T20 World Cup Cricket Data Pre Processing

by Dewansh Vishwakarma

Importing Necessary Library

In [1]:
```
import pandas as pd
import json
```

(1) Process Match Results

In [3]:
```
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()
```

Out[3]:

	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	Geelong	Oct 18, 2022	T20I # 1830

In [4]:
```
df_match.shape
```

Out[4]:
```
(45, 7)
```

Use scorecard as a match id to link with other tables

In [5]:
```
df_match.rename({'scorecard':'match_id'}, axis = 1, inplace = True)
df_match.head()
```

Out[5]:

	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 [18]:
```
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"]

match_ids_dict
```

Out[18]:

```
{'Namibia Vs Sri Lanka': 'T20I # 1823',
 'Sri Lanka Vs Namibia': 'T20I # 1823',
 'Netherlands Vs U.A.E.': 'T20I # 1825',
 'U.A.E. Vs Netherlands': 'T20I # 1825',
 'Scotland Vs West Indies': 'T20I # 1826',
 'West Indies Vs Scotland': 'T20I # 1826',
 'Ireland Vs Zimbabwe': 'T20I # 1828',
 'Zimbabwe Vs Ireland': 'T20I # 1828',
 'Namibia Vs Netherlands': 'T20I # 1830',
 'Netherlands Vs Namibia': 'T20I # 1830',
 'Sri Lanka Vs U.A.E.': 'T20I # 1832',
 'U.A.E. Vs Sri Lanka': 'T20I # 1832',
 'Ireland Vs Scotland': 'T20I # 1833',
 'Scotland Vs Ireland': 'T20I # 1833',
 'West Indies Vs Zimbabwe': 'T20I # 1834',
 'Zimbabwe Vs West Indies': 'T20I # 1834',
 'Netherlands Vs Sri Lanka': 'T20I # 1835',
 'Sri Lanka Vs Netherlands': 'T20I # 1835',
 'Namibia Vs U.A.E.': 'T20I # 1836',
 'U.A.E. Vs Namibia': 'T20I # 1836',
 'Ireland Vs West Indies': 'T20I # 1837',
 'West Indies Vs Ireland': 'T20I # 1837',
 'Scotland Vs Zimbabwe': 'T20I # 1838',
 'Zimbabwe Vs Scotland': 'T20I # 1838',
 'Australia Vs New Zealand': 'T20I # 1839',
 'New Zealand Vs Australia': 'T20I # 1839',
 'Afghanistan Vs England': 'T20I # 1840',
 'England Vs Afghanistan': 'T20I # 1840',
 'Ireland Vs Sri Lanka': 'T20I # 1841',
 'Sri Lanka Vs Ireland': 'T20I # 1841',
 'India Vs Pakistan': 'T20I # 1842',
 'Pakistan Vs India': 'T20I # 1842',
 'Bangladesh Vs Netherlands': 'T20I # 1843',
 'Netherlands Vs Bangladesh': 'T20I # 1843',
 'South Africa Vs Zimbabwe': 'T20I # 1844',
 'Zimbabwe Vs South Africa': 'T20I # 1844',
 'Australia Vs Sri Lanka': 'T20I # 1845',
 'Sri Lanka Vs Australia': 'T20I # 1845',
 'England Vs Ireland': 'T20I # 1846',
 'Ireland Vs England': 'T20I # 1846',
 'Afghanistan Vs New Zealand': 'T20I # 1846a',
 'New Zealand Vs Afghanistan': 'T20I # 1846a',
 'Bangladesh Vs South Africa': 'T20I # 1847',
 'South Africa Vs Bangladesh': 'T20I # 1847',
 'India Vs Netherlands': 'T20I # 1848',
 'Netherlands Vs India': 'T20I # 1848',
 'Pakistan Vs Zimbabwe': 'T20I # 1849',
 'Zimbabwe Vs Pakistan': 'T20I # 1849',
 'Afghanistan Vs Ireland': 'T20I # 1849a',
 'Ireland Vs Afghanistan': 'T20I # 1849a',
 'Australia Vs England': 'T20I # 1849b',
 'England Vs Australia': 'T20I # 1849b',
 'New Zealand Vs Sri Lanka': 'T20I # 1850',
 'Sri Lanka Vs New Zealand': 'T20I # 1850',
 'Bangladesh Vs Zimbabwe': 'T20I # 1851',
 'Zimbabwe Vs Bangladesh': 'T20I # 1851',
 'Netherlands Vs Pakistan': 'T20I # 1852',
 'Pakistan Vs Netherlands': 'T20I # 1852',
 'India Vs South Africa': 'T20I # 1853',
 'South Africa Vs India': 'T20I # 1853',
```

```
 'Australia Vs Ireland': 'T20I # 1855',
 'Ireland Vs Australia': 'T20I # 1855',
 'Afghanistan Vs Sri Lanka': 'T20I # 1856',
 'Sri Lanka Vs Afghanistan': 'T20I # 1856',
 'England Vs New Zealand': 'T20I # 1858',
 'New Zealand Vs England': 'T20I # 1858',
 'Netherlands Vs Zimbabwe': 'T20I # 1859',
 'Zimbabwe Vs Netherlands': 'T20I # 1859',
 'Bangladesh Vs India': 'T20I # 1860',
 'India Vs Bangladesh': 'T20I # 1860',
 'Pakistan Vs South Africa': 'T20I # 1861',
 'South Africa Vs Pakistan': 'T20I # 1861',
 'Ireland Vs New Zealand': 'T20I # 1862',
 'New Zealand Vs Ireland': 'T20I # 1862',
 'Australia Vs Afghanistan': 'T20I # 1864',
 'Afghanistan Vs Australia': 'T20I # 1864',
 'England Vs Sri Lanka': 'T20I # 1867',
 'Sri Lanka Vs England': 'T20I # 1867',
 'Netherlands Vs South Africa': 'T20I # 1871',
 'South Africa Vs Netherlands': 'T20I # 1871',
 'Bangladesh Vs Pakistan': 'T20I # 1872',
 'Pakistan Vs Bangladesh': 'T20I # 1872',
 'India Vs Zimbabwe': 'T20I # 1873',
 'Zimbabwe Vs India': 'T20I # 1873',
 'New Zealand Vs Pakistan': 'T20I # 1877',
 'Pakistan Vs New Zealand': 'T20I # 1877',
 'England Vs India': 'T20I # 1878',
 'India Vs England': 'T20I # 1878',
 'England Vs Pakistan': 'T20I # 1879',
 'Pakistan Vs England': 'T20I # 1879'}
```

In [20]:
```
df_match.to_csv('T20 csv Files/dim_match_summary.csv', index = False)
```

(2) Process Batting Summary

In [9]:
```
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)
```

Out[9]:

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

In [12]:
```
df_batting["out/not_out"] = df_batting.dismissal.apply(lambda x:"out" if len(x)>
df_batting.head(11)
```

Out[12]:

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

In [13]:
```
df_batting.drop(columns=["dismissal"], inplace=True)
df_batting.head(11)
```

Out[13]:

	match	teamInnings	battingPos	batsmanName	runs	balls	4s	6s	SR	out/
0	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	3	6	0	0	50.00	
1	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	9	9	1	0	100.00	
2	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	20	12	1	2	166.66	
3	Namibia Vs Sri Lanka	Namibia	4	Stephan Baard	26	24	2	0	108.33	
4	Namibia Vs Sri Lanka	Namibia	5	Gerhard Erasmus(c)	20	24	0	0	83.33	
5	Namibia Vs Sri Lanka	Namibia	6	Jan Frylinck	44	28	4	0	157.14	
6	Namibia Vs Sri Lanka	Namibia	7	David Wiese	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	9	10	1	0	90.00	
9	Namibia Vs Sri Lanka	Sri Lanka	2	Kusal Mendisâ€	6	6	0	0	100.00	
10	Namibia Vs Sri Lanka	Sri Lanka	3	Dhananjaya de Silva	12	11	1	0	109.09	

Cleanup weird characters

In [15]:
```
df_batting['batsmanName']= df_batting['batsmanName'].apply(lambda x: x.replace('
df_batting['batsmanName']= df_batting['batsmanName'].apply(lambda x: x.replace('
df_batting.head(11)
```

Out[15]:

	match	teamInnings	battingPos	batsmanName	runs	balls	4s	6s	SR	out/
0	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	3	6	0	0	50.00	
1	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	9	9	1	0	100.00	
2	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	20	12	1	2	166.66	
3	Namibia Vs Sri Lanka	Namibia	4	Stephan Baard	26	24	2	0	108.33	
4	Namibia Vs Sri Lanka	Namibia	5	Gerhard Erasmus(c)	20	24	0	0	83.33	
5	Namibia Vs Sri Lanka	Namibia	6	Jan Frylinck	44	28	4	0	157.14	
6	Namibia Vs Sri Lanka	Namibia	7	David Wiese	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	9	10	1	0	90.00	
9	Namibia Vs Sri Lanka	Sri Lanka	2	Kusal Mendis	6	6	0	0	100.00	
10	Namibia Vs Sri Lanka	Sri Lanka	3	Dhananjaya de Silva	12	11	1	0	109.09	

In [19]:
```
df_batting["match_id"] = df_batting["match"].map(match_ids_dict)

df_batting.head()
```

Out[19]:

	match	teamInnings	battingPos	batsmanName	runs	balls	4s	6s	SR	out/n
0	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	3	6	0	0	50.00	
1	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	9	9	1	0	100.00	
2	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	20	12	1	2	166.66	
3	Namibia Vs Sri Lanka	Namibia	4	Stephan Baard	26	24	2	0	108.33	
4	Namibia Vs Sri Lanka	Namibia	5	Gerhard Erasmus(c)	20	24	0	0	83.33	

In [21]:
```
df_batting.shape
```

Out[21]:
```
(699, 11)
```

In [22]:
```
df_batting.to_csv('T20 csv files/fact_bating_summary.csv', index = False)
```

(3) Process Bowling Summary

In [23]:
```
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]
```

Out[23]:

```
[{'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'}]
```


In [24]:
```
df_bowling = pd.DataFrame(all_records)
print(df_bowling.shape)
df_bowling.head()
```

```
(500, 13)
```

Out[24]:

	match	bowlingTeam	bowlerName	overs	maiden	runs	wickets	economy	0s
0	Namibia Vs Sri Lanka	Sri Lanka	Maheesh Theekshana	4	0	23	1	5.75	7
1	Namibia Vs Sri Lanka	Sri Lanka	Dushmantha Chameera	4	0	39	1	9.75	6
2	Namibia Vs Sri Lanka	Sri Lanka	Pramod Madushan	4	0	37	2	9.25	6
3	Namibia Vs Sri Lanka	Sri Lanka	Chamika Karunaratne	4	0	36	1	9.00	7
4	Namibia Vs Sri Lanka	Sri Lanka	Wanindu Hasaranga de Silva	4	0	27	1	6.75	8

In [25]:
```
df_bowling['match_id'] = df_bowling['match'].map(match_ids_dict)
df_bowling.head()
```

Out[25]:

	match	bowlingTeam	bowlerName	overs	maiden	runs	wickets	economy	0s
0	Namibia Vs Sri Lanka	Sri Lanka	Maheesh Theekshana	4	0	23	1	5.75	7
1	Namibia Vs Sri Lanka	Sri Lanka	Dushmantha Chameera	4	0	39	1	9.75	6
2	Namibia Vs Sri Lanka	Sri Lanka	Pramod Madushan	4	0	37	2	9.25	6
3	Namibia Vs Sri Lanka	Sri Lanka	Chamika Karunaratne	4	0	36	1	9.00	7
4	Namibia Vs Sri Lanka	Sri Lanka	Wanindu Hasaranga de Silva	4	0	27	1	6.75	8

In [26]:
```
df_bowling.to_csv('T20 csv files/fact_bowling_summary.csv', index = False)
```

(4) Process Players Information

In [27]:
```
with open('t20_json_files/t20_wc_player_info.json') as f:
    data = json.load(f)
```

In [28]:
```
df_players = pd.DataFrame(data)

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

```
(219, 6)
```

Out[28]:

	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 [29]:

```
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[29]:

	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 [30]:
```
df_players[df_players['team'] == 'India']
```

Out[30]:

	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,

	name	team	battingStyle	bowlingStyle	playingRole	description
						swashbuckling batter-keeper i...

In [31]:
```
df_players.to_csv('T20 csv files/dim_players_no_images.csv', index = False)
```