T20 World Cup Cricket Data Pre Processing

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

(1) Process Match Results

In [24]:
```
#reading the match results csv
match_df = pd.read_csv('wc_match_results.csv')
```

Use scorecard as a match id to link with other tables

In [27]:
```
#renaming the columns
match_df.rename({'Team 1':'team1','Team 2':'team2','Winner':'winner','Margin':'marg
match_df.head()
```

Out[27]:

	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 [22]:
```
match_ids_dict = {}

for index, row in match_df.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 [28]:
```
match_df.to_csv('wc_match_results.csv', index = False)
```

(2) Process Batting Summary

In [58]:
```
#reading the batting summary csv
batting_df = pd.read_csv('batting_summary.csv')
batting_df
```

Out[58]:

	match	teamInnings	battingPos	batsmanName	runs	balls	4s	6s	SR	out/no
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	
...	...	...	...	...	...	...	...	...	...	
694	Pakistan Vs England	England	3	Phil Salt	10	9	2	0	111.11	
695	Pakistan Vs England	England	4	Ben Stokes	52	49	5	1	106.12	nc
696	Pakistan Vs England	England	5	Harry Brook	20	23	1	0	86.95	
697	Pakistan Vs England	England	6	Moeen Ali	19	13	3	0	146.15	
698	Pakistan Vs England	England	7	Liam Livingstone	1	1	0	0	100.00	nc

699 rows × 11 columns

In [30]:
```
#removing blank spaces and replacing 'not out'
batting_df.rename(columns=lambda x: x.strip(), inplace=True)
batting_df['dismissal'] = batting_df['dismissal'].apply(lambda x: x.strip())
batting_df['dismissal'] = batting_df['dismissal'].apply(lambda x: x.replace('not ou
batting_df['batsmanName'] = batting_df['batsmanName'].apply(lambda x: x.strip())
batting_df['teamInnings'] = batting_df['teamInnings'].apply(lambda x: x.strip())
```

In [31]:
```
#adding out/not_out column
batting_df["out/not_out"] = batting_df.dismissal.apply(lambda x :"out" if len(x)>0
```

```
batting_df
```

Out[31]:

	match	teamInnings	battingPos	batsmanName	dismissal	runs	balls	hidden	4
0	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	c Pramod Madushan b Chameera	3	6	7	
1	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	c Shanaka b Pramod Madushan	9	9	15	
2	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	c †Mendis b Karunaratne	20	12	18	
3	Namibia Vs Sri Lanka	Namibia	4	Stephan Baard	c DM de Silva b Pramod Madushan	26	24	49	
4	Namibia Vs Sri Lanka	Namibia	5	Gerhard Erasmus (c)	c Gunathilaka b PWH de Silva	20	24	30	
...	...	...	...	...	...	...	...	...	
694	Pakistan Vs England	England	3	Phil Salt	c Iftikhar Ahmed b Haris Rauf	10	9	16	
695	Pakistan Vs England	England	4	Ben Stokes		52	49	81	
696	Pakistan Vs England	England	5	Harry Brook	c Shaheen Shah Afridi b Shadab Khan	20	23	36	
697	Pakistan Vs England	England	6	Moeen Ali	b Mohammad Wasim	19	13	30	
698	Pakistan Vs England	England	7	Liam Livingstone		1	1	3	

699 rows × 12 columns

In [32]:
```
batting_df['match_id'] = batting_df['match'].map(match_ids_dict)
batting_df.head()
```

Out[32]:

	match	teamInnings	battingPos	batsmanName	dismissal	runs	balls	hidden	4s
0	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	c Pramod Madushan b Chameera	3	6	7	0
1	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	c Shanaka b Pramod Madushan	9	9	15	1
2	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	c †Mendis b Karunaratne	20	12	18	1
3	Namibia Vs Sri Lanka	Namibia	4	Stephan Baard	c DM de Silva b Pramod Madushan	26	24	49	2
4	Namibia Vs Sri Lanka	Namibia	5	Gerhard Erasmus (c)	c Gunathilaka b PWH de Silva	20	24	30	0

In [36]:
```
batting_df.drop(columns=["dismissal"], inplace=True)
batting_df.drop(columns=["hidden"], inplace=True)
batting_df.head(10)
```

Out[36]:

	match	teamInnings	battingPos	batsmanName	runs	balls	4s	6s	SR	out/not_o
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	not_o
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	

Cleanup weird characters

In [37]:
```
batting_df['batsmanName'] = batting_df['batsmanName'].apply(lambda x: x.replace('â€
batting_df['batsmanName'] = batting_df['batsmanName'].apply(lambda x: x.replace('\x
batting_df.head()
```

Out[37]:

	match	teamInnings	battingPos	batsmanName	runs	balls	4s	6s	SR	out/not_o
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 [38]:
```
batting_df.shape
```

Out[38]:
```
(699, 11)
```

In [40]:
```
batting_df.to_csv('batting_summary.csv', index = False)
```

(3) Process Bowling Summary

In [59]:
```
#reading the bowling summary csv
bowling_df = pd.read_csv('bowling_summary.csv')
bowling_df
```

Out[59]:

	match	bowlingTeam	bowlerName	overs	maiden	runs	wickets	economy	0s	4s
0	Namibia Vs Sri Lanka	Sri Lanka	Maheesh Theekshana	4.0	0	23	1	5.75	7	0
1	Namibia Vs Sri Lanka	Sri Lanka	Dushmantha Chameera	4.0	0	39	1	9.75	6	3
2	Namibia Vs Sri Lanka	Sri Lanka	Pramod Madushan	4.0	0	37	2	9.25	6	3
3	Namibia Vs Sri Lanka	Sri Lanka	Chamika Karunaratne	4.0	0	36	1	9.00	7	3
4	Namibia Vs Sri Lanka	Sri Lanka	Wanindu Hasaranga	4.0	0	27	1	6.75	8	1
...	...	...	...	...	...	...	...	...	...	...
495	Pakistan Vs England	Pakistan	Naseem Shah	4.0	0	30	0	7.50	15	3
496	Pakistan Vs England	Pakistan	Haris Rauf	4.0	0	23	2	5.75	13	3
497	Pakistan Vs England	Pakistan	Shadab Khan	4.0	0	20	1	5.00	10	1
498	Pakistan Vs England	Pakistan	Mohammad Wasim	4.0	0	38	1	9.50	5	5
499	Pakistan Vs England	Pakistan	Iftikhar Ahmed	0.5	0	13	0	15.60	0	1

500 rows × 14 columns

In [42]:
```
bowling_df['match_id'] = bowling_df['match'].map(match_ids_dict)
bowling_df.head()
```

Out[42]:

	match	bowlingTeam	bowlerName	overs	maiden	runs	wickets	economy	0s	4s
0	Namibia Vs Sri Lanka	Sri Lanka	Maheesh Theekshana	4.0	0	23	1	5.75	7	0
1	Namibia Vs Sri Lanka	Sri Lanka	Dushmantha Chameera	4.0	0	39	1	9.75	6	3
2	Namibia Vs Sri Lanka	Sri Lanka	Pramod Madushan	4.0	0	37	2	9.25	6	3
3	Namibia Vs Sri Lanka	Sri Lanka	Chamika Karunaratne	4.0	0	36	1	9.00	7	3
4	Namibia Vs Sri Lanka	Sri Lanka	Wanindu Hasaranga	4.0	0	27	1	6.75	8	1

In [43]:
```
bowling_df.to_csv('bowling_summary.csv', index = False)
```

(4) Process Players Information

In [60]:
```
#reading players information csv
player_df = pd.read_csv('player_info.csv')
player_df
```

Out[60]:

	Name	Team	Batting Style	Bowling Style	Playing Role	Description	name
0	Michael van Lingen	Namibia	Left hand Bat	Left arm Medium, Slow Left arm Orthodox	Bowling Allrounder	NaN	Michael van Lingen
1	Divan la Cock	Namibia	Right hand Bat	Legbreak	Opening Batter	NaN	Divan la Cock
2	Jan Nicol Loftie-Eaton	Namibia	Left hand Bat	Right arm Medium, Legbreak	Batter	NaN	Jan Nicol Loftie-Eaton
3	Stephan Baard	Namibia	Right hand Bat	Right arm Medium fast	Batter	NaN	Stephan Baard
4	Gerhard Erasmus (c)	Namibia	Right hand Bat	Right arm Offbreak	Allrounder	NaN	Gerhard Erasmus(c)
...	...	...	...	...	...	...	...
1194	Naseem Shah	Pakistan	Right hand Bat	Right arm Fast	Bowler	Zarai Taraqiati Bank Limited may not be an est...	Naseem Shah
1195	Haris Rauf	Pakistan	Right hand Bat	Right arm Fast	Bowler	NaN	Haris Rauf
1196	Shadab Khan	Pakistan	Right hand Bat	Legbreak	Allrounder	A prodigious turner of the ball, teenage legsp...	Shadab Khan
1197	Mohammad Wasim	Pakistan	Right hand Bat	Right arm Fast medium	Bowling Allrounder	NaN	Mohammad Wasim
1198	Iftikhar Ahmed	Pakistan	Right hand Bat	Right arm Offbreak	Middle order Batter	NaN	Iftikhar Ahmed

1199 rows × 7 columns

In [52]:
```
player_df['Description'] = player_df['Description'].fillna('')
```

In [55]:
```
player_df['name'] = player_df['Name'].apply(lambda x: x.replace('â€', ''))
player_df['name'] = player_df['Name'].apply(lambda x: x.replace('†', ''))
```

```
player_df['name'] = player_df['Name'].apply(lambda x: x.replace('\xa0', ''))
player_df.head(10)
```

Out[55]:

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

In [56]:
```
player_df[player_df['Team'] == 'India']
```

Out[56]:

	Name	Team	Batting Style	Bowling Style	Playing Role	Description	name
440	KL Rahul	India	Right hand Bat	NaN	Wicketkeeper Batter	A tall, elegant right-hand batsman who can kee...	KL Rahul
441	Rohit Sharma (c)	India	Right hand Bat	Right arm Offbreak	Top order Batter	Languid and easy on the eye, Rohit Sharma owne...	Rohit Sharma(c)
442	Virat Kohli	India	Right hand Bat	Right arm Medium	Top order Batter	India has given to the world many a great cric...	Virat Kohli
443	Suryakumar Yadav	India	Right hand Bat	Right arm Medium, Right arm Offbreak	Batter	Hard-hitting 360-degree batter Suryakumar Yada...	Suryakumar Yadav
444	Axar Patel	India	Left hand Bat	Slow Left arm Orthodox	Bowling Allrounder	Left-arm spinner Axar Patel has been increasin...	Axar Patel
...	...	...	...	...	...	...	...
1165	Arshdeep Singh	India	Left hand Bat	Left arm Medium fast	Bowler		Arshdeep Singh
1166	Axar Patel	India	Left hand Bat	Slow Left arm Orthodox	Bowling Allrounder	Left-arm spinner Axar Patel has been increasin...	Axar Patel
1167	Mohammed Shami	India	Right hand Bat	Right arm Fast	Bowler	Mohammed Shami was India's leading fast bowler...	Mohammed Shami
1168	Ravichandran Ashwin	India	Right hand Bat	Right arm Offbreak	Bowling Allrounder	R Ashwin took the tricks and skills he learned...	Ravichandran Ashwin

	Name	Team	Batting Style	Bowling Style	Playing Role	Description	name
1169	Hardik Pandya	India	Right hand Bat	Right arm Medium fast	Allrounder	Allrounder Hardik Pandya's calling cards brisk...	Hardik Pandya

80 rows × 7 columns

In [57]:
```
player_df.to_csv('player_info.csv', index = False)
```