# CRICKET

*Cricket Project Team*
*ICC Men’s T20 World Cup*

Padma Danturty, Ian Loree, Bryce Carson, Aryan Shah, Sam Dorfman, and Maddie Coe

# What is Cricket?

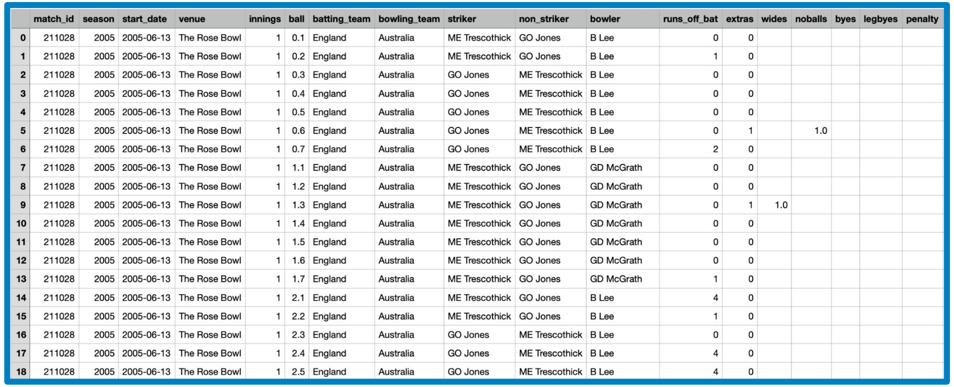
- Cricket is a game between two teams
  - Both teams comprise of 11 players
  - All teams have specialist batters, specialist bowlers, and all-rounders
- The main goal of the game is to score more runs than the opposing team in a set number of balls that are bowled by the other team, or until the team all gets out
- Runs are scored by running after hitting the ball or hitting a boundary
  - A boundary is set by a rope which makes the outermost circumference of the field
  - If the ball is hit into the boundary after it bounces then it is counted as 4 runs,
  - If the ball is hit directly passed the boundary, then it is counted as 6 runs

# Goals for the Project

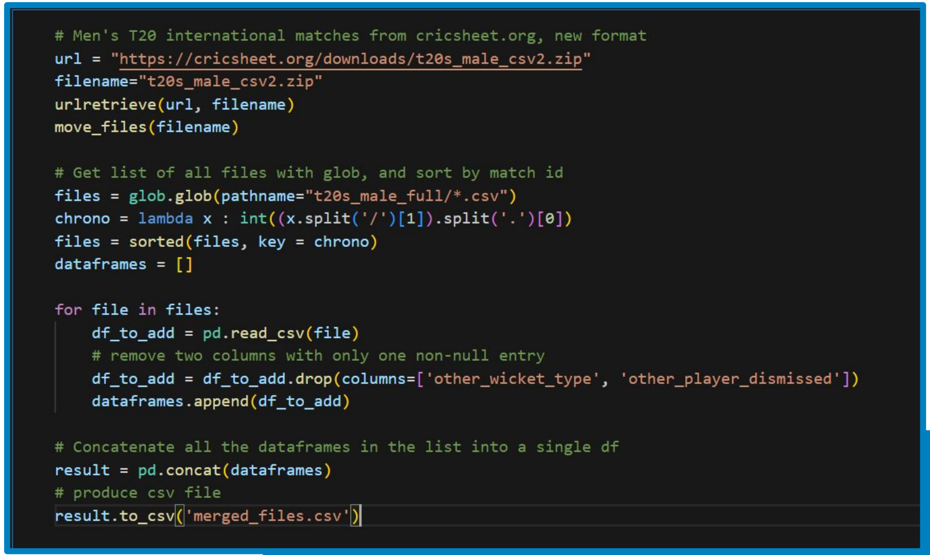
- To predict the outcome of all 55 matches in this summer’s ICC Men’s T20 World Cup
- Predict the run difference between the teams per match based on per-player statistics
- Predict the overall winner of the World Cup

# Data Collection

- We found a website[1] that includes ball-by-ball stats from every cricket international and club match
- We used a dataset incorporating every International T20 match since June 13, 2005
    - For our project, we used only matches since 2020
    - Updates after every match, so our data is current through the beginning of April



| | match_id | season | start_date | venue | innings | ball | batting_team | bowling_team | striker | non_striker | bowler | runs_off_bat | extras | wides | noballs | byes | legbyes | penalty |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 0.1 | England | Australia | ME Trescothick | GO Jones | B Lee | 0 | 0 | | | | | |
| 1 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 0.2 | England | Australia | ME Trescothick | GO Jones | B Lee | 1 | 0 | | | | | |
| 2 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 0.3 | England | Australia | GO Jones | ME Trescothick | B Lee | 0 | 0 | | | | | |
| 3 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 0.4 | England | Australia | GO Jones | ME Trescothick | B Lee | 0 | 0 | | | | | |
| 4 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 0.5 | England | Australia | GO Jones | ME Trescothick | B Lee | 0 | 0 | | | | | |
| 5 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 0.6 | England | Australia | GO Jones | ME Trescothick | B Lee | 0 | 1 | | 1.0 | | | |
| 6 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 0.7 | England | Australia | GO Jones | ME Trescothick | B Lee | 2 | 0 | | | | | |
| 7 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 1.1 | England | Australia | ME Trescothick | GO Jones | GD McGrath | 0 | 0 | | | | | |
| 8 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 1.2 | England | Australia | ME Trescothick | GO Jones | GD McGrath | 0 | 0 | | | | | |
| 9 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 1.3 | England | Australia | ME Trescothick | GO Jones | GD McGrath | 0 | 1 | 1.0 | | | | |
| 10 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 1.4 | England | Australia | ME Trescothick | GO Jones | GD McGrath | 0 | 0 | | | | | |
| 11 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 1.5 | England | Australia | ME Trescothick | GO Jones | GD McGrath | 0 | 0 | | | | | |
| 12 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 1.6 | England | Australia | ME Trescothick | GO Jones | GD McGrath | 0 | 0 | | | | | |
| 13 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 1.7 | England | Australia | ME Trescothick | GO Jones | GD McGrath | 1 | 0 | | | | | |
| 14 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 2.1 | England | Australia | ME Trescothick | GO Jones | B Lee | 4 | 0 | | | | | |
| 15 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 2.2 | England | Australia | ME Trescothick | GO Jones | B Lee | 1 | 0 | | | | | |
| 16 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 2.3 | England | Australia | GO Jones | ME Trescothick | B Lee | 0 | 0 | | | | | |
| 17 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 2.4 | England | Australia | GO Jones | ME Trescothick | B Lee | 4 | 0 | | | | | |
| 18 | 211028 | 2005 | 2005-06-13 | The Rose Bowl | 1 | 2.5 | England | Australia | GO Jones | ME Trescothick | B Lee | 4 | 0 | | | | | |



```python
# Men's T20 international matches from cricsheet.org, new format
url = "https://cricsheet.org/downloads/t20s_male_csv2.zip"
filename="t20s_male_csv2.zip"
urlretrieve(url, filename)
move_files(filename)

# Get list of all files with glob, and sort by match id
files = glob.glob(pathname="t20s_male_full/*.csv")
chrono = lambda x : int((x.split('/')[1]).split('.')[0])
files = sorted(files, key = chrono)
dataframes = []

for file in files:
    df_to_add = pd.read_csv(file)
    # remove two columns with only one non-null entry
    df_to_add = df_to_add.drop(columns=['other_wicket_type', 'other_player_dismissed'])
    dataframes.append(df_to_add)

# Concatenate all the dataframes in the list into a single df
result = pd.concat(dataframes)
# produce csv file
result.to_csv('merged_files.csv')
```

[1] https://cricsheet.org

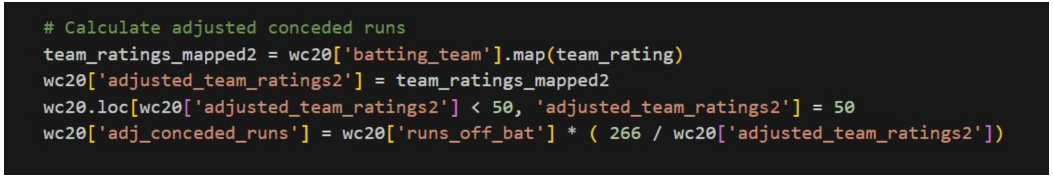
# Implementation

- We decided that bowler-batter matchups were important.
- So, we created a score for each team using 7 players that we thought were likely to make the squad for the World Cup.
  - Three bowlers
  - Three batters
  - One all-rounder
- We calculated the score based on runs/bowl and runs conceded/bowl, and adjusted based on the T20I overall ratings[2] due to the skill gaps between the countries.

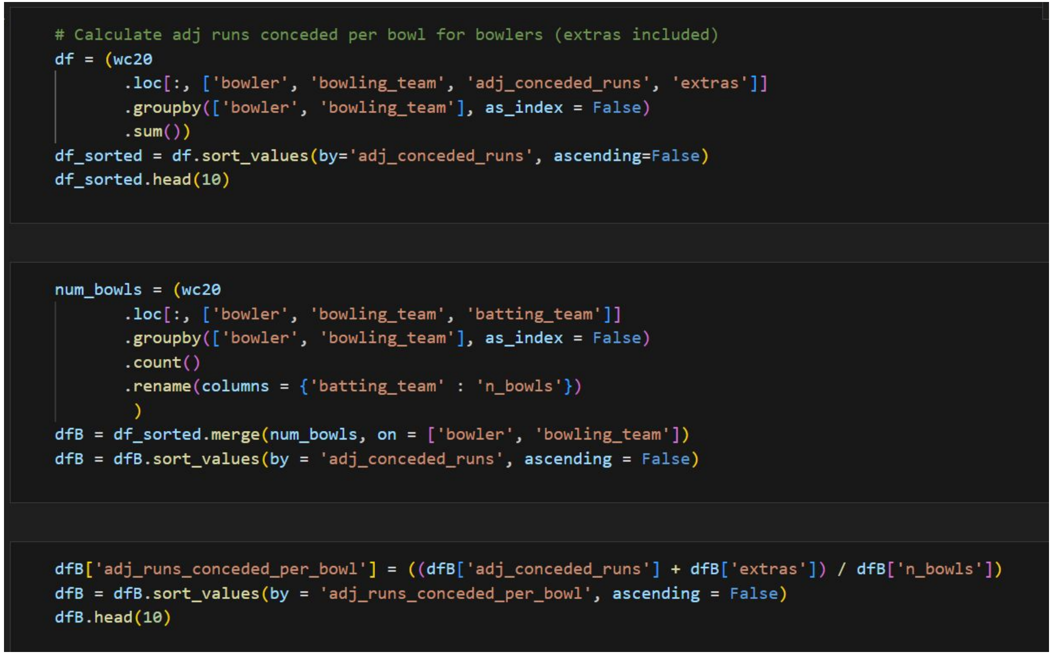
| Example: India Squad of 7 | |
|---|---|
| **Bowler 1** | Jasprit Bumrah |
| **Bowler 2** | Kuldeep Yadav |
| **Bowler 3** | Ravindra Jadeja |
| **Batter 1** | Rohit Sharma |
| **Batter 2** | Virat Kohli |
| **Batter 3** | Suryakumar Yadav |
| **All-Rounder** | Hardik Pandya |

[2] https://www.icc-cricket.com/rankings/team-rankings/mens/t20i

# Calculating Stats



```
# Calculate adjusted conceded runs
team_ratings_mapped2 = wc20['batting_team'].map(team_rating)
wc20['adjusted_team_ratings2'] = team_ratings_mapped2
wc20.loc[wc20['adjusted_team_ratings2'] < 50, 'adjusted_team_ratings2'] = 50
wc20['adj_conceded_runs'] = wc20['runs_off_bat'] * ( 266 / wc20['adjusted_team_ratings2'])
```



```
# Calculate adj runs conceded per bowl for bowlers (extras included)
df = (wc20
       .loc[:, ['bowler', 'bowling_team', 'adj_conceded_runs', 'extras']]
       .groupby(['bowler', 'bowling_team'], as_index = False)
       .sum())
df_sorted = df.sort_values(by='adj_conceded_runs', ascending=False)
df_sorted.head(10)
```

```
num_bowls = (wc20
       .loc[:, ['bowler', 'bowling_team', 'batting_team']]
       .groupby(['bowler', 'bowling_team'], as_index = False)
       .count()
       .rename(columns = {'batting_team' : 'n_bowls'})
        )
dfB = df_sorted.merge(num_bowls, on = ['bowler', 'bowling_team'])
dfB = dfB.sort_values(by = 'adj_conceded_runs', ascending = False)
```

```
dfB['adj_runs_conceded_per_bowl'] = ((dfB['adj_conceded_runs'] + dfB['extras']) / dfB['n_bowls'])
dfB = dfB.sort_values(by = 'adj_runs_conceded_per_bowl', ascending = False)
dfB.head(10)
```

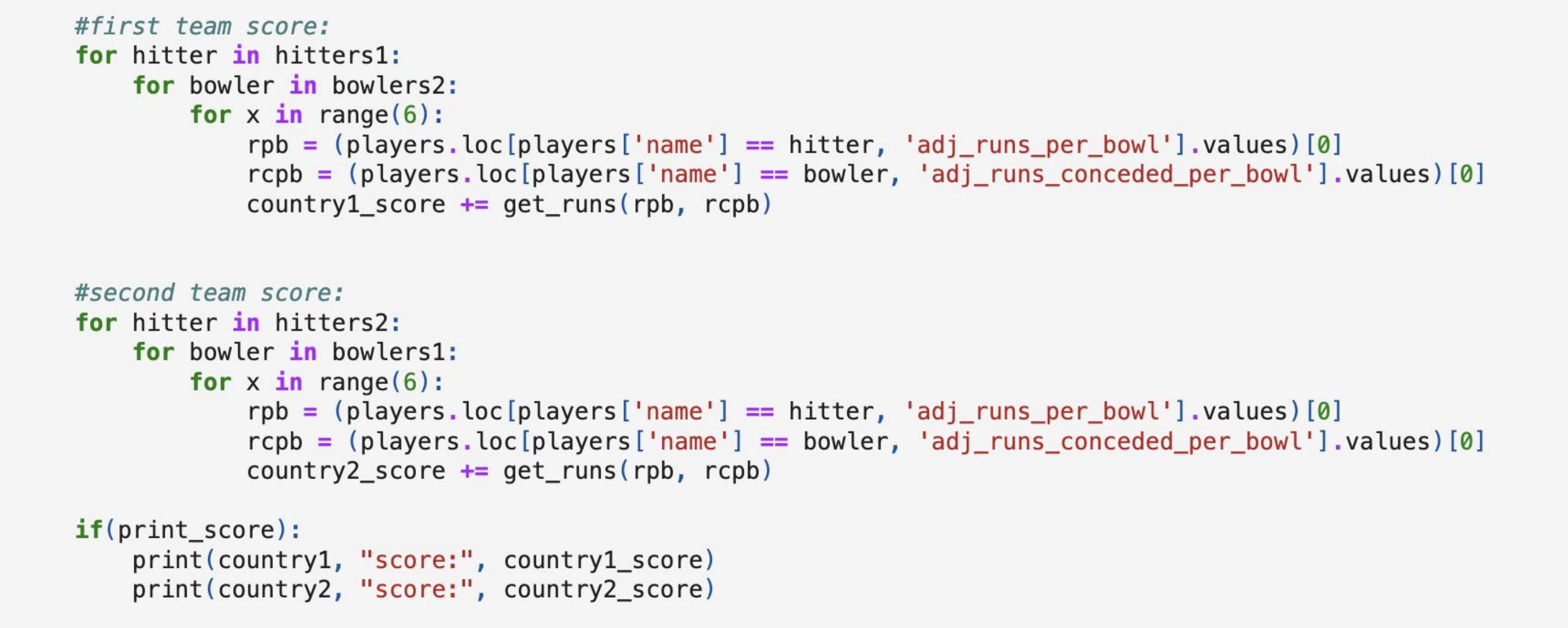
How we calculated the average runs conceded per bowl for each bowler:

1. Calculate adjusted conceded runs for each row
2. Aggregate total runs + bowls (balls thrown)
3. Calculate runs conceded per bowl

Similar process for runs/bowl

# Implementation - One Match

```
# 2 x 4 hitters x 4 bowlers x 6 bowls = 192 total bowls
```

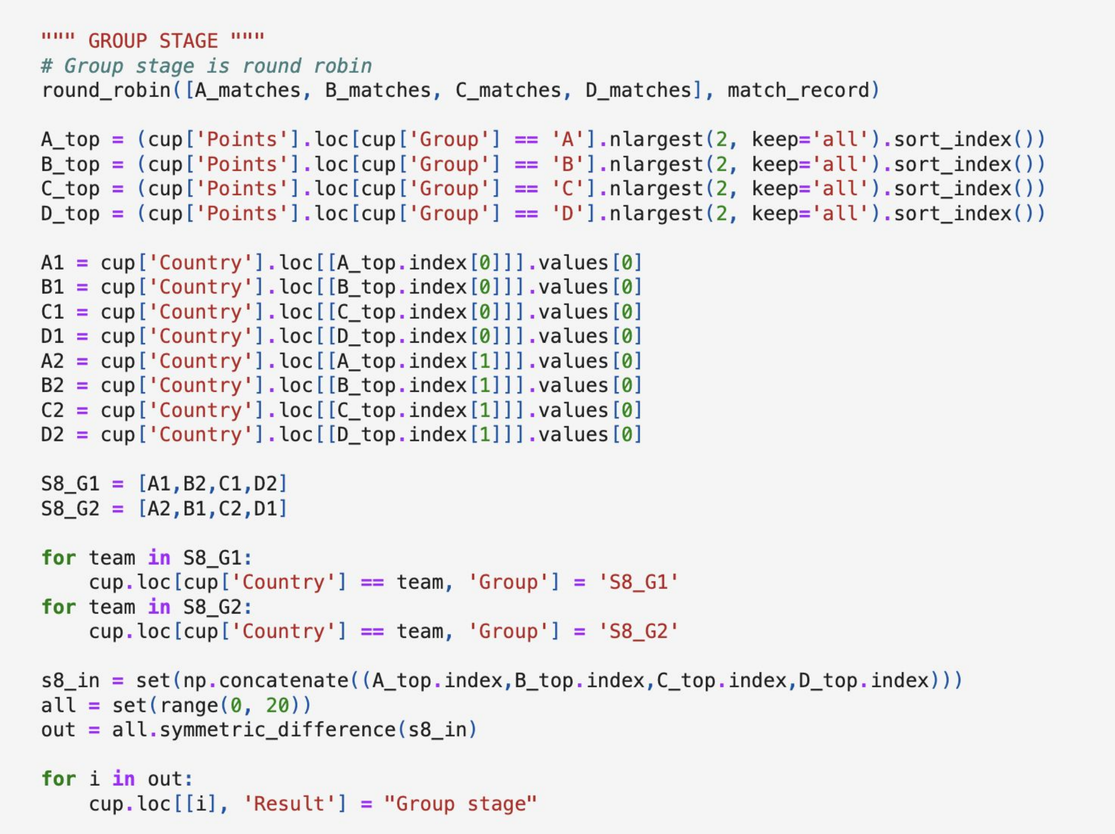


```
#first team score:
for hitter in hitters1:
    for bowler in bowlers2:
        for x in range(6):
            rpb = (players.loc[players['name'] == hitter, 'adj_runs_per_bowl'].values)[0]
            rcpb = (players.loc[players['name'] == bowler, 'adj_runs_conceded_per_bowl'].values)[0]
            country1_score += get_runs(rpb, rcpb)


#second team score:
for hitter in hitters2:
    for bowler in bowlers1:
        for x in range(6):
            rpb = (players.loc[players['name'] == hitter, 'adj_runs_per_bowl'].values)[0]
            rcpb = (players.loc[players['name'] == bowler, 'adj_runs_conceded_per_bowl'].values)[0]
            country2_score += get_runs(rpb, rcpb)

if(print_score):
    print(country1, "score:", country1_score)
    print(country2, "score:", country2_score)
```

# Implementation - Cup Simulation



```
""" GROUP STAGE """
# Group stage is round robin
round_robin([A_matches, B_matches, C_matches, D_matches], match_record)

A_top = (cup['Points'].loc[cup['Group'] == 'A'].nlargest(2, keep='all').sort_index())
B_top = (cup['Points'].loc[cup['Group'] == 'B'].nlargest(2, keep='all').sort_index())
C_top = (cup['Points'].loc[cup['Group'] == 'C'].nlargest(2, keep='all').sort_index())
D_top = (cup['Points'].loc[cup['Group'] == 'D'].nlargest(2, keep='all').sort_index())

A1 = cup['Country'].loc[[A_top.index[0]]].values[0]
B1 = cup['Country'].loc[[B_top.index[0]]].values[0]
C1 = cup['Country'].loc[[C_top.index[0]]].values[0]
D1 = cup['Country'].loc[[D_top.index[0]]].values[0]
A2 = cup['Country'].loc[[A_top.index[1]]].values[0]
B2 = cup['Country'].loc[[B_top.index[1]]].values[0]
C2 = cup['Country'].loc[[C_top.index[1]]].values[0]
D2 = cup['Country'].loc[[D_top.index[1]]].values[0]

S8_G1 = [A1,B2,C1,D2]
S8_G2 = [A2,B1,C2,D1]

for team in S8_G1:
    cup.loc[cup['Country'] == team, 'Group'] = 'S8_G1'
for team in S8_G2:
    cup.loc[cup['Country'] == team, 'Group'] = 'S8_G2'

s8_in = set(np.concatenate((A_top.index,B_top.index,C_top.index,D_top.index)))
all = set(range(0, 20))
out = all.symmetric_difference(s8_in)

for i in out:
    cup.loc[[i], 'Result'] = "Group stage"
```

# Group Stage Results - Group A

| Country | Points |
|---|---|
| India (A1) | 8 |
| Pakistan (A2) | 6 |
| Ireland | 4 |
| United States | 2 |
| Canada | 0 |

| Date | Winner | Score (W) | Loser | Score (L) |
|---|---|---|---|---|
| 6/1 | US | 153.186 | Canada | 97.569 |
| 6/5 | India | 138.925 | Ireland | 105.048 |
| 6/6 | Pakistan | 119.270 | US | 92.401 |
| 6/7 | Ireland | 179.190 | Canada | 96.883 |
| 6/9 | India | 128.767 | Pakistan | 113.669 |
| 6/11 | Pakistan | 179.653 | Canada | 94.462 |
| 6/12 | India | 118.593 | US | 88.964 |
| 6/14 | Ireland | 104.944 | US | 103.358 |
| 6/15 | India | 176.121 | Canada | 83.537 |
| 6/16 | Pakistan | 135.154 | Ireland | 111.420 |

# Group Stage Results - Group B

| Country | Points |
|---|---|
| England (B1) | 8 |
| Australia (B2) | 6 |
| Scotland | 4 |
| Namibia | 2 |
| Oman | 0 |

| Date | Winner | Score (W) | Loser | Score (L) |
|---|---|---|---|---|
| 6/2 | Namibia | 130.097 | Oman | 114.992 |
| 6/4 | England | 155.746 | Scotland | 106.191 |
| 6/5 | Australia | 163.500 | Oman | 90.715 |
| 6/6 | Scotland | 124.429 | Namibia | 122.168 |
| 6/8 | England | 134.839 | Australia | 130.385 |
| 6/9 | Scotland | 134.186 | Oman | 119.145 |
| 6/11 | Australia | 147.310 | Namibia | 97.018 |
| 6/13 | England | 162.266 | Oman | 92.204 |
| 6/15 | England | 145.145 | Namibia | 99.402 |
| 6/15 | Australia | 149.253 | Scotland | 105.288 |

# Group Stage Results - Group C

| Country | Points |
|---|---|
| New Zealand (C1) | 8 |
| West Indies (C2) | 6 |
| Afghanistan | 4 |
| Papua New Guinea | 2 |
| Uganda | 0 |

| Date | Winner | Score (W) | Loser | Score (L) |
|---|---|---|---|---|
| 6/2 | West Indies | 180.838 | PNG | 109.322 |
| 6/3 | Afghanistan | 172.464 | Uganda | 123.540 |
| 6/5 | PNG | 175.139 | Uganda | 145.353 |
| 6/7 | New Zealand | 160.863 | Afghanistan | 112.553 |
| 6/8 | West Indies | 212.850 | Uganda | 90.173 |
| 6/12 | New Zealand | 138.977 | West Indies | 124.374 |
| 6/13 | Afghanistan | 150.664 | PNG | 146.295 |
| 6/14 | New Zealand | 199.123 | Uganda | 85.651 |
| 6/17 | New Zealand | 171.238 | PNG | 89.578 |
| 6/17 | West Indies | 167.865 | Afghanistan | 122.551 |

# Group Stage Results - Group D

| Country | Points |
|---|---|
| South Africa (D1) | 8 |
| Bangladesh (D2) | 6 |
| Sri Lanka | 4 |
| Netherlands | 2 |
| Nepal | 0 |

| Date | Winner | Score (W) | Loser | Score (L) |
|---|---|---|---|---|
| 6/3 | South Africa | 130.292 | Sri Lanka | 122.275 |
| 6/4 | Netherlands | 152.571 | Nepal | 124.319 |
| 6/7 | Bangladesh | 121.576 | Sri Lanka | 115.186 |
| 6/6 | Scotland | 123.022 | Namibia | 120.206 |
| 6/8 | South Africa | 142.883 | Netherlands | 115.286 |
| 6/10 | South Africa | 126.987 | Bangladesh | 118.807 |
| 6/11 | Sri Lanka | 160.007 | Nepal | 108.431 |
| 6/13 | Bangladesh | 128.357 | Netherlands | 114.937 |
| 6/14 | South Africa | 172.825 | Nepal | 106.850 |
| 6/16 | Bangladesh | 139.477 | Nepal | 91.778 |

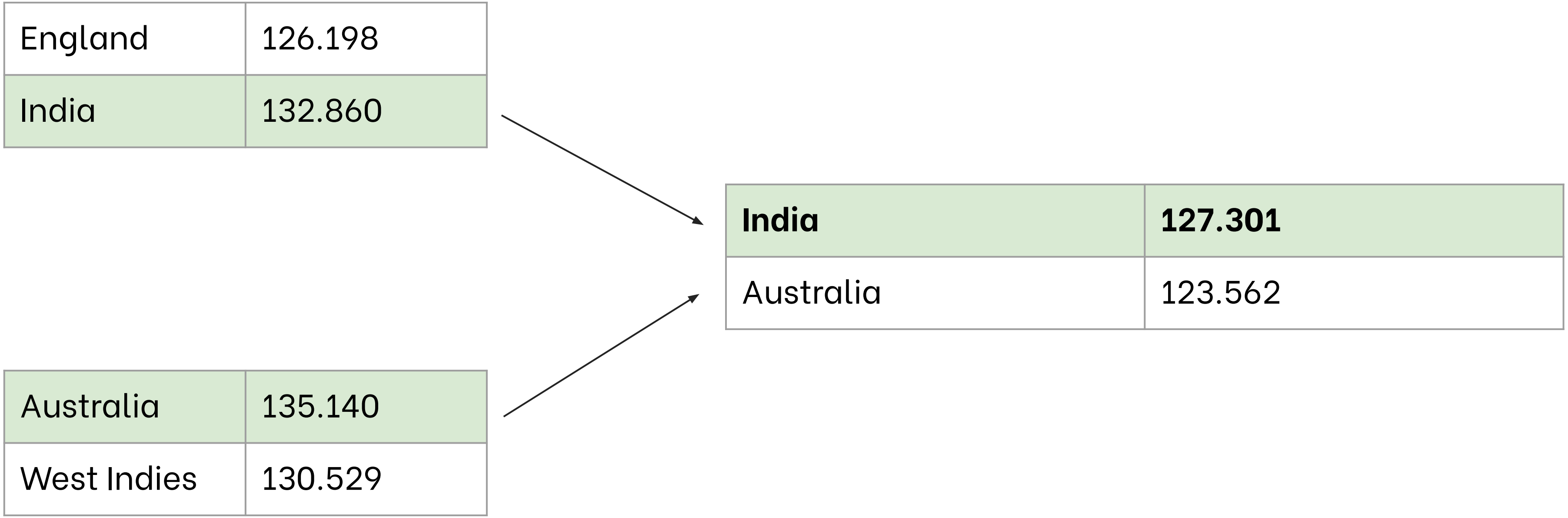
# Super Eight Groups

| Country | Points |
|---|---|
| Australia (B2) | 6 |
| India (A1) | 4 |
| New Zealand (C1) | 2 |
| Bangladesh (D2) | 0 |

| Country | Points |
|---|---|
| England (B1) | 6 |
| West Indies (C2) | 4 |
| Pakistan (A2) | 2 |
| South Africa (D1) | 0 |

| Date | Matchup | Winner | Score (W) | Loser | Score (L) |
|---|---|---|---|---|---|
| 6/19 | A2 v D1 | Pakistan | 131.895 | South Africa | 128.705 |
| 6/19 | B1 v C2 | England | 142.742 | West Indies | 136.773 |
| 6/20 | C1 v A1 | India | 125.348 | New Zealand | 123.034 |
| 6/20 | B2 v D2 | Australia | 133.880 | Bangladesh | 120.435 |
| 6/21 | B1 v D1 | England | 140.750 | South Africa | 130.676 |
| 6/21 | A2 v C2 | West Indies | 129.200 | Pakistan | 124.708 |
| 6/22 | A1 v D2 | India | 129.704 | Bangladesh | 111.409 |
| 6/22 | C1 v B2 | Australia | 131.554 | New Zealand | 128.109 |
| 6/23 | A2 v B1 | England | 132.469 | Pakistan | 124.453 |
| 6/23 | C2 v D1 | West Indies | 133.647 | South Africa | 132.109 |
| 6/24 | B2 v A1 | Australia | 127.070 | India | 125.643 |
| 6/24 | C1 v D2 | New Zealand | 125.287 | Bangladesh | 110.273 |

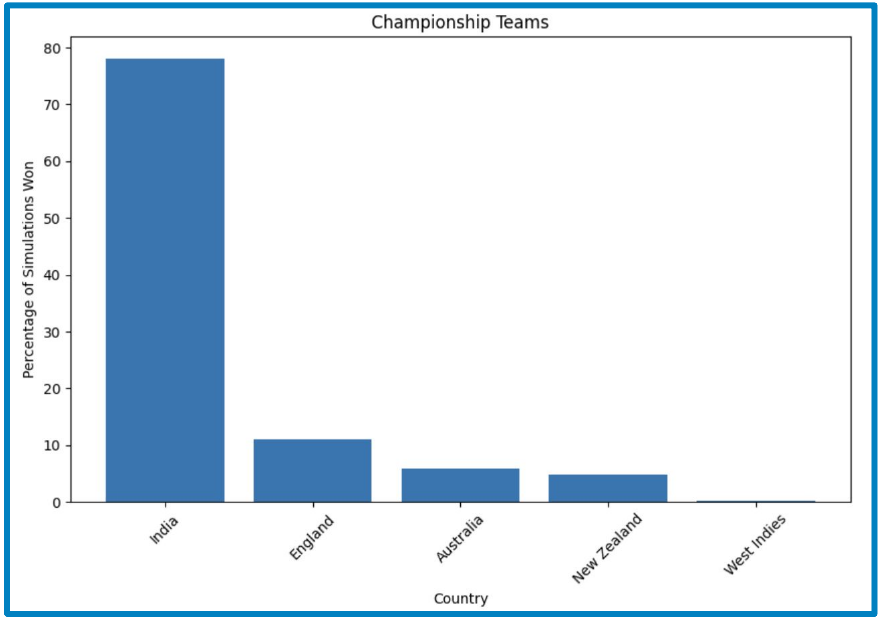
# Semi-Finals and Finals



| | |
|---|---|
| England | 126.198 |
| India | 132.860 |

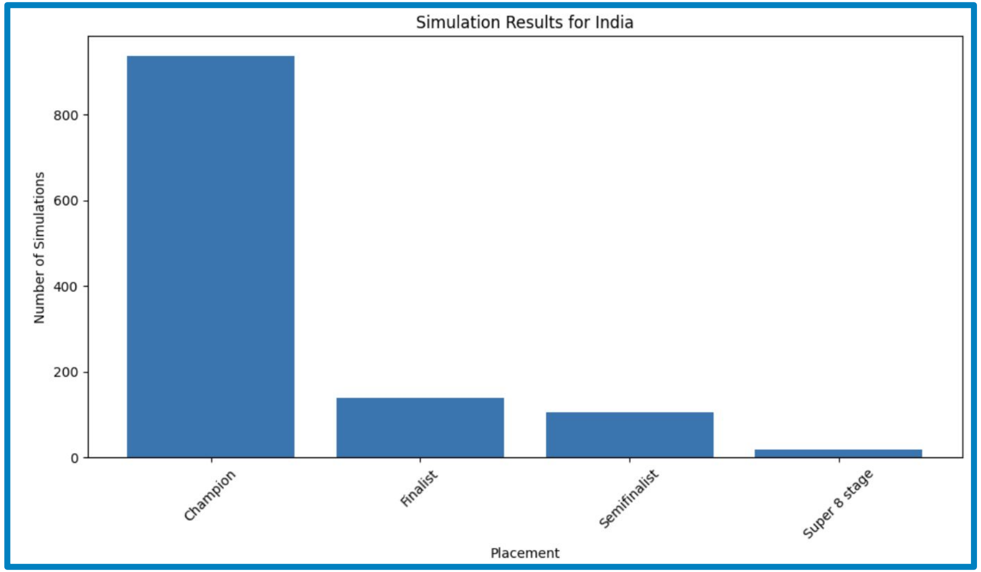
| | |
|---|---|
| Australia | 135.140 |
| West Indies | 130.529 |

| | |
|---|---|
| **India** | **127.301** |
| Australia | 123.562 |



India won the WC about 80% of the time

Where India ended in each simulation



# Notes and Conclusions

- Group Stages: The model favored Bangladesh (9th ICC) over Sri Lanka (8th ICC), and the US beating Canada!
- Super Eights: Half of the countries moving on were second in their group stages
- Semi-Finals: India, Australia, England, West Indies, and New Zealand are the favorites to make it to the Semi-Finals.
  - West Indies is ranked 7th, jumping over Pakistan (5th ICC) and South Africa (6th ICC).
- Finals: The final projection is India v Australia, with India winning the WC.
  - These are not the top two seeds, though India is the highest-rated country according to the ICC rankings.

# Areas for Improvement

- Data is currently limited for some teams
    - Needed to adjust player statistics to account for differences in team skill levels
    - We could incorporate match data from this year's World Cup into a future version of the project!
- Could expand model to predict on additional variables
    - Venue, wicket type, byes, penalties, etc.
- Could experiment with different types of distributions for getting runs