untitled22

July 24, 2024

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

[2]:
```
# load the data
df = pd.read_csv('ipl2024.csv')
df
```

[2]:
```
    id           date      team1      team2 toss_winner decision  first_score  \
0    1  March 22,2024   Banglore    Chennai    Banglore      Bat          173
1    2  March 23,2024      Delhi     Punjab      Punjab    Field          174
2    3  March 23,2024    Kolkata  Hyderabad   Hyderabad    Field          208
3    4  March 24,2024  Rajasthan    Lucknow   Rajasthan      Bat          193
4    5  March 24,2024    Gujarat     Mumbai      Mumbai    Field          168
..  ..            …        …        …         …      …          …
69  70    May 19,2024  Rajasthan    Kolkata     Kolkata      NaN            0
70  71    May 21,2024  Hyderabad    Kolkata   Hyderabad      Bat          159
71  72    May 22,2024   Banglore  Rajasthan   Rajasthan    Field          172
72  73    May 24,2024  Hyderabad  Rajasthan   Rajasthan    Field          175
73  74    May 26,2024  Hyderabad    Kolkata   Hyderabad      Bat          113

    first_wkts  second_score  second_wkts     winner  player_of_the_match  \
0            6           176            4    Chennai    Mustafizur Rahman
1            9           177            6     Punjab           Sam Curran
2            7           204            7    Kolkata        Andre Russell
3            4           173            6  Rajasthan         Sanju Samson
4            6           162            9    Gujarat        Sai Sudharsan
..         …           …          …        …                  …
69           0             0            0  Abandoned                  NaN
70          10           164            2    Kolkata       Mitchell Starc
71           8           174            6  Rajasthan  Ravichandran Ashwin
72           9           139            7  Hyderabad        Shahbaz Ahmed
73          10           114            2    Kolkata       Mitchell Starc

           most_runs          most_wkts
0         Anuj Rawat  Mustafizur Rahman
1         Sam Curran      Kuldeep Yadav
2      Andre Russell        T Natarajan
3       Sanju Samson        Trent Boult
```

```
4      Dewald Brevis     Jasprit Bumrah
..               …                …
69             NaN              NaN
70      Shreyas Iyer     Mitchell Starc
71  Yashasvi Jaiswal         Avesh Khan
72       Dhruv Jurel      Shahbaz Ahmed
73    Venkatesh Iyer      Andre Russell

[74 rows x 14 columns]
```

[64]:

```
# exporing the data
df.info()
df.describe().T
df.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 74 entries, 0 to 73
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   id                   74 non-null     int64
 1   date                 74 non-null     datetime64[ns]
 2   team1                74 non-null     object
 3   team2                74 non-null     object
 4   toss_winner          74 non-null     object
 5   decision             71 non-null     object
 6   first_score          74 non-null     int64
 7   first_wkts           74 non-null     int64
 8   second_score         74 non-null     int64
 9   second_wkts          74 non-null     int64
 10  winner               74 non-null     object
 11  player_of_the_match  71 non-null     object
 12  most_runs            71 non-null     object
 13  most_wkts            71 non-null     object
 14  Year                 74 non-null     int32
 15  Month                74 non-null     object
 16  Day                  74 non-null     object
 17  score comparison     74 non-null     object
dtypes: datetime64[ns](1), int32(1), int64(5), object(11)
memory usage: 10.2+ KB
```

[3]:

```
print(df.head)
```

```
<bound method NDFrame.head of     id            date      team1      team2
toss_winner decision  first_score  \
0    1  March 22,2024   Banglore    Chennai    Banglore      Bat          173
1    2  March 23,2024      Delhi     Punjab      Punjab    Field          174
2    3  March 23,2024    Kolkata  Hyderabad   Hyderabad    Field          208
```

```
3    4  March 24,2024  Rajasthan    Lucknow   Rajasthan     Bat          193
4    5  March 24,2024    Gujarat     Mumbai      Mumbai   Field          168
..  ..            …          …          …          …      …             …
69  70    May 19,2024  Rajasthan    Kolkata     Kolkata     NaN            0
70  71    May 21,2024  Hyderabad    Kolkata   Hyderabad     Bat          159
71  72    May 22,2024   Banglore  Rajasthan   Rajasthan   Field          172
72  73    May 24,2024  Hyderabad  Rajasthan   Rajasthan   Field          175
73  74    May 26,2024  Hyderabad    Kolkata   Hyderabad     Bat          113

    first_wkts  second_score  second_wkts     winner  player_of_the_match  \
0            6           176            4    Chennai     Mustafizur Rahman
1            9           177            6     Punjab            Sam Curran
2            7           204            7    Kolkata         Andre Russell
3            4           173            6  Rajasthan          Sanju Samson
4            6           162            9    Gujarat         Sai Sudharsan
..         …           …          …        …                 …
69           0             0            0  Abandoned                   NaN
70          10           164            2    Kolkata        Mitchell Starc
71           8           174            6  Rajasthan  Ravichandran Ashwin
72           9           139            7  Hyderabad         Shahbaz Ahmed
73          10           114            2    Kolkata        Mitchell Starc

             most_runs          most_wkts
0          Anuj Rawat  Mustafizur Rahman
1          Sam Curran       Kuldeep Yadav
2       Andre Russell         T Natarajan
3        Sanju Samson         Trent Boult
4       Dewald Brevis      Jasprit Bumrah
..                …                 …
69                NaN                NaN
70        Shreyas Iyer     Mitchell Starc
71  Yashasvi Jaiswal          Avesh Khan
72        Dhruv Jurel      Shahbaz Ahmed
73     Venkatesh Iyer      Andre Russell

[74 rows x 14 columns]>
```

[4]:

```
df.isnull().sum()
```

[4]:

```
id                     0
date                   0
team1                  0
team2                  0
toss_winner            0
decision               3
first_score            0
first_wkts             0
```

```
second_score          0
second_wkts           0
winner                0
player_of_the_match   3
most_runs             3
most_wkts             3
dtype: int64
```

```
[5]: # CONVERTING DATE COLUMN TO DATETIME FORMAT FOR MAKING NEW COLUMNS FOR␣
       ↪DAYS,MONTHS AND YEAR
     df['date']=pd.to_datetime(df['date'],errors='coerce')
```

```
C:\Users\Microsoft\AppData\Local\Temp\ipykernel_5212\3169875093.py:1:
UserWarning: Could not infer format, so each element will be parsed
individually, falling back to `dateutil`. To ensure parsing is consistent and
as-expected, please specify a format.
  df['date']=pd.to_datetime(df['date'],errors='coerce')
```

```
[6]: df['date'].fillna(df['date'].mean(),inplace=True)
```

```
[7]: df['Year']=df['date'].dt.year
```

```
[8]: df['Month']=df['date'].dt.month_name()
```

```
[9]: df['Day']=df['date'].dt.day_name()
```

```
[10]: # MAKING SCORE COMPARISON COLUMN TO CHECK WHICH TEAM SCORE MORE
      def comparision(row):
          if row['first_score']>row['second_score']:
              return 'team first scored more'
          elif row['first_score']<row['second_score']:
              return 'team second scored more'
          else:
              return 'same'
```

```
[11]: df['score comparison']=df.apply(comparision,axis=1)
```

```
[12]: df.head()
```

```
[12]:    id                date      team1      team2 toss_winner decision  \
      0   1 2024-04-20 02:24:00   Banglore    Chennai    Banglore      Bat
      1   2 2024-04-20 02:24:00      Delhi     Punjab      Punjab    Field
      2   3 2024-04-20 02:24:00    Kolkata  Hyderabad   Hyderabad    Field
      3   4 2024-04-20 02:24:00  Rajasthan    Lucknow   Rajasthan      Bat
      4   5 2024-04-20 02:24:00    Gujarat     Mumbai      Mumbai    Field

         first_score  first_wkts  second_score  second_wkts     winner  \
      0          173           6           176           4    Chennai
      1          174           9           177           6     Punjab
      2          208           7           204           7    Kolkata
      3          193           4           173           6  Rajasthan
      4          168           6           162           9    Gujarat

        player_of_the_match      most_runs          most_wkts  Year  Month  \
      0   Mustafizur Rahman     Anuj Rawat  Mustafizur Rahman  2024  April
      1          Sam Curran     Sam Curran      Kuldeep Yadav  2024  April
      2       Andre Russell  Andre Russell        T Natarajan  2024  April
      3        Sanju Samson   Sanju Samson        Trent Boult  2024  April
      4       Sai Sudharsan  Dewald Brevis     Jasprit Bumrah  2024  April

              Day         score comparison
      0  Saturday  team second scored more
      1  Saturday  team second scored more
      2  Saturday   team first scored more
      3  Saturday   team first scored more
      4  Saturday   team first scored more
```

[13]:

```python
df['first_score'].duplicated().sum()
```

[13]:

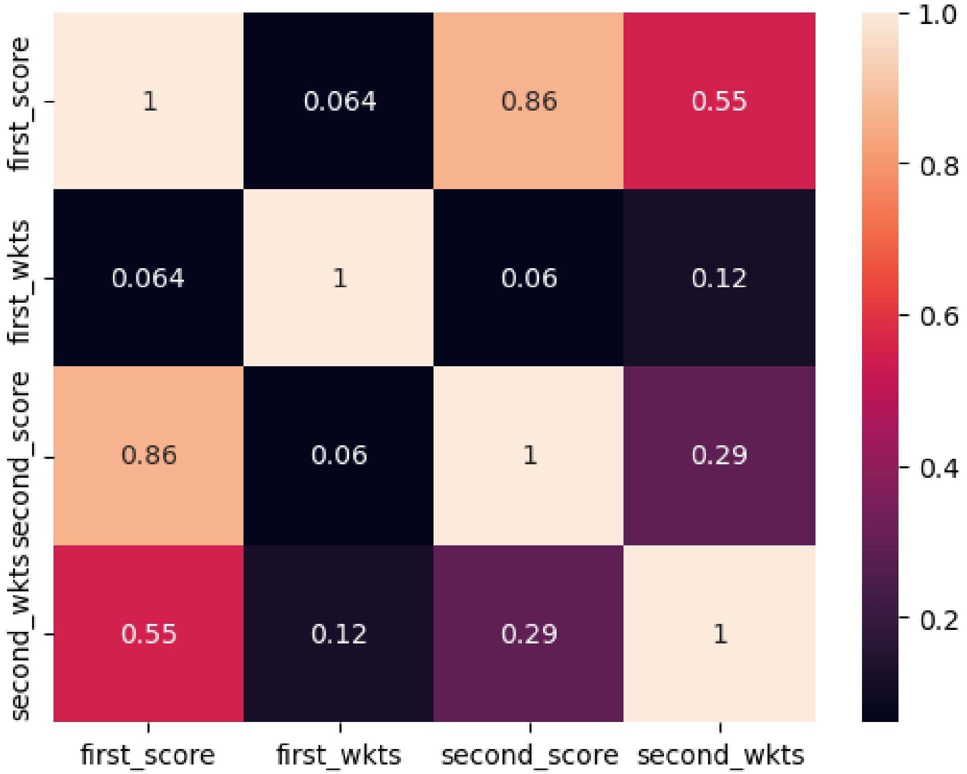
```
18
```

[14]:

```python
#importinf libraries
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
import plotly.graph_objects as go
import warnings
from scipy.stats import pearsonr
warnings.filterwarnings('ignore')
import os
```

[15]:

```python
# Checking correlation¶
df1=df[['first_score','first_wkts','second_score','second_wkts']]
cor=df1.corr()
sns.heatmap(cor,annot=True)
plt.show()
```



[16]:
```
# Insights From Data¶
# Check when teams bat or field first they win more or lose more
df.head()
des=df['decision']=='Field'
res=df[des].groupby('winner').size()
a=res.sum()
res.reset_index(name='wins')
a
```

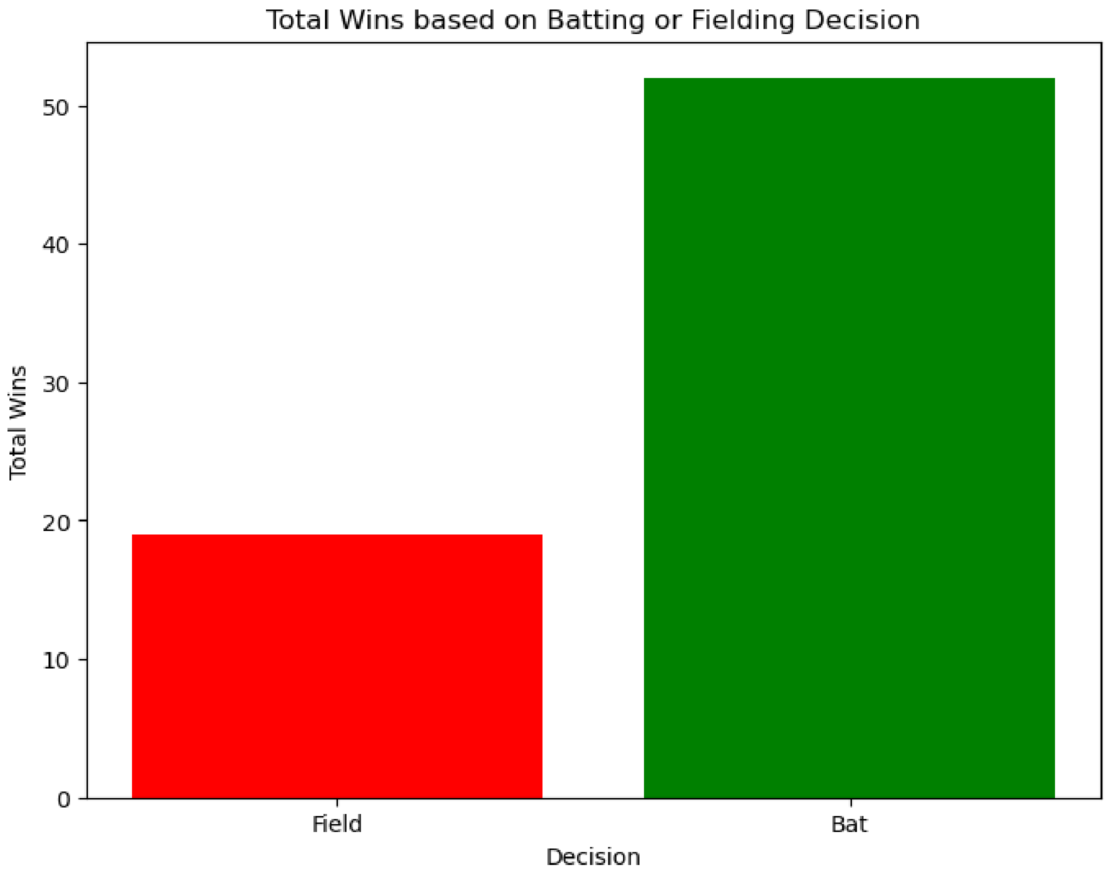
[16]:
```
52
```

[17]:
```
df.head()
des=df['decision']=='Bat'
res=df[des].groupby('winner').size()
b=res.sum()
res.reset_index(name='wins')
b
```

[17]:
```
19
```

[18]:
```
decisions=['Field' , 'Bat']
totals=[b, a]
```

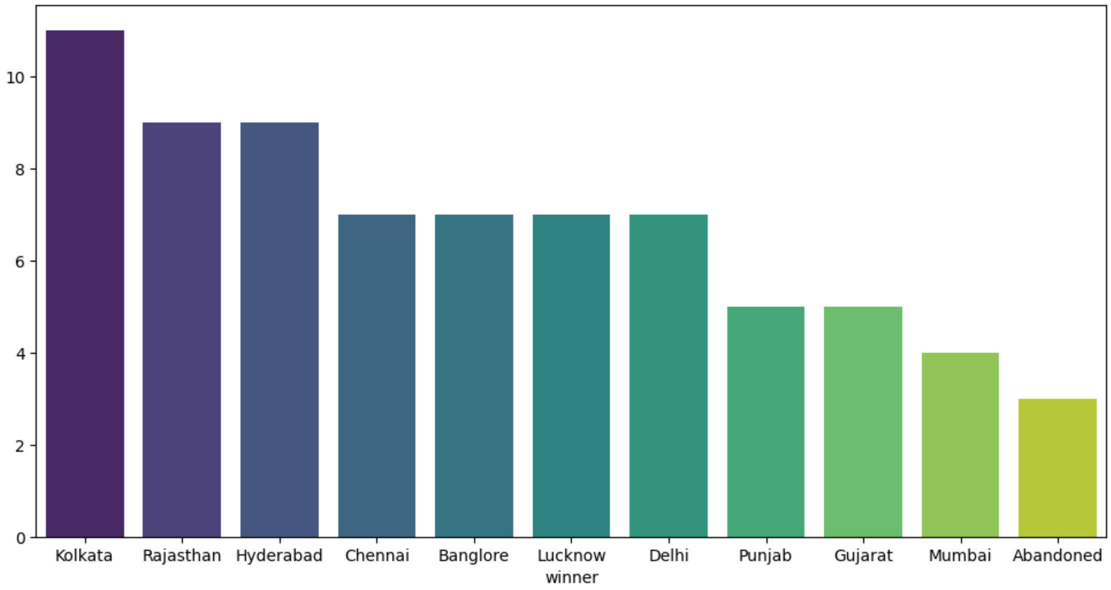
```
plt.figure(figsize=(8, 6))
plt.bar(decisions, totals, color=['red', 'green'])
plt.xlabel('Decision')
plt.ylabel('Total Wins')
plt.title('Total Wins based on Batting or Fielding Decision')
plt.show()
```



```
[19]: # which team win maximum matches.
      win=df['winner'].value_counts()
```

```
[20]: plt.figure(figsize=(12, 6))
      sns.barplot(x=win.index, y=win.values, palette='viridis')
```

```
[20]: <Axes: xlabel='winner'>
```



[21]:

```
# which team lost the most matches¶
def lost(row):
    if row['first_score']<row['second_score']:
        return row['team1']
    elif row['second_score']<row['first_score']:
        return row['team2']
    else:
        return 'tie'
```

[22]:

```
res=df.apply(lost,axis=1)
res1=res.value_counts()
res1
```

[22]:

```
Mumbai       10
Punjab        9
Hyderabad     8
Banglore      7
Delhi         7
Lucknow       7
Gujarat       7
Chennai       7
Rajasthan     6
Kolkata       3
tie           3
Name: count, dtype: int64
```

[23]:

```
# team win most tossess
toss=df['toss_winner'].value_counts().reset_index()
toss
```

[23]:

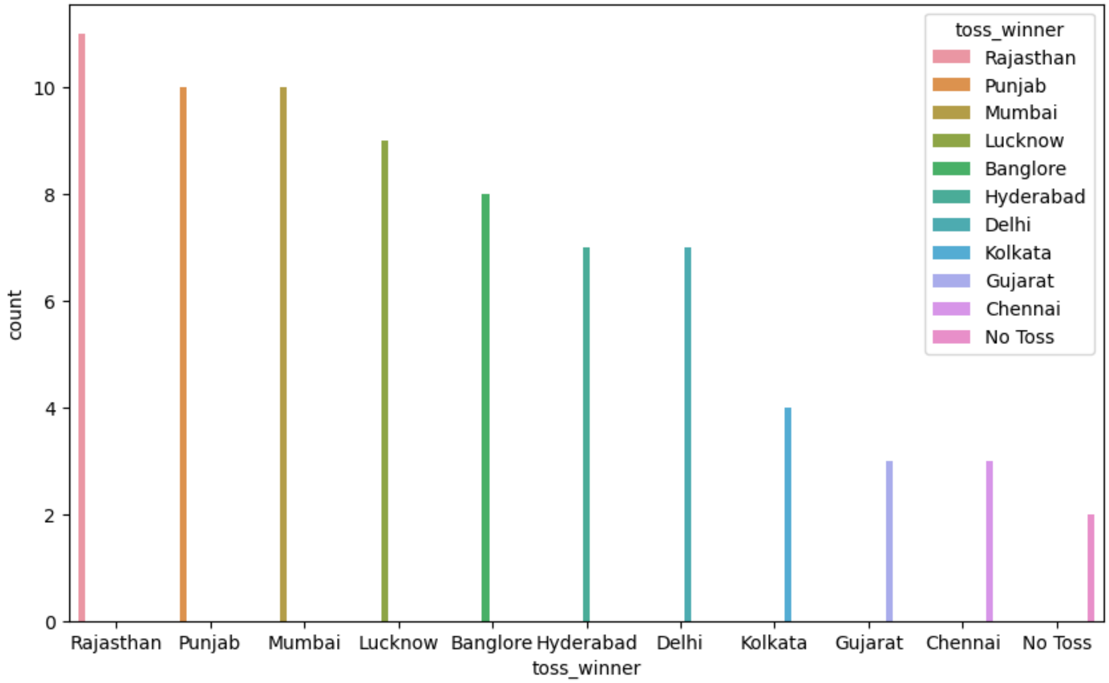
```
   toss_winner  count
0    Rajasthan     11
1       Punjab     10
2       Mumbai     10
3      Lucknow      9
4     Banglore      8
5    Hyderabad      7
6        Delhi      7
7      Kolkata      4
8      Gujarat      3
9      Chennai      3
10     No Toss      2
```

[24]:

```
plt.figure(figsize= (10,6))
sns.barplot(toss,x='toss_winner', y='count' , hue='toss_winner')
```

[24]:

```
<Axes: xlabel='toss_winner', ylabel='count'>
```



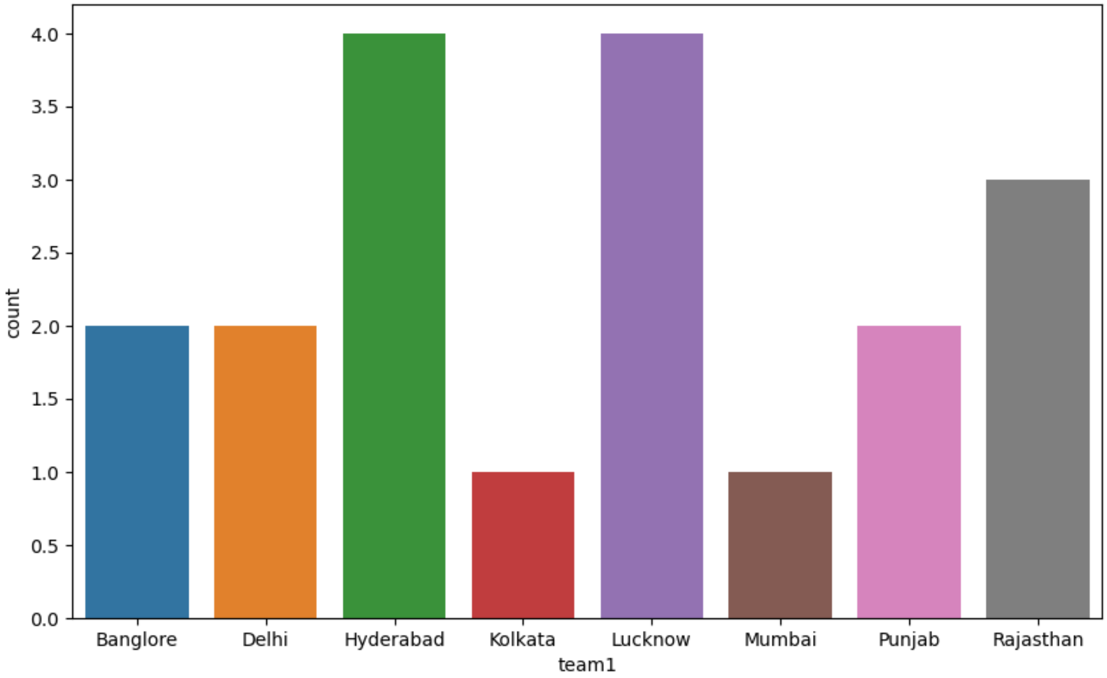
[25]:

```
des = df['decision'] == 'Bat'
team1 = df[des].groupby('team1').size().reset_index(name ='count')
team2 = df[des].groupby('team2').size().reset_index(name ='counts')
```

```
team1
```

[25]:
```
       team1  count
0   Banglore      2
1      Delhi      2
2  Hyderabad      4
3    Kolkata      1
4    Lucknow      4
5     Mumbai      1
6     Punjab      2
7  Rajasthan      3
```

[26]:
```
plt.figure(figsize=(10,6))
sns.barplot(team1,x = 'team1', y = 'count')
plt.show()
```

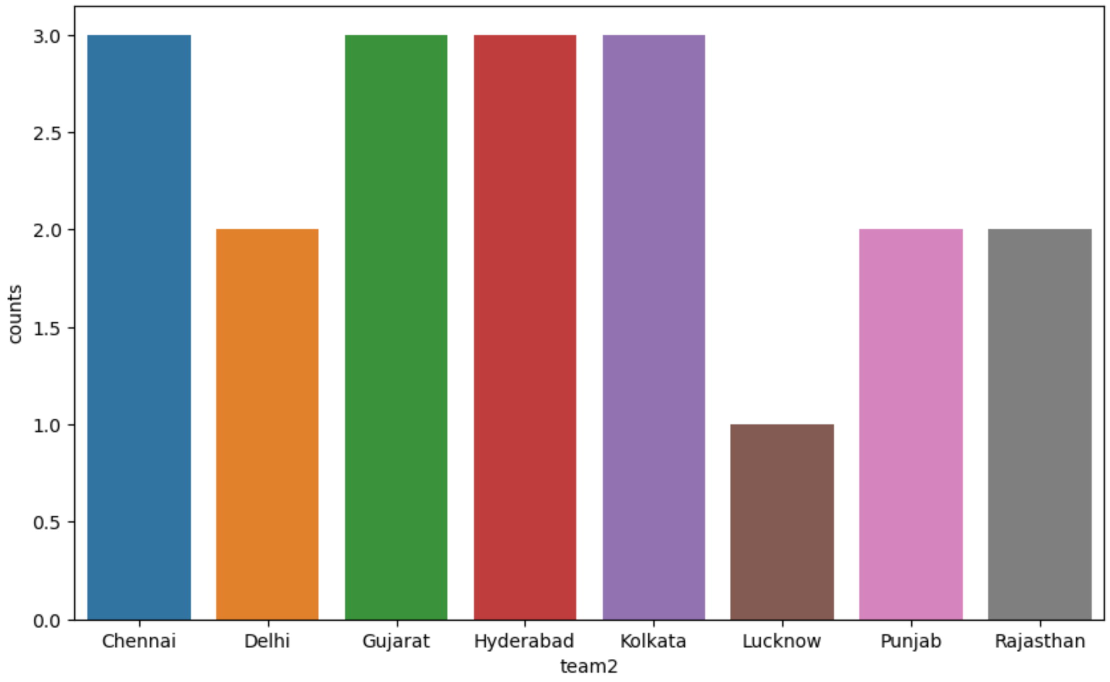


[27]:
```
team2
```

[27]:
```
       team2  counts
0    Chennai       3
1      Delhi       2
2    Gujarat       3
3  Hyderabad       3
4    Kolkata       3
5    Lucknow       1
```

```
6     Punjab       2
7  Rajasthan       2
```

[28]:
```
plt.figure(figsize = (10,6))
sns.barplot(team2, x = 'team2', y ='counts')
plt.show()
```

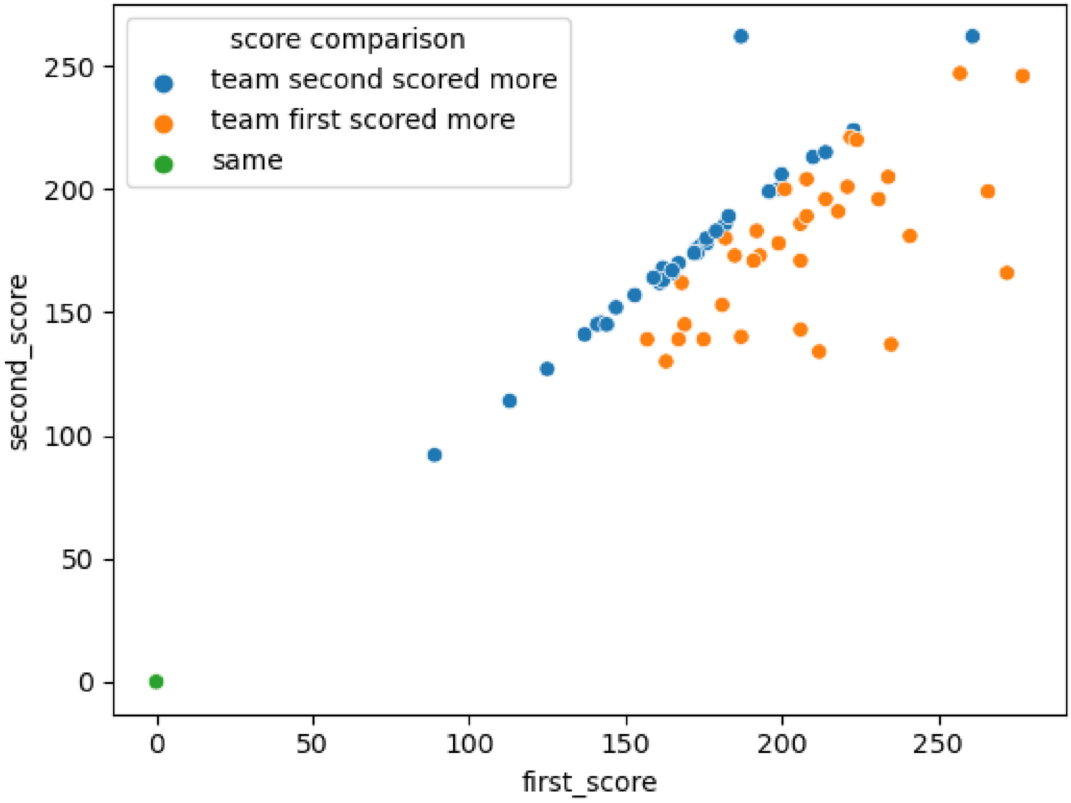


[29]:
```
# On which day the highest matches are played
day = df['Day'].value_counts().sort_values()
day
```

[29]:
```
Day
Friday        2
Monday        3
Tuesday       3
Wednesday     3
Thursday      3
Sunday        4
Saturday     56
Name: count, dtype: int64
```

[31]:
```
# In each match which team score more¶
sns.scatterplot(df,x = 'first_score', y='second_score', hue = 'score␣
 ↪comparison')
```

```
plt.show()
```



[35]:

```
# which team has highest wickets¶
team1 = df.groupby('team1')['first_wkts'].sum().sort_values(ascending = False)
team2 = df.groupby('team2')['second_wkts'].sum().reset_index(name = 'counts')
```

[36]:

```
merged_df = pd.merge(team1, team2, left_index=True, right_on='team2',␣
 ↪how='outer').fillna(0)
merged_df.columns = ['team1_wickets', 'team', 'team2_wickets']
```

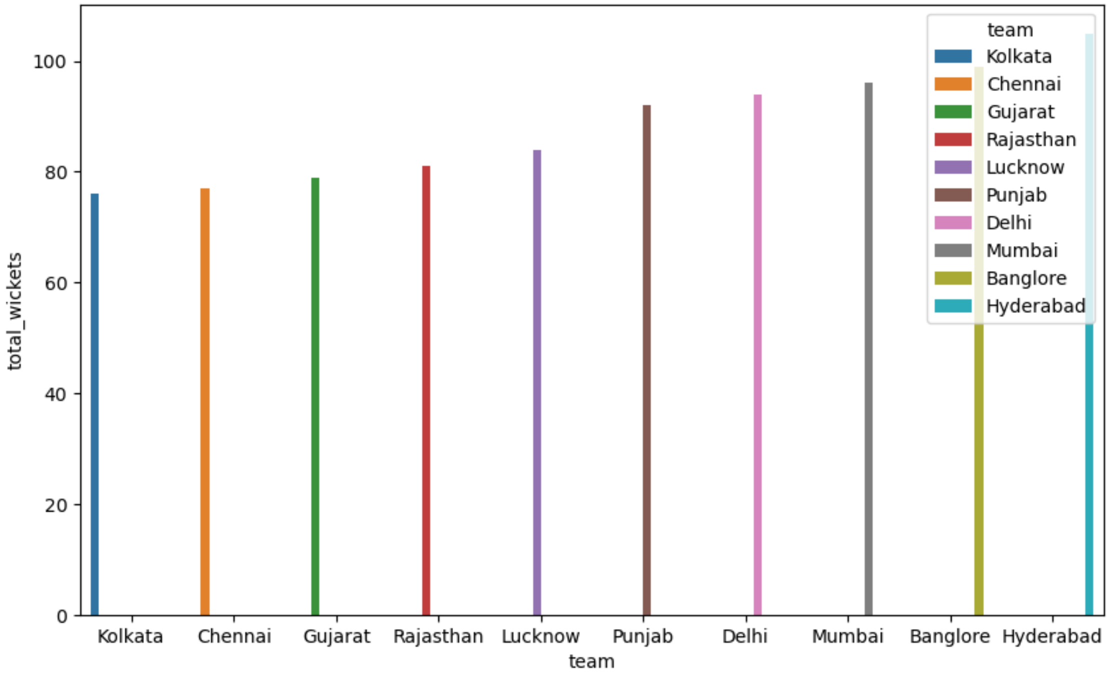
[37]:

```
merged_df['total_wickets'] = merged_df['team1_wickets'] +␣
 ↪merged_df['team2_wickets']
print(merged_df)
sort=merged_df.sort_values(by='total_wickets')
```

```
   team1_wickets       team  team2_wickets  total_wickets
4             70  Hyderabad             35            105
5             64    Kolkata             12             76
0             59   Banglore             40             99
6             47    Lucknow             37             84
1             44    Chennai             33             77
```

```
2            43      Delhi            51             94
7            37     Mumbai            59             96
3            36    Gujarat            43             79
8            29     Punjab            63             92
9            26  Rajasthan            55             81
```

[40]:
```
plt.figure(figsize=(10,6))
sns.barplot(sort, x= 'team', y = 'total_wickets', hue = 'team')
plt.show()
```



[48]:
```
# which team has high score¶
team1 = df.groupby('team1')['first_score'].sum().sort_values(ascending = False)
team2 = df.groupby('team2')['second_score'].sum().reset_index(name = 'counts')
```

[49]:
```
merge = pd.merge(team1, team2, left_index=True, right_on='team2', how='outer').
 ↪fillna(0)
merge.columns = ['team1_score', 'team', 'team2_score']
```

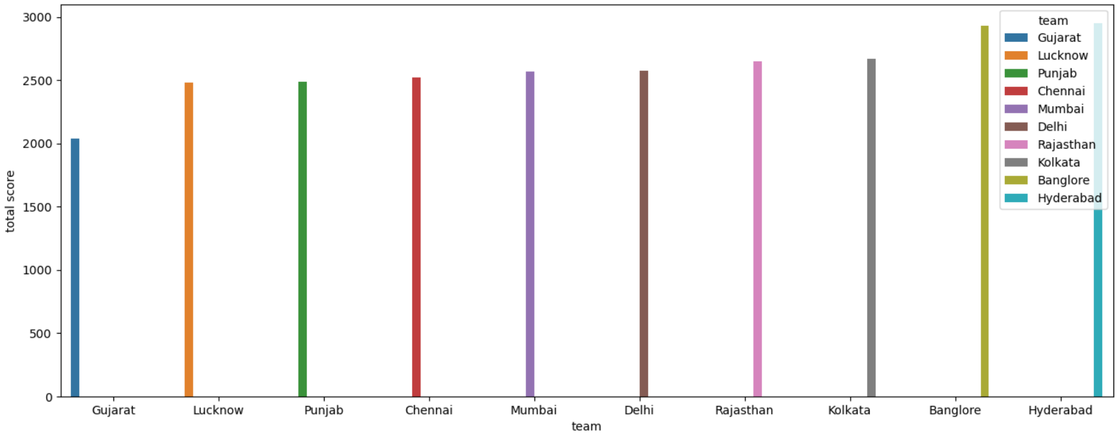
[50]:
```
merge['total score']=merge['team1_score'] + merge['team2_score']
```

[52]:
```
sol=merge.sort_values(by='total score')
sol
```

[52]:
```
   team1_score      team  team2_score  total score
3         1034   Gujarat         1006         2040
```

```
6          1446     Lucknow          1037         2483
8           679      Punjab          1808         2487
1          1504     Chennai          1020         2524
7           874      Mumbai          1694         2568
2          1428       Delhi          1145         2573
9           859   Rajasthan          1788         2647
5          1884     Kolkata           783         2667
0          1758    Banglore          1172         2930
4          1895   Hyderabad          1057         2952
```

[63]:
```
plt.figure(figsize= (16,6))
sns.barplot(sol,x='team',y= 'total score', hue ='team')
plt.show()
```



[ ]:
```
# CONCLUSION in the datset as i found most the matches won by Kolkata and total␣
↪won matches count is 11, number of tosses won by Rajastan and count is 11,4␣
↪times bat first by Lucknow team and hydrabad team in team 1. number of day␣
↪when most matches played on saturday which is 56,

# overall insight from the datset of IPL 2024.
```

[ ]: