Project Report: Exploratory Data Analysis on Cricket World Cups (1975-2019)

Objective

The primary objective of this project is to conduct an exploratory data analysis on Cricket World Cup data from 1975 to 2019. The dataset consists of information from each World Cup, including match details, teams, venues, and outcomes. The project aims to import the data into a Jupyter Notebook, transfer it to a MySQL database, and perform various analyses and visualizations to uncover trends and patterns.

1. Data Import and Database Setup:

The initial step involved importing data from CSV files into the Jupyter Notebook. The dataset is divided into separate files for each World Cup year (1975-2019). After importing, the data was transferred to a MySQL database using the SQLAlchemy library.

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
from sqlalchemy import create_engine
import pymysql
```

In [3]:
```
database_connection = create_engine("mysql+pymysql://root:Yaadnhihai_1629@localhost/cric
```

In [4]:
```
conn = database_connection.connect()
```

In [5]:
```
WC_1975 = pd.read_csv("E:\sql\Cric_analysis\world_cup_1975.csv", encoding='latin1')
```

In [6]:
```
WC_1979 = pd.read_csv("E:\sql\Cric_analysis\world_cup_1979.csv", encoding='latin1')
```

In [7]:
```
WC_1983 = pd.read_csv("E:\sql\Cric_analysis\world_cup_1983.csv", encoding='latin1')
```

In [8]:
```
WC_1987 = pd.read_csv("E:\sql\Cric_analysis\world_cup_1987.csv", encoding='latin1')
```

In [9]:
```
WC_1992 = pd.read_csv("E:/sql/Cric_analysis/world_cup_1991.csv", encoding='latin1')
```

In [10]:
```
WC_1996 = pd.read_csv("E:\sql\Cric_analysis\world_cup_1995.csv", encoding='latin1')
```

In [11]:
```
WC_1999 = pd.read_csv("E:\sql\Cric_analysis\world_cup_1999.csv", encoding='latin1')
```

In [12]:
```
WC_2003 = pd.read_csv("E:\sql\Cric_analysis\world_cup_2003.csv", encoding='latin1')
```

In [13]:
```
WC_2007 = pd.read_csv("E:\sql\Cric_analysis\world_cup_2007.csv", encoding='latin1')
```

In [14]:
```
WC_2015 = pd.read_csv("E:\sql\Cric_analysis\world_cup_2015.csv", encoding='latin1')
```

```
WC_2019 = pd.read_csv("E:\sql\Cric_analysis\world_cup_2019.csv", encoding='latin1')
```

```
In [15]:
```

```
In [16]: WC_1975.head(30)
```

Out[16]:

	Id	Date	Country _1	Country_2	Result	Winner	Margin
0	1	07-06-1975	England	India	England won by 202 runs	England	202 runs
1	2	07-06-1975	East Africa	New Zealand	New Zealand won by 181 runs	New Zealand	181 runs
2	3	07-06-1975	Australia	Pakistan	Australia won by 73 runs	Australia	73 runs
3	4	07-06-1975	Sri Lanka	West Indies	West Indies won by 9 wickets	West Indies	9 wickets
4	5	11-06-1975	England	New Zealand	England won by 80 runs	England	80 runs
5	6	11-06-1975	East Africa	India	India won by 10 wickets	India	10 wickets
6	7	11-06-1975	Australia	Sri Lanka	Australia won by 52 runs	Australia	52 runs
7	8	11-06-1975	Pakistan	West Indies	West Indies won by 1 wicket	West Indies	1 wicket
8	9	14-06-1975	England	East Africa	England won by 196 runs	England	196 runs
9	10	14-06-1975	India	New Zealand	New Zealand won by 4 wickets	New Zealand	4 wickets
10	11	14-06-1975	Australia	West Indies	West Indies won by 7 wickets	West Indies	7 wickets
11	12	14-06-1975	Pakistan	Sri Lanka	Pakistan won by 192 runs	Pakistan	192 runs
12	13	18-06-1975	England	Australia	Australia won by 4 wickets	Australia	4 wickets
13	14	18-06-1975	New Zealand	West Indies	West Indies won by 5 wickets	West Indies	5 wickets
14	15	21-06-1975	Australia	West Indies	West Indies won by 17 runs	West Indies	17 runs

```
In [17]: WC_1975.to_sql("wc_1975",conn,if_exists = "replace",index="False")
         WC_1979.to_sql("wc_1979",conn,if_exists = "replace",index="False")
         WC_1983.to_sql("wc_1983",conn,if_exists = "replace",index="False")
         WC_1987.to_sql("wc_1987",conn,if_exists = "replace",index="False")
         WC_1992.to_sql("wc_1992",conn,if_exists = "replace",index="False")
         WC_1996.to_sql("wc_1996",conn,if_exists = "replace",index="False")
         WC_1999.to_sql("wc_1999",conn,if_exists = "replace",index="False")
         WC_2003.to_sql("wc_2003",conn,if_exists = "replace",index="False")
         WC_2007.to_sql("wc_2007",conn,if_exists = "replace",index="False")
         WC_2015.to_sql("wc_2015",conn,if_exists = "replace",index="False")
         WC_2019.to_sql("wc_2019",conn,if_exists = "replace",index="False")
```

```
Out[17]: 48
```

```
In [18]: Show_all_tables = pd.read_sql("SHOW TABLES;",con=database_connection)
```

```
In [19]: Show_all_tables
```

Out[19]:

	Tables_in_cric_stats
0	virat_kohli
1	wc_1975
2	wc_1979
3	wc_1983
4	wc_1987
5	wc_1992

6	wc_1996
7	wc_1999
8	wc_2003
9	wc_2007
10	wc_2011
11	wc_2015
12	wc_2019
13	world_cup_stats

2. Data Cleaning and Transformation:

The dataset underwent cleaning and transformation processes to ensure consistency and accuracy. Date columns were converted to the datetime format, and data types were adjusted as needed. This step ensures the reliability of subsequent analyses. Rest the Data cleaning was done in MS EXCEL to speed up the data cleaning process like stripping of columns

In [20]:

```
WC_1975.head()
```

Out[20]:

	Id	Date	Country _1	Country_2	Result	Winner	Margin
0	1	07-06-1975	England	India	England won by 202 runs	England	202 runs
1	2	07-06-1975	East Africa	New Zealand	New Zealand won by 181 runs	New Zealand	181 runs
2	3	07-06-1975	Australia	Pakistan	Australia won by 73 runs	Australia	73 runs
3	4	07-06-1975	Sri Lanka	West Indies	West Indies won by 9 wickets	West Indies	9 wickets
4	5	11-06-1975	England	New Zealand	England won by 80 runs	England	80 runs

In [21]:

```
WC_1975["Date"]=pd.to_datetime(WC_1975["Date"],format='%d-%m-%Y')
WC_1979["Date"]=pd.to_datetime(WC_1979["Date"],format='%d-%m-%Y')
WC_1983["Date"]=pd.to_datetime(WC_1983["Date"],format='%d-%m-%Y')
WC_1987["Date"]=pd.to_datetime(WC_1987["Date"],format='%d-%m-%Y')
WC_1992["Date"]=pd.to_datetime(WC_1992["Date"],format='%d-%m-%Y')
WC_1996["Date"]=pd.to_datetime(WC_1996["Date"],format='%d-%m-%Y')
WC_1999["Date"]=pd.to_datetime(WC_1999["Date"],format='%d-%m-%Y')
WC_2003["Date"]=pd.to_datetime(WC_2003["Date"],format='%d-%m-%Y')
WC_2007["Date"]=pd.to_datetime(WC_2007["Date"],format='%d-%m-%Y')
WC_2015["Date"]=pd.to_datetime(WC_2015["Date"],format='%d-%m-%Y')
WC_2019["Date"]=pd.to_datetime(WC_2019["Date"],format='%d-%m-%Y')
```

In [22]:

```
WC_1987.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27 entries, 0 to 26
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Id           27 non-null     int64
 1   Date         27 non-null     datetime64[ns]
 2   Counttry_1   27 non-null     object
 3   Country_2)   27 non-null     object
 4   Ground       27 non-null     object
 5   Result       27 non-null     object
 6   Winner       27 non-null     object
 7   Margin       27 non-null     object
dtypes: datetime64[ns](1), int64(1), object(6)
memory usage: 1.8+ KB
```

I have dumped the data into sql and changed the required datatypes. Now first of all we will run some basic queries of sql in jupyter notebook by using pd.read_sql command and find some of the trends and details of each world cup

3. SQL Queries and Analysis:

QUERY 1 - Total Matches In 1975 World Cup

In [23]:
```
Total_number_of_matches = pd.read_sql(""" SELECT
COUNT(*)
AS total_matches
FROM wc_1975""",con = database_connection)
```

In [24]:
```
Total_number_of_matches
```

Out[24]:

	total_matches
0	15

Query 2 - Total matches played at each Ground in 1979 WC

In [25]:
```
Matches_at_each_ground = pd.read_sql("""
SELECT Ground, COUNT(*) AS Matches_at_each_ground FROM wc_1979 GROUP BY Ground ORDER BY
```

In [26]:
```
Matches_at_each_ground
```

Out[26]:

	Ground	Matches_at_each_ground
0	Trent Bridge	3
1	Headingley	3
2	Old Trafford	3
3	Edgbaston	2
4	Lord's	2
5	Kennington Oval	2

QUERY 3 - Details of The final Match of 1983 WC

In [27]:
```
Final_match = pd.read_sql("""
SELECT * FROM wc_1983 ORDER BY Date DESC LIMIT 1 """, con = database_connection)
```

In [28]:
```
Final_match
```

Out[28]:

	False	Id	Date	Country 1	Country_2	Result	Winner	Margin

0	26	27	25-06-1983	India	West Indies	India won by 43 runs	India	43 runs

Query 4 - Matches Won By Each Team In 2015

In [29]:
```
Matches_won_by_each_team = pd.read_sql(""" SELECT Winner , COUNT(*) As Total_matches_won
wc_2015 WHERE Winner IS NOT  NULL GROUP BY Winner ORDER BY Total_matches_won_by_each_tea
""",con = database_connection)
```

In [30]:
```
Matches_won_by_each_team
```

Out[30]:

	Winner	Total_matches_won_by_each_team
0	New Zealand	8
1	Australia	7
2	India	7
3	South Africa	5
4	Pakistan	4
5	Sri Lanka	4
6	Bangladesh	3
7	Ireland	3
8	West Indies	3
9	England	2
10	Afghanistan	1
11	Zimbabwe	1

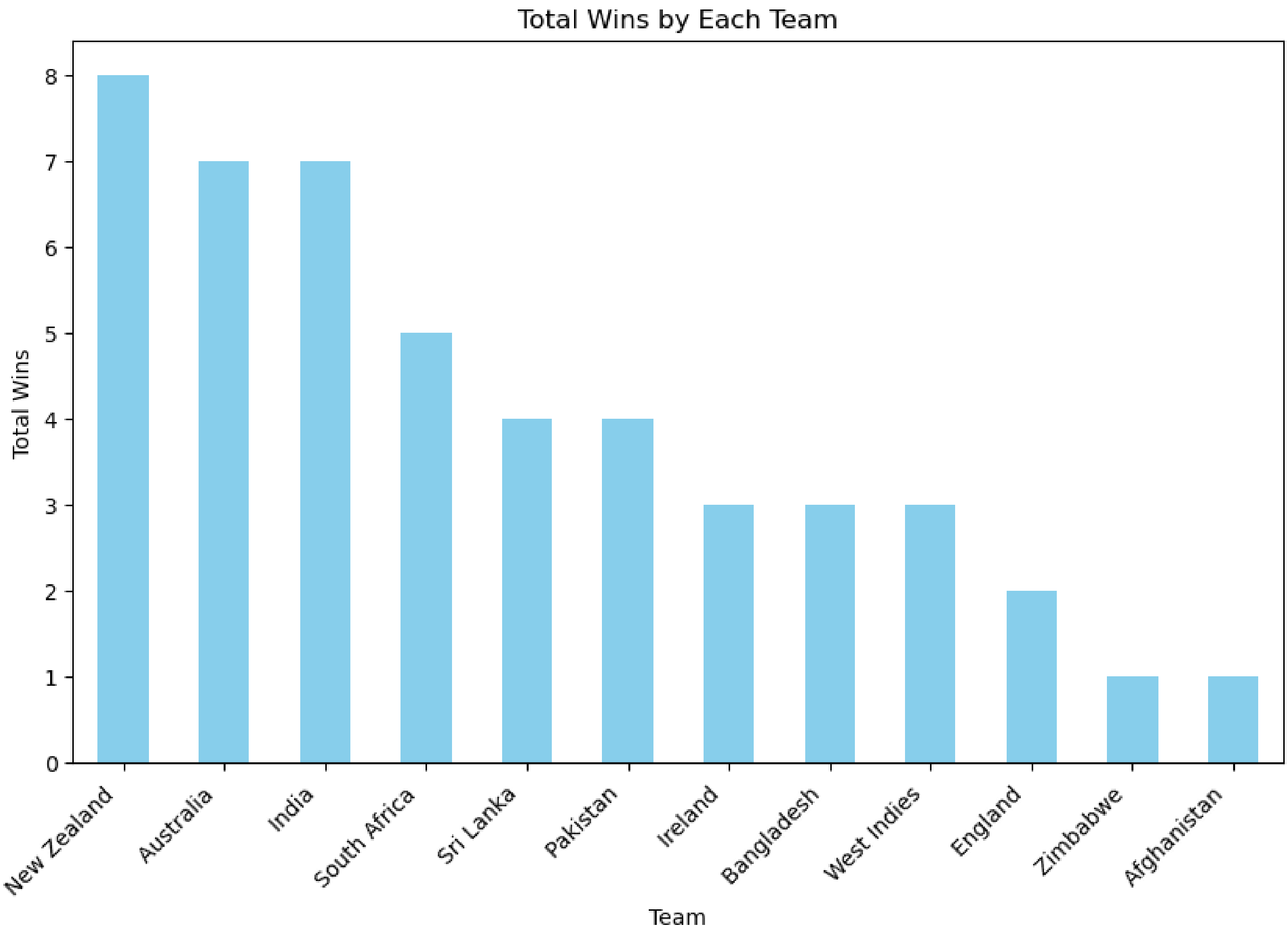
We have ran some of the basic queries for each dataset. Now we will be running some of the visualisation code for the above dataset and find out some trends.

4. Data Visualisation

Visualizations were created using Matplotlib and Seaborn to illustrate trends and patterns in the data. Examples include bar charts depicting total wins by each team, win percentages, heatmaps of match results, and distribution of matches for each country at different grounds.

Visualisation 1 - Total Wins By Each Team In WC_2015

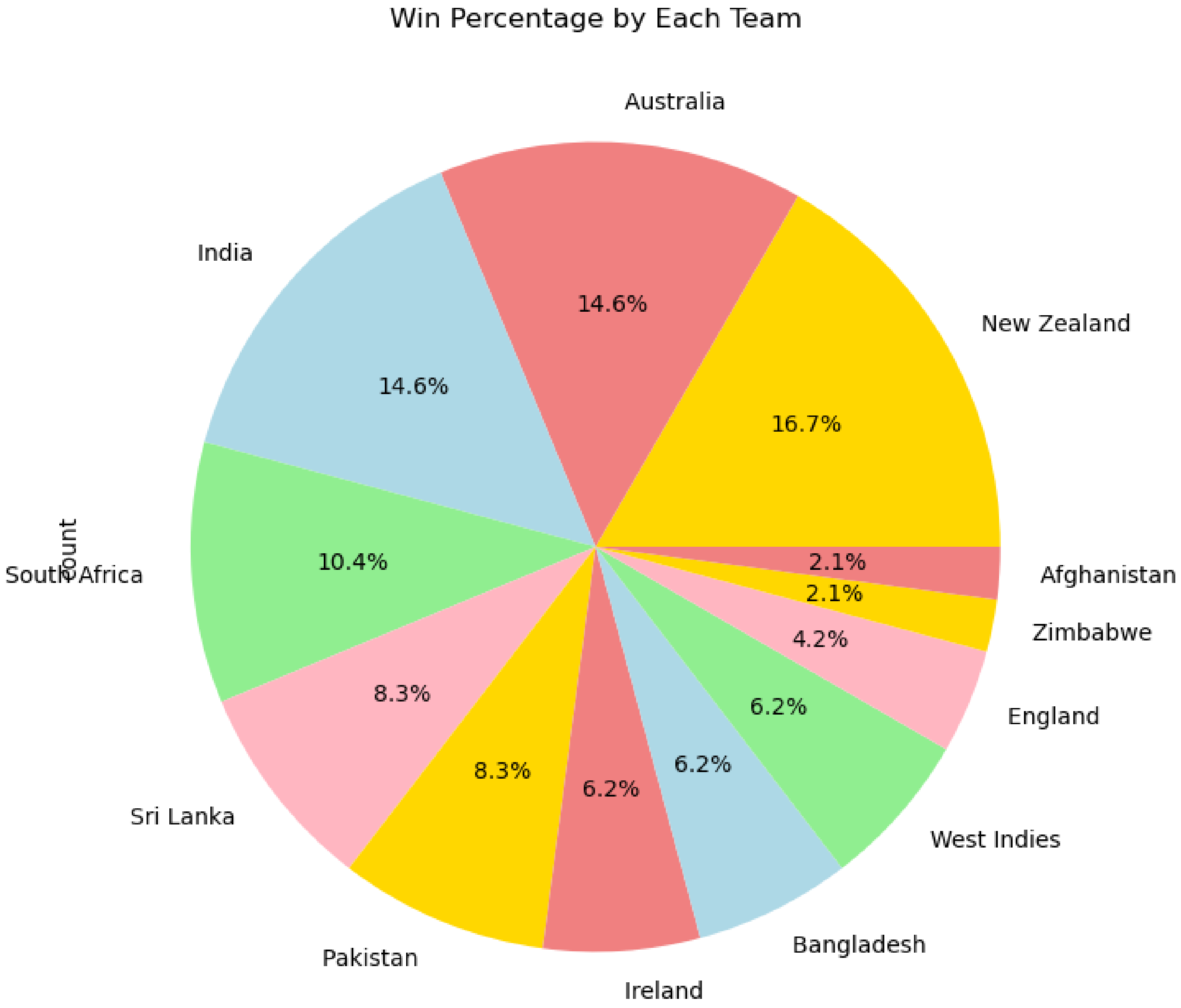
In [31]:
```
total_wins = WC_2015['Winner'].value_counts()

plt.figure(figsize=(10, 6))
total_wins.sort_values(ascending = False).plot(kind='bar', color='skyblue')
plt.title('Total Wins by Each Team')
plt.xlabel('Team')
plt.ylabel('Total Wins')
plt.xticks(rotation=45, ha='right')  # Rotate x-axis labels for better readability
plt.show()
```



Visualisation 2 - Win Percentage of Each team In WC_2015

In [32]:

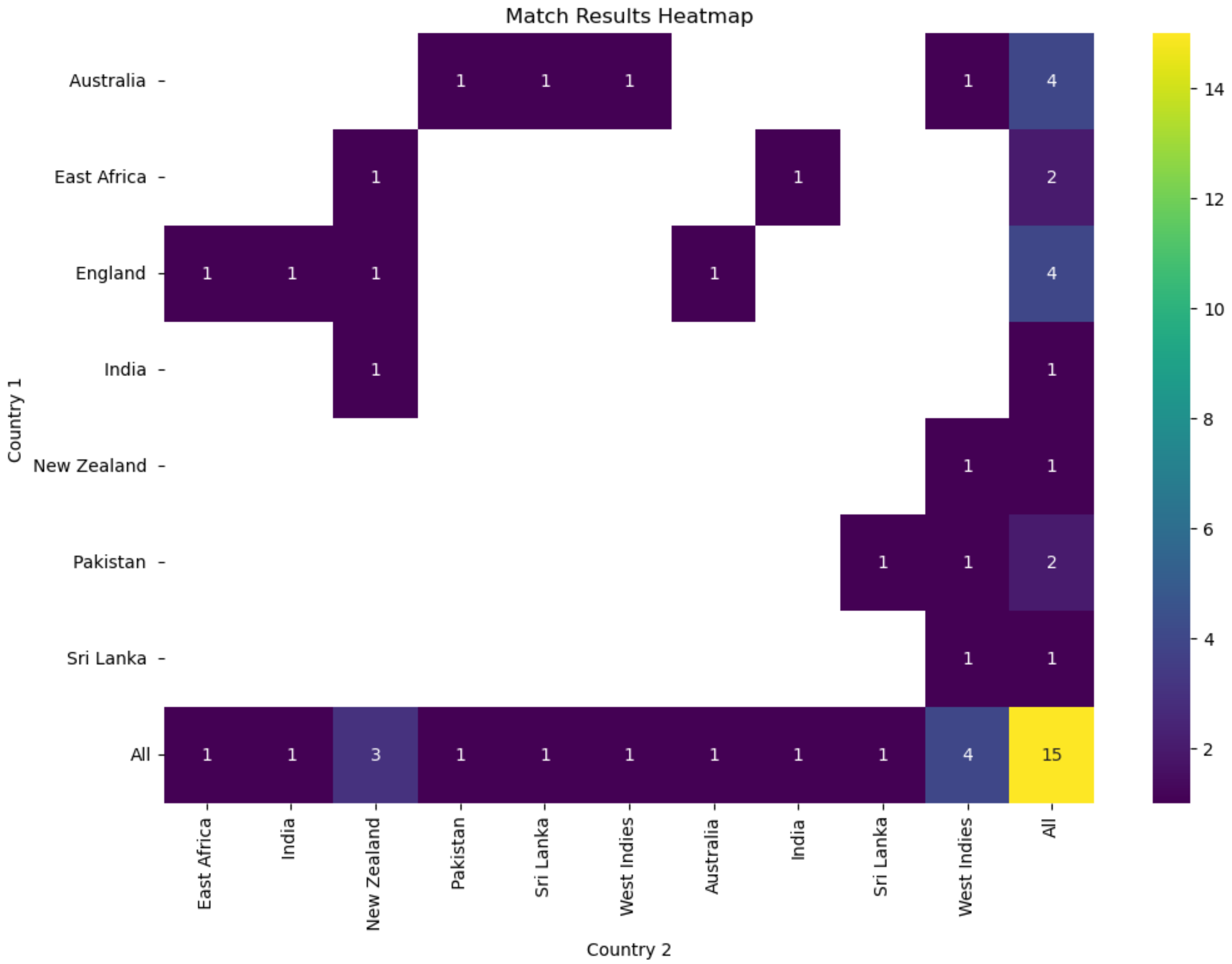
```
plt.figure(figsize=(8, 8))
win_percentage = total_wins / total_wins.sum() * 100
win_percentage.plot(kind='pie', autopct='%1.1f%%', colors=['gold', 'lightcoral', 'lightb
plt.title('Win Percentage by Each Team')
plt.show()
```



Visualisation 3 - HeatMap of every match for every team in WC 1975

In [33]:

```
result_matrix = pd.crosstab(WC_1975['Country _1'], WC_1975['Country_2'], values=WC_1975[

plt.figure(figsize=(12, 8))
sns.heatmap(result_matrix, annot=True, cmap='viridis', fmt='g', cbar=True)
plt.title('Match Results Heatmap')
plt.xlabel('Country 2')
plt.ylabel('Country 1')
plt.show()
```

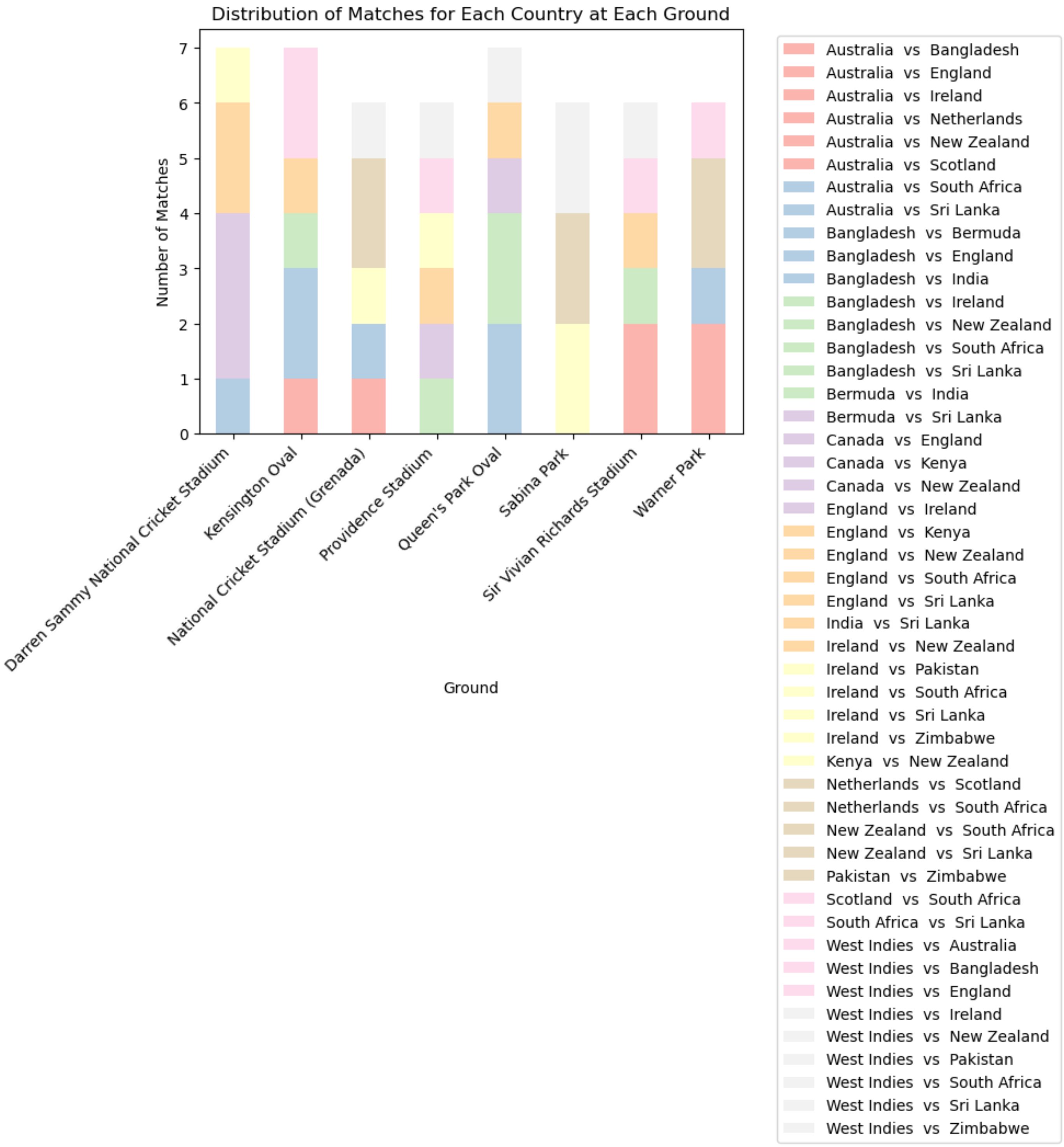


Visualisation 4 Matches played by each team at each ground in 2007 World Cup

In [34]:

```
WC_2007['Country'] = WC_2007['Country_1'] + ' vs ' + WC_2007['Country_2']

# Create a DataFrame for plotting
plot_df = WC_2007.groupby(['Ground', 'Country']).size().unstack().fillna(0)

# Plotting
plt.figure(figsize=(15, 10))
plot_df.plot(kind='bar', stacked=True, colormap='Pastel1')
plt.title('Distribution of Matches for Each Country at Each Ground')
plt.xlabel('Ground')
plt.ylabel('Number of Matches')
plt.xticks(rotation=45, ha='right')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```

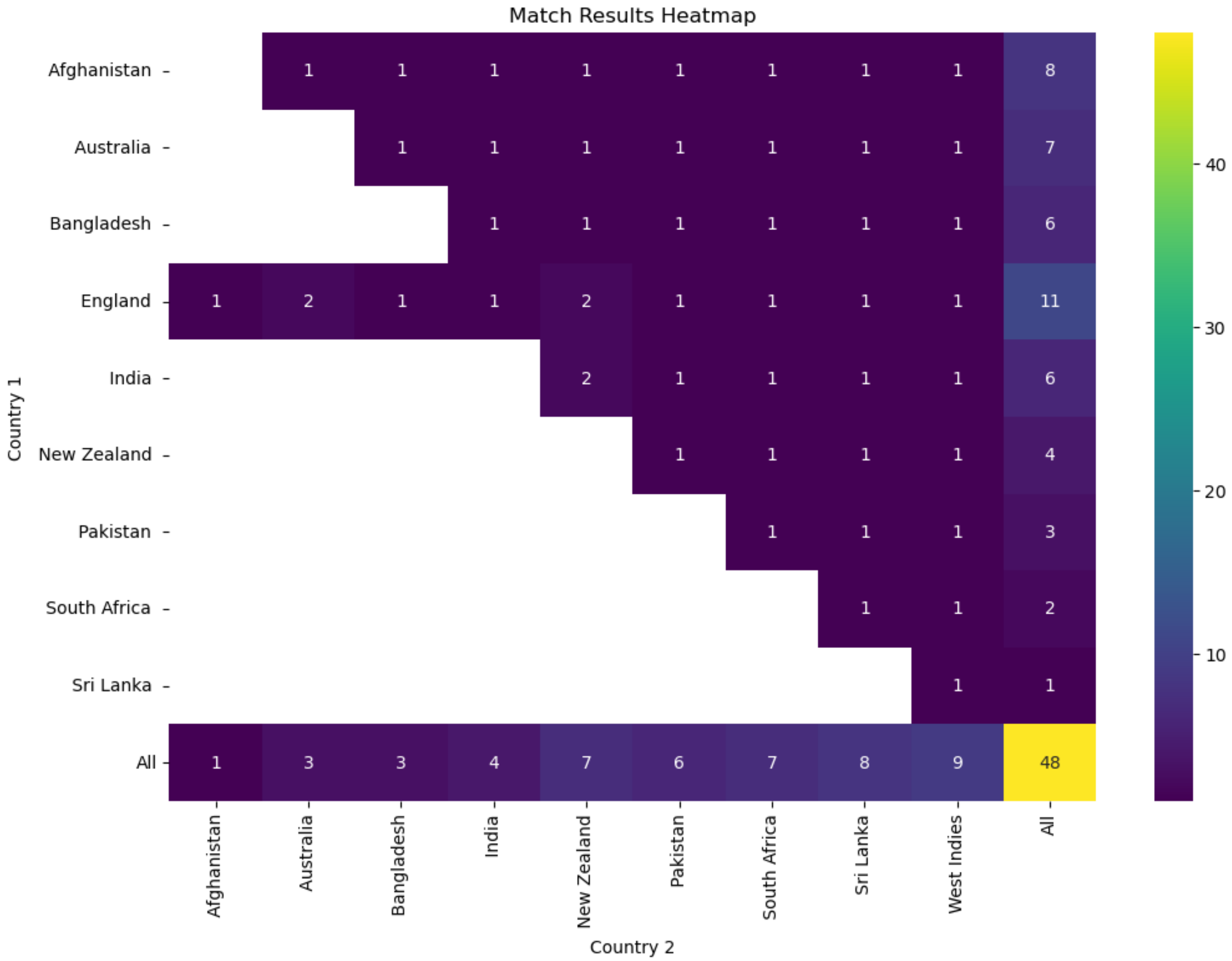
```
<Figure size 1500x1000 with 0 Axes>
```



Visualisation 5 - HeatMap of every match for every team in WC 2019

In [35]:

```
result_matrix = pd.crosstab(WC_2019['Country_1'], WC_2019['Country_2'], values=WC_2019['

plt.figure(figsize=(12, 8))
sns.heatmap(result_matrix, annot=True, cmap='viridis', fmt='g', cbar=True)
plt.title('Match Results Heatmap')
plt.xlabel('Country 2')
plt.ylabel('Country 1')
plt.show()
```

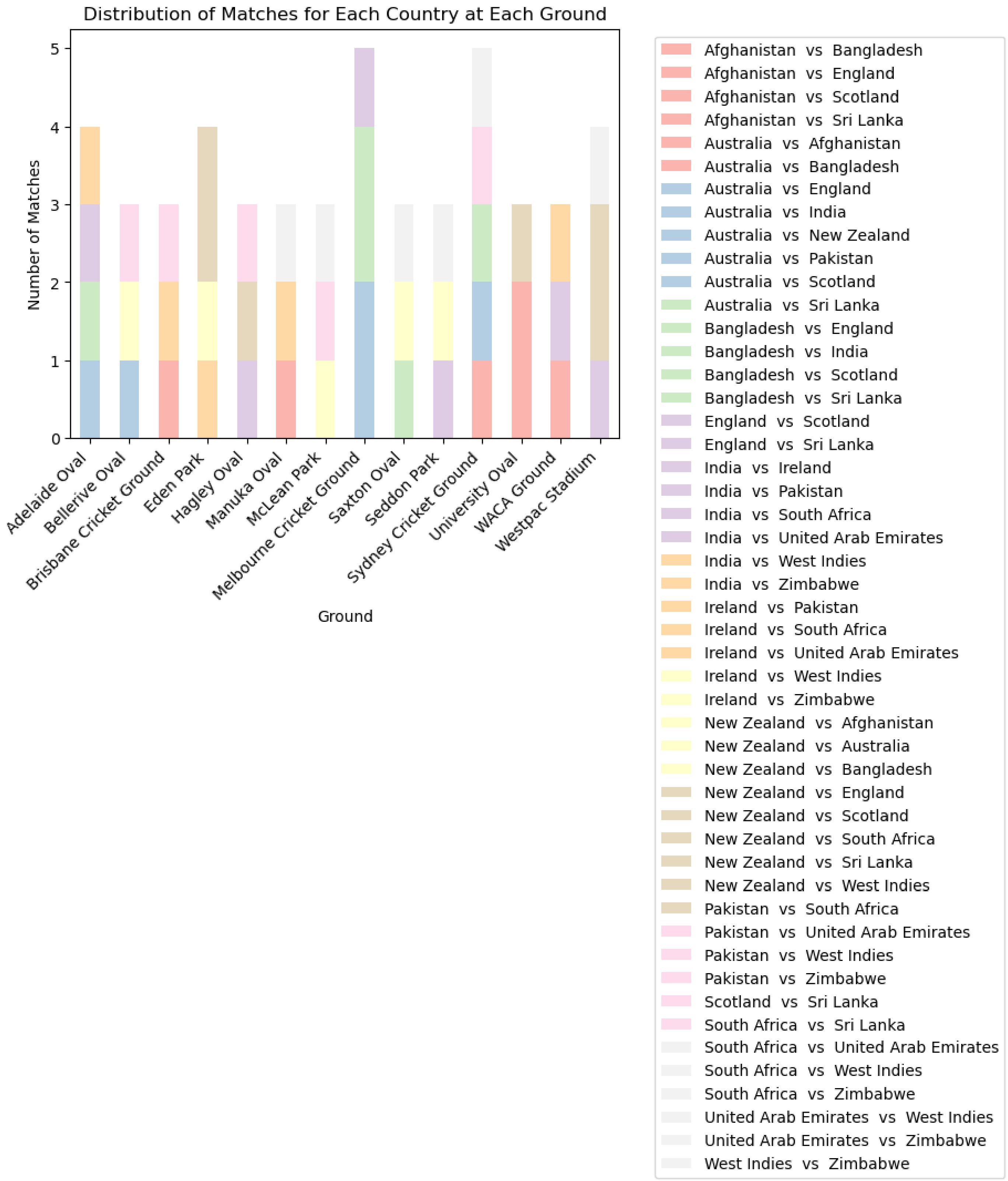


Visualisation 6 - Matches played by each team at each ground in 2015 World Cup

In [36]:

```
WC_2015['Country'] = WC_2015['Country_1'] + ' vs ' + WC_2015['Country_2']

# Create a DataFrame for plotting
plot_df = WC_2015.groupby(['Ground', 'Country']).size().unstack().fillna(0)

# Plotting
plt.figure(figsize=(15, 10))
plot_df.plot(kind='bar', stacked=True, colormap='Pastel1')
plt.title('Distribution of Matches for Each Country at Each Ground')
plt.xlabel('Ground')
plt.ylabel('Number of Matches')
plt.xticks(rotation=45, ha='right')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```

```
<Figure size 1500x1000 with 0 Axes>
```



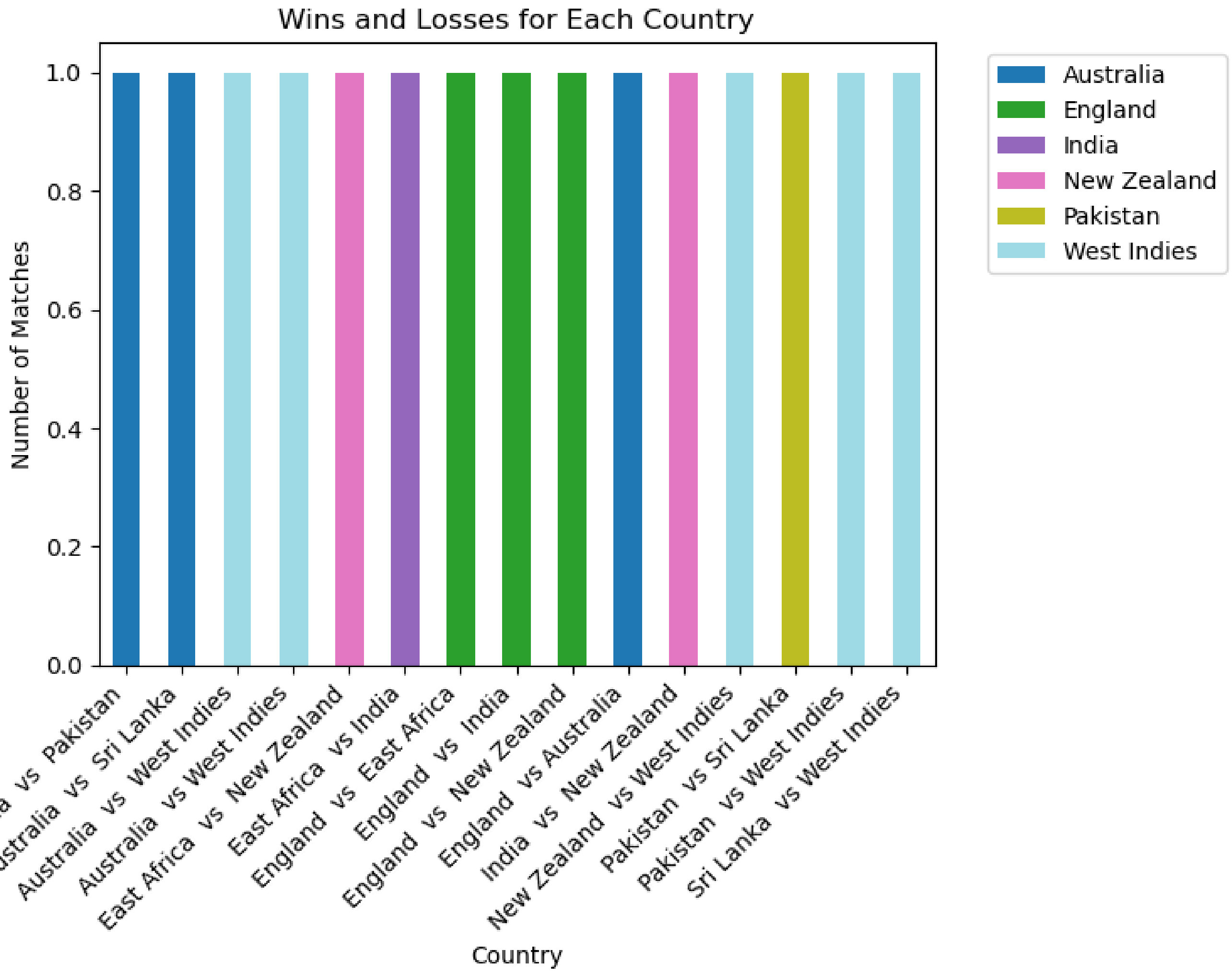
Visualisation 7 - Win and loss for each country in WC 1975

In [37]:

```
# Combine 'Country_1' and 'Country_2' into a single column 'Country'
WC_1975['Country'] = WC_1975['Country _1'] + ' vs ' + WC_1975['Country_2']

# Create a DataFrame for plotting
win_loss_df = WC_1975.groupby(['Country', 'Winner']).size().unstack().fillna(0)

# Plotting
plt.figure(figsize=(15, 10))
```

```
win_loss_df.plot(kind='bar', stacked=True, colormap=plt.cm.get_cmap('tab20'))
plt.title('Wins and Losses for Each Country')
plt.xlabel('Country')
plt.ylabel('Number of Matches')
plt.xticks(range(len(win_loss_df.index)), win_loss_df.index, rotation=45, ha='right')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_9108\631437697.py:9: MatplotlibDeprecationWar
ning: The get_cmap function was deprecated in Matplotlib 3.7 and will be removed two min
or releases later. Use ``matplotlib.colormaps[name]`` or ``matplotlib.colormaps.get_cmap
(obj)`` instead.
  win_loss_df.plot(kind='bar', stacked=True, colormap=plt.cm.get_cmap('tab20'))
<Figure size 1500x1000 with 0 Axes>
```



5. Key Findings:

Total Matches: The total number of matches played in each World Cup. Matches at Each Ground: Distribution of matches across different grounds. Final Match of 1983 WC: Details of the final match, including teams and outcomes. Matches Won by Each Team: Analysis of the number of matches won by each team.

6. Conclusion:

The exploratory data analysis provides valuable insights into the Cricket World Cup data from 1975 to 2019. The project successfully imported and transformed the data, executed SQL queries to extract relevant information, and visualized trends through various charts. The findings contribute to a better understanding of the tournament's history and the performance of participating teams.

7. Future Work:

Future work may involve more in-depth analyses, including player statistics, team performance over specific periods, and correlation analyses between different variables. Additionally, machine learning models could be implemented to predict match outcomes based on historical data.

This project serves as a foundation for further exploration and analysis of Cricket World Cup data, offering a comprehensive view of the tournament's evolution over the years.