T20 World Cup Cricket Data Collection

Match results collection

In [1]:
```
#import libraries
import requests
from bs4 import BeautifulSoup
import pandas as pd

#specify the url we want to scrape from
Link = "https://www.espncricinfo.com/records/tournament/team-match-results/icc-men-
Link_text = requests.get(Link).text
soup = BeautifulSoup(Link_text, 'html.parser')
soup.prettify()

#fetch the table for match summary
our_table = soup.find('table', class_= 'ds-w-full ds-table ds-table-xs ds-table-aut
```

In [2]:
```
#match summary details
match_data = []
for row in our_table.find_all('tr'):
    row_data = []
    for cell in row.find_all('td'):
        row_data.append(cell.text)
    match_data.append(row_data)
```

In [3]:
```
#match summary dataframe
match_df = pd.DataFrame(match_data)
match_df
```

Out[3]:

	0	1	2	3	4	5	6
0	Team 1	Team 2	Winner	Margin	Ground	Match Date	Scorecard
1	Namibia	Sri Lanka	Namibia	55 runs	Geelong	Oct 16, 2022	T20I # 1823
2	Netherlands	U.A.E.	Netherlands	3 wickets	Geelong	Oct 16, 2022	T20I # 1825
3	Scotland	West Indies	Scotland	42 runs	Hobart	Oct 17, 2022	T20I # 1826
4	Ireland	Zimbabwe	Zimbabwe	31 runs	Hobart	Oct 17, 2022	T20I # 1828
5	Namibia	Netherlands	Netherlands	5 wickets	Geelong	Oct 18, 2022	T20I # 1830
6	Sri Lanka	U.A.E.	Sri Lanka	79 runs	Geelong	Oct 18, 2022	T20I # 1832
7	Ireland	Scotland	Ireland	6 wickets	Hobart	Oct 19, 2022	T20I # 1833
8	West Indies	Zimbabwe	West Indies	31 runs	Hobart	Oct 19, 2022	T20I # 1834
9	Netherlands	Sri Lanka	Sri Lanka	16 runs	Geelong	Oct 20, 2022	T20I # 1835
10	Namibia	U.A.E.	U.A.E.	7 runs	Geelong	Oct 20, 2022	T20I # 1836
11	Ireland	West Indies	Ireland	9 wickets	Hobart	Oct 21, 2022	T20I # 1837
12	Scotland	Zimbabwe	Zimbabwe	5 wickets	Hobart	Oct 21, 2022	T20I # 1838
13	Australia	New Zealand	New Zealand	89 runs	Sydney	Oct 22, 2022	T20I # 1839
14	Afghanistan	England	England	5 wickets	Perth	Oct 22, 2022	T20I # 1840
15	Ireland	Sri Lanka	Sri Lanka	9 wickets	Hobart	Oct 23, 2022	T20I # 1841
16	India	Pakistan	India	4 wickets	Melbourne	Oct 23, 2022	T20I # 1842
17	Bangladesh	Netherlands	Bangladesh	9 runs	Hobart	Oct 24, 2022	T20I # 1843
18	South Africa	Zimbabwe	no result	-	Hobart	Oct 24, 2022	T20I # 1844

	0	1	2	3	4	5	6
19	Australia	Sri Lanka	Australia	7 wickets	Perth	Oct 25, 2022	T20I # 1845
20	England	Ireland	Ireland	5 runs	Melbourne	Oct 26, 2022	T20I # 1846
21	Bangladesh	South Africa	South Africa	104 runs	Sydney	Oct 27, 2022	T20I # 1847
22	India	Netherlands	India	56 runs	Sydney	Oct 27, 2022	T20I # 1848
23	Pakistan	Zimbabwe	Zimbabwe	1 run	Perth	Oct 27, 2022	T20I # 1849
24	New Zealand	Sri Lanka	New Zealand	65 runs	Sydney	Oct 29, 2022	T20I # 1850
25	Bangladesh	Zimbabwe	Bangladesh	3 runs	Brisbane	Oct 30, 2022	T20I # 1851
26	Netherlands	Pakistan	Pakistan	6 wickets	Perth	Oct 30, 2022	T20I # 1852
27	India	South Africa	South Africa	5 wickets	Perth	Oct 30, 2022	T20I # 1853
28	Australia	Ireland	Australia	42 runs	Brisbane	Oct 31, 2022	T20I # 1855
29	Afghanistan	Sri Lanka	Sri Lanka	6 wickets	Brisbane	Nov 1, 2022	T20I # 1856
30	England	New Zealand	England	20 runs	Brisbane	Nov 1, 2022	T20I # 1858
31	Netherlands	Zimbabwe	Netherlands	5 wickets	Adelaide	Nov 2, 2022	T20I # 1859
32	Bangladesh	India	India	5 runs	Adelaide	Nov 2, 2022	T20I # 1860
33	Pakistan	South Africa	Pakistan	33 runs	Sydney	Nov 3, 2022	T20I # 1861
34	Ireland	New Zealand	New Zealand	35 runs	Adelaide	Nov 4, 2022	T20I # 1862
35	Australia	Afghanistan	Australia	4 runs	Adelaide	Nov 4, 2022	T20I # 1864
36	England	Sri Lanka	England	4 wickets	Sydney	Nov 5, 2022	T20I # 1867
37	Netherlands	South Africa	Netherlands	13 runs	Adelaide	Nov 6, 2022	T20I # 1871

	0	1	2	3	4	5	6
38	Bangladesh	Pakistan	Pakistan	5 wickets	Adelaide	Nov 6, 2022	T20I # 1872
39	India	Zimbabwe	India	71 runs	Melbourne	Nov 6, 2022	T20I # 1873
40	New Zealand	Pakistan	Pakistan	7 wickets	Sydney	Nov 9, 2022	T20I # 1877
41	England	India	England	10 wickets	Adelaide	Nov 10, 2022	T20I # 1878
42	England	Pakistan	England	5 wickets	Melbourne	Nov 13, 2022	T20I # 1879

In [ ]:
```
#converting match summary dataframe to csv
match_df.to_csv('wc_match_results.csv',header=False, index=False)
```

In [4]:
```
#collecting match summary score links
score_links = []
for row in our_table.tbody.find_all('tr'):
    columns = row.find_all('td')[6]
    for td in columns:
        if td.a:
            link ="https://www.espncricinfo.com"+td.a.get('href')
            score_links.append(link)
```

Batting Summary Collection

In [5]:
```
#collecting match battting summary
battingSummary = [["match","teamInnings","battingPos","batsmanName","dismissal","ru
for link in score_links:
    detail = requests.get(link).text
    soup = BeautifulSoup(detail, 'html.parser')
    soup.prettify()
    team_info = soup.find_all('span', class_= 'ds-text-title-xs ds-font-bold ds-cap
    team1 = team_info[0].text
    team2 = team_info[1].text
    match_info = team1 +" Vs "+team2
    table = soup.find_all('table', class_= 'ci-scorecard-table')
    first_innings = table[0].tbody.find_all('tr')
    second_innings = table[1].tbody.find_all('tr')

    first_innings_index = 1
    for i in first_innings:
        row_data = []
        row_data.append(match_info)
        row_data.append(team1)
        row_data.append(first_innings_index)
        row_length = len(i.find_all('td'))
        if row_length == 8:
            for cell in i.find_all('td'):
                row_data.append(cell.text)
```

```
            battingSummary.append(row_data)
            first_innings_index = first_innings_index+1

    second_innings_index = 1
    for j in second_innings:
        row_data = []
        row_data.append(match_info)
        row_data.append(team2)
        row_data.append(second_innings_index)
        row_length = len(j.find_all('td'))
        if row_length == 8:
            for cell in j.find_all('td'):
                row_data.append(cell.text)
            battingSummary.append(row_data)
            second_innings_index = second_innings_index+1
```

In [6]:
```
#batting summary dataframe
batting_df = pd.DataFrame(battingSummary)
batting_df
```

Out[6]:

	0	1	2	3	4	5	6	7	8
0	match	teamInnings	battingPos	batsmanName	dismissal	runs	balls	hidden	4s
1	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	c Pramod Madushan b Chameera	3	6	7	0
2	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	c Shanaka b Pramod Madushan	9	9	15	1
3	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	c †Mendis b Karunaratne	20	12	18	1
4	Namibia Vs Sri Lanka	Namibia	4	Stephan Baard	c DM de Silva b Pramod Madushan	26	24	49	2
...	...	...	...	...	...	...	...	...	...
695	Pakistan Vs England	England	3	Phil Salt	c Iftikhar Ahmed b Haris Rauf	10	9	16	2
696	Pakistan Vs England	England	4	Ben Stokes	not out	52	49	81	5
697	Pakistan Vs England	England	5	Harry Brook	c Shaheen Shah Afridi b Shadab Khan	20	23	36	1
698	Pakistan Vs England	England	6	Moeen Ali	b Mohammad Wasim	19	13	30	3
699	Pakistan Vs England	England	7	Liam Livingstone	not out	1	1	3	0

700 rows × 11 columns

In [72]:
```
#converting the batting summary dataframe to csv
batting_df.to_csv("batting_summary.csv",header=False, index=False)
```

Bowling Summary Collection

In [7]:
```
#collecting match bowling summary
bowlingSummary = [["match","bowlingTeam","bowlerName","overs","maiden","runs","wick
for link in score_links:
    detail = requests.get(link).text
```

```
    soup = BeautifulSoup(detail, 'html.parser')
    soup.prettify()
    team_info = soup.find_all('span', class_= 'ds-text-title-xs ds-font-bold ds-cap
    team1 = team_info[0].text
    team2 = team_info[1].text
    match_info = team1 +" Vs "+team2
    table = soup.find_all('table')
    first_innings = table[1].tbody.find_all('tr')
    second_innings = table[3].tbody.find_all('tr')

    for i in first_innings:
        row_data = []
        row_data.append(match_info)
        row_data.append(team2)
        row_length = len(i.find_all('td'))
        if row_length == 11:
            for cell in i.find_all('td'):
                row_data.append(cell.text)
            bowlingSummary.append(row_data)

    for j in second_innings:
        row_data = []
        row_data.append(match_info)
        row_data.append(team1)
        row_length = len(j.find_all('td'))
        if row_length == 11:
            for cell in j.find_all('td'):
                row_data.append(cell.text)
            bowlingSummary.append(row_data)
```

In [8]:
```
#bowling summary dataframe
bowling_df = pd.DataFrame(bowlingSummary)
bowling_df
```

Out[8]:

	0	1	2	3	4	5	6	7	8	9	1
0	match	bowlingTeam	bowlerName	overs	maiden	runs	wickets	economy	0s	4s	6
1	Namibia Vs Sri Lanka	Sri Lanka	Maheesh Theekshana	4	0	23	1	5.75	7	0	
2	Namibia Vs Sri Lanka	Sri Lanka	Dushmantha Chameera	4	0	39	1	9.75	6	3	
3	Namibia Vs Sri Lanka	Sri Lanka	Pramod Madushan	4	0	37	2	9.25	6	3	
4	Namibia Vs Sri Lanka	Sri Lanka	Chamika Karunaratne	4	0	36	1	9.00	7	3	
...	...	...	...	...	...	...	...	...	...	...	
496	Pakistan Vs England	Pakistan	Naseem Shah	4	0	30	0	7.50	15	3	
497	Pakistan Vs England	Pakistan	Haris Rauf	4	0	23	2	5.75	13	3	
498	Pakistan Vs England	Pakistan	Shadab Khan	4	0	20	1	5.00	10	1	
499	Pakistan Vs England	Pakistan	Mohammad Wasim	4	0	38	1	9.50	5	5	
500	Pakistan Vs England	Pakistan	Iftikhar Ahmed	0.5	0	13	0	15.60	0	1	

501 rows × 13 columns

In [75]:
```
#converting bowling summary dataframe to csv
bowling_df.to_csv("bowling_summary.csv",header=False, index=False)
```

Player Information Collection

In [ ]:
```
#collecting the player information
player = [["Name","Team","Batting Style","Bowling Style","Playing Role","Descriptio
count =1
for link in score_links:
    count = count+1
    detail = requests.get(link).text
```

```
    soup = BeautifulSoup(detail, 'html.parser')
    soup.prettify()
    team_info = soup.find_all('span', class_= 'ds-text-title-xs ds-font-bold ds-cap
    team1 = team_info[0].text
    team2 = team_info[1].text
    #batting players
    table_batting = soup.find_all('table',class_= 'ci-scorecard-table')
    first_innings_batting = table_batting[0].tbody.find_all('tr')
    second_innings_batting = table_batting[1].tbody.find_all('tr')

    #bowling
    table_bowling = soup.find_all('table')
    first_innings_bowling = table_bowling[1].tbody.find_all('tr')
    second_innings_bowling = table_bowling[3].tbody.find_all('tr')

    for first_batting in first_innings_batting:
        # Find player url
        if len(first_batting) == 8:
            columns = first_batting.find_all('td')[0]
            if columns.find('a'):
                a_tag = columns.find('a')
                player_name = a_tag.text
                player_link = "https://www.espncricinfo.com"+a_tag['href']
                player_detail = requests.get(player_link).text
                soup = BeautifulSoup(player_detail, 'html.parser')
                soup.prettify()
                player_info = soup.find('div',class_='ds-grid').find_all('div')
                batting_style = ""
                bowling_style = ""
                playing_role = ""
                for data in player_info:
                    if data.p and data.p.text == "Batting Style":
                        batting_style = data.span.text
                    if data.p and data.p.text == "Bowling Style":
                        bowling_style = data.span.text
                    if data.p and data.p.text == "Playing Role":
                        playing_role = data.span.text
                player_content = ""
                if soup.find('div',class_='ci-player-bio-content'):
                    player_content = soup.find('div',class_='ci-player-bio-content'
                info = []
                info.append(player_name)
                info.append(team1)
                info.append(batting_style)
                info.append(bowling_style)
                info.append(playing_role)
                info.append(player_content)
                player.append(info)

    for second_batting in second_innings_batting:
        if len(second_batting) == 8:
            columns = second_batting.find_all('td')[0]
            if columns.find('a'):
                a_tag = columns.find('a')
                player_name = a_tag.text
                player_link = "https://www.espncricinfo.com"+a_tag['href']
```

```
                player_detail = requests.get(player_link).text
                soup = BeautifulSoup(player_detail, 'html.parser')
                soup.prettify()
                player_info = soup.find('div',class_='ds-grid').find_all('div')
                batting_style = ""
                bowling_style = ""
                playing_role = ""
                for data in player_info:
                    if data.p and data.p.text == "Batting Style":
                        batting_style = data.span.text
                    if data.p and data.p.text == "Bowling Style":
                        bowling_style = data.span.text
                    if data.p and data.p.text == "Playing Role":
                        playing_role = data.span.text
                player_content = ""
                if soup.find('div',class_='ci-player-bio-content'):
                    player_content = soup.find('div',class_='ci-player-bio-content'
                info = []
                info.append(player_name)
                info.append(team2)
                info.append(batting_style)
                info.append(bowling_style)
                info.append(playing_role)
                info.append(player_content)
                player.append(info)

    #bowling players
    for first_bowling in first_innings_bowling:
        # Find player url
        if len(first_bowling) == 11:
            columns = first_bowling.find_all('td')[0]
            if columns.find('a'):
                a_tag = columns.find('a')
                player_name = a_tag.text
                player_link = "https://www.espncricinfo.com"+a_tag['href']
                player_detail = requests.get(player_link).text
                soup = BeautifulSoup(player_detail, 'html.parser')
                soup.prettify()
                player_info = soup.find('div',class_='ds-grid').find_all('div')
                batting_style = ""
                bowling_style = ""
                playing_role = ""
                for data in player_info:
                    if data.p and data.p.text == "Batting Style":
                        batting_style = data.span.text
                    if data.p and data.p.text == "Bowling Style":
                        bowling_style = data.span.text
                    if data.p and data.p.text == "Playing Role":
                        playing_role = data.span.text
                player_content = ""
                if soup.find('div',class_='ci-player-bio-content'):
                    player_content = soup.find('div',class_='ci-player-bio-content'
                info = []
                info.append(player_name)
                info.append(team2)
                info.append(batting_style)
```

```
                info.append(bowling_style)
                info.append(playing_role)
                info.append(player_content)
                player.append(info)

    for second_bowling in second_innings_bowling:
        # Find player url
        if len(second_bowling) == 11:
            columns = second_bowling.find_all('td')[0]
            if columns.find('a'):
                a_tag = columns.find('a')
                player_name = a_tag.text
                player_link = "https://www.espncricinfo.com"+a_tag['href']
                player_detail = requests.get(player_link).text
                soup = BeautifulSoup(player_detail, 'html.parser')
                soup.prettify()
                player_info = soup.find('div',class_='ds-grid').find_all('div')
                batting_style = ""
                bowling_style = ""
                playing_role = ""
                for data in player_info:
                    if data.p and data.p.text == "Batting Style":
                        batting_style = data.span.text
                    if data.p and data.p.text == "Bowling Style":
                        bowling_style = data.span.text
                    if data.p and data.p.text == "Playing Role":
                        playing_role = data.span.text
                player_content = ""
                if soup.find('div',class_='ci-player-bio-content'):
                    player_content = soup.find('div',class_='ci-player-bio-content'
                info = []
                info.append(player_name)
                info.append(team1)
                info.append(batting_style)
                info.append(bowling_style)
                info.append(playing_role)
                info.append(player_content)
                player.append(info)
```

In [76]:

```
#player information dataframe
player_df = pd.DataFrame(player)
player_df
```

Out[76]:

	0	1	2	3	4	5
0	Name	Team	Batting Style	Bowling Style	Playing Role	Description
1	Michael van Lingen	Namibia	Left hand Bat	Left arm Medium, Slow Left arm Orthodox	Bowling Allrounder	
2	Divan la Cock	Namibia	Right hand Bat	Legbreak	Opening Batter	
3	Jan Nicol Loftie-Eaton	Namibia	Left hand Bat	Right arm Medium, Legbreak	Batter	
4	Stephan Baard	Namibia	Right hand Bat	Right arm Medium fast	Batter	
...	...	...	...	...	...	...
1195	Naseem Shah	Pakistan	Right hand Bat	Right arm Fast	Bowler	Zarai Taraqiati Bank Limited may not be an est...
1196	Haris Rauf	Pakistan	Right hand Bat	Right arm Fast	Bowler	
1197	Shadab Khan	Pakistan	Right hand Bat	Legbreak	Allrounder	A prodigious turner of the ball, teenage legsp...
1198	Mohammad Wasim	Pakistan	Right hand Bat	Right arm Fast medium	Bowling Allrounder	
1199	Iftikhar Ahmed	Pakistan	Right hand Bat	Right arm Offbreak	Middle order Batter	

1200 rows × 6 columns

In [61]:
```
#converting player information dataframe to csv
player_df.to_csv("player_info.csv",header=False, index=False)
```