ipl-capstone-project

August 7, 2025

0.0.1 IPL 2022 Capstone Project

The Indian Premier League (IPL) is a professional T20 cricket league in India, featuring franchises representing cities. This project explores IPL 2022 match-level data to derive meaningful insights and understand match outcomes, player performances, and team dynamics.

These are some of the important columns that we'll focus on for meaningful insights in this project.

column names: Variable Type * date : string
* venue : string
* stage : string
* team1 : string
* team2 : string
* toss_winner : string
* toss_decision : string
* first_ings_score : integer
* second_ings_score : integer
* match_winner : string
* won_by : string
* margin : integer
* player_of_the_match : string
* top_scorer : string
* highscore : integer
* best_bowling : string
* best_bowling_fgure : string
gure : string

0.0.2 Loading the Libraries and Dataset

[98]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings

warnings.filterwarnings("ignore")
```

```
df = pd.read_csv('ipl.csv')
df
```

[98]:

	match_id	date	venue	\
0	1	March 26,2022	Wankhede Stadium, Mumbai	
1	2	March 27,2022	Brabourne Stadium, Mumbai	
2	3	March 27,2022	Dr DY Patil Sports Academy, Mumbai	
3	4	March 28,2022	Wankhede Stadium, Mumbai	
4	5	March 29,2022	Maharashtra Cricket Association Stadium,Pune	
..	…	…	…	
69	70	May 22,2022	Wankhede Stadium, Mumbai	
70	71	May 24,2022	Eden Gardens, Kolkata	
71	72	May 25,2022	Eden Gardens, Kolkata	
72	73	May 27,2022	Narendra Modi Stadium, Ahmedabad	
73	74	May 29,2022	Narendra Modi Stadium, Ahmedabad	

	team1	team2	stage	toss_winner	toss_decision	first_ings_score	\
0	Chennai	Kolkata	Group	Kolkata	Field	131	
1	Delhi	Mumbai	Group	Delhi	Field	177	
2	Banglore	Punjab	Group	Punjab	Field	205	
3	Gujarat	Lucknow	Group	Gujarat	Field	158	
4	Hyderabad	Rajasthan	Group	Hyderabad	Field	210	
..	…	…	…	…	…	…	
69	Hyderabad	Punjab	Group	Hyderabad	Bat	157	
70	Gujarat	Rajasthan	Playoff	Gujarat	Field	188	
71	Banglore	Lucknow	Playoff	Lucknow	Field	207	
72	Banglore	Rajasthan	Playoff	Rajasthan	Field	157	
73	Gujarat	Rajasthan	Final	Rajasthan	Bat	130	

	first_ings_wkts	second_ings_score	second_ings_wkts	match_winner	\
0	5	133	4	Kolkata	
1	5	179	6	Delhi	
2	2	208	5	Punjab	
3	6	161	5	Gujarat	
4	6	149	7	Rajasthan	
..	…	…	…	…	
69	8	160	5	Punjab	
70	6	191	3	Gujarat	
71	4	193	6	Banglore	
72	8	161	3	Rajasthan	
73	9	133	3	Gujarat	

	won_by	margin	player_of_the_match	top_scorer	highscore	\
0	Wickets	6	Umesh Yadav	MS Dhoni	50	
1	Wickets	4	Kuldeep Yadav	Ishan Kishan	81	
2	Wickets	5	Odean Smith	Faf du Plessis	88	
3	Wickets	5	Mohammed Shami	Deepak Hooda	55	

```
4      Runs      61        Sanju Samson     Aiden Markram         57
..      …     …                 …               …        …
69  Wickets       5       Harpreet Brar  Liam Livingstone         49
70  Wickets       7        David Miller       Jos Buttler         89
71     Runs      14       Rajat Patidar     Rajat Patidar        112
72  Wickets       7         Jos Buttler       Jos Buttler        106
73  Wickets       7       Hardik Pandya      Shubman Gill         45

        best_bowling best_bowling_figure
0       Dwayne Bravo               3--20
1      Kuldeep Yadav               3--18
2     Mohammed Siraj               2--59
3     Mohammed Shami               3--25
4   Yuzvendra Chahal               3--22
..               …                  …
69     Harpreet Brar               3--26
70     Hardik Pandya               1--14
71    Josh Hazlewood               3--43
72   Prasidh Krishna               3--22
73     Hardik Pandya               3--17

[74 rows x 20 columns]
```

0.0.3 Basic Information

[101]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 74 entries, 0 to 73
Data columns (total 20 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   match_id             74 non-null     int64
 1   date                 74 non-null     object
 2   venue                74 non-null     object
 3   team1                74 non-null     object
 4   team2                74 non-null     object
 5   stage                74 non-null     object
 6   toss_winner          74 non-null     object
 7   toss_decision        74 non-null     object
 8   first_ings_score     74 non-null     int64
 9   first_ings_wkts      74 non-null     int64
 10  second_ings_score    74 non-null     int64
 11  second_ings_wkts     74 non-null     int64
 12  match_winner         74 non-null     object
 13  won_by               74 non-null     object
 14  margin               74 non-null     int64
```

```
 15  player_of_the_match  74 non-null     object
 16  top_scorer           74 non-null     object
 17  highscore            74 non-null     int64
 18  best_bowling         74 non-null     object
 19  best_bowling_figure  74 non-null     object
dtypes: int64(7), object(13)
memory usage: 11.7+ KB
```

Check the size of rows and columns of the dataset

[104]:
```
print(f"Rows are {df.shape[0]},and columns are {df.shape[1]}")
```

```
Rows are 74,and columns are 20
```

Now let's see how many columns have null values in total.

[107]:
```
df.isnull().sum()
```

[107]:
```
match_id              0
date                  0
venue                 0
team1                 0
team2                 0
stage                 0
toss_winner           0
toss_decision         0
first_ings_score      0
first_ings_wkts       0
second_ings_score     0
second_ings_wkts      0
match_winner          0
won_by                0
margin                0
player_of_the_match   0
top_scorer            0
highscore             0
best_bowling          0
best_bowling_figure   0
dtype: int64
```

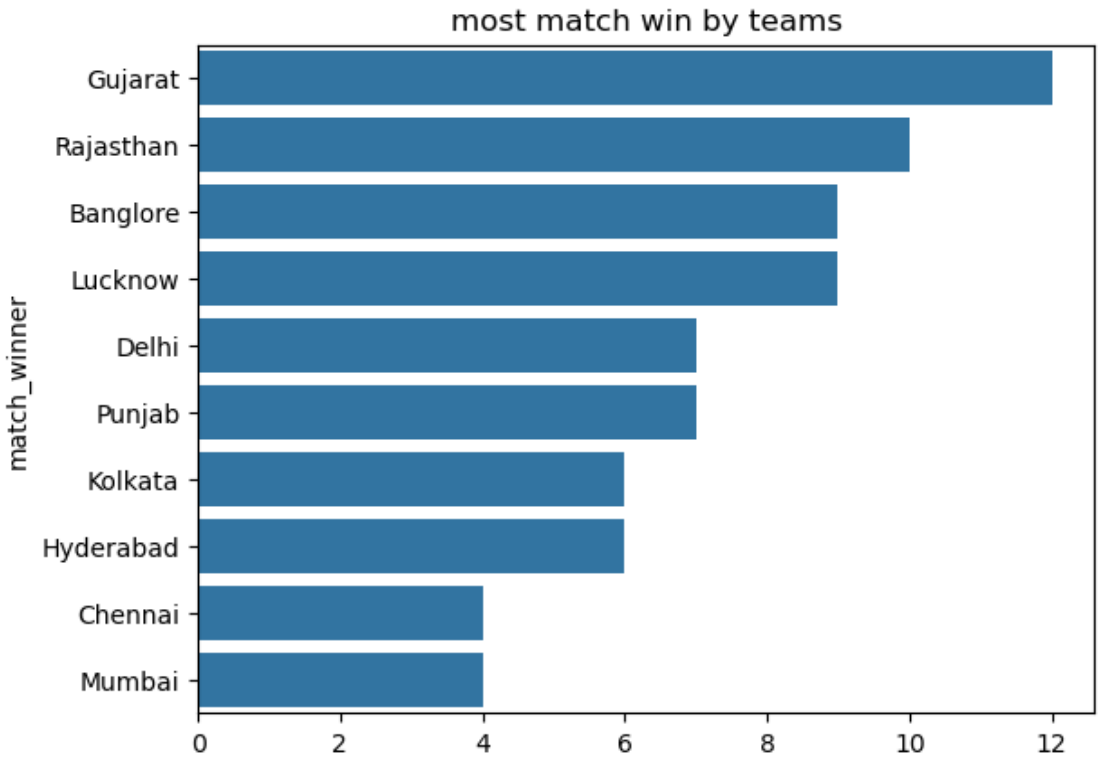
Now, Here comes some Basic Questions

1. Which team won the most matches?

[111]:
```
match_wins = df['match_winner'].value_counts()
```

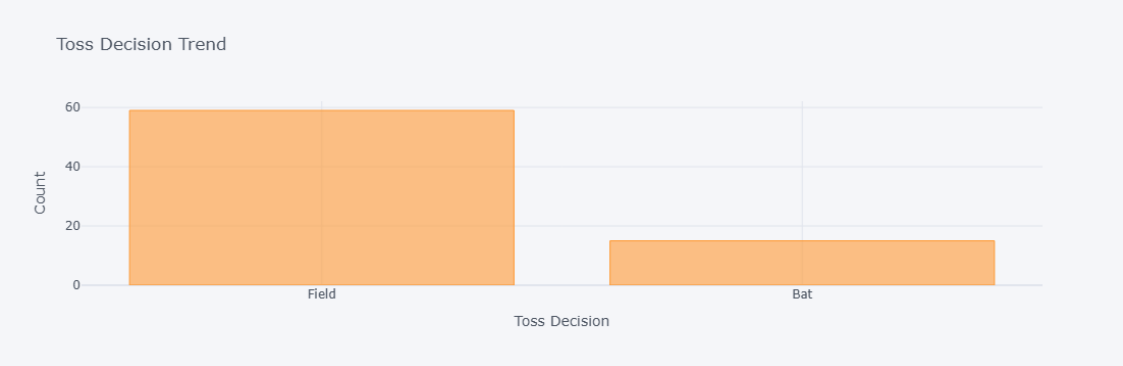
[113]:
```
sns.barplot(y = match_wins.index , x = match_wins.values)
plt.title("most match win by teams")
```

[113]:
```
Text(0.5, 1.0, 'most match win by teams')
```



2. Toss Decision Trends

[116]:
```
import cufflinks as cf
import plotly.offline as pyo

cf.go_offline()
df['toss_decision'].value_counts().iplot(
    kind='bar',
    xTitle='Toss Decision',
    yTitle='Count',
    title='Toss Decision Trend',
    color='orange'
)
```



[ ]:

[ ]:

3. Toss Winner vs Match Winner

[121]:
```
count = df[df['toss_winner'] == df['match_winner']]['match_id'].count()
percent = (count * 100)/df.shape[0]
percent.round(2)
```

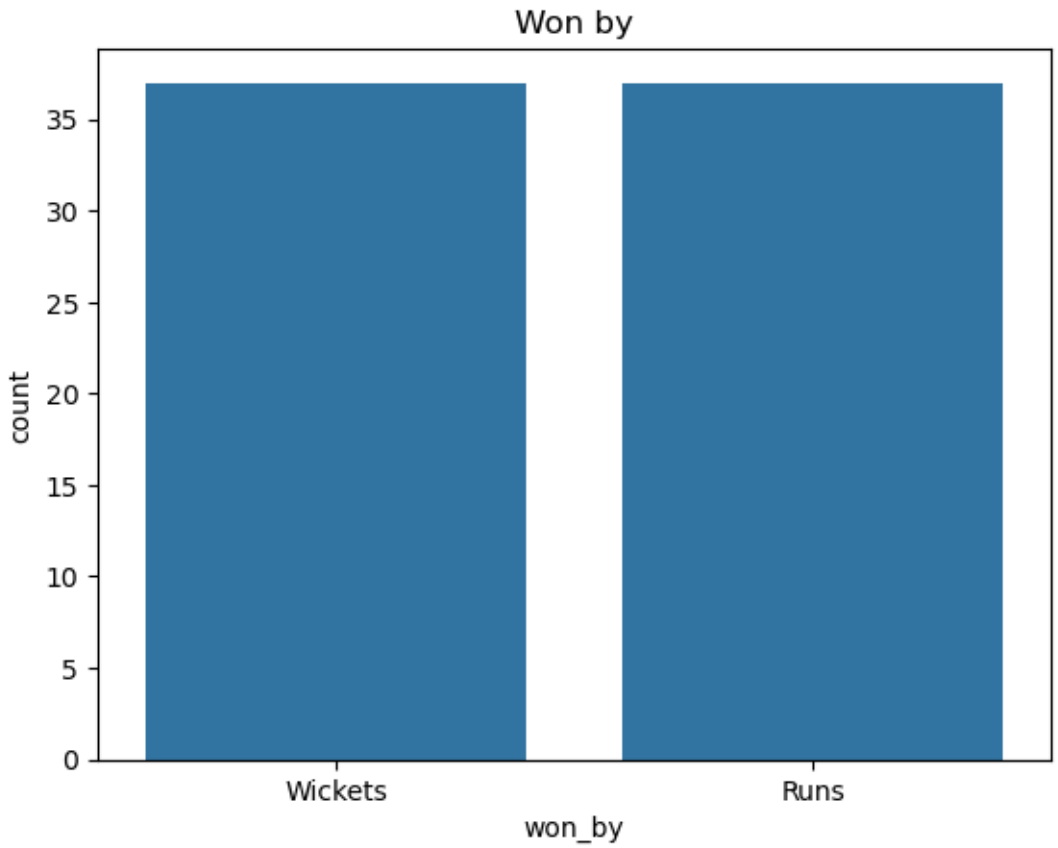
[121]: 48.65

[ ]:

4. How do teams win? (Runs vs Wickets)

[281]:
```
sns.countplot(x = df['won_by'])
plt.title("Won by")
```

[281]: Text(0.5, 1.0, 'Won by')



0.0.4 Key Player Performances

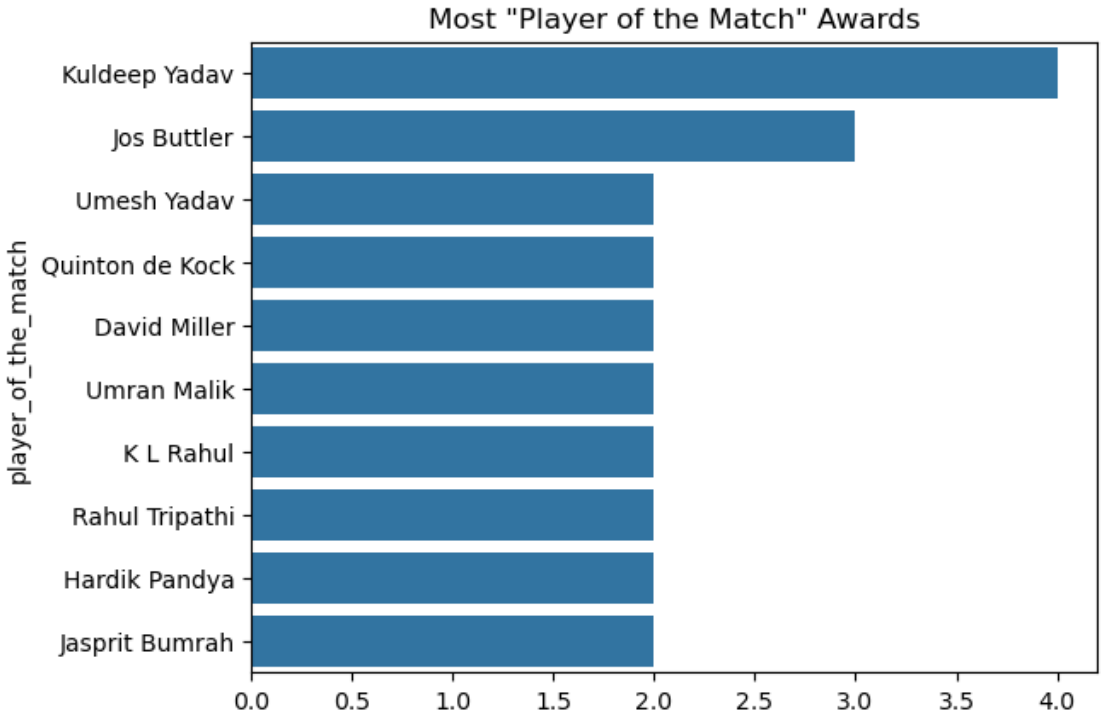
1 Most "Player of the Match" Awards

[257]:
```
count = df['player_of_the_match'].value_counts().head(10)
count
```

[257]:
```
player_of_the_match
Kuldeep Yadav      4
Jos Buttler        3
Umesh Yadav        2
Quinton de Kock    2
David Miller       2
Umran Malik        2
K L Rahul          2
Rahul Tripathi     2
Hardik Pandya      2
Jasprit Bumrah     2
Name: count, dtype: int64
```

[263]:
```
sns.barplot( x = count.values, y = count.index)
plt.title('Most "Player of the Match" Awards')
```

[263]:
```
Text(0.5, 1.0, 'Most "Player of the Match" Awards')
```

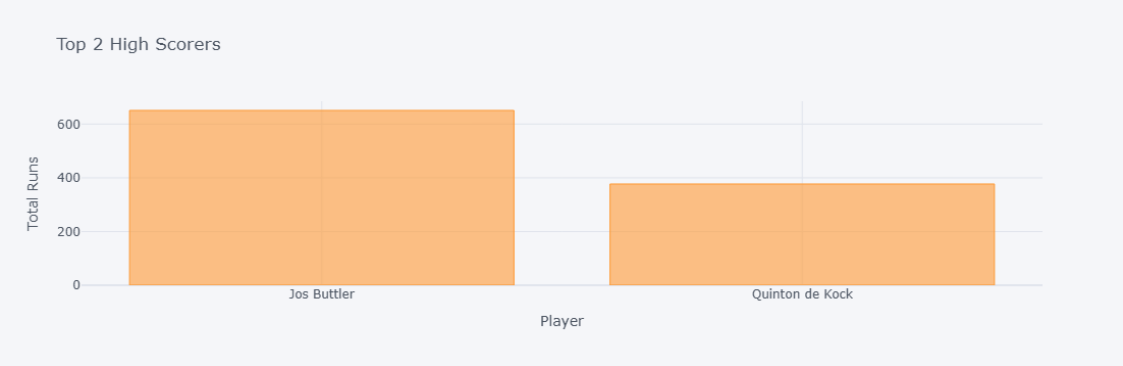


2 Top Scorers

[209]:
```
count = df.groupby('top_scorer')['highscore'].sum().
 ↪sort_values(ascending=False).head(2)
count
```

[209]:
```
top_scorer
Jos Buttler        651
Quinton de Kock    377
Name: highscore, dtype: int64
```

[213]:
```
import cufflinks as cf
import plotly.offline as pyo
import pandas as pd

cf.go_offline()
count_df.columns = ['Top Scorer', 'Total Highscore']
count_df.iplot(kind='bar', x='Top Scorer', y='Total Highscore', title='Top 2␣
 ↪High Scorers', xTitle='Player', yTitle='Total Runs')
```



[ ]:

10 Best Bowling Figures

[141]:
```
df.head()
```

[141]:

	match_id	date	venue	team1	team2	stage	toss_winner	toss_decision	first_ings_score	first_ings_wkts	second_ings_score	second_ings_wkts	match_winner	won_by	margin	player_of_the_match	top_scorer	highscore	best_bowling
0	1	March 26,2022	Wankhede Stadium, Mumbai	Chennai	Kolkata	Group	Kolkata	Field	131	5	133	4	Kolkata	Wickets	6	Umesh Yadav	MS Dhoni	50	Dwayne Bravo
1	2	March 27,2022	Brabourne Stadium, Mumbai	Delhi	Mumbai	Group	Delhi	Field	177	5	179	6	Delhi	Wickets	4	Kuldeep Yadav	Ishan Kishan	81	Kuldeep Yadav
2	3	March 27,2022	Dr DY Patil Sports Academy, Mumbai	Banglore	Punjab	Group	Punjab	Field	205	2	208	5	Punjab	Wickets	5	Odean Smith	Faf du Plessis	88	Mohammed Siraj
3	4	March 28,2022	Wankhede Stadium, Mumbai	Gujarat	Lucknow	Group	Gujarat	Field	158	6	161	5	Gujarat	Wickets	5	Mohammed Shami	Deepak Hooda	55	Mohammed Shami
4	5	March 29,2022	Maharashtra Cricket Association Stadium,Pune	Hyderabad	Rajasthan	Group	Hyderabad	Field	210	6	149	7	Rajasthan	Runs	61	Sanju Samson	Aiden Markram	57	Yuzvendra Chahal

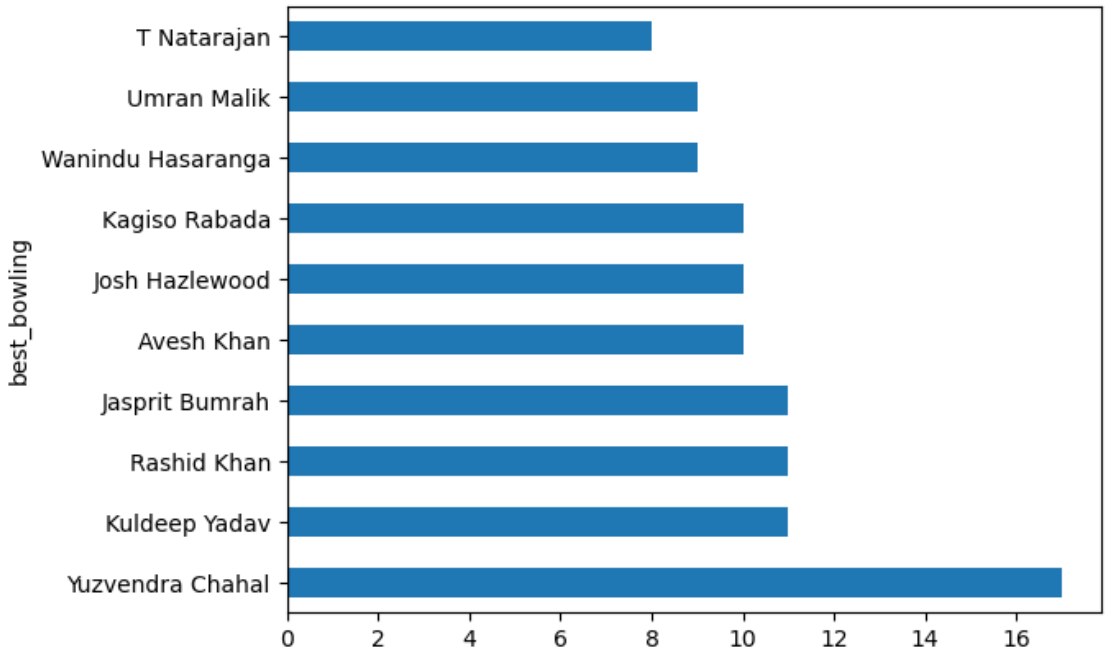
```
  best_bowling_figure
0               3--20
1               3--18
2               2--59
3               3--25
4               3--22
```

[243]:
```
df['hightest_wickets'] = df['best_bowling_figure'].apply(lambda x : x.
 ↪split('--')[0])
df['hightest_wickets'] = df['hightest_wickets'].astype(int)
top_bowlers = df.groupby('best_bowling')['hightest_wickets'].sum().
 ↪sort_values(ascending=False).head(10)
top_bowlers
```

[243]:
```
best_bowling
Yuzvendra Chahal     17
Kuldeep Yadav        11
Rashid Khan          11
Jasprit Bumrah       11
Avesh Khan           10
Josh Hazlewood       10
Kagiso Rabada        10
Wanindu Hasaranga     9
Umran Malik           9
T Natarajan           8
Name: hightest_wickets, dtype: int32
```

[245]:
```
top_bowlers.plot(kind = 'barh')
```

[245]:
```
<Axes: ylabel='best_bowling'>
```



0.0.5 Venue Analysis

Most Matches Played by Venue

[149]:
```
df.head()
```

[149]:

	match_id	date	venue	team1	team2	stage	toss_winner	toss_decision	first_ings_score	first_ings_wkts	…	second_ings_wkts	match_winner	won_by	margin
0	1	March 26,2022	Wankhede Stadium, Mumbai	Chennai	Kolkata	Group	Kolkata	Field	131	5	…	4	Kolkata	Wickets	6
1	2	March 27,2022	Brabourne Stadium, Mumbai	Delhi	Mumbai	Group	Delhi	Field	177	5	…	6	Delhi	Wickets	4
2	3	March 27,2022	Dr DY Patil Sports Academy, Mumbai	Banglore	Punjab	Group	Punjab	Field	205	2	…	5	Punjab	Wickets	5
3	4	March 28,2022	Wankhede Stadium, Mumbai	Gujarat	Lucknow	Group	Gujarat	Field	158	6	…	5	Gujarat	Wickets	5
4	5	March 29,2022	Maharashtra Cricket Association Stadium,Pune	Hyderabad	Rajasthan	Group	Hyderabad	Field	210						

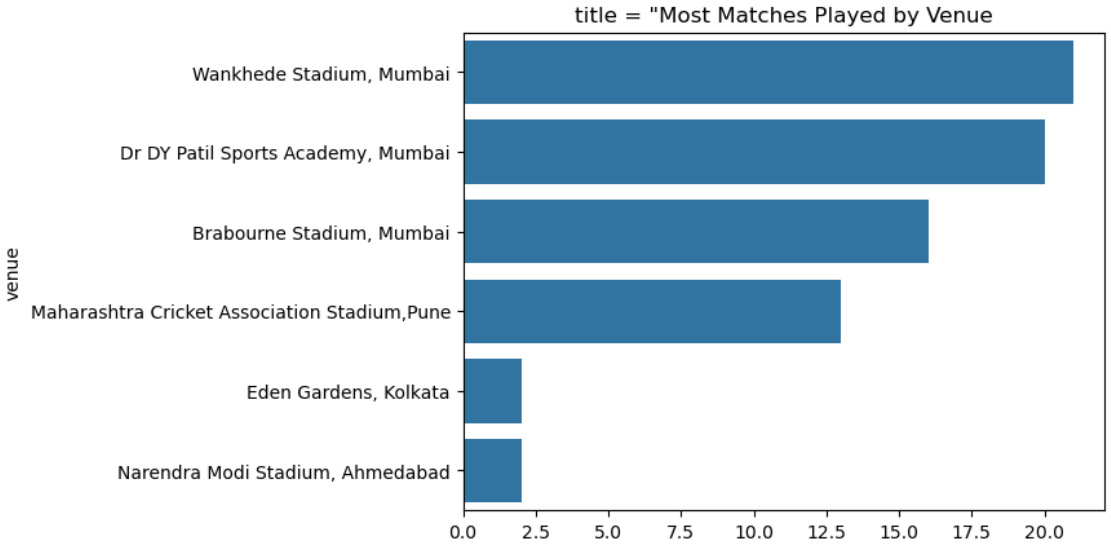
```
4                6  …                 7     Rajasthan     Runs     61

   player_of_the_match      top_scorer highscore      best_bowling  \
0          Umesh Yadav        MS Dhoni        50      Dwayne Bravo
1        Kuldeep Yadav    Ishan Kishan        81     Kuldeep Yadav
2          Odean Smith  Faf du Plessis        88    Mohammed Siraj
3       Mohammed Shami    Deepak Hooda        55    Mohammed Shami
4         Sanju Samson   Aiden Markram        57  Yuzvendra Chahal

  best_bowling_figure hightest_wickets
0               3--20                3
1               3--18                3
2               2--59                2
3               3--25                3
4               3--22                3

[5 rows x 21 columns]
```

[151]:
```
count = df['venue'].value_counts()
count
```

[151]:
```
venue
Wankhede Stadium, Mumbai                          21
Dr DY Patil Sports Academy, Mumbai                20
Brabourne Stadium, Mumbai                         16
Maharashtra Cricket Association Stadium,Pune      13
Eden Gardens, Kolkata                              2
Narendra Modi Stadium, Ahmedabad                   2
Name: count, dtype: int64
```

[172]:
```
new = sns.barplot( y = count.index , x = count.values)
plt.title('title = "Most Matches Played by Venue')
new
```

[172]:
```
<Axes: title={'center': 'title = "Most Matches Played by Venue'},
ylabel='venue'>
```



0.0.6 Custom Questions & Insights

Q1: Who won the highest margin by runs?

[187]:
```
df[df['won_by'] == 'Runs'].sort_values(by = 'margin',ascending= False).
 ↪head(1)[['match_winner','margin']]
```

[187]:

	match_winner	margin
54	Chennai	91

Q2: Which player had the highest individual score?

[191]:
```
df[df['highscore'] == df['highscore'].max()][['top_scorer','highscore']]
```

[191]:

	top_scorer	highscore
65	Quinton de Kock	140

Q3: Which bowler had the best bowling figures?

[3]:
```
df[df['hightest_wickets'] == df['hightest_wickets'].
 ↪max()][['best_bowling','best_bowling_figure']]
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[3], line 1
----> 1 df[df['hightest_wickets'] == df['hightest_wickets'].
 ↪max()][['best_bowling','best_bowling_figure']]

NameError: name 'df' is not defined
```

[ ]:

1 Good Work