24  0  0   83.33         out

      match_id
0  T20I # 1823
1  T20I # 1823
2  T20I # 1823
3  T20I # 1823
4  T20I # 1823
```

[94]:

```
df_batting.drop(columns=["dismissal"], inplace=True)
df_batting.head(10)
```

[94]:

```
                  match teamInnings  battingPos             batsmanName runs  \
0  Namibia Vs Sri Lanka     Namibia           1      Michael van Lingen    3
1  Namibia Vs Sri Lanka     Namibia           2           Divan la Cock    9
2  Namibia Vs Sri Lanka     Namibia           3  Jan Nicol Loftie-Eaton   20
3  Namibia Vs Sri Lanka     Namibia           4           Stephan Baard   26
4  Namibia Vs Sri Lanka     Namibia           5      Gerhard Erasmus(c)   20
5  Namibia Vs Sri Lanka     Namibia           6            Jan Frylinck   44
6  Namibia Vs Sri Lanka     Namibia           7             David Wiese    0
7  Namibia Vs Sri Lanka     Namibia           8                 JJ Smit   31
8  Namibia Vs Sri Lanka   Sri Lanka           1         Pathum Nissanka    9
9  Namibia Vs Sri Lanka   Sri Lanka           2         Kusal Mendisâ€      6

  balls 4s 6s      SR out/not_out     match_id
0     6  0  0   50.00         out  T20I # 1823
1     9  1  0  100.00         out  T20I # 1823
2    12  1  2  166.66         out  T20I # 1823
3    24  2  0  108.33         out  T20I # 1823
4    24  0  0   83.33         out  T20I # 1823
5    28  4  0  157.14         out  T20I # 1823
6     1  0  0    0.00         out  T20I # 1823
7    16  2  2  193.75     not_out  T20I # 1823
8    10  1  0   90.00         out  T20I # 1823
9     6  0  0  100.00         out  T20I # 1823
```

Cleanup weird characters

[95]:
```
df_batting['batsmanName'] = df_batting['batsmanName'].apply(lambda x: x.
 ↪replace('â€', ''))
df_batting['batsmanName'] = df_batting['batsmanName'].apply(lambda x: x.
 ↪replace('\xa0', ''))
df_batting.head()
```

[95]:
```
                  match teamInnings  battingPos            batsmanName runs  \
0  Namibia Vs Sri Lanka     Namibia           1      Michael van Lingen    3
1  Namibia Vs Sri Lanka     Namibia           2           Divan la Cock    9
2  Namibia Vs Sri Lanka     Namibia           3  Jan Nicol Loftie-Eaton   20
3  Namibia Vs Sri Lanka     Namibia           4           Stephan Baard   26
4  Namibia Vs Sri Lanka     Namibia           5      Gerhard Erasmus(c)   20

  balls 4s 6s      SR out/not_out     match_id
0     6  0  0   50.00         out  T20I # 1823
1     9  1  0  100.00         out  T20I # 1823
2    12  1  2  166.66         out  T20I # 1823
3    24  2  0  108.33         out  T20I # 1823
4    24  0  0   83.33         out  T20I # 1823
```

[96]:
```
df_batting.shape
```

[96]:
```
(699, 11)
```

[97]:
```
df_batting.to_csv('t20_csv_files/fact_bating_summary.csv', index = False)
```

(3) Process Bowling Summary

[98]:
```
with open('t20_json_files/t20_wc_bowling_summary.json') as f:
    data = json.load(f)
    all_records = []
    for rec in data:
        all_records.extend(rec['bowlingSummary'])
all_records[:2]
```

[98]:
```
[{'match': 'Namibia Vs Sri Lanka',
  'bowlingTeam': 'Sri Lanka',
  'bowlerName': 'Maheesh Theekshana',
  'overs': '4',
  'maiden': '0',
  'runs': '23',
  'wickets': '1',
  'economy': '5.75',
  '0s': '7',
  '4s': '0',
  '6s': '0',
  'wides': '2',
  'noBalls': '0'},
 {'match': 'Namibia Vs Sri Lanka',
  'bowlingTeam': 'Sri Lanka',
  'bowlerName': 'Dushmantha Chameera',
  'overs': '4',
  'maiden': '0',
  'runs': '39',
  'wickets': '1',
  'economy': '9.75',
  '0s': '6',
  '4s': '3',
  '6s': '1',
  'wides': '2',
  'noBalls': '0'}]
```

[99]:

```
df_bowling = pd.DataFrame(all_records)
print(df_bowling.shape)
df_bowling.head()
```

```
(500, 13)
```

[99]:

```
                  match bowlingTeam                 bowlerName overs maiden  \
0  Namibia Vs Sri Lanka   Sri Lanka          Maheesh Theekshana     4      0
1  Namibia Vs Sri Lanka   Sri Lanka         Dushmantha Chameera     4      0
2  Namibia Vs Sri Lanka   Sri Lanka             Pramod Madushan     4      0
3  Namibia Vs Sri Lanka   Sri Lanka         Chamika Karunaratne     4      0
4  Namibia Vs Sri Lanka   Sri Lanka  Wanindu Hasaranga de Silva     4      0

  runs wickets economy 0s 4s 6s wides noBalls
0   23       1    5.75  7  0  0     2       0
1   39       1    9.75  6  3  1     2       0
2   37       2    9.25  6  3  1     0       0
3   36       1    9.00  7  3  1     1       0
4   27       1    6.75  8  1  1     0       0
```

[100]:

```
df_bowling['match_id'] = df_bowling['match'].map(match_ids_dict)
df_bowling.head()
```

[100]:

```
                  match bowlingTeam                 bowlerName overs maiden  \
0  Namibia Vs Sri Lanka   Sri Lanka          Maheesh Theekshana     4      0
1  Namibia Vs Sri Lanka   Sri Lanka         Dushmantha Chameera     4      0
2  Namibia Vs Sri Lanka   Sri Lanka             Pramod Madushan     4      0
3  Namibia Vs Sri Lanka   Sri Lanka         Chamika Karunaratne     4      0
4  Namibia Vs Sri Lanka   Sri Lanka  Wanindu Hasaranga de Silva     4      0

  runs wickets economy 0s 4s 6s wides noBalls    match_id
```

```
0   23         1     5.75   7   0   0      2        0   T20I # 1823
1   39         1     9.75   6   3   1      2        0   T20I # 1823
2   37         2     9.25   6   3   1      0        0   T20I # 1823
3   36         1     9.00   7   3   1      1        0   T20I # 1823
4   27         1     6.75   8   1   1      0        0   T20I # 1823
```

```
[101]: df_bowling.to_csv('t20_csv_files/fact_bowling_summary.csv', index = False)
```

(4) Process Players Information

```
[102]: with open('t20_json_files/t20_wc_player_info.json') as f:
           data = json.load(f)
```

```
[103]: df_players = pd.DataFrame(data)

       print(df_players.shape)
       df_players.head(10)
```

```
(219, 6)
```

```
[103]:                      name       team    battingStyle  \
       0      Michael van Lingen    Namibia   Left hand Bat
       1           Divan la Cock    Namibia  Right hand Bat
       2  Jan Nicol Loftie-Eaton    Namibia   Left hand Bat
       3           Stephan Baard    Namibia  Right hand Bat
       4      Gerhard Erasmus(c)    Namibia  Right hand Bat
       5            Jan Frylinck    Namibia   Left hand Bat
       6             David Wiese    Namibia  Right hand Bat
       7                 JJ Smit    Namibia  Right hand Bat
       8         Pathum Nissanka  Sri Lanka  Right hand Bat
       9         Kusal Mendisâ€   Sri Lanka  Right hand Bat

                      bowlingStyle          playingRole  \
       0             Left arm Medium   Bowling Allrounder
       1                    Legbreak       Opening Batter
       2  Right arm Medium, Legbreak               Batter
       3       Right arm Medium fast               Batter
       4          Right arm Offbreak           Allrounder
       5        Left arm Fast medium           Allrounder
       6       Right arm Medium fast           Allrounder
       7        Left arm Medium fast   Bowling Allrounder
       8                                 Top order Batter
       9                    Legbreak  Wicketkeeper Batter

                                                description
       0
       1
```

```
2
3
4
5
6  David Wiese joined a marked outflow of South A…
7
8
9  Blessed with a compact technique, an aggressiv…
```

Cleanup weird characters

[104]:

```
df_players['name'] = df_players['name'].apply(lambda x: x.replace('â€', ''))
df_players['name'] = df_players['name'].apply(lambda x: x.replace('†', ''))
df_players['name'] = df_players['name'].apply(lambda x: x.replace('\xa0', ''))
df_players.head(10)
```

[104]:

```
                      name       team    battingStyle  \
0      Michael van Lingen    Namibia   Left hand Bat
1           Divan la Cock    Namibia  Right hand Bat
2  Jan Nicol Loftie-Eaton    Namibia   Left hand Bat
3           Stephan Baard    Namibia  Right hand Bat
4      Gerhard Erasmus(c)    Namibia  Right hand Bat
5             Jan Frylinck    Namibia   Left hand Bat
6             David Wiese    Namibia  Right hand Bat
7                  JJ Smit    Namibia  Right hand Bat
8          Pathum Nissanka  Sri Lanka  Right hand Bat
9             Kusal Mendis  Sri Lanka  Right hand Bat

                bowlingStyle          playingRole  \
0           Left arm Medium  Bowling Allrounder
1                  Legbreak       Opening Batter
2  Right arm Medium, Legbreak               Batter
3      Right arm Medium fast               Batter
4         Right arm Offbreak           Allrounder
5        Left arm Fast medium           Allrounder
6      Right arm Medium fast           Allrounder
7        Left arm Medium fast  Bowling Allrounder
8                               Top order Batter
9                  Legbreak  Wicketkeeper Batter

                                         description
0
1
2
3
4
5
```

```
6  David Wiese joined a marked outflow of South A…
7
8
9  Blessed with a compact technique, an aggressiv…
```

[105]:

```
df_players[df_players['team'] == 'India']
```

[105]:

```
                    name   team    battingStyle  \
127             KL Rahul  India  Right hand Bat
128      Rohit Sharma(c)  India  Right hand Bat
129          Virat Kohli  India  Right hand Bat
130     Suryakumar Yadav  India  Right hand Bat
131           Axar Patel  India   Left hand Bat
132        Hardik Pandya  India  Right hand Bat
133       Dinesh Karthik  India  Right hand Bat
134  Ravichandran Ashwin  India  Right hand Bat
135    Bhuvneshwar Kumar  India  Right hand Bat
136       Arshdeep Singh  India   Left hand Bat
137       Mohammed Shami  India  Right hand Bat
192         Deepak Hooda  India  Right hand Bat
211         Rishabh Pant  India   Left hand Bat

                                 bowlingStyle          playingRole  \
127                                                 Opening Batter
128                        Right arm Offbreak     Top order Batter
129                          Right arm Medium     Top order Batter
130  Right arm Medium, Right arm Offbreak                   Batter
131                    Slow Left arm Orthodox   Bowling Allrounder
132                     Right arm Medium fast           Allrounder
133                        Right arm Offbreak  Wicketkeeper Batter
134                        Right arm Offbreak   Bowling Allrounder
135                          Right arm Medium               Bowler
136                      Left arm Medium fast               Bowler
137                            Right arm Fast               Bowler
192                        Right arm Offbreak           Allrounder
211                                            Wicketkeeper Batter

                                           description
127  A tall, elegant right-hand batsman who can kee…
128  Languid and easy on the eye, Rohit Sharma owne…
129  India has given to the world many a great cric…
130  Hard-hitting 360-degree batter Suryakumar Yada…
131  Left-arm spinner Axar Patel has been increasin…
132  Hardik Pandya swears by living life king size …
133  Not many would forget the sight of Dinesh Kart…
134  R Ashwin took the tricks and skills he learned…
135  At the time of his India debut in 2012, Bhuvne…
```

```
136
137  Mohammed Shami was India's leading fast bowler…
192  An allrounder who can bat in any position, Dee…
211  A match-turning, swashbuckling batter-keeper i…
```

[106]:

```
df_players.to_csv('t20_csv_files/dim_players_no_images.csv', index = False)
```