Capstone Project - 1

IPL T20 Cricket Analysis

Team Members

- S. Sravya Sri
- Ankit Bansal
- Bhwesh Gaur

Table Of Content

Problem Statements

- Provided six files which includes data of players played, teams played, batsman's performance details, teams performance in home ground and away ground, each ball deliveries and each match details in entire IPL from 2008 to 2019.

- **Explore and analyze the data to discover results and statistics for different teams playing in Indian Premier League.**

Indian Premier League - IPL

The Indian Premier League (IPL) is an annual T20 cricket tournament that began in 2008.

It was founded by then BCCI vice-president Lalit Modi as a result of the popularity boom of the T20 format in India when the national team won the 2007 World T20.

The IPL takes place in March, April and May of every year and lasts for approximately seven weeks.

The IPL is a franchise-based competition with eight teams, each representing an Indian city.

Data Summary (Given)

Data Frame	Shape	Important Columns
players_df	(566, 5)	['Player_Name', 'DOB', 'Batting_Hand', 'Bowling_Skill', 'Country']
matches_df	(756, 19)	['id', 'Season', 'city', 'date', 'team1', 'team2', 'toss_winner', 'toss_decision', 'result', 'dl_applied', 'winner', 'win_by_runs', 'win_by_wickets', 'player_of_match', 'venue', 'umpire1', 'umpire2', 'umpire3', 'match_number']
deliveries_df	(179078, 21)	['match_id', 'inning', 'batting_team', 'bowling_team', 'over', 'ball', 'batsman', 'non_striker', 'bowler', 'is_super_over', 'wide_runs', 'bye_runs', 'legbye_runs', 'noball_runs', 'penalty_runs', 'batsman_runs', 'extra_runs', 'total_runs', 'player_dismissed', 'dismissal_kind', 'fielder']
most_runs_avg_sr_df	(516, 6)	['batsman', 'total_runs', 'out', 'numberofballs', 'average', 'strikerate']
teamwise_home_and_away_df	(14, 7)	['team', 'home_wins', 'away_wins', 'home_matches', 'away_matches', 'home_win_percentage', 'away_win_percentage']
teams_df	(15, 1)	['team1']

Data Summary (Added)

We added two shapefiles in-order to perform visualization on the map of India.

Data Frame	Shape	Important Columns
`map_state_df`	`(36, 2)`	`['st_nm', 'geometry']`
`map_district_df`	`(594, 13)`	`['NAME_2', 'geometry']`

Data Wrangling

Data type Conversion

- In matches_df, season column is converted from string to integer.
- In matches_df, date column is converted from string to datetime.

Sorting and Reordering

- Sorted matches_df based on date column.

Addressing missing or invalid data

- In players_df, changed NaN to ‘NA’.
- In most_runs_avg_sr_df changed NaN to 0.
- In matches_df, dropped NaN on winners column.
- In matches_df, filled NaN values of city with ‘Dubai’.

Data Wrangling (Contd.)

Data Transformation

- Merging columns: In matches_df, venue and city are merged.

Data Enrichment

- Added season wise match number in matches_df.
- Divided total runs by 20 in most_runs_avg_sr_df.

Deduplicating data

- In matches_df, venue and city are deduplicated.
- Team names are deduplicated from every dataset, wherever present.

Analysis Approach

We divided the whole analysis into three parts:

01 Match Wise Team Analysis

02 Season Wise Team Analysis

03 Entire IPL Analysis

Match Wise Team Analysis

Select a Match Number

[167] Match_No: 352

Scorecard

```
[169] 1 scorecard(inning = 1)
```

```
*******************************************
                Inning: 1
*******************************************
Mumbai Indians
*******************************************
| Score: 194/7 | Target: 195 | Overs: 20.0 |
```

Scorecard:

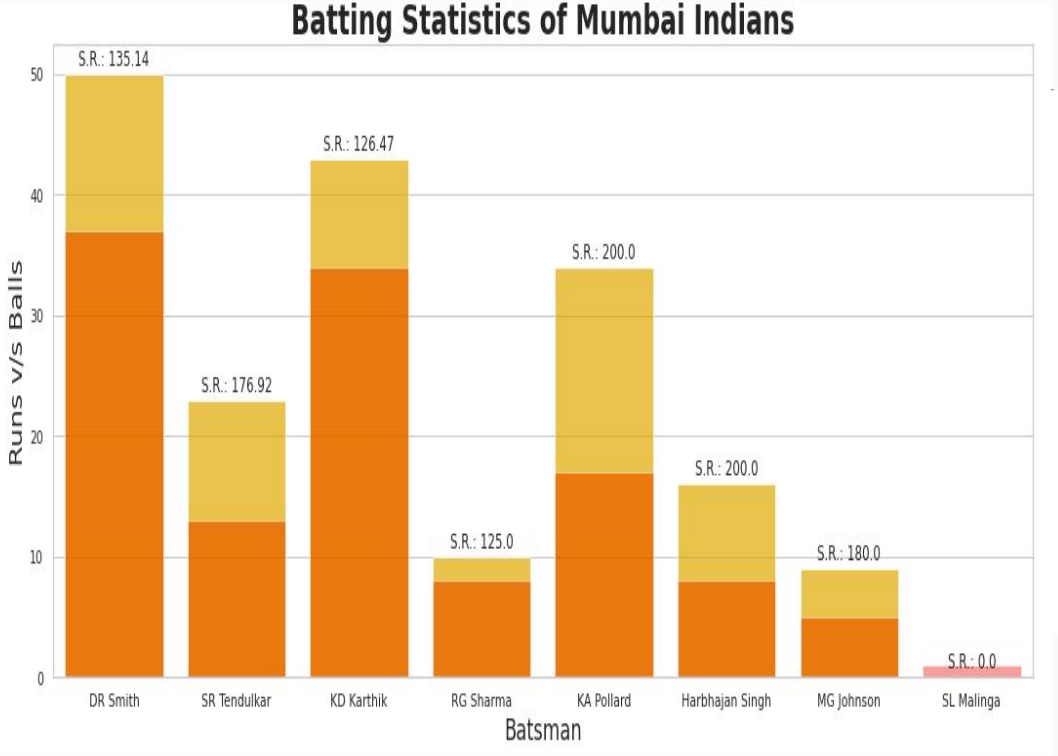
	batsman	runs	ball	strike_rate	bowler	over	runs	wickets	economy
0	DR Smith	50	37	135.14	R Rampaul	4.0	35	1	8.75
1	SR Tendulkar	23	13	176.92	RP Singh	4.2	41	1	9.46
2	KD Karthik	43	34	126.47	JD Unadkat	4.0	32	0	8
3	RG Sharma	10	8	125.00	R Vinay Kumar	4.1	43	1	10.32
4	KA Pollard	34	17	200.00	J Syed Mohammad	4.0	43	1	10.75
5	Harbhajan Singh	16	8	200.00					
6	MG Johnson	9	5	180.00					
7	SL Malinga	0	1	0.00					

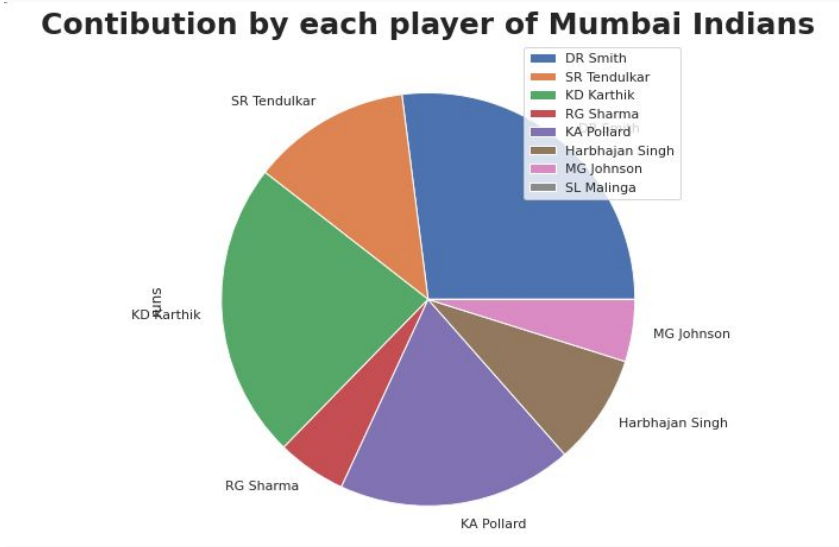
Match Result

```
##################################
      <<<< MATCH RESULT >>>>
**********************************
```

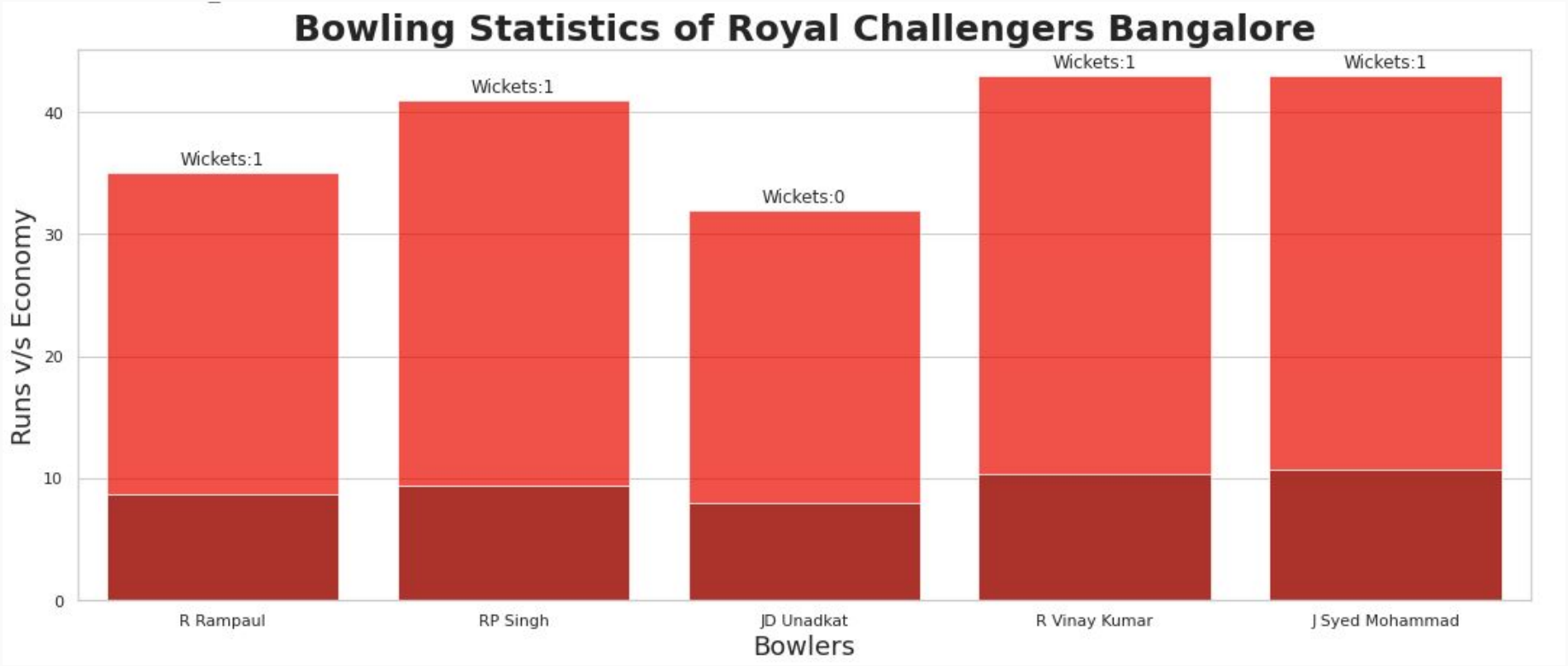
	Result
DL Applied	No
Winner	Mumbai Indians
Loser	Royal Challengers Bangalore
Win By	58 Runs
Tied	No

Inning Statistics

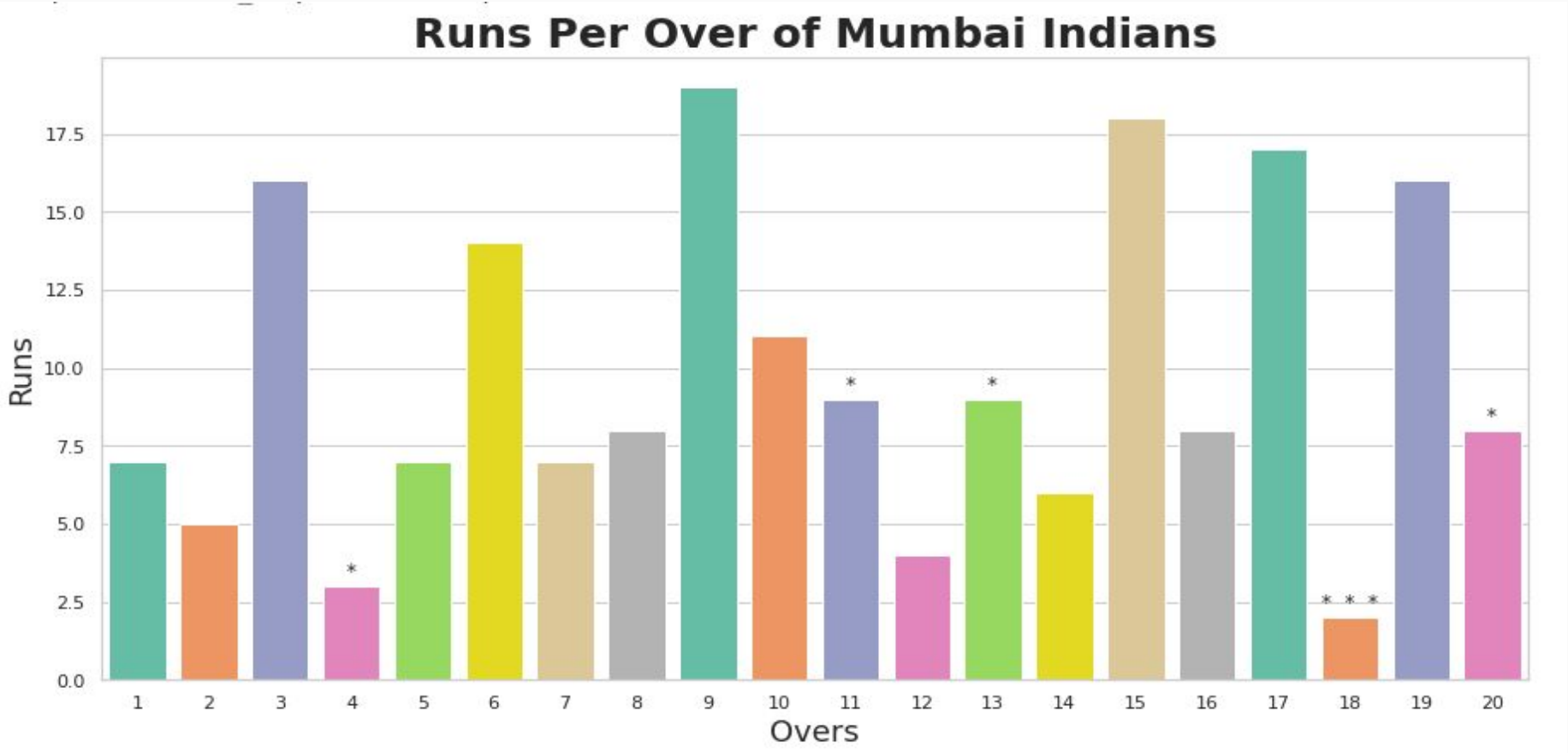




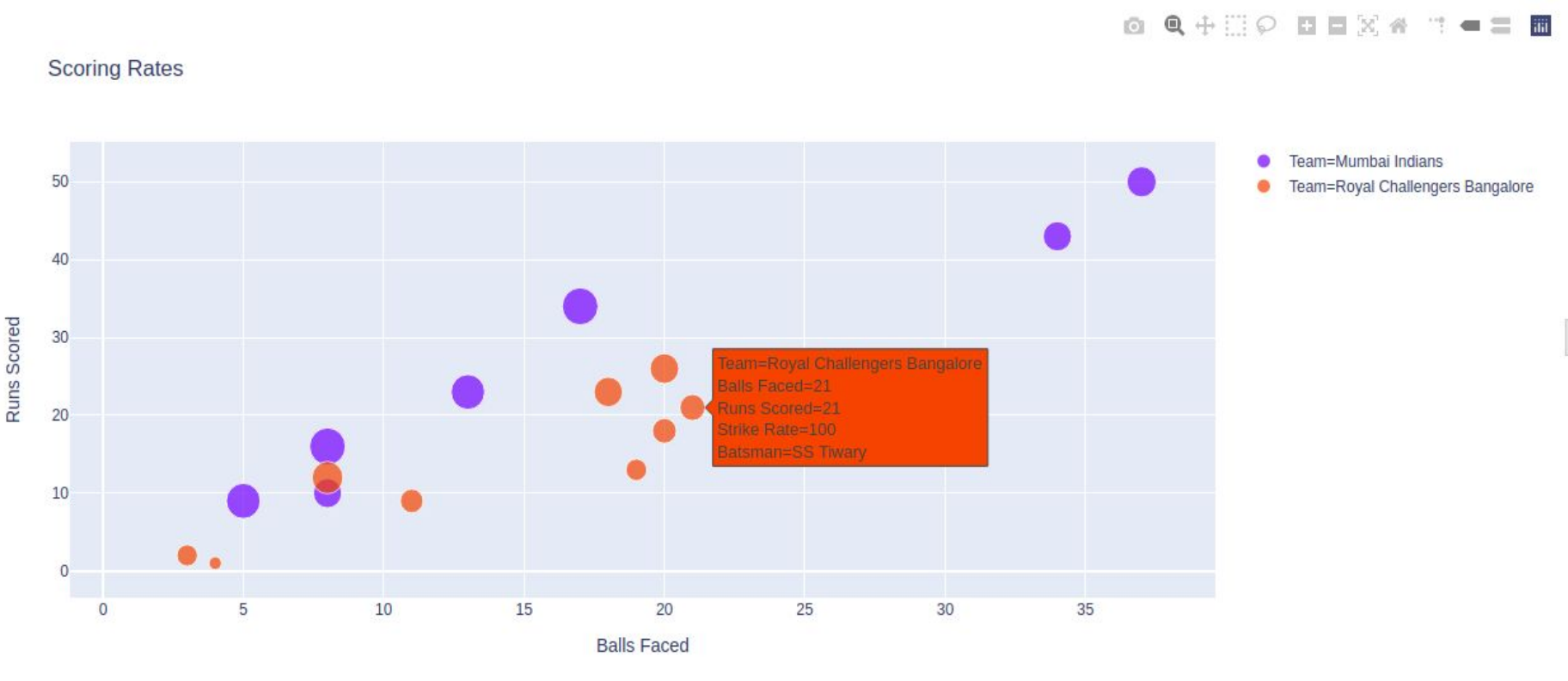
Inning Statistics



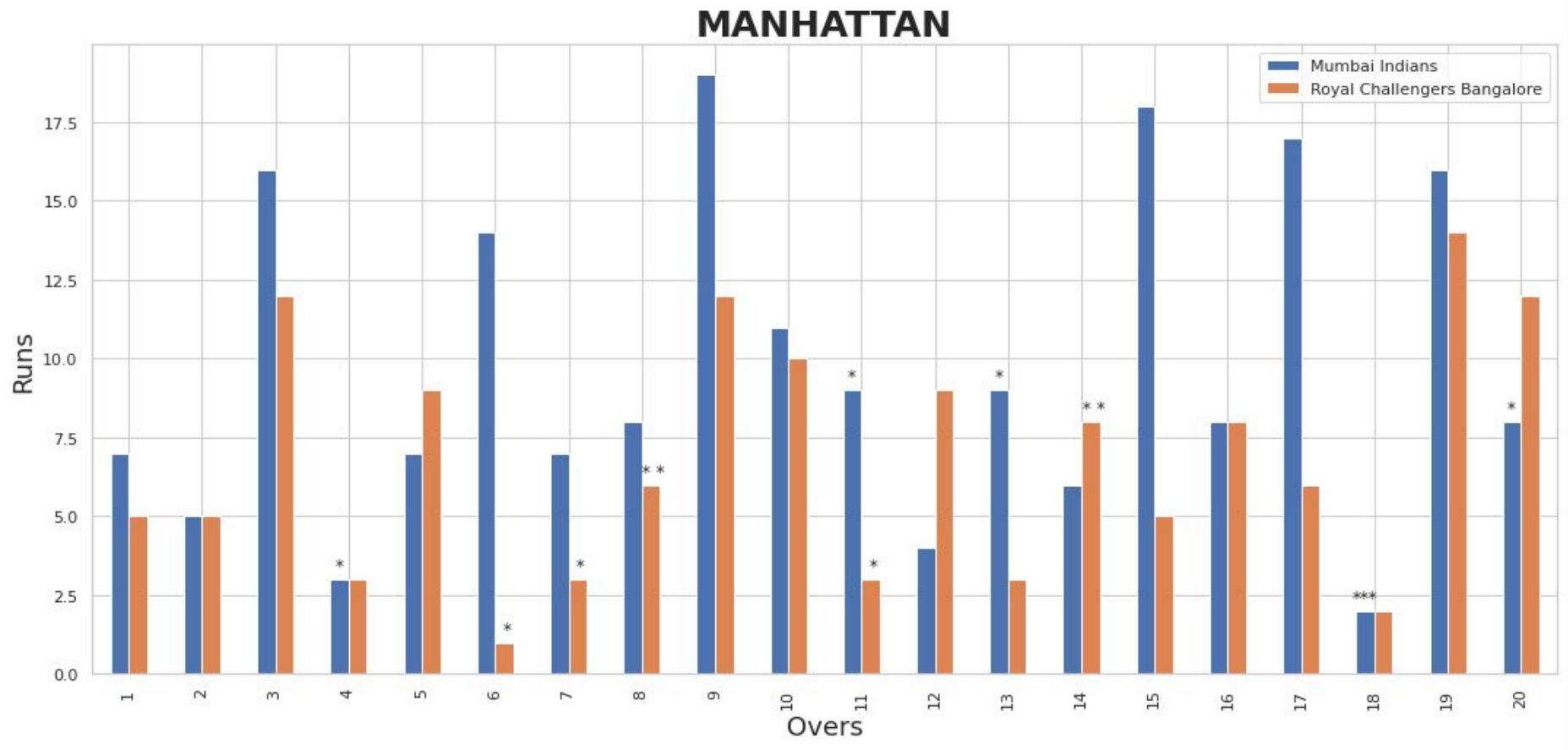
Inning Statistics



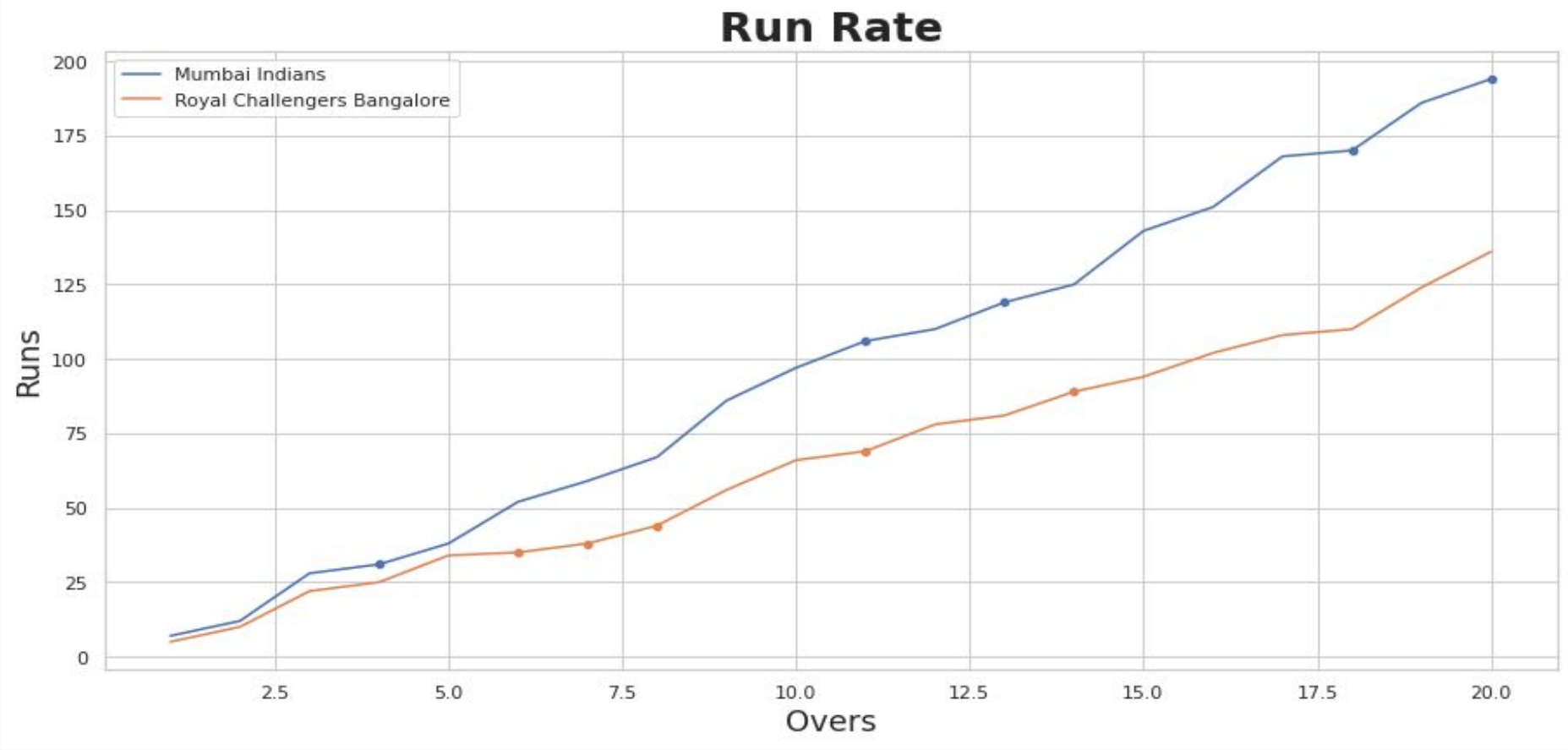
Inning Statistics



Match Statistics



Match Statistics

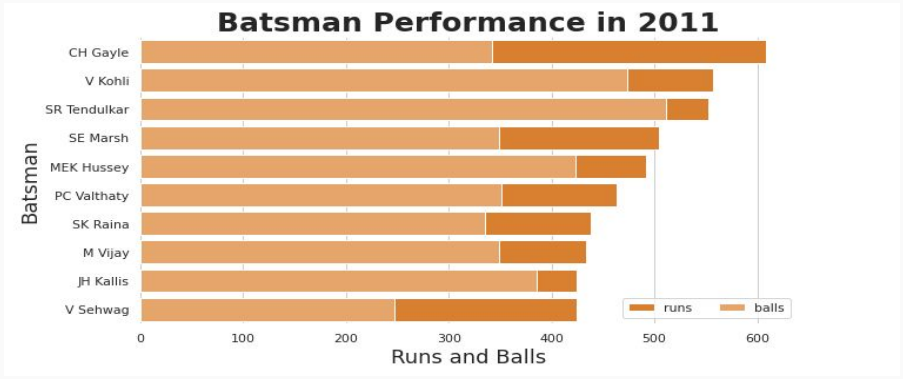


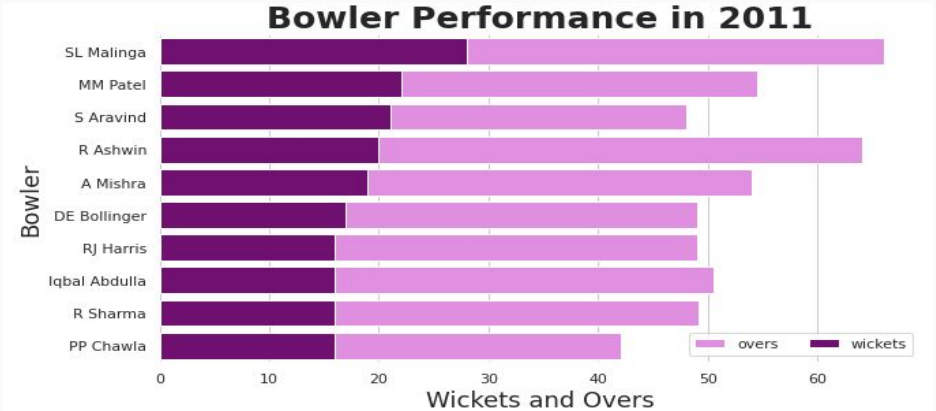
Season Wise Team Analysis

Season: 2011

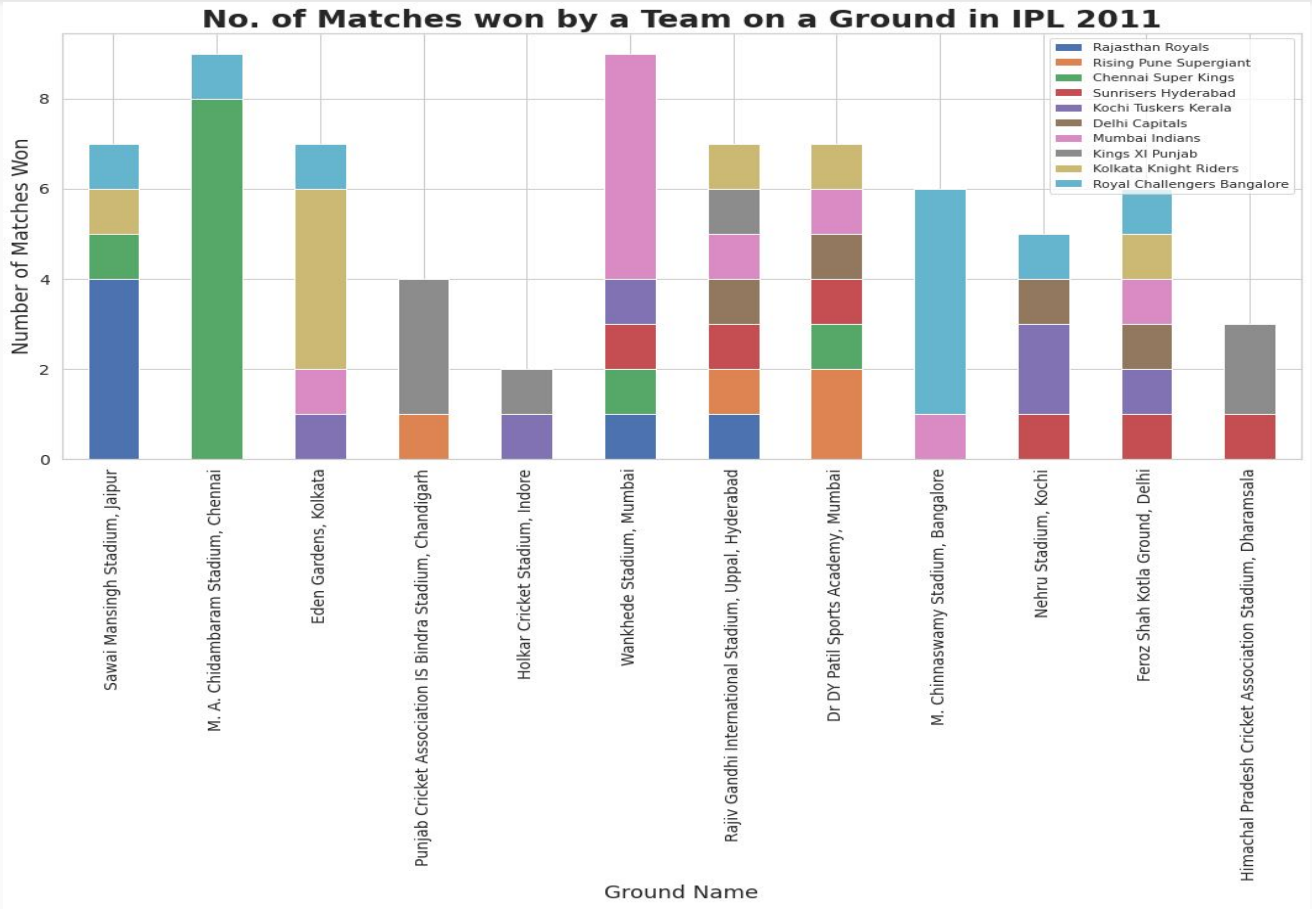
IPL Seaon: 2011

	team_name	matches_won	points
0	Chennai Super Kings	10	20
1	Delhi Capitals	4	8
2	Kings XI Punjab	7	14
3	Kochi Tuskers Kerala	6	12
4	Kolkata Knight Riders	8	16
5	Mumbai Indians	9	18
6	Rajasthan Royals	6	12
7	Rising Pune Supergiant	4	8
8	Royal Challengers Bangalore	10	20
9	Sunrisers Hyderabad	6	12

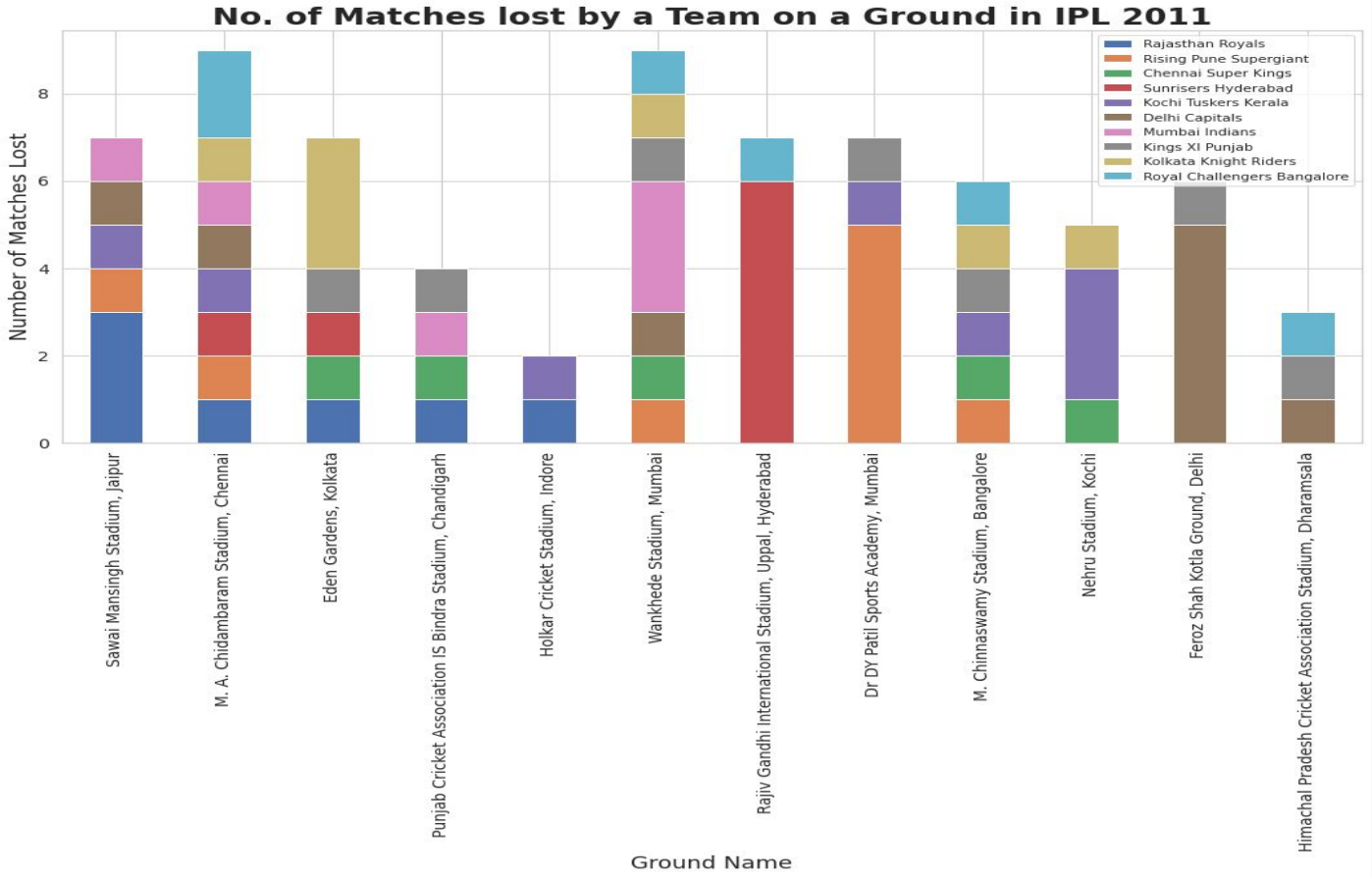




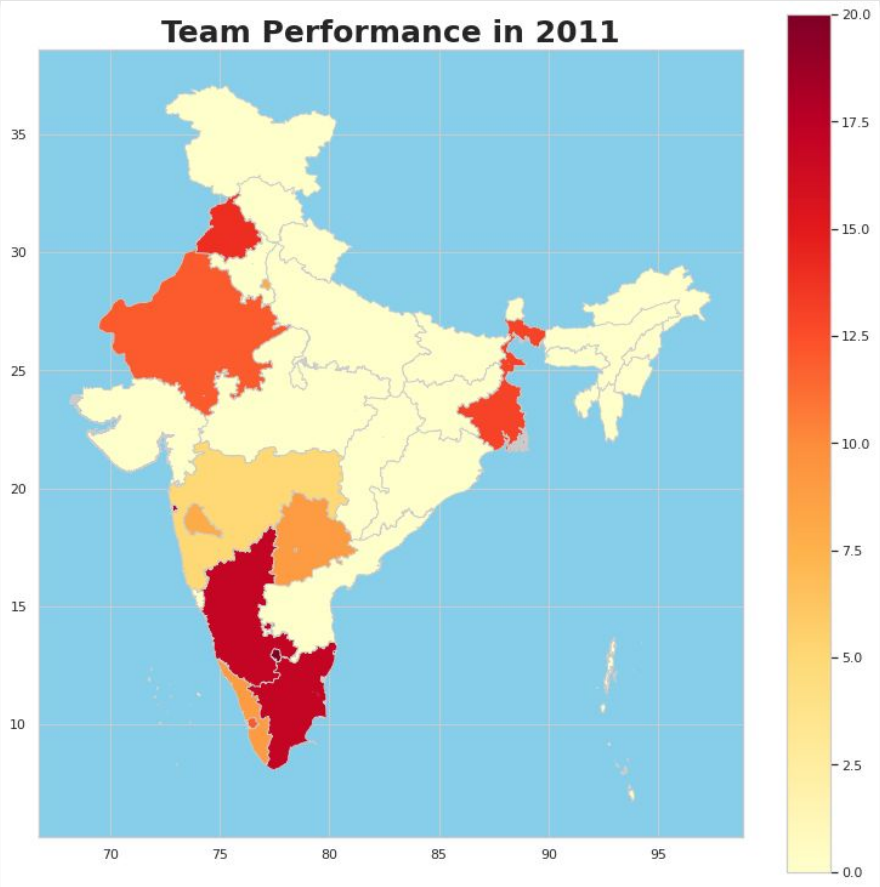
Season Wise Team Analysis



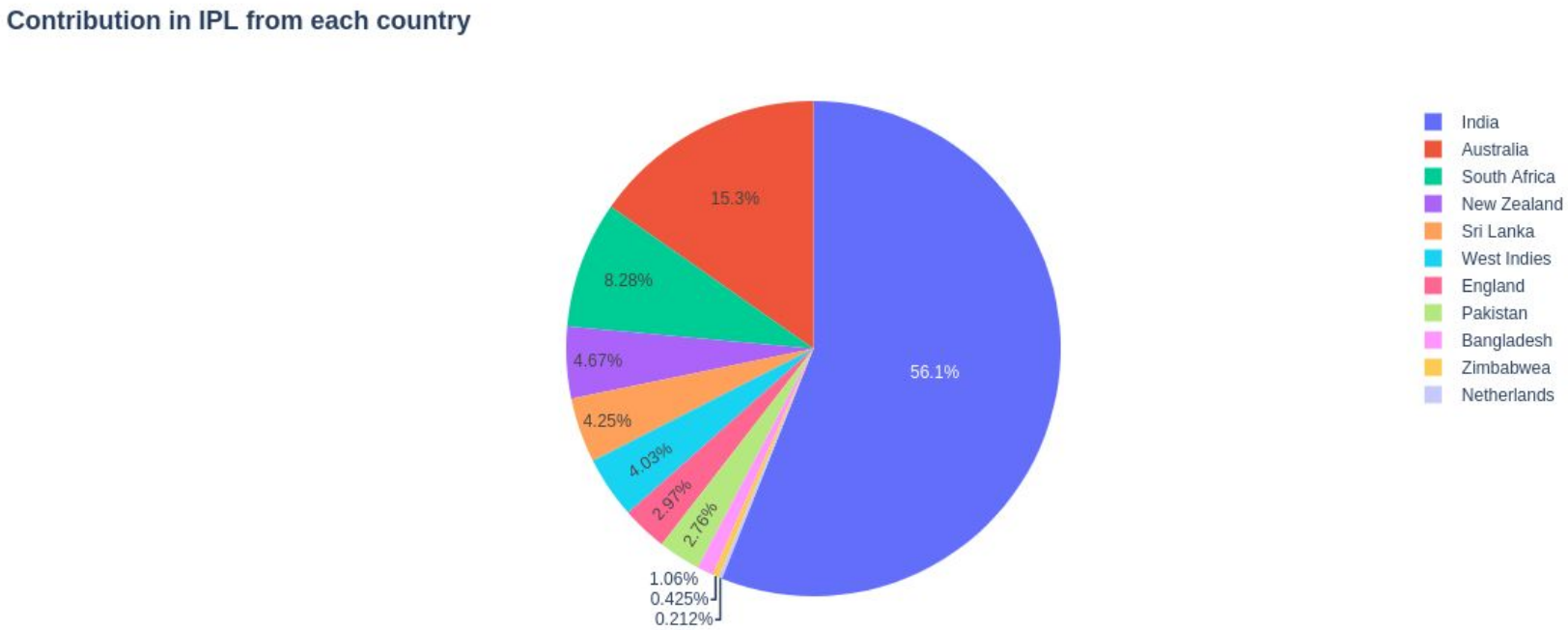
Season Wise Team Analysis



Season Wise Team Analysis



Entire IPL Analysis



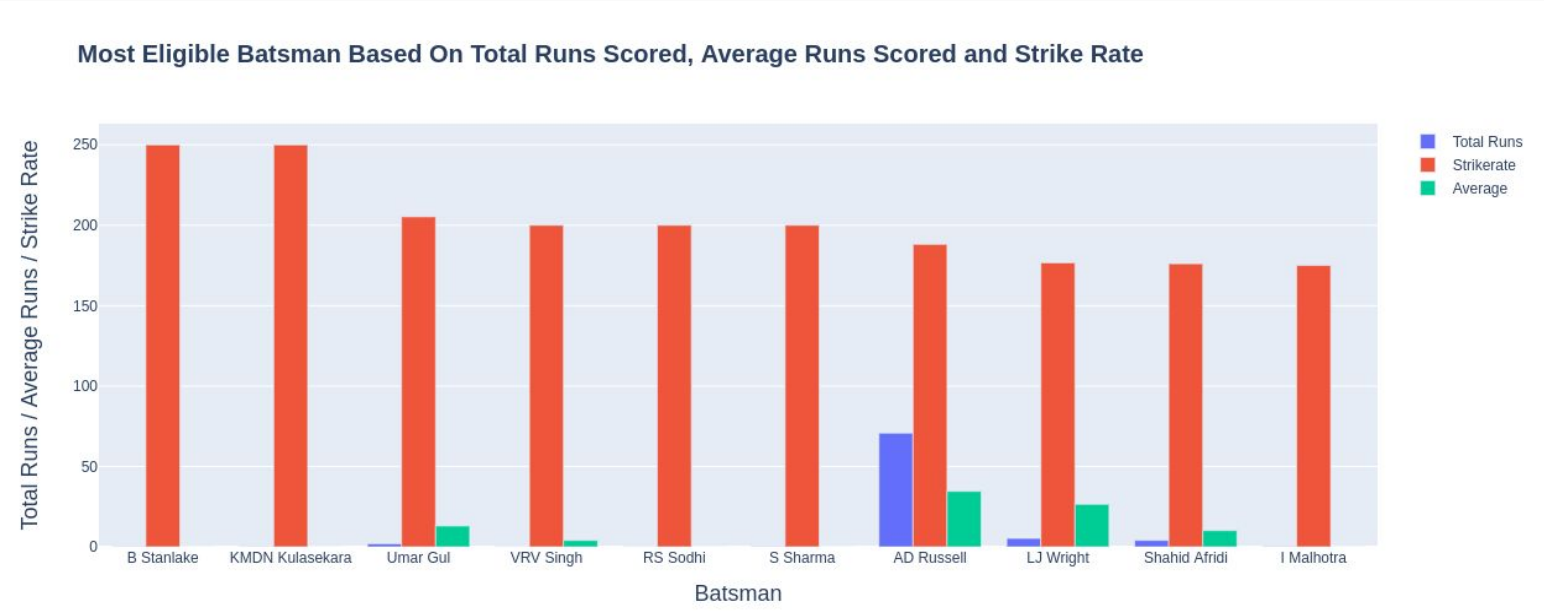
Entire IPL Analysis

Total_Runs: ☐

Strikerate: ☑

Average: ☐

Top: 10



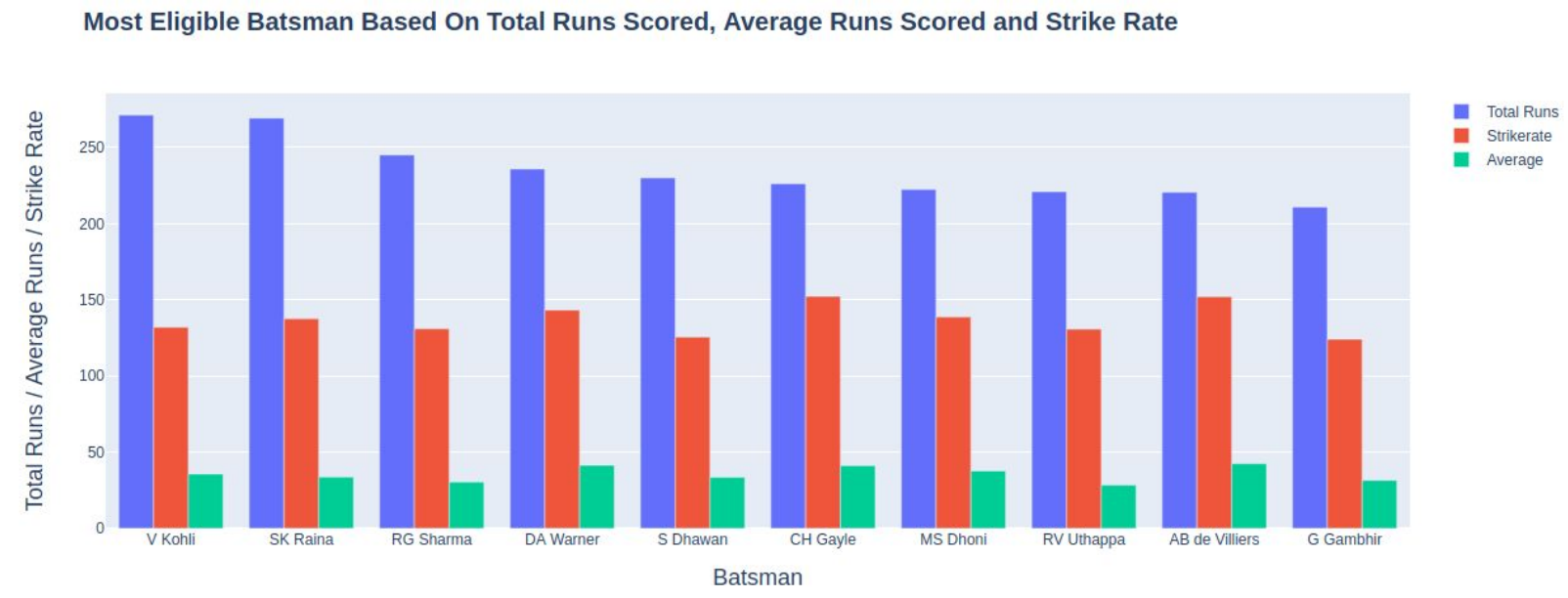
Entire IPL Analysis

Total_Runs: ☑

Strikerate: ☐

Average: ☐

Top:



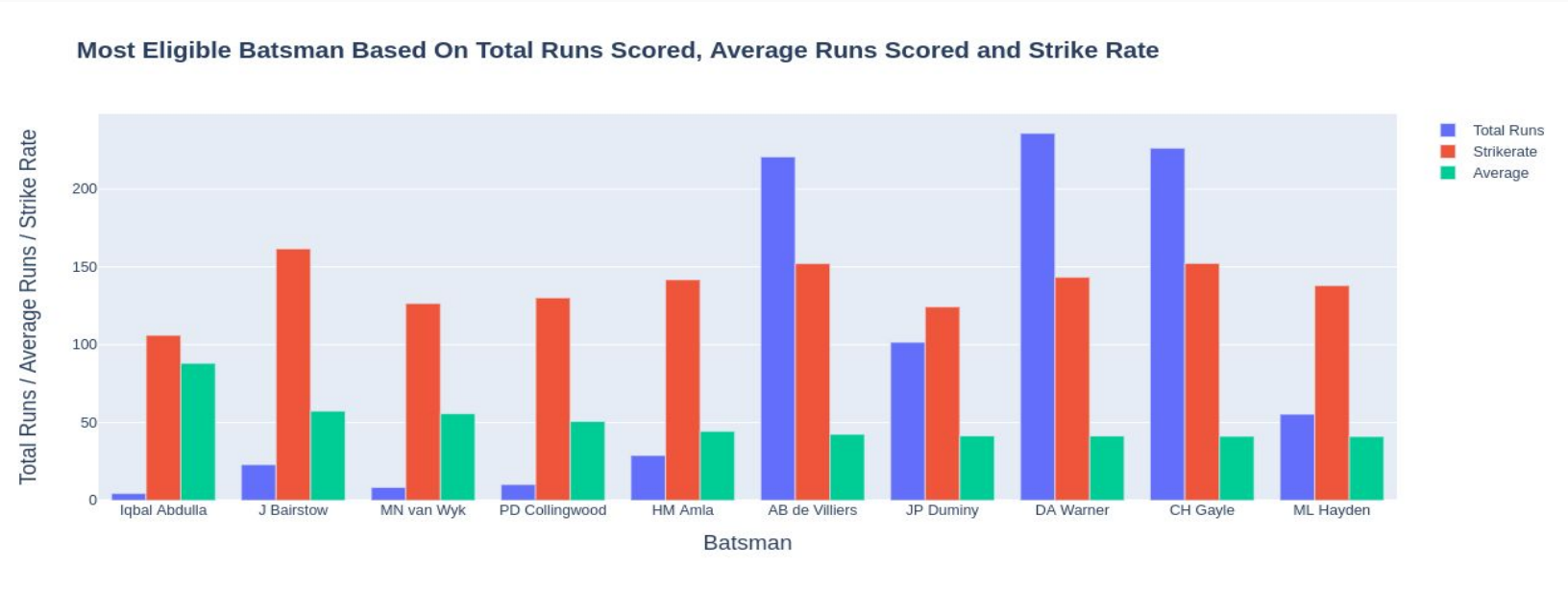
Entire IPL Analysis

Total_Runs: ☐

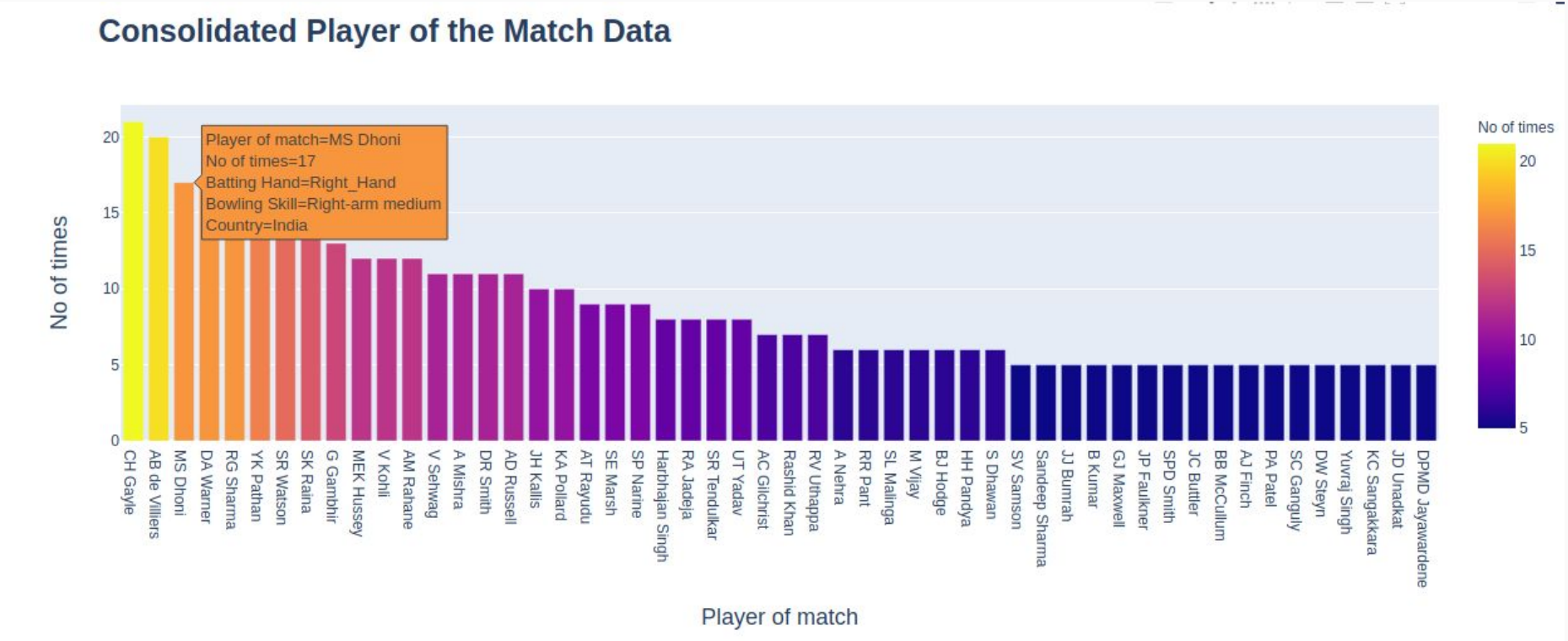
Strikerate: ☐

Average: ☑

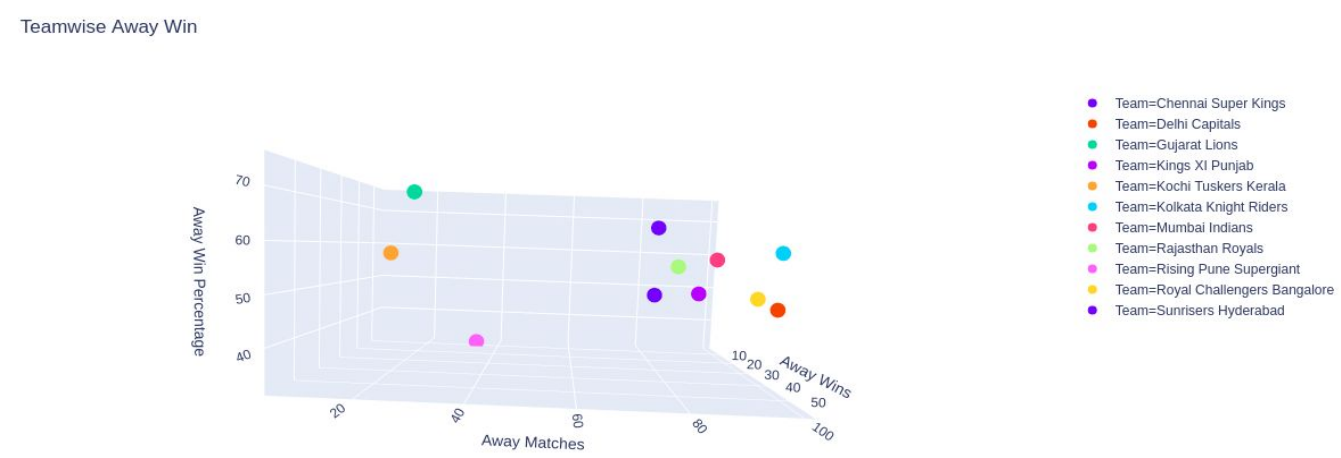
Top: 10

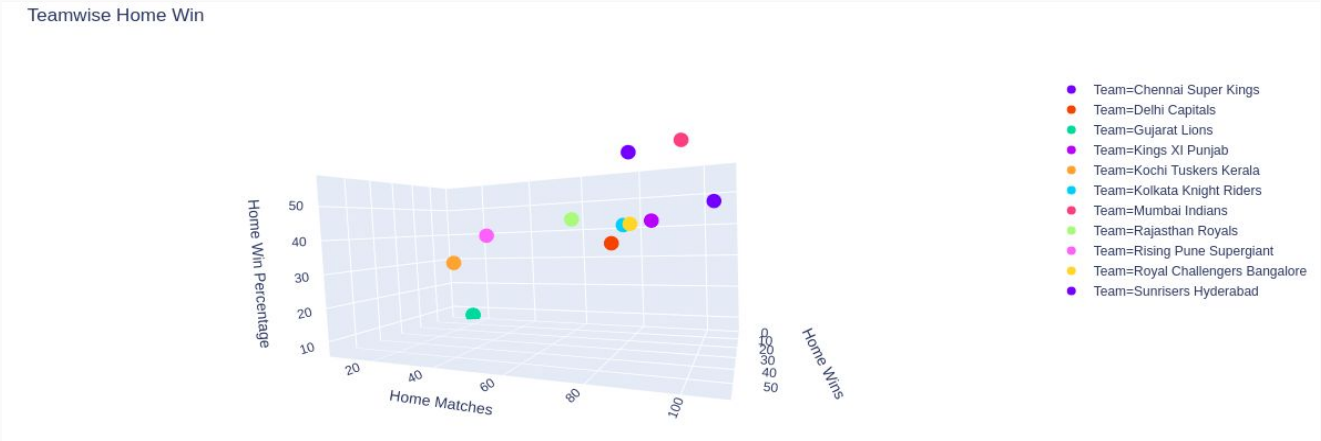


Entire IPL Analysis

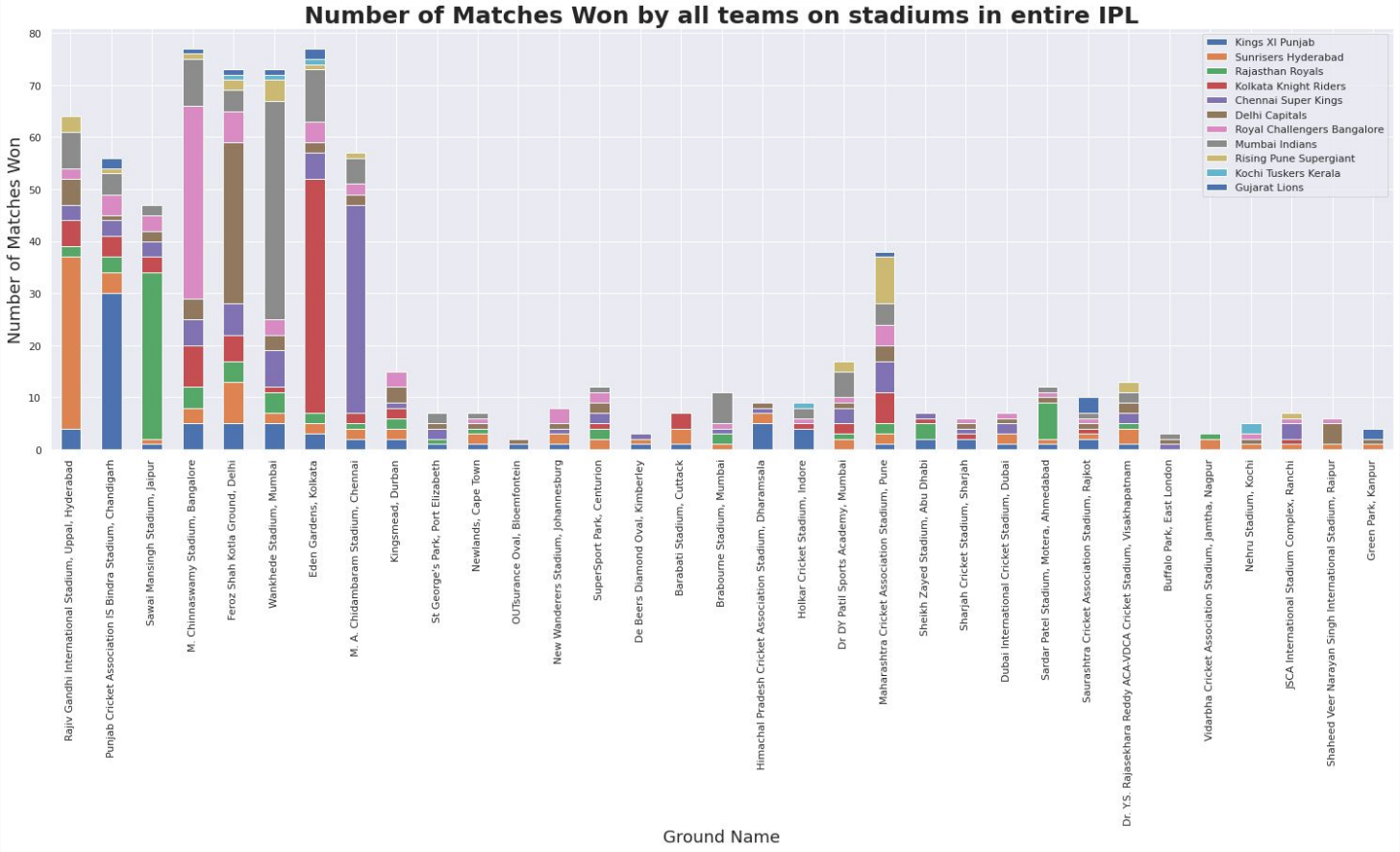


Entire IPL Analysis

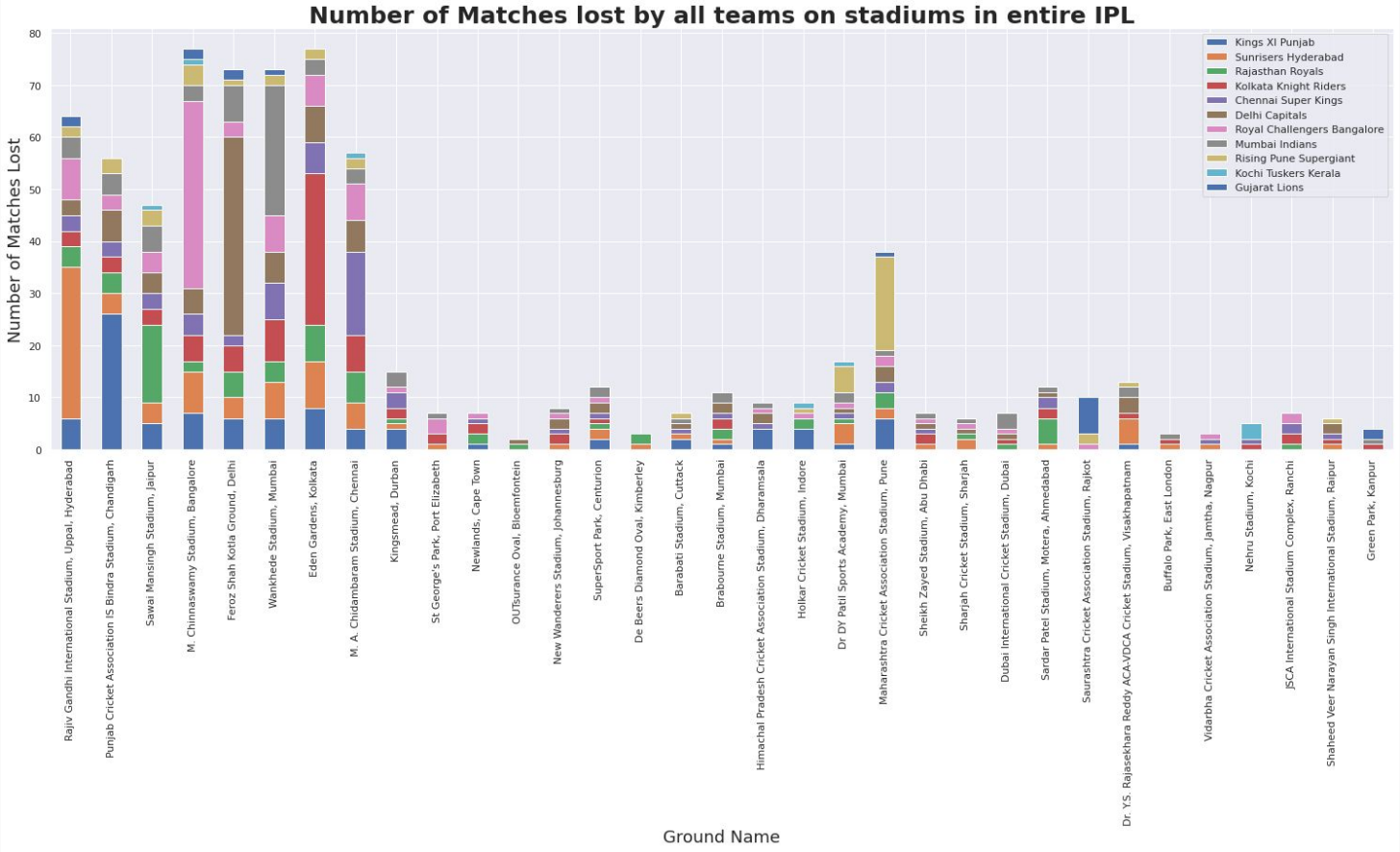




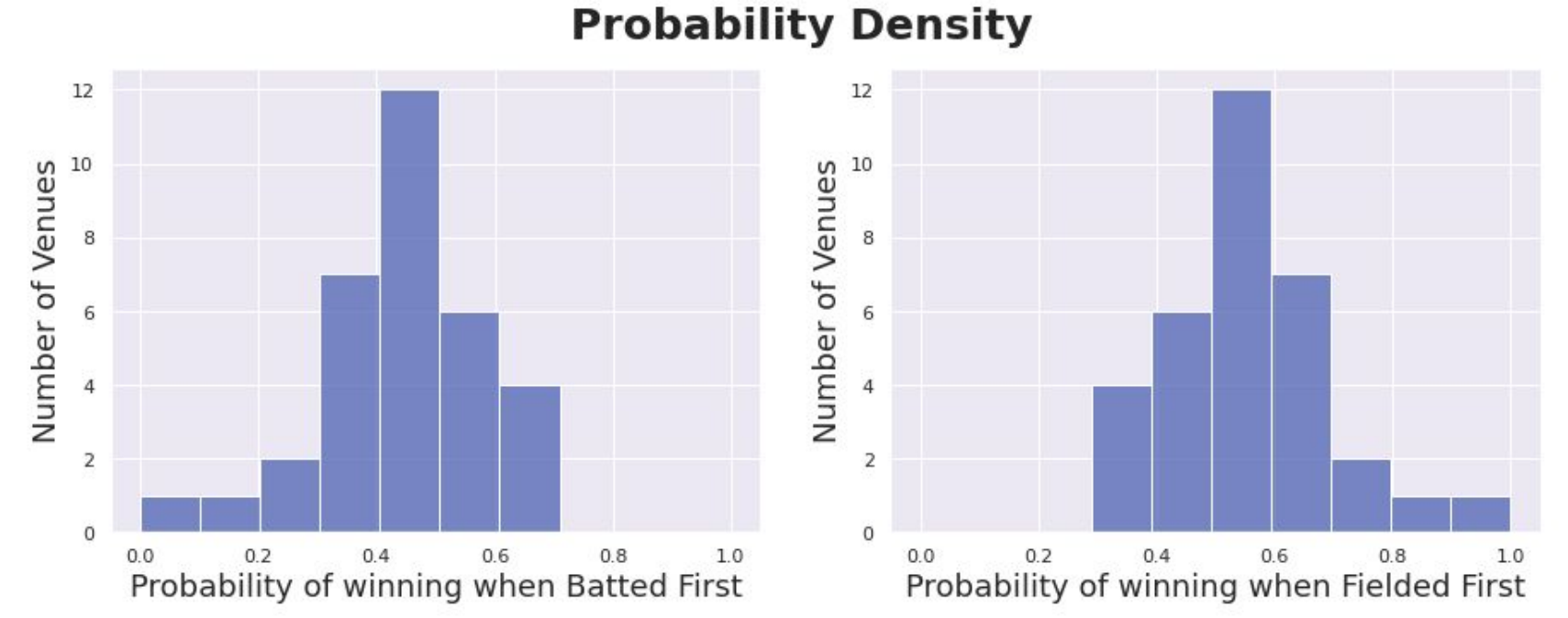
Entire IPL Analysis



Entire IPL Analysis



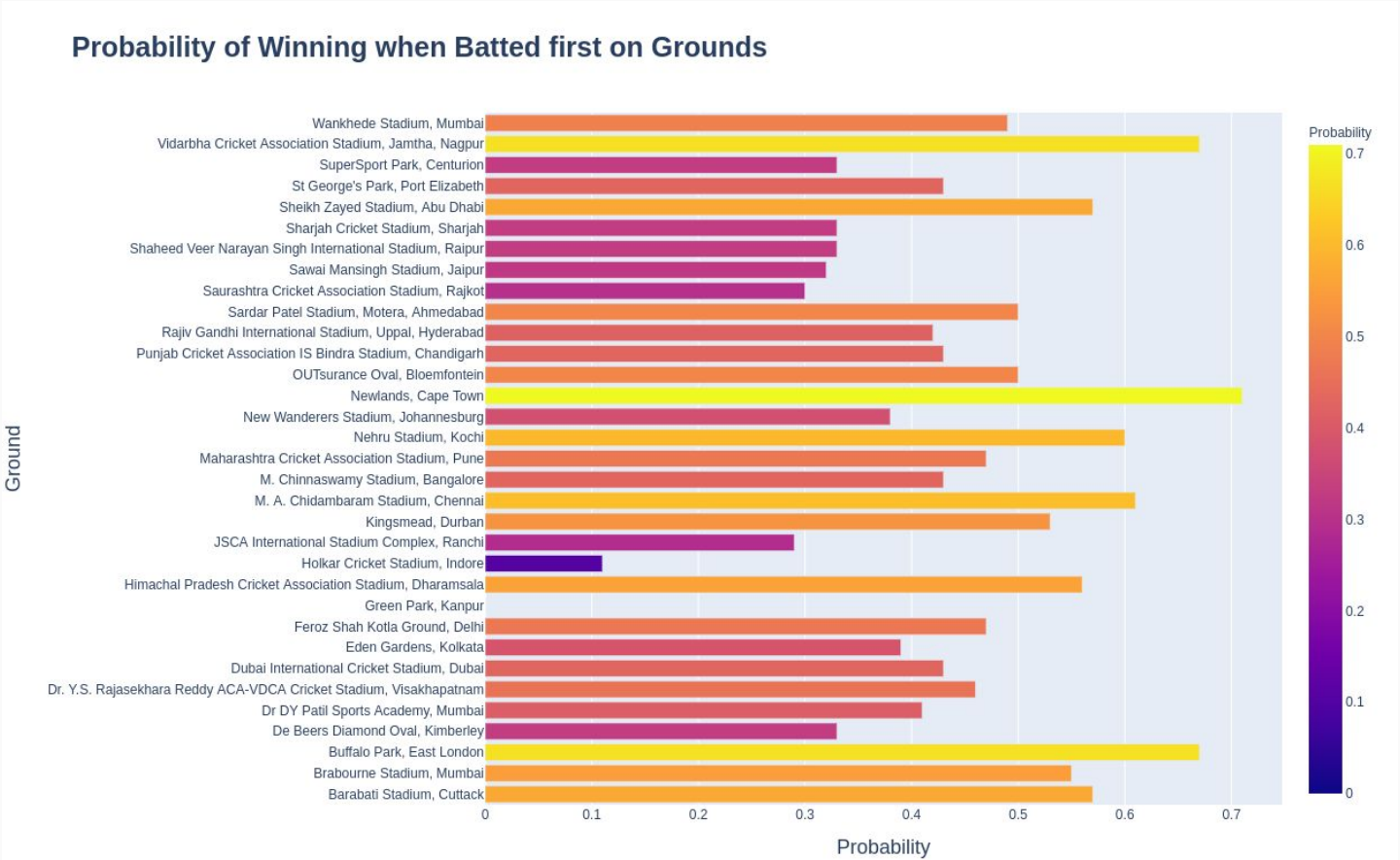
Entire IPL Analysis



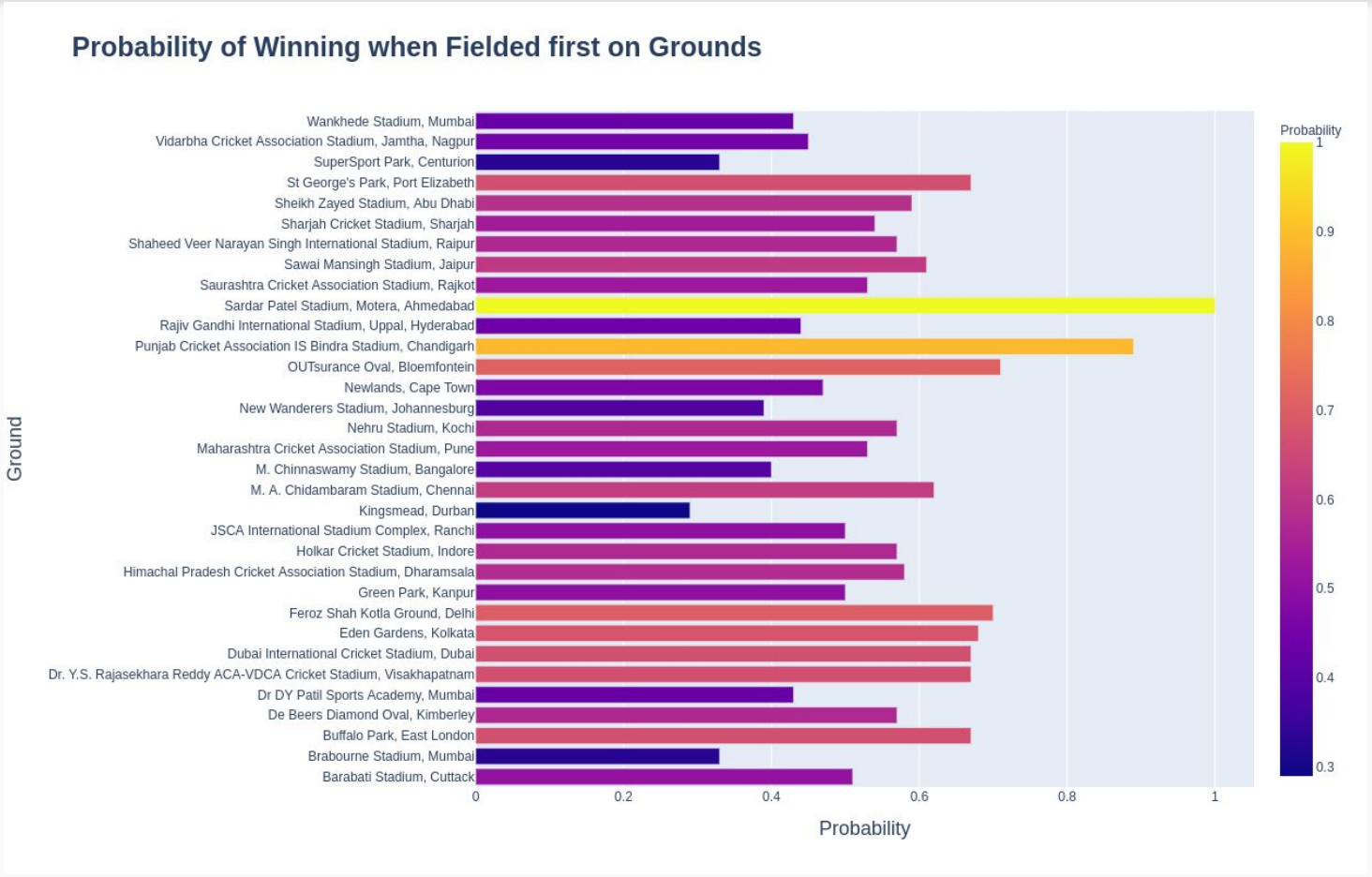

```
venue_probability_df.loc[venue,'prob_win_batfirst'] = round(bat_first_win/venue_probability_df.loc[venue,'total_matches_played'], 2)
```

```
venue_probability_df.loc[venue,'prob_win_fieldfirst'] = round(field_first_win/venue_probability_df.loc[venue,'total_matches_played'],2)
```

Entire IPL Analysis

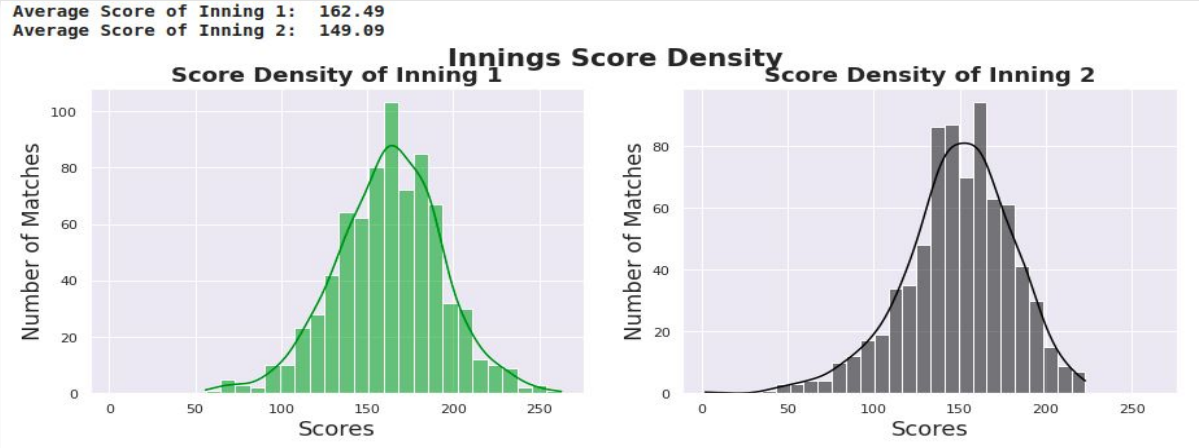


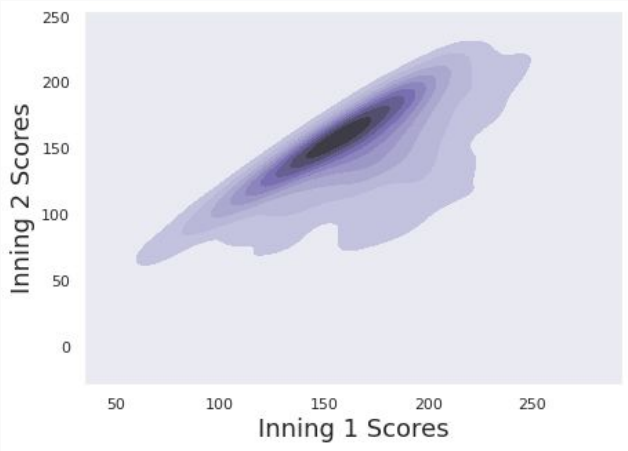
Entire IPL Analysis



Entire IPL Analysis

```
Average Score of Inning 1:  162.49
Average Score of Inning 2:  149.09
```



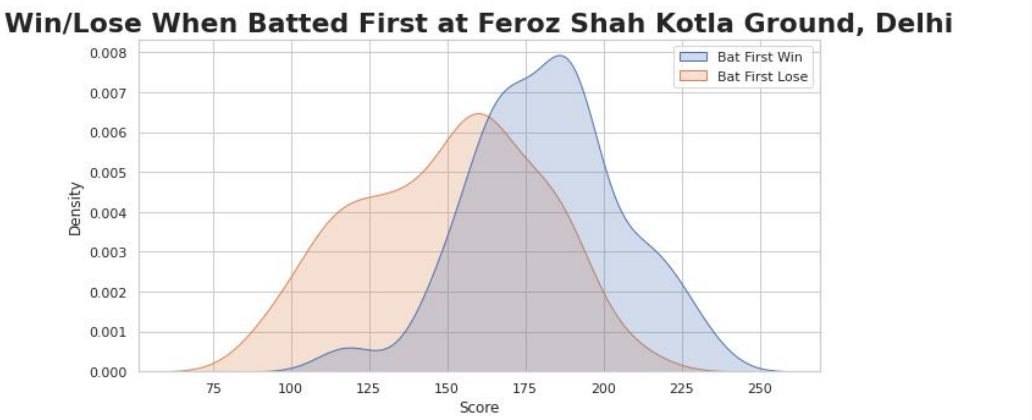


Entire IPL Analysis

Venue: Feroz Shah Kotla Ground, Delhi

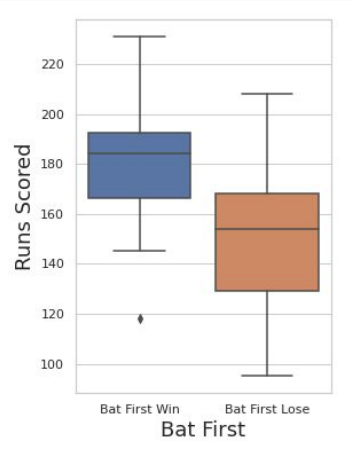
Inning: 1

```
==========================================================================================
Average Runs Scored in Feroz Shah Kotla Ground, Delhi in First Inning is 165.81
Which implies that this Ground's pitch is a NORMAL pitch.
==========================================================================================
```




```
# Average Runs Scored
avg_score_inng =
((req_match_id_df_1['bat_first_win'].sum()/len(req_
match_id_df_1.index)) +
(req_match_id_df_2['bat_first_lose'].sum()/len(req_
match_id_df_2.index)))/2
```

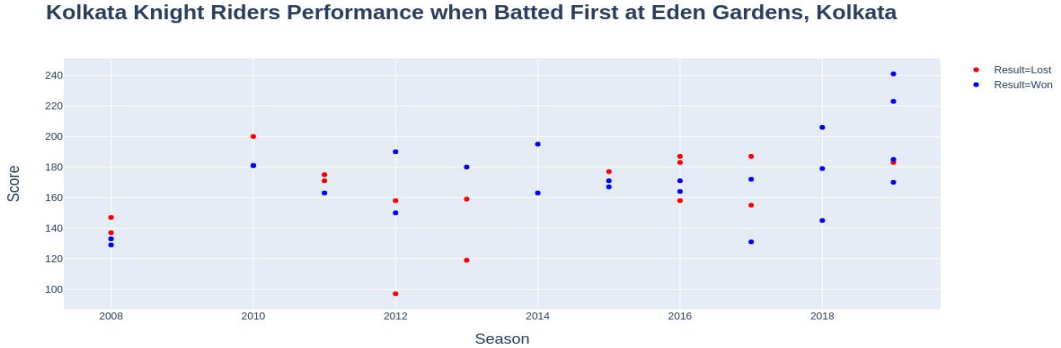
```
for id in req_match_id_df_1.id:
        temp_df = deliveries_df[deliveries_df.match_id ==
id][['match_id','total_runs','inning']]
         temp_df.where(temp_df['inning']==1, inplace= True)
         temp_df.dropna(inplace = True)

req_match_id_df_1['bat_first_win'][req_match_id_df_1.id == id] =
temp_df[temp_df.match_id == id]['total_runs'].sum()
```



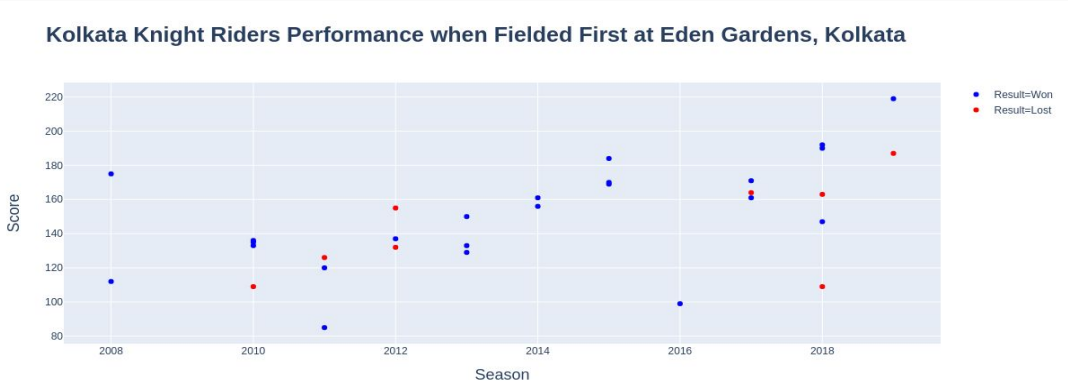
Entire IPL Analysis

Venue: Eden Gardens, Kolkata

Team: Kolkata Knight Riders



	season	team1	team1_score	team1_wickets	team2	team2_score	team2_wickets	result	result_type
0	2008	opposite_team	110	10	Kolkata Knight Riders	112	5	Won by 5 wickets	Won
1	2008	Kolkata Knight Riders	137	8	opposite_team	138	3	Lost	Lost
2	2008	Kolkata Knight Riders	133	6	opposite_team	110	10	Won by 23 runs	Won
3	2008	Kolkata Knight Riders	147	8	opposite_team	150	4	Lost	Lost
4	2008	opposite_team	174	6	Kolkata Knight Riders	175	7	Won by 3 wickets	Won
...	...	...	...	...	...	...	...	...	...




```
if
deliveries_df[deliveries_df['match_id']==id]['batting_team'].v
alues[0] == team:
            team_a = team
            team_b = 'opposite_team'
```

Challenges

- Implementation of map needs two different shapefiles, one for state boundaries and another for city boundaries, using the two in a way that only required city and state is visible, was a difficult task.
- Creation of scorecard involves multiple dataframe to work with, and dataframe like deliveries_df contains columns like leg-bye runs, penalty runs, dismissal kind etc which needs to handle before showing the score card.
- Duplicated data like stadium names required research before de-duplication because one stadium is named in two different ways and that needs to be handled.
- We found many discrepancies in the datasets like matches tied (incorrect data), trash deliveries, irrelevant rows in matches_df, etc. which took a lot of time to resolve.
- Handling NaN values in the datasets.

Conclusion

➔ **Match-wise team analysis**
- Demonstrate the inning statistics of players and teams in each match.
- We formed out a conclusion over the runs made by each team per over and compared teams run rate.

➔ **Season-wise team analysis**
- Draws an idea regarding the various batsmen and bowlers' performance in a particular season.
- Provided a comparison of the matches won or lost by teams on the ground in a season.
- Identified the leaderboard position of teams through which we plotted a heat map that showed us the top team for the season.

➔ **Entire IPL analysis**
- Through probability density, we have identified the winning probability on a ground which could help a team to decide whether to bat first or field first provided they won the toss.
- Through the score density graph, we have estimated the minimum and maximum targets.
- Through match result analysis, we can estimate the winning and losing chances in a match, and the boxplot signifies the minimum, maximum average, and exceptional(outliers) score at a ground.

THANK - YOU