Name: Kuntesh Kothawade

PRN No: 202401120063

Roll No: 50

Division: CS8

```
# Cricket World Cup 2023/24 Data Analysis using NumPy and Pandas

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


# Load the data
df = pd.read_csv('/content/matches.csv')

# Display the first few rows and the structure of the dataset
print("Dataset Preview:")
display(df.head())

print("\nDataset Information:")
display(df.info())

print("\nDataset Statistical Summary:")
display(df.describe())

print("\nColumn Names:")
display(df.columns.tolist())

# Now let's solve 20 problem statements

# Problem 1: How many matches were played in total?
print("\nProblem 1: How many matches were played in total?")
total_matches = len(df)
print(f"Total matches played: {total_matches}")

# Problem 2: List all unique teams that participated in the tournament
print("\nProblem 2: List all unique teams that participated in the
tournament")
team1 = set(df['team1'].unique())
team2 = set(df['team2'].unique())
```

```
all_teams = sorted(list(team1.union(team2)))
print(f"Teams that participated: {all_teams}")
print(f"Total number of teams: {len(all_teams)}")

# Problem 3: Which team won the most matches?
print("\nProblem 3: Which team won the most matches?")
team_wins = df['winner'].value_counts()
print(team_wins)
most_wins_team = team_wins.idxmax()
print(f"Team with most wins: {most_wins_team} with {team_wins.max()}
wins")

# Problem 4: What percentage of matches were won by the team that won
the toss?
print("\nProblem 4: What percentage of matches were won by the team
that won the toss?")
toss_and_match_winner = df[df['toss_winner'] == df['winner']]
toss_win_percentage = (len(toss_and_match_winner) / total_matches) *
100
print(f"Percentage of matches won by the toss winner:
{toss_win_percentage:.2f}%")

# Problem 5: Which venue hosted the most matches?
print("\nProblem 5: Which venue hosted the most matches?")
venue_counts = df['venue'].value_counts()
print(venue_counts.head())
most_used_venue = venue_counts.idxmax()
print(f"Venue that hosted the most matches: {most_used_venue} with
{venue_counts.max()} matches")

# Problem 6: What was the most common toss decision (bat or field)?
print("\nProblem 6: What was the most common toss decision?")
toss_decision_counts = df['toss_decision'].value_counts()
print(toss_decision_counts)
most_common_decision = toss_decision_counts.idxmax()
print(f"Most common toss decision: {most_common_decision}
({toss_decision_counts.max()} times)")

# Problem 7: Which player won the most 'Player of the Match' awards?
print("\nProblem 7: Which player won the most 'Player of the Match'
awards?")
pom_counts = df['player_of_match'].value_counts()
print(pom_counts.head())
most_pom = pom_counts.idxmax()
print(f"Player with most Player of the Match awards: {most_pom} with
{pom_counts.max()} awards")
```

```
# Problem 8: What was the average winning margin (in runs) for teams
batting first?
print("\nProblem 8: What was the average winning margin (in runs) for
teams batting first?")
teams_batting_first = df[df['winner_runs'].notna()]
avg_winning_runs = np.mean(teams_batting_first['winner_runs'])
print(f"Average winning margin in runs: {avg_winning_runs:.2f}")

# Problem 9: What was the average winning margin (in wickets) for teams
batting second?
print("\nProblem 9: What was the average winning margin (in wickets)
for teams batting second?")
teams_batting_second = df[df['winner_wickets'].notna()]
avg_winning_wickets = np.mean(teams_batting_second['winner_wickets'])
print(f"Average winning margin in wickets: {avg_winning_wickets:.2f}")

# Problem 10: How many matches were played in each city?
print("\nProblem 10: How many matches were played in each city?")
city_counts = df['city'].value_counts()
print(city_counts)

# Problem 11: Which team won by the largest margin (in runs)?
print("\nProblem 11: Which team won by the largest margin (in runs)?")
max_run_margin_idx = df['winner_runs'].idxmax()
max_run_match = df.loc[max_run_margin_idx]
print(f"{max_run_match['winner']} won by {max_run_match['winner_runs']}
runs against {max_run_match['team1'] if max_run_match['team2'] ==
max_run_match['winner'] else max_run_match['team2']}")

# Problem 12: Which team won by the largest margin (in wickets)?
print("\nProblem 12: Which team won by the largest margin (in
wickets)?")
max_wicket_margin_idx = df['winner_wickets'].idxmax()
max_wicket_match = df.loc[max_wicket_margin_idx]
print(f"{max_wicket_match['winner']} won by
{max_wicket_match['winner_wickets']} wickets against
{max_wicket_match['team1'] if max_wicket_match['team2'] ==
max_wicket_match['winner'] else max_wicket_match['team2']}")

# Problem 13: Create a crosstab of team1 vs team2 to see which teams
played against each other
print("\nProblem 13: Create a crosstab of team1 vs team2 to see which
teams played against each other")
team_matchups = pd.crosstab(df['team1'], df['team2'])
print(team_matchups)

# Problem 14: What was the win-loss record of each team?
print("\nProblem 14: What was the win-loss record of each team?")
```

```
# First, create a function to count matches played by each team
def get_matches_played(team, dataframe):
    return len(dataframe[(dataframe['team1'] == team) |
(dataframe['team2'] == team)])

# Create a DataFrame to store team statistics
team_stats = pd.DataFrame(index=all_teams)
team_stats['Matches_Played'] = [get_matches_played(team, df) for team
in all_teams]
team_stats['Matches_Won'] = [len(df[df['winner'] == team]) for team in
all_teams]
team_stats['Win_Percentage'] = (team_stats['Matches_Won'] /
team_stats['Matches_Played'] * 100).round(2)
team_stats = team_stats.sort_values('Win_Percentage', ascending=False)
print(team_stats)

# Problem 15: Which umpire officiated the most matches?
print("\nProblem 15: Which umpire officiated the most matches?")
# Combine umpire1 and umpire2
all_umpires = pd.concat([df['umpire1'], df['umpire2']])
umpire_counts = all_umpires.value_counts()
print(umpire_counts.head())
most_matches_umpire = umpire_counts.idxmax()
print(f"Umpire who officiated the most matches: {most_matches_umpire}
with {umpire_counts.max()} matches")

# Problem 16: Were there any trends in toss decisions over time?
print("\nProblem 16: Were there any trends in toss decisions over
time?")
df['date'] = pd.to_datetime(df['date'])
toss_decisions_over_time = df.groupby([pd.Grouper(key='date',
freq='7D'), 'toss_decision']).size().unstack()
print(toss_decisions_over_time)

# Problem 17: Which match had the closest finish (smallest winning
margin)?
print("\nProblem 17: Which match had the closest finish?")
# For teams that won batting second (winning by wickets)
wicket_wins = df[df['winner_wickets'].notna()]
min_wicket_idx = wicket_wins['winner_wickets'].idxmin()
min_wicket_match = df.loc[min_wicket_idx]

# For teams that won batting first (winning by runs)
run_wins = df[df['winner_runs'].notna()]
min_run_idx = run_wins['winner_runs'].idxmin()
min_run_match = df.loc[min_run_idx]
```

```
print(f"Closest finish (wickets): {min_wicket_match['winner']} won by
{min_wicket_match['winner_wickets']} wickets against
{min_wicket_match['team1'] if min_wicket_match['winner'] ==
min_wicket_match['team2'] else min_wicket_match['team2']}")
print(f"Closest finish (runs): {min_run_match['winner']} won by
{min_run_match['winner_runs']} runs against {min_run_match['team1'] if
min_run_match['winner'] == min_run_match['team2'] else
min_run_match['team2']}")

# Problem 18: How many matches did each team play at each venue?
print("\nProblem 18: How many matches did each team play at each
venue?")
# Create a new DataFrame with team and venue columns
team_venue_df = pd.DataFrame()
for index, row in df.iterrows():
    team_venue_df = pd.concat([team_venue_df, pd.DataFrame({
        'team': [row['team1'], row['team2']],
        'venue': [row['venue'], row['venue']]
    })], ignore_index=True)

team_venue_counts = pd.crosstab(team_venue_df['team'],
team_venue_df['venue'])
print(team_venue_counts)

# Problem 19: What was the average number of runs scored by teams
batting first?
print("\nProblem 19: What was the average number of runs scored by
teams batting first?")
# We don't have exact scores, but we can estimate from winning margins
# For matches where team batting second won, we can calculate:
runs_scored_by_batting_first = runs_scored_by_batting_second -
winner_wickets
# This is just a demonstration of the concept, not actual calculation
without full score data
print("Cannot calculate with the current dataset as it lacks total
scores information")

# Problem 20: What was the win percentage of teams that chose to bat
first after winning the toss?
print("\nProblem 20: What was the win percentage of teams that chose to
bat first after winning the toss?")
bat_first_teams = df[df['toss_decision'] == 'bat']
bat_first_wins = bat_first_teams[bat_first_teams['toss_winner'] ==
bat_first_teams['winner']]
bat_first_win_pct = (len(bat_first_wins) / len(bat_first_teams)) * 100
print(f"Win percentage of teams choosing to bat first:
{bat_first_win_pct:.2f}%")
```

```
print("\nData analysis complete!")
```

Dataset Preview:

	season	team1	team2	date	match_number	venue	city	toss_winner	toss_decision	player_of_match	umpire1
0	2023/24	England	New Zealand	2023/10/05	1	Narendra Modi Stadium	Ahmedabad	New Zealand	field	R Ravindra	HDPK Dharmasena
1	2023/24	Pakistan	Netherlands	2023/10/06	2	Rajiv Gandhi International Stadium	Hyderabad	Netherlands	field	Saud Shakeel	AT Holdstock
2	2023/24	Afghanistan	Bangladesh	2023/10/07	3	Himachal Pradesh Cricket Association Stadium	Dharamsala	Bangladesh	field	Mehedi Hasan Miraz	JS Wilson
3	2023/24	South Africa	Sri Lanka	2023/10/07	4	Arun Jaitley Stadium	Delhi	Sri Lanka	field	AK Markram	RK Illingworth
4	2023/24	Australia	India	2023/10/08	5	MA Chidambaram Stadium	Chennai	Australia	bat	KL Rahul	CB Gaffaney

memory usage: 6.9+ KB
None

Dataset Statistical Summary:

	match_number	winner_runs	winner_wickets
count	48.00	24.000000	24.000000
mean	24.50	125.916667	5.916667
std	14.00	82.242968	2.019829
min	1.00	5.000000	1.000000
25%	12.75	69.750000	5.000000
50%	24.50	101.000000	6.000000
75%	36.25	160.000000	7.250000
max	48.00	309.000000	9.000000

Column Names:

```
['season',
 'team1',
 'team2',
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48 entries, 0 to 47
e cell output actions  total 18 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   season           48 non-null     object
 1   team1            48 non-null     object
 2   team2            48 non-null     object
 3   date             48 non-null     object
 4   match_number     48 non-null     int64
 5   venue            48 non-null     object
 6   city             48 non-null     object
 7   toss_winner      48 non-null     object
 8   toss_decision    48 non-null     object
 9   player_of_match  48 non-null     object
 10  umpire1          48 non-null     object
 11  umpire2          48 non-null     object
 12  reserve_umpire   47 non-null     object
 13  match_referee    48 non-null     object
 14  winner           48 non-null     object
 15  winner_runs      24 non-null     float64
 16  winner_wickets   24 non-null     float64
 17  match_type       48 non-null     object
dtypes: float64(2), int64(1), object(15)
memory usage: 6.9+ KB
None
```

```
 'match_type']

Problem 1: How many matches were played in total?
Total matches played: 48

Problem 2: List all unique teams that participated in the tournament
Teams that participated: ['Afghanistan', 'Australia', 'Bangladesh', 'England', 'India', 'Netherlands', 'New Zealand', 'Pakistan', 'South Africa
Total number of teams: 10

Problem 3: Which team won the most matches?
winner
India          10
Australia       9
South Africa    7
New Zealand     5
Pakistan        4
Afghanistan     4
England         3
Bangladesh      2
Netherlands     2
Sri Lanka       2
```

```
Problem 4: What percentage of matches were won by the team that won the toss?
Percentage of matches won by the toss winner: 39.58%

Problem 5: Which venue hosted the most matches?
venue
Narendra Modi Stadium                          5
Himachal Pradesh Cricket Association Stadium    5
Maharashtra Cricket Association Stadium         5
Arun Jaitley Stadium                            5
MA Chidambaram Stadium                          5
Name: count, dtype: int64
Venue that hosted the most matches: Narendra Modi Stadium with 5 matches

Problem 6: What was the most common toss decision?
toss_decision
field    26
bat      22
Name: count, dtype: int64
Most common toss decision: field (26 times)

Problem 7: Which player won the most 'Player of the Match' awards?
player_of_match
Mohammed Shami    3
TM Head           3
A Zampa           2
RG Sharma         2
V Kohli           2
Name: count, dtype: int64
Player with most Player of the Match awards: Mohammed Shami with 3 awards
```

```
Problem 8: What was the average winning margin (in runs) for teams batting first?
Average winning margin in runs: 125.92

Problem 9: What was the average winning margin (in wickets) for teams batting second?
Average winning margin in wickets: 5.92

Problem 10: How many matches were played in each city?
city
Ahmedabad     5
Dharamsala    5
Pune          5
Delhi         5
Chennai       5
Lucknow       5
Mumbai        5
Bengaluru     5
Kolkata       5
Hyderabad     3
Name: count, dtype: int64

Problem 11: Which team won by the largest margin (in runs)?
Australia won by 309.0 runs against Netherlands

Problem 12: Which team won by the largest margin (in wickets)?
New Zealand won by 9.0 wickets against England

Problem 13: Create a crosstab of team1 vs team2 to see which teams played against each other
team2         Afghanistan  Australia  Bangladesh  England  India  Netherlands  \
team1
```

Code cell output actions

```
Problem 13: Create a crosstab of team1 vs team2 to see which teams played against each other
team2         Afghanistan  Australia  Bangladesh  England  India  Netherlands  \
team1
Afghanistan             0          1           1        1      1            0
Australia               0          0           0        1      1            1
Bangladesh              0          1           0        0      1            0
England                 0          0           1        0      0            1
India                   0          1           0        1      0            1
Netherlands             1          0           1        0      0            0
New Zealand             1          0           0        0      1            1
Pakistan                1          0           0        0      1            1
South Africa            0          2           1        1      0            0
Sri Lanka               1          1           1        0      0            0

team2         New Zealand  Pakistan  South Africa  Sri Lanka
team1
Afghanistan             0         0             1          0
Australia               1         1             0          0
Bangladesh              1         1             0          0
England                 1         1             0          1
India                   1         0             1          1
Netherlands             0         0             1          1
New Zealand             0         1             0          0
Pakistan                0         0             1          0
South Africa            1         0             0          1
Sri Lanka               1         1             0          0

Problem 14: What was the win-loss record of each team?
             Matches_Played  Matches_Won  Win_Percentage
```

```
Problem 14: What was the win-loss record of each team?
              Matches_Played  Matches_Won  Win_Percentage
India                     11           10           90.91
Australia                 11            9           81.82
South Africa              10            7           70.00
New Zealand               10            5           50.00
Afghanistan                9            4           44.44
Pakistan                   9            4           44.44
England                    9            3           33.33
Bangladesh                 9            2           22.22
Netherlands                9            2           22.22
Sri Lanka                  9            2           22.22

Problem 15: Which umpire officiated the most matches?
RK Illingworth      8
JS Wilson           7
Nitin Menon         7
RJ Tucker           7
RA Kettleborough    7
Name: count, dtype: int64
Umpire who officiated the most matches: RK Illingworth with 8 matches

Problem 16: Were there any trends in toss decisions over time?
toss_decision  bat  field
date
2023-10-05       3      6
2023-10-12       1      6
2023-10-19       5      3
2023-10-26       4      4
2023-11-02       4      4
```

```
Problem 17: Which match had the closest finish?
Closest finish (wickets): South Africa won by 1.0 wickets against Pakistan
Closest finish (runs): Australia won by 5.0 runs against New Zealand

Problem 18: How many matches did each team play at each venue?
venue         Arun Jaitley Stadium  \
team
Afghanistan                      2
Australia                        1
Bangladesh                       1
England                          1
India                            1
Netherlands                      1
New Zealand                      0
Pakistan                         0
South Africa                     1
Sri Lanka                        2

venue         Bharat Ratna Shri Atal Bihari Vajpayee Ekana Cricket Stadium  \
team
Afghanistan                                                      1
Australia                                                        2
Bangladesh                                                       0
England                                                          1
India                                                            1
Netherlands                                                      2
New Zealand                                                      0
Pakistan                                                         0
South Africa                                                     1
```

✓ 0s completed at 11:51 PM

```
New Zealand                                1                1
Pakistan                                   2                0
South Africa                               0                2
Sri Lanka                                  1                1

Problem 19: What was the average number of runs scored by teams batting first?
Cannot calculate with the current dataset as it lacks total scores information

Problem 20: What was the win percentage of teams that chose to bat first after winning the toss?
Win percentage of teams choosing to bat first: 36.36%

Data analysis complete!
```