IPL T20 Analysis EDA Capstone Project

Akash K,
Data science Trainee,
AlmaBetter, Bangalore.

Abstract:

The Indian Premier League (IPL) is a professional Twenty20 cricket league, contested by eight teams based out of eight different Indian cities/states. In this EDA Project, we were provided with six datasets: players, matches, deliveries, most runs_avg_strikerate, teams and teamwise_home_and_away_win.

Keywords: EDA, Data Correlation and Trends in Data

1.Problem Statement

Indian Premier League(IPL) is a professional Twenty20 cricket league in India contested during March or April and May of every year by eight teams representing eight different cities in India. The league was founded by the Board of Control for Cricket in India(BCCI) in 2008.

Explore and analyze the data to discover results and statistics for different teams playing in IPL.

2. Introduction

The problem statement comprises six datasets each with different columns and different information.

These six datasets contain almost ~52500 null entries and it contains data of 13 seasons of IPL T20 which consists of every delivery bowled in all the seasons.

2.1 Delivery Dataset

#	Column	Non-Null Count	Dtype
0	match_id	179078 non-null	int64
1	inning	179078 non-null	int64
2	batting_team	179078 non-null	object
3	bowling_team	179078 non-null	object
4	over	179078 non-null	int64
5	ball	179078 non-null	int64
6	batsman	179078 non-null	object
7	non_striker	179078 non-null	object
8	bowler	179078 non-null	object
9	is_super_over	179078 non-null	int64
10	wide_runs	179078 non-null	int64
11	bye_runs	179078 non-null	int64
12	legbye_runs	179078 non-null	int64
13	noball_runs	179078 non-null	int64
14	penalty_runs	179078 non-null	int64
15	batsman_runs	179078 non-null	int64
16	extra_runs	179078 non-null	int64
17	total_runs	179078 non-null	int64
18	player_dismissed	179078 non-null	object
19	dismissal_kind	179078 non-null	object
20	fielder	179078 non-null	object

The above picture represents the columns of the Delivery Dataset after processing the data.

2.2 Matches Dataset

The matches dataset consists of the season, match details, umpires, player of the match, venue and a lot more.

The data is processed as shown in the picture below.

#	Column	Non-Null Count	Dtype
0	id	756 non-null	int64
1	Season	756 non-null	object
2	city	756 non-null	object
3	date	756 non-null	object
4	team1	756 non-null	object
5	team2	756 non-null	object
6	toss_winner	756 non-null	object
7	toss_decision	756 non-null	object
8	result	756 non-null	object
9	dl_applied	756 non-null	int64
10	winner	756 non-null	object
11	win_by_runs	756 non-null	int64
12	win_by_wickets	756 non-null	int64
13	player_of_match	756 non-null	object
14	venue	756 non-null	object
15	umpire1	756 non-null	object
16	umpire2	756 non-null	object

The umpire1 and umpire2 null values are filled with the values of NaN.

2.3 Players Dataset

The data has been processed as below to reduce the null values as shown in the figure.

#	Column	Non-Null Count	Dtype
0	Player_Name	566 non-null	object
1	DOB	566 non-null	object
2	Batting_Hand	566 non-null	object
3	Bowling_Skill	566 non-null	object
4	Country	566 non-null	object

2.4 Players Stats Dataset:

The data has been processed as below to reduce the null values as shown in the figure.

#	Column	Non-Null Count	Dtype
0	batsman	516 non-null	object
1	total_runs	516 non-null	int64
2	out	516 non-null	int64
3	numberofballs	516 non-null	int64
4	average	516 non-null	float64
5	strikerate	516 non-null	float64

2.5 Teams Dataset:

The data has been processed as below to reduce the null values as shown in the figure.

#	Column	Non-Null Count	Dtype
0	team1	15 non-null	object

2.6 Teams Home and Away Dataset:

The data has been processed as below to reduce the null values as shown in the figure.

#	Column	Non-Null Count	Dtype
0	team	14 non-null	object
1	home_wins	14 non-null	int64
2	away_wins	14 non-null	int64
3	home_matches	14 non-null	int64
4	away_matches	14 non-null	int64
5	home_win_percentage	14 non-null	float64
6	away_win_percentage	14 non-null	float64

3. Data Handling and Removal of null values:

All the six datasets consists of many null values those are replaced in the following method:

3.1 Delivery Dataset:

Following are the null values present in the Dataset:

```
player_dismissed   8834 non-null
dismissal_kind     8834 non-null
fielder            6448 non-null
```

- Player_dismissed column is filled with 0s as the column contains null values where the players aren't dismissed.
- Dismissal_kind too have been filled with 0.
- Fielder too carries the same intuition.

3.2 Matches Dataset:

Following are the null values present in the Dataset:

```
city              749 non-null
winner            752 non-null
player_of_match   752 non-null
umpire1           754 non-null
umpire2           754 non-null
umpire3           119 non-null
```

- City is replaced with 'Dubai' .
- Winner and player of the match is replaced with 'Not Declared' as the games are tied.
- Umpire1 and Umpire2 are filled with the respective umpires on the respective matches through the match list.

3.3 Players Dataset:

Following are the null values present in the Dataset:

```
DOB             471 non-null
Batting_Hand    563 non-null
Bowling_Skill   502 non-null
Country         471 non-null
```

- Batting hand's null values are replaced with 'Right Hand' and Bowling Skill with 'Doesn't Bowl'
- Country is replaced in the same distribution as the non null values
- DOB has been replaced with the average age.

3.4 Players Stats Dataset:

Following are the null values present in the Dataset:

```
average         482 non-null
```

- Average is filled with the average by giving the value of runs scored. As the batsman is not out and divisibility by zero is NaN, runs scored are filled.

3.5 Teams Dataset:

```
team1    15 non-null
```

There are no null values and the values are sane.

3.6 Team home and away wins Dataset:

```
team                  14 non-null
home_wins             14 non-null
away_wins             14 non-null
home_matches          14 non-null
away_matches          14 non-null
home_win_percentage   14 non-null
away_win_percentage   14 non-null
```

There are no null values and the values are sane.

4.Data Correlation and Exploratory data analysis:

4.1 By Teams:

4.1.1 Highest Runs:

	batting_team	total_runs
0	Mumbai Indians	29809
1	Royal Challengers Bangalore	28126
2	Kings XI Punjab	27893
3	Kolkata Knight Riders	27419
4	Chennai Super Kings	26418
5	Delhi Daredevils	24388
6	Rajasthan Royals	22431
7	Sunrisers Hyderabad	17059
8	Deccan Chargers	11463

4.1.2 Lowest Runs:

	batting_team	total_runs
0	Kochi Tuskers Kerala	1901
1	Rising Pune Supergiants	2063
2	Rising Pune Supergiant	2470
3	Delhi Capitals	2630
4	Gujarat Lions	4862
5	Pune Warriors	6358
6	Deccan Chargers	11463
7	Sunrisers Hyderabad	17059
8	Rajasthan Royals	22431

4.1.3 Most Wins:

	team	total_wins
0	Mumbai Indians	109
1	Chennai Super Kings	100
2	Kolkata Knight Riders	92
3	Royal Challengers Bangalore	84
4	Kings XI Punjab	82
5	Rajasthan Royals	75
6	Delhi Daredevils	67
7	Sunrisers Hyderabad	58
8	Deccan Chargers	29
9	Gujarat Lions	13
10	Pune Warriors	12
11	Rising Pune Supergiant	10
12	Delhi Capitals	10
13	Kochi Tuskers Kerala	6

4.1.4 Least Wins:

	team	total_wins
0	Kochi Tuskers Kerala	6
1	Rising Pune Supergiant	10
2	Delhi Capitals	10
3	Pune Warriors	12
4	Gujarat Lions	13
5	Deccan Chargers	29
6	Sunrisers Hyderabad	58
7	Delhi Daredevils	67
8	Rajasthan Royals	75
9	Kings XI Punjab	82
10	Royal Challengers Bangalore	84
11	Kolkata Knight Riders	92
12	Chennai Super Kings	100
13	Mumbai Indians	109

4.1.5 Most Winning Percentage:

	team	total_win_percentage
0	Rising Pune Supergiant	62.500000
1	Chennai Super Kings	61.318352
2	Delhi Capitals	60.000000
3	Mumbai Indians	58.364034
4	Sunrisers Hyderabad	54.920635
5	Kolkata Knight Riders	51.008244
6	Rajasthan Royals	50.391791
7	Kings XI Punjab	46.761474
8	Royal Challengers Bangalore	46.377709
9	Kochi Tuskers Kerala	42.857143
10	Gujarat Lions	41.071429
11	Delhi Daredevils	40.956617
12	Deccan Chargers	38.117733
13	Pune Warriors	26.538462

4.1.6 Least Winning Percentage:

	team	total_win_percentage
0	Pune Warriors	26.538462
1	Deccan Chargers	38.117733
2	Delhi Daredevils	40.956617
3	Gujarat Lions	41.071429
4	Kochi Tuskers Kerala	42.857143
5	Royal Challengers Bangalore	46.377709
6	Kings XI Punjab	46.761474
7	Rajasthan Royals	50.391791
8	Kolkata Knight Riders	51.008244
9	Sunrisers Hyderabad	54.920635
10	Mumbai Indians	58.364034
11	Delhi Capitals	60.000000
12	Chennai Super Kings	61.318352
13	Rising Pune Supergiant	62.500000

4.2 By Players:

4.2.1 Most Runs:

	batsman	total_runs	out	numberofballs	average	strikerate
0	V Kohli	5426	152	4111	35.697368	131.987351
1	SK Raina	5386	160	3916	33.662500	137.538304
2	RG Sharma	4902	161	3742	30.447205	130.999466
3	DA Warner	4717	114	3292	41.377193	143.286756
4	S Dhawan	4601	137	3665	33.583942	125.538881

4.2.2 Most Sixes:

	batsman	Number_of_sixes
0	CH Gayle	1962
1	AB de Villiers	1284
2	MS Dhoni	1242
3	SK Raina	1170
4	RG Sharma	1164

4.2.3 Most Fours:

	batsman	Number_of_fours
0	S Dhawan	2104
1	SK Raina	1980
2	G Gambhir	1968
3	V Kohli	1928
4	DA Warner	1836

4.2.4 Highest Strike Rate:

	batsman	total_runs	out	numberofballs	average	strikerate
0	V Kohli	5426	152	4111	35.697368	131.987351
1	SK Raina	5386	160	3916	33.662500	137.538304
2	RG Sharma	4902	161	3742	30.447205	130.999466
3	DA Warner	4717	114	3292	41.377193	143.286756
4	S Dhawan	4601	137	3665	33.583942	125.538881

4.2.5 Highest Average:

	batsman	average
0	J Bairstow	57.375000
1	HM Amla	44.384615
2	AB de Villiers	42.442308
3	JP Duminy	41.448980
4	DA Warner	41.377193

5. Conclusion:

Therefore, the null values are removed from the dataset and the exploratory data analysis is carried out in a simplistic way due to the time constraint.