T20 World Cup Cricket Data Pre Processing

```
In [1]: #import necessary libraries
  import pandas as pd
  import json
```

(1) Process Match Results

```
In [2]: with open('t20_json_files/t20_json_files/t20_wc_match_results.json') as f:
    data = json.load(f)

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

Out[2]:		team1	team2	winner	margin	ground	matchDate	scorecard
	0	Namibia	Sri Lanka	Namibia	55 runs	Geelong	Oct 16, 2022	T20I # 1823
	1	Netherlands	U.A.E.	Netherlands	3 wickets	Geelong	Oct 16, 2022	T20I # 1825
	2	Scotland	West Indies	Scotland	42 runs	Hobart	Oct 17, 2022	T20I # 1826
	3	Ireland	Zimbabwe	Zimbabwe	31 runs	Hobart	Oct 17, 2022	T20I # 1828
	4	Namibia	Netherlands	Netherlands	5 wickets	Geelona	Oct 18, 2022	T20I # 1830

```
In [3]: df_match.shape
Out[3]: (45, 7)
```

Use scorecard as a match id to link with other tables

```
In [4]: df_match.rename({'scorecard': 'match_id'}, axis = 1, inplace = True)
    df_match.head()
```

Out[4]:		team1	team2	winner	margin	ground	matchDate	match_id
	0	Namibia	Sri Lanka	Namibia	55 runs	Geelong	Oct 16, 2022	T20I # 1823
	1	Netherlands	U.A.E.	Netherlands	3 wickets	Geelong	Oct 16, 2022	T20I # 1825
	2	Scotland	West Indies	Scotland	42 runs	Hobart	Oct 17, 2022	T20I # 1826
	3	Ireland	Zimbabwe	Zimbabwe	31 runs	Hobart	Oct 17, 2022	T20I # 1828
	4	Namibia	Netherlands	Netherlands	5 wickets	Geelong	Oct 18, 2022	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

```
In [5]: 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']
```

```
In [6]: df_match.to_csv('csv_files/dim_match_summary.csv', index = False)
```

(2) Process Batting Summary

```
In [7]: with open('t20_json_files/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)
```

ut[7]:		match	teamInnings	battingPos	batsmanName	dismissal	runs	balls	4s	6s	SR
•	0	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	c Pramod Madushan b Chameera	3	6	0	0	50.00
	1	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	c Shanaka b Pramod Madushan	9	9	1	0	100.00
	2	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	c †Mendis b Karunaratne	20	12	1	2	166.66
	3	Namibia Vs Sri Lanka	Namibia	4	Stephan Baard	c DM de Silva b Pramod Madushan	26	24	2	0	108.33
	4	Namibia Vs Sri Lanka	Namibia	5	Gerhard Erasmus(c)	c Gunathilaka b PWH de Silva	20	24	0	0	83.33
	5	Namibia Vs Sri Lanka	Namibia	6	Jan Frylinck	run out (Gunathilaka/ †Mendis)	44	28	4	0	157.14
	6	Namibia Vs Sri Lanka	Namibia	7	David Wiese	c †Mendis b Theekshana	0	1	0	0	0.00
	7	Namibia Vs Sri Lanka	Namibia	8	JJ Smit		31	16	2	2	193.75
	8	Namibia Vs Sri Lanka	Sri Lanka	1	Pathum Nissanka	c Smit b Shikongo	9	10	1	0	90.00
	9	Namibia Vs Sri Lanka	Sri Lanka	2	Kusal Mendisâ€	c †Green b Wiese	6	6	0	0	100.00
	10	Namibia Vs Sri Lanka	Sri Lanka	3	Dhananjaya de Silva	c Shikongo b Frylinck	12	11	1	0	109.09

```
In [8]: df_batting['out/not_out'] = df_batting.dismissal.apply(lambda x: "out" if len(x)>0
    df_batting.head(11)
```

Lanka Lingen Chameera	ut[8]:		match	teamInnings	battingPos	batsmanName	dismissal	runs	balls	4s	6s	S
1 Vs Sri Lanka Namibia 2 Divan la Cock Madushan Pramod Madushan 9 9 1 0 100 Namibia Namibia 3 Jan Nicol Loftie-Eaton Câé Mendis Karunaratne 20 12 1 2 16 Namibia Namibia 4 Stephan Baard Silva b Pramod Madushan 26 24 2 0 100 Namibia 4 Stephan Baard Silva b Pramod Madushan 20 24 2 0 100 Namibia 4 Stephan Baard Erasmus(c) Gunathilaka b Pramod Madushan 20 24 2 0 100 Namibia Namibia 5 Gerhard Erasmus(c) Gunathilaka b Pramod Madushan 20 24 0 0 8 Namibia Namibia 6 Jan Frylinck (Gunathilaka) 44 28 4 0 15 Namibia 7 David Wiese 6 4 28 4 0 15 Namibia 7 David Wiese 6 0 1 0 0 0 Namibia		0	Vs Sri	Namibia	1		Madushan b	3	6	0	0	50.0
2 Vs Sri Lanka Namibia 3 Jan Nicol Loftie-Eaton Karunaratne b Karunaratne 20 12 1 2 166 Namibia Namibia 4 Stephan Baard Silva b Pramod Madushan 26 24 2 0 100 Namibia Namibia 5 Gerhard Erasmus(c) Gunathilaka b PVWH de Silva 20 24 0 0 8: Namibia Namibia 6 Jan Frylinck (Gunathilaka/ †Mendis) 44 28 4 0 15: Namibia Namibia 7 David Wiese b D 1 0 0 0 Namibia Namibia 8 JJ Smit 31 16 2 2 19: Namibia Namibia 8 JJ Smit 31 16 2 2 19: Namibia Sri Lanka 1 Pathum Nissanka C Smit b Shikongo 9 10 1 0 90 Namibia 9 Vs Sri Sri Lanka 2 Kusal C †Green b Mendisâ€ Wiese 6 6 0		1	Vs Sri	Namibia	2	Divan la Cock	Pramod	9	9	1	0	100.0
Namibia Vs Sri Lanka Namibia 4 Stephan Baard Silva b Pramod Madushan 26 24 2 0 108 Namibia Namibia 5 Gerhard Erasmus(c) Gunathilaka b PWH de Silva 20 24 0 0 83 Namibia Namibia 6 Jan Frylinck (Gunathilaka/ šé Mendis) 44 28 4 0 15 Namibia Namibia 7 David Wiese 6 Mendis 0 1 0 0 0 0 Namibia Namibia 8 JJ Smit 31 16 2 2 2 193 Namibia Namibia 8 JS Smit 31 16 2 2 193 Namibia 8 Vs Sri Sri Lanka 1 Pathum Shikongo 9 10 1 0 90 Namibia 9 Vs Sri Sri Lanka 2 Kusal Mendisâ€ Green b Mendisâ€ 6 6 0 0 0 100 Namibia Namibia 2 Mendisâ€ Obananjaya de Calikongo b Silva 12 11 1 0 100 Namibia 10 Vs Sri Sri Lanka Dhananjaya de Calikongo b Silva 1 2 11 1 0 100		2	Vs Sri	Namibia	3		b	20	12	1	2	166.6
Namibia Vs Sri Namibia 5 Gerhard Erasmus(c) Gunathilaka b PWH de Silva 20 24 0 0 83 Namibia Namibia 6 Jan Frylinck (Gunathilaka/ (Gunathilaka/ †Mendis) 44 28 4 0 15 Namibia Namibia 7 David Wiese 6 0 1 0 0 15 Namibia Namibia 8 JJ Smit 31 16 2 2 193 Namibia Namibia 8 JJ Smit 31 16 2 2 193 Namibia Namibia 1 Pathum Nissanka C Smit b Shikongo 9 10 1 0 90 Namibia 9 Vs Sri Sri Lanka 2 Kusal Mendisâ€ C Shikongo b Wiese 6 6 0 0 10 Namibia 10 Vs Sri Sri Lanka 3 Dhananjaya de Silva C Shikongo b Shikongo b Silva 12 11 1		3	Vs Sri	Namibia	4	Stephan Baard	Silva b Pramod	26	24	2	0	108.3
5 Vs Sri Lanka Namibia 6 Jan Frylinck (Gunathilaka/ †Mendis) 44 28 4 0 157 Namibia 6 Vs Sri 		4	Vs Sri	Namibia	5		Gunathilaka b PWH de	20	24	0	0	83.3
6 Vs Sri Lanka Namibia 7 David Wiese b 0 1 0 <t< td=""><td></td><td>5</td><td>Vs Sri</td><td>Namibia</td><td>6</td><td>Jan Frylinck</td><td>(Gunathilaka/</td><td>44</td><td>28</td><td>4</td><td>0</td><td>157.1</td></t<>		5	Vs Sri	Namibia	6	Jan Frylinck	(Gunathilaka/	44	28	4	0	157.1
7 Vs Sri Lanka Namibia 8 JJ Smit 31 16 2 2 199 Namibia Ns Sri Lanka 1 Pathum c Smit b Nissanka 9 10 1 0 90 Namibia Vs Sri Lanka 2 Kusal c †Green b Mendisâ€ 6 6 0 0 100 Namibia Namibia 3 Dhananjaya de c Shikongo b Silva Envlinck 12 11 1 0 100		6	Vs Sri	Namibia	7	David Wiese	b	0	1	0	0	0.0
8 Vs Sri Lanka 1 Pathum C Smit b Nissanka 9 10 1 0 90 Namibia 9 Vs Sri Lanka 2 Kusal C †Green b Mendisâ€ 6 6 0 0 100 Namibia Namibia 3 Dhananjaya de C Shikongo b Silva 12 11 1 0 100		7	Vs Sri	Namibia	8	JJ Smit		31	16	2	2	193.7
9 Vs Sri Lanka Sri Lanka 2 Kusal Mendisâ€ c a€ Green b Wiese 6 6 0 0 100 Namibia Namibia Dhananjaya de 		8	Vs Sri	Sri Lanka	1			9	10	1	0	90.0
10 Vs Sri Sri Lanka 3 Silva Envlinch 12 11 1 0 109		9	Vs Sri	Sri Lanka	2			6	6	0	0	100.0
Lanka	1	0		Sri Lanka	3		c Shikongo b Frylinck	12	11	1	0	109.0

Out[9]:		match	teamInnings	battingPos	batsmanName	dismissal	runs	balls	4s	6s	SR	out,
	0	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	c Pramod Madushan b Chameera	3	6	0	0	50.00	
	1	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	c Shanaka b Pramod Madushan	9	9	1	0	100.00	
	2	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	c †Mendis b Karunaratne	20	12	1	2	166.66	
	3	Namibia Vs Sri Lanka	Namibia	4	Stephan Baard	c DM de Silva b Pramod Madushan	26	24	2	0	108.33	
	4	Namibia Vs Sri Lanka	Namibia	5	Gerhard Erasmus(c)	c Gunathilaka b PWH de Silva	20	24	0	0	83.33	
4												•
In [10]:			g.drop(colum g.head(10)	ns=["dismi	ssal"], inpla	ce= True)						

localhost:8888/nbconvert/html/Downloads/data-analytics-project-for-beginners/t20_data_preprocessing.ipynb?download=false

Out[10]:		match	teamInnings	battingPos	batsmanName	runs	balls	4s	6s	SR	out/not_out	ma
	0	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	3	6	0	0	50.00	out	
	1	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	9	9	1	0	100.00	out	
	2	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	20	12	1	2	166.66	out	
	3	Namibia Vs Sri Lanka	Namibia	4	Stephan Baard	26	24	2	0	108.33	out	
	4	Namibia Vs Sri Lanka	Namibia	5	Gerhard Erasmus(c)	20	24	0	0	83.33	out	
	5	Namibia Vs Sri Lanka	Namibia	6	Jan Frylinck	44	28	4	0	157.14	out	
	6	Namibia Vs Sri Lanka	Namibia	7	David Wiese	0	1	0	0	0.00	out	
	7	Namibia Vs Sri Lanka	Namibia	8	JJ Smit	31	16	2	2	193.75	not_out	
	8	Namibia Vs Sri Lanka	Sri Lanka	1	Pathum Nissanka	9	10	1	0	90.00	out	
	9	Namibia Vs Sri Lanka	Sri Lanka	2	Kusal Mendisâ€	6	6	0	0	100.00	out	

Cleanup weird characters

```
In [11]: df_batting['batsmanName'] = df_batting['batsmanName'].apply(lambda x: x.replace('âf
    df_batting['batsmanName'] = df_batting['batsmanName'].apply(lambda x: x.replace('\>
    df_batting.head()
```

```
match teamInnings battingPos batsmanName runs balls 4s 6s
Out[11]:
                                                                                   SR out/not_out ma
             Namibia
                                                 Michael van
                                            1
                                                                3
          0
                Vs Sri
                          Namibia
                                                                         0
                                                                             0
                                                                                 50.00
                                                                                               out
                                                      Lingen
                Lanka
              Namibia
          1
                                           2
                Vs Sri
                          Namibia
                                                Divan la Cock
                                                                             0 100.00
                                                                                               out
                Lanka
              Namibia
                                                   Jan Nicol
                                           3
          2
                                                               20
                                                                             2 166.66
                Vs Sri
                          Namibia
                                                                     12
                                                                                               out
                                                 Loftie-Eaton
                Lanka
              Namibia
          3
                Vs Sri
                          Namibia
                                               Stephan Baard
                                                               26
                                                                             0 108.33
                                                                                               out
                Lanka
              Namibia
                                                    Gerhard
                                            5
                Vs Sri
                          Namibia
                                                               20
                                                                     24
                                                                                 83.33
                                                                                               out
                                                  Erasmus(c)
                Lanka
          df_batting.shape
In [12]:
          (699, 11)
Out[12]:
          df_batting.to_csv('csv_files/fact_bating_summary.csv', index = False)
In [13]:
          (3) Process Bowling Summary
          with open('t20_json_files/t20_json_files/t20_wc_bowling_summary.json') as f:
In [14]:
               data = json.load(f)
               all_records = []
               for rec in data:
                   all_records.extend(rec['bowlingSummary'])
          all_records[:2]
```

```
[{'match': 'Namibia Vs Sri Lanka',
Out[14]:
             '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'}]
In [15]: df_bowling = pd.DataFrame(all_records)
          print(df_bowling.shape)
          df_bowling.head()
          (500, 13)
Out[15]:
               match bowlingTeam bowlerName overs maiden runs wickets economy 0s 4s 6s wid
             Namibia
                                       Maheesh
          0
                Vs Sri
                                                             0
                                                                 23
                                                                                  5.75
                                                                                        7
                                                                                                0
                          Sri Lanka
                                                    4
                                                                           1
                                                                                            0
                                     Theekshana
               Lanka
             Namibia
                                    Dushmantha
          1
                                                             0
               Vs Sri
                          Sri Lanka
                                                    4
                                                                 39
                                                                           1
                                                                                  9.75
                                                                                            3
                                                                                               1
                                                                                        6
                                      Chameera
               Lanka
             Namibia
                                        Pramod
          2
                                                             0
                                                                           2
                Vs Sri
                                                    4
                                                                 37
                                                                                  9.25
                                                                                       6
                                                                                            3
                                                                                               1
                          Sri Lanka
                                      Madushan
               Lanka
             Namibia
                                        Chamika
          3
                Vs Sri
                          Sri Lanka
                                                    4
                                                             0
                                                                 36
                                                                           1
                                                                                  9.00
                                                                                        7
                                                                                            3
                                                                                               1
                                     Karunaratne
               Lanka
             Namibia
                                        Wanindu
                                                             0
                                                                 27
                                                                           1
                Vs Sri
                          Sri Lanka
                                      Hasaranga
                                                    4
                                                                                  6.75
                                                                                       8
                                                                                            1
               Lanka
                                        de Silva
          df_bowling['match_id'] = df_bowling['match'].map(match_ids_dict)
In [16]:
          df_bowling.head()
```

Out[16]:		match	bowlingTeam	bowlerName	overs	maiden	runs	wickets	economy	0s	4s	6s	wid
	0	Namibia Vs Sri Lanka	Sri Lanka	Maheesh Theekshana	4	0	23	1	5.75	7	0	0	
	1	Namibia Vs Sri Lanka	Sri Lanka	Dushmantha Chameera	4	0	39	1	9.75	6	3	1	
	2	Namibia Vs Sri Lanka	Sri Lanka	Pramod Madushan	4	0	37	2	9.25	6	3	1	
	3	Namibia Vs Sri Lanka	Sri Lanka	Chamika Karunaratne	4	0	36	1	9.00	7	3	1	
	4	Namibia Vs Sri Lanka	Sri Lanka	Wanindu Hasaranga de Silva	4	0	27	1	6.75	8	1	1	
4													•
In [17]:	d1	_bowling	:to_csv('csv	_files/fact_	_bowlir	ng_summa	ry.cs	v', inde	x = False	!)			
	(4) Process Players Information												
In [18]:	wi	<pre>with open('t20_json_files/t20_json_files/t20_wc_player_info.json') as f: data = json.load(f)</pre>											
In [19]:	d1	_players	= pd.DataFr	ame(data)									

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

(219, 6)

Out[19]:		name	team	battingStyle	bowlingStyle	playingRole	description
	0	Michael van Lingen	Namibia	Left hand Bat	Left arm Medium	Bowling Allrounder	
	1	Divan la Cock	Namibia	Right hand Bat	Legbreak	Opening Batter	
	2	Jan Nicol Loftie-Eaton	Namibia	Left hand Bat	Right arm Medium, Legbreak	Batter	
	3	Stephan Baard	Namibia	Right hand Bat	Right arm Medium fast	Batter	
	4	Gerhard Erasmus(c)	Namibia	Right hand Bat	Right arm Offbreak	Allrounder	
	5	Jan Frylinck	Namibia	Left hand Bat	Left arm Fast medium	Allrounder	
	6	David Wiese	Namibia	Right hand Bat	Right arm Medium fast	Allrounder	David Wiese joined a marked outflow of South A
	7	JJ Smit	Namibia	Right hand Bat	Left arm Medium fast	Bowling Allrounder	
	8	Pathum Nissanka	Sri Lanka	Right hand Bat		Top order Batter	
	9	Kusal Mendisâ€	Sri Lanka	Right hand Bat	Legbreak	Wicketkeeper Batter	Blessed with a compact technique, an aggressiv

Cleanup weird characters

```
In [20]: 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)
```

Out[20]

	name	team	battingStyle	bowlingStyle	playingRole	description
0	Michael van Lingen	Namibia	Left hand Bat	Left arm Medium	Bowling Allrounder	
1	Divan la Cock	Namibia	Right hand Bat	Legbreak	Opening Batter	
2	Jan Nicol Loftie-Eaton	Namibia	Left hand Bat	Right arm Medium, Legbreak	Batter	
3	Stephan Baard	Namibia	Right hand Bat	Right arm Medium fast	Batter	
4	Gerhard Erasmus(c)	Namibia	Right hand Bat	Right arm Offbreak	Allrounder	
5	Jan Frylinck	Namibia	Left hand Bat	Left arm Fast medium	Allrounder	
6	David Wiese	Namibia	Right hand Bat	Right arm Medium fast	Allrounder	David Wiese joined a marked outflow of South A
7	JJ Smit	Namibia	Right hand Bat	Left arm Medium fast	Bowling Allrounder	
8	Pathum Nissanka	Sri Lanka	Right hand Bat		Top order Batter	
9	Kusal Mendis	Sri Lanka	Right hand Bat	Legbreak	Wicketkeeper Batter	Blessed with a compact technique, an aggressiv

In [21]: df_players[df_players['team'] == 'India']

Out[21]:		name	team	battingStyle	bowlingStyle	playingRole	description
	127	KL Rahul	India	Right hand Bat		Opening Batter	A tall, elegant right- hand batsman who can kee
	128	Rohit Sharma(c)	India	Right hand Bat	Right arm Offbreak	Top order Batter	Languid and easy on the eye, Rohit Sharma owne
	129	Virat Kohli	India	Right hand Bat	Right arm Medium	Top order Batter	India has given to the world many a great cric
	130	Suryakumar Yadav	India	Right hand Bat	Right arm Medium, Right arm Offbreak	Batter	Hard-hitting 360- degree batter Suryakumar Yada
	131	Axar Patel	India	Left hand Bat	Slow Left arm Orthodox	Bowling Allrounder	Left-arm spinner Axar Patel has been increasin
	132	Hardik Pandya	India	Right hand Bat	Right arm Medium fast	Allrounder	Hardik Pandya swears by living life king size
	133	Dinesh Karthik	India	Right hand Bat	Right arm Offbreak	Wicketkeeper Batter	Not many would forget the sight of Dinesh Kart
	134	Ravichandran Ashwin	India	Right hand Bat	Right arm Offbreak	Bowling Allrounder	R Ashwin took the tricks and skills he learned
	135	Bhuvneshwar Kumar	India	Right hand Bat	Right arm Medium	Bowler	At the time of his India debut in 2012, Bhuvne
	136	Arshdeep Singh	India	Left hand Bat	Left arm Medium fast	Bowler	
	137	Mohammed Shami	India	Right hand Bat	Right arm Fast	Bowler	Mohammed Shami was India's leading fast bowler
	192	Deepak Hooda	India	Right hand Bat	Right arm Offbreak	Allrounder	An allrounder who can bat in any position, Dee
	211	Rishabh Pant	India	Left hand Bat		Wicketkeeper Batter	A match-turning, swashbuckling batter-keeper i
In [22]:	df_p	olayers.to_cs	v('csv	_files/dim_p	olayers_no_images	s.csv', index	= False)