

# Data Mining on IPL Dataset

Group No :- 26

## Group Members

- |    |                    |        |  |
|----|--------------------|--------|--|
| 1. | Aditya Raghuwanshi | 170052 | <a href="mailto:adityarg@iitk.ac.in">adityarg@iitk.ac.in</a> |
| 2. | Shashi Kumar       | 160646 | <a href="mailto:shashik@iitk.ac.in">shashik@iitk.ac.in</a>   |
| 3. | Ritesh Naik        | 170579 | <a href="mailto:riteshn@iitk.ac.in">riteshn@iitk.ac.in</a>   |

# **Abstract**

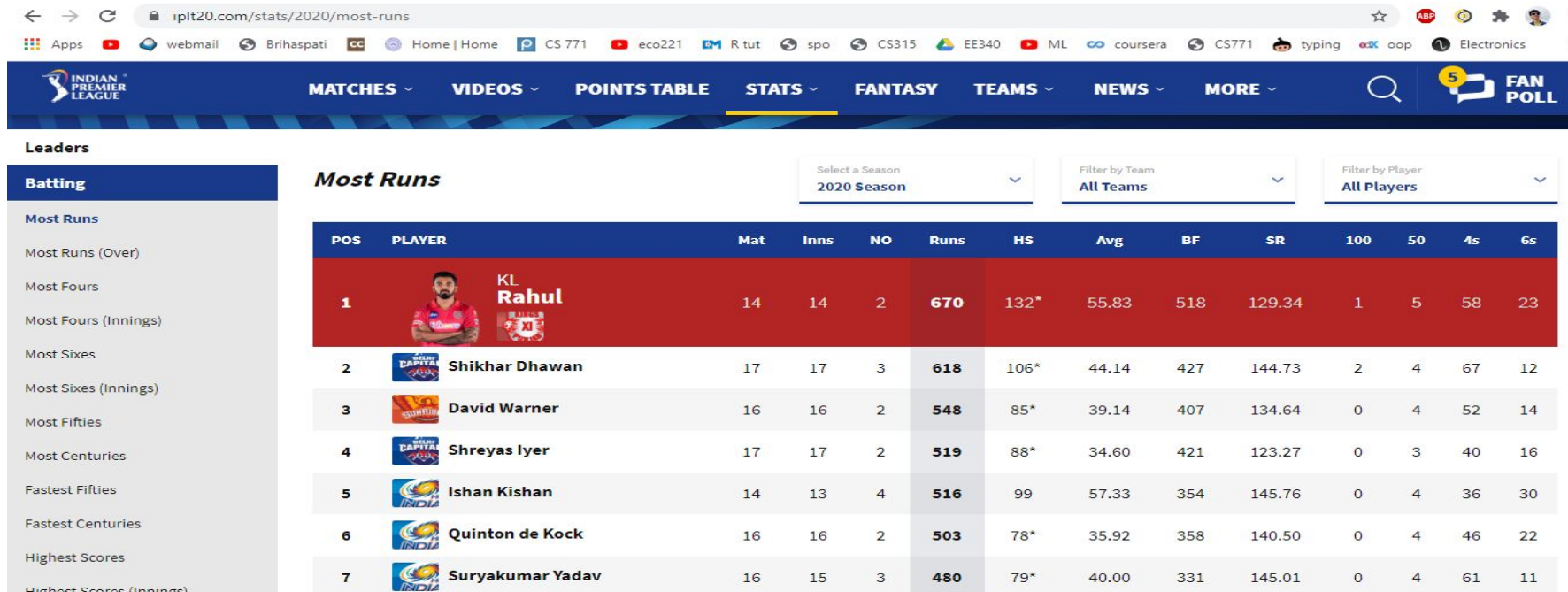
- Now a days , analysis of useful data is helping us to take some crucial decision for future . This process can also be extended to the field of sports and activities . Keeping this point in mind , We have taken the data of IPL (Indian Premiere League) , one of the most popular cricket tournament and attempted to learn some interesting stuffs from it . The results of our analysis can be a considerable asset for players and teams to see the places where they want to improve .

# Objectives








- To visualize different stats of players and teams throughout the seasons
- Classify players according to their positions in certain groups
- Find the scores of each player(or rank them) according to their performance in the IPL
- Create a best 11 team for the IPL
- Study data to find the relation of unconventional factors on each team results like Home-Away ground , Toss etc .
- Predicting expected number of boundarie using previous season stats

# Dataset used and the Source

- 1) We have taken data from the official website of IPL . Extracted data using the code “scraper.py” .



The screenshot shows the 'Most Runs' section of the IPL 2020 stats page. The table lists the top 7 batsmen based on runs scored. KL Rahul is at the top with 670 runs. The table includes columns for position, player name, matches, innings, not outs, runs, highest score, average, balls faced, strike rate, and various milestones like 100s, 50s, 4s, and 6s.

POS	PLAYER	Mat	Inns	NO	Runs	HS	Avg	BF	SR	100	50	4s	6s
1	 <b>KL Rahul</b>	14	14	2	<b>670</b>	132*	55.83	518	129.34	1	5	58	23
2	 <b>Shikhar Dhawan</b>	17	17	3	<b>618</b>	106*	44.14	427	144.73	2	4	67	12
3	 <b>David Warner</b>	16	16	2	<b>548</b>	85*	39.14	407	134.64	0	4	52	14
4	 <b>Shreyas Iyer</b>	17	17	2	<b>519</b>	88*	34.60	421	123.27	0	3	40	16
5	 <b>Ishan Kishan</b>	14	13	4	<b>516</b>	99	57.33	354	145.76	0	4	36	30
6	 <b>Quinton de Kock</b>	16	16	2	<b>503</b>	78*	35.92	358	140.50	0	4	46	22
7	 <b>Suryakumar Yadav</b>	16	15	3	<b>480</b>	79*	40.00	331	145.01	0	4	61	11

- 2) We have also used some data from Kaggle .

# Clustering of players

- We have used K-means clustering to the “player-points” data to place specific type of player in a group . It is able to classify between starting order batsman , middle order batsman and bowlers .

- startingBatsmen      MiddleOrder      Bowlers

```
['KL Rahul',  
'Quinton de Kock',  
'Shikhar Dhawan',  
'Suryakumar Yadav',  
'David Warner',  
'Ishan Kishan',  
'AB de Villiers',  
'Faf du Plessis',  
'Mayank Agarwal',  
'Devdutt Padikkal',  
'Eoin Morgan',  
'Shreyas Iyer',  
'Manish Pandey',  
'Ben Stokes',  
'Shubman Gill',  
'Nitish Rana',  
'Rohit Sharma',  
'Jonny Bairstow',
```

```
['Rahul Tewatia',  
'Marcus Stoinis',  
'Sam Curran',  
'Ravindra Jadeja',  
'Kieron Pollard',  
'Sanju Samson',  
'Nicholas Pooran',  
'Sunil Narine',  
'Hardik Pandya',  
'Andre Russell',  
'Prithvi Shaw',  
'MS Dhoni',  
'Virat Kohli',  
'Rahul Tripathi',  
'Shimron Hetmyer',  
'Dinesh Karthik',  
'Robin Uthappa',  
'Shivam Dube',
```

```
['Jofra Archer',  
'Kagiso Rabada',  
'Jasprit Bumrah',  
'Rashid Khan',  
'Trent Boult',  
'Anrich Nortje',  
'Pat Cummins',  
'Mohammad Shami',  
'T Natarajan',  
'Yuzvendra Chahal',  
'Axar Patel',  
'Varun Chakravarthy',  
'Washington Sundar',  
'Ravi Bishnoi',  
'Krunal Pandya',  
'Sandeep Sharma',  
'Deepak Chahar',  
'Rahul Chahar',  
'Ravichandran Ashwin',
```

# Ranking of players

- We have calculated scores of each players from their performance statistics and ranked them accordingly . We have used different stats for different types of players as grouped in clustering method .
- Batsman – runs , strike rate , not out
- Middle order – no of 6s , strike rate , average
- Bowlers – wickets , economy

# Best 11

- From every cluster group , we took the players with best scores and made the best 11 . Our code can do this for every IPL season .
- For IPL 2020 season , inferred Best 11 is :

```
[('KL Rahul', 2.888638095238095), ('Mayank Agarwal', 2.849348484848485), ('Ishan Kishan', 2.6861714285714284), ('Shikhar Dhawan', 2.6590647058823524)]  
[('Kieron Pollard', 2.647533333333333), ('Hardik Pandya', 2.623133333333333), ('Nicholas Pooran', 2.530433333333333)]  
[('Rashid Khan', 1.9702048417132216), ('Jasprit Bumrah', 1.940118870728083), ('Kagiso Rabada', 1.839328537170264), ('Jofra Archer', 1.7353689567430024)]
```

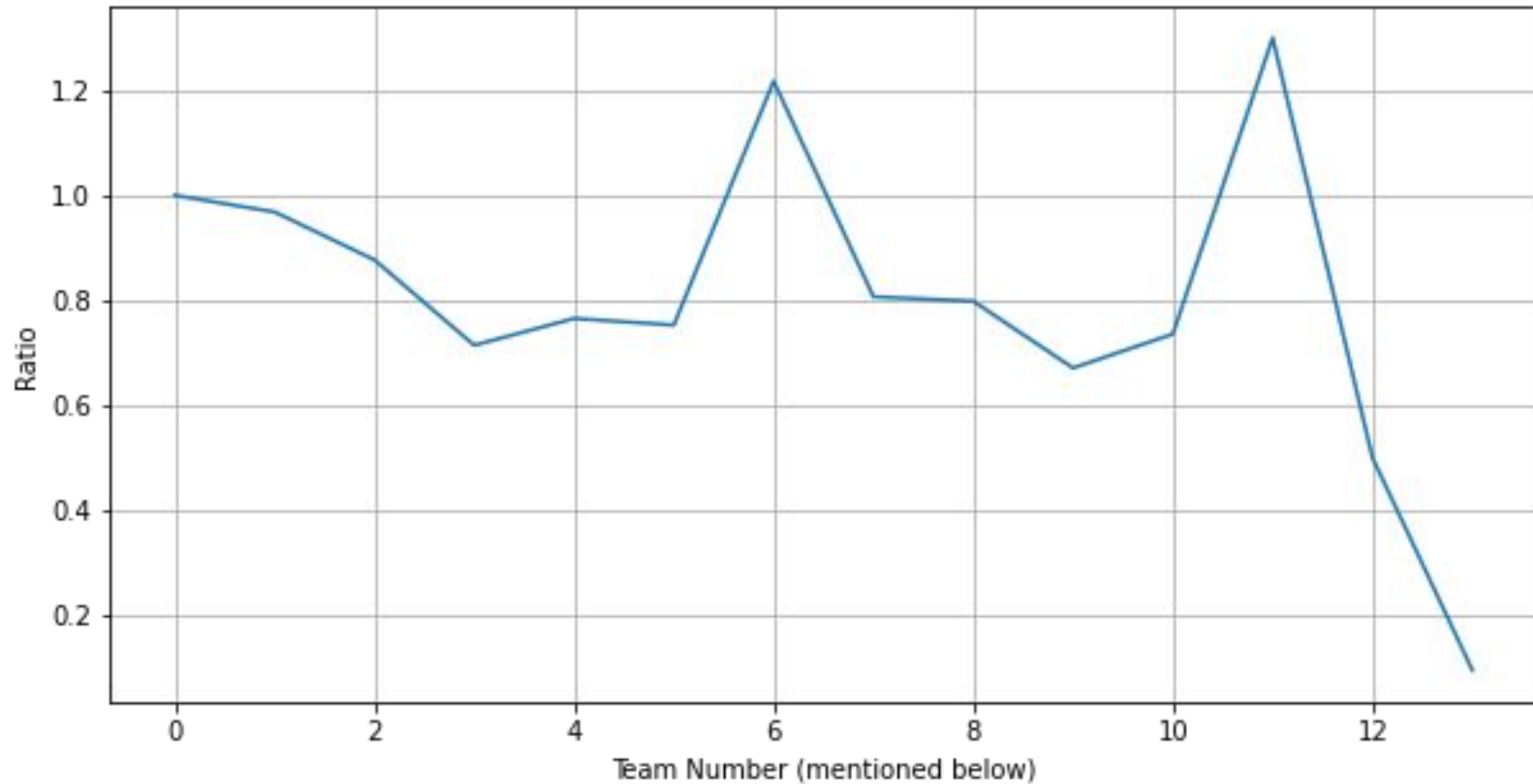
# Home-Away Ground Effect

- Over many seasons, we have observed that some team performs different at different grounds.
- We tried to observe this affect for each team.
- We found the ratio of matches won at home to matches won away.

$$\text{Ratio} = \frac{\text{Home Wins Percent}}{\text{Away Wins Percent}}$$

- If  $\text{ratio} \gg 1$ , it implies good ground home performance
- If  $\text{ratio} \ll 1$ , it implies away ground home performance
- If  $\text{ratio} \approx 1$ , no effect of the ground.





- Most of the teams performs better at away matches.
- Only one team (0) has almost no effect.
- Two teams (6,11) are showing better home performance, but they have less data than other teams.

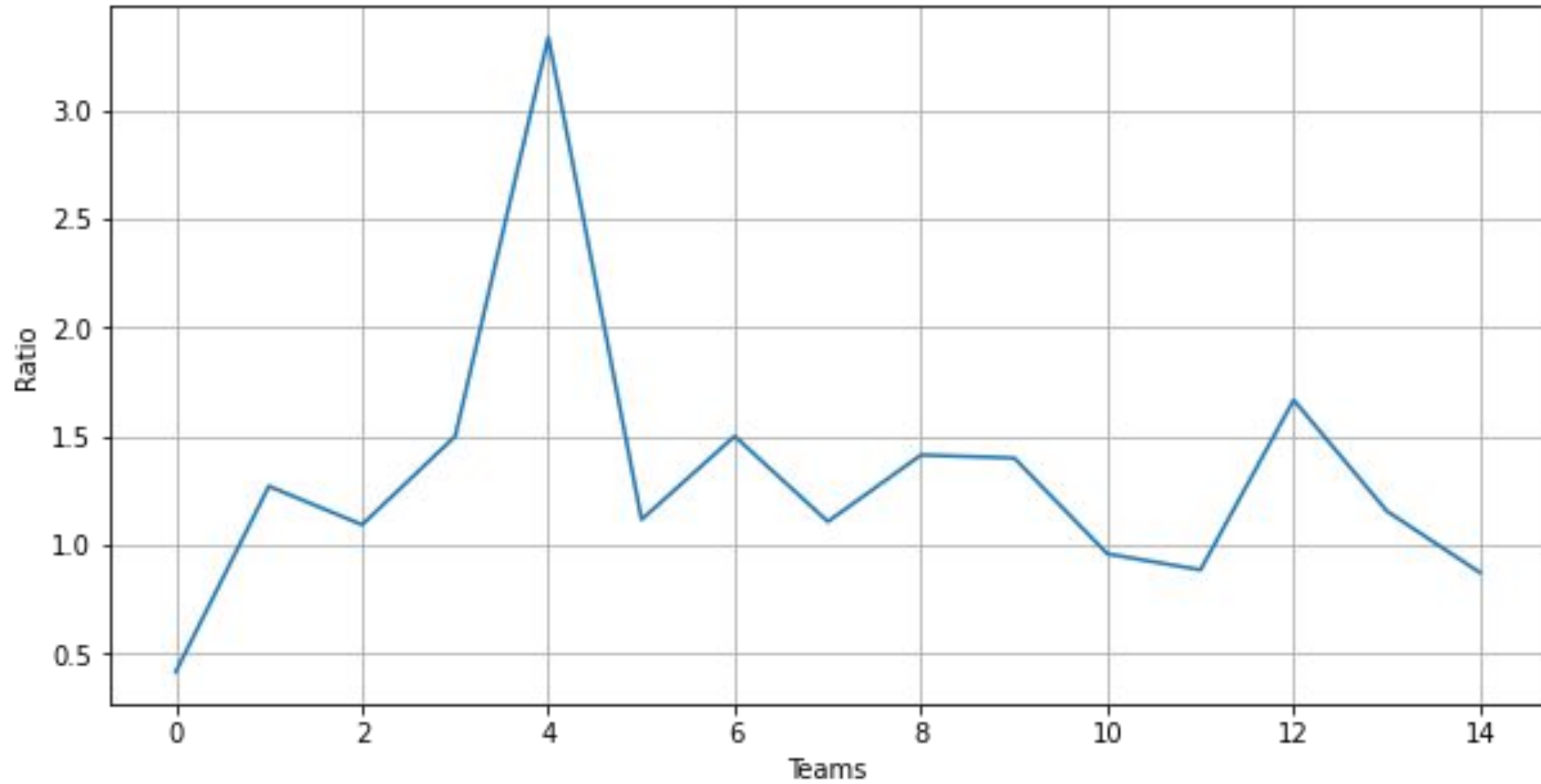
# Toss Winning Effect

- This is also an unconventional factor that can affect team's performance.
- We found the ratios of matches won at home to matches won away.

$$r1 = \frac{\text{No of (Toss Lost + Match Won)}}{\text{Total Toss Won}} \quad r2 = \frac{\text{No of (Toss Lost + Match Won)}}{\text{Total Toss Lost}}$$

$$\text{final ratio} = \frac{r1}{r2}$$

- If final ratio  $\gg 1$  or  $\ll 1$ , then toss winning effect the team significantly.
- If final ratio  $\approx 1$ , then the toss winning has no effect.



- Almost all teams performs better when they won the toss.
- Almost all teams have slight effect.
- Only teams (0,4) have large effect.

# Boundaries (4s & 6s)

- Did separate but similar analysis for fours and sixes.
- Find the total & average boundaries for each season.
- Then use “Moving Average” method for prediction.

assume  $\mu$  = avg of data  
find  $y_t = \mu + \Phi \text{ error}_{t-1}$ , where  $\text{error}_t = y_t - a_t$

- Prediction:

	TOTAL	AVERAGE
4s fours	1646	205
6s sixes	771	96

# Boundaries analysis

