Massive Data Storage And Retrieval
CS 543 - Final report
Date: 12/13/2018

Teammates:
Vivek Dhandha (vd264)
Akshay Sovani (as3041)

Indian Premier League Data Analysis

Project Goals:

- To analyze the Indian Premier League (IPL) data over all the seasons (2008-2017) till date and find insights from it.
- To find 'Top 10 Most Valuable Players' in IPL using most valuable player index (MVPI) calculation strategy.
- To apply machine learning algoruthm to it and predict the

Project Problem Description:

Board of Control for Cricket in India (BCCI) started a T20 cricket league named Indian Premier League (IPL) in 2008. From a pool of available players which are selected by BCCI according to their rules, the franchises select their players through competitive bidding. Every year BCCI has been organizing the IPL tournament. 11 tournaments have been held till date.

The main objective of the project is to come up with a list of top players IPL history. This will be achieved by using MVPI strategy. MVPI is in general an index calculated for a player. It is nothing but a wishful proposition to quantify some of the metrics of that player depending on its type (i.e. Batsmen/Bowler). This index shows the potential value of that player to the team.

Approach for solution to the problem:

The heart of the solution to this problem is the MVPI index calculation. This index will be calculated in step by step manner as follows:

1. Identify the metrics for each player.
2. Calculate absolute values for matrics and normalize them to remove any metric bias.
3. Carry out feature selection on these matrics by using recursive feature elimination technique. This technique retains features by recursively narrowing down to smaller and smaller sets of attributes. Now, calculate weights for each and every selected feature.
4. The last step involves multiplying the importance (weight) of each feature with its value and then aggregating the weighted value set for each player.

This way, we find index for each player which are then sorted in decreasing order to get the top 10 Most Valuable Players.

Some basic analysis:

Total number of matches : 636
Total number of deliveries : 150460

Different Venues matches were played at:
Dubai International Cricket Stadium, Himachal Pradesh Cricket Association Stadium, Sardar Patel Stadium, Motera, Punjab Cricket Association Stadium, Mohali, Barabati Stadium, Punjab Cricket Association IS Bindra Stadium, Mohali, Nehru Stadium, Maharashtra Cricket Association Stadium, Eden Gardens, OUTsurance Oval, M Chinnaswamy Stadium, Feroz Shah Kotla, Rajiv Gandhi International Stadium, Uppal, Brabourne Stadium, Vidarbha Cricket Association Stadium, Jamtha, Green Park, Holkar Cricket Stadium, Shaheed Veer Narayan Singh International Stadium, Sheikh Zayed Stadium, Sharjah Cricket Stadium, St George's Park, Wankhede Stadium, Newlands, JSCA International Stadium Complex, Saurashtra Cricket Association Stadium, Dr DY Patil Sports Academy, Buffalo Park, New Wanderers Stadium, SuperSport Park, Sawai Mansingh Stadium, Dr. Y.S. Rajasekhara Reddy ACA-VDCA Cricket Stadium, MA Chidambaram Stadium, Chepauk, De Beers Diamond Oval, Kingsmead, Subrata Roy Sahara Stadium,

Different Cities matches were played in:
Bangalore, Kochi, Chennai, None, Centurion, Ranchi, Mumbai, Ahmedabad, Durban, Kolkata, Cape Town, Dharamsala, Sharjah, Pune, Johannesburg, Kimberley, Delhi, Raipur, Chandigarh, Nagpur, Abu Dhabi, Bloemfontein, Kanpur, Hyderabad, Rajkot, Port Elizabeth, Indore, Cuttack, East London, Jaipur, Visakhapatnam

Teams :
Sunrisers Hyderabad, Chennai Super Kings, Rising Pune Supergiant, Deccan Chargers, Kochi Tuskers Kerala, Rajasthan Royals, Gujarat Lions, Royal Challengers Bangalore, Kolkata Knight Riders, Rising Pune Supergiants, Kings XI Punjab, Pune Warriors, Delhi Daredevils, Mumbai Indians

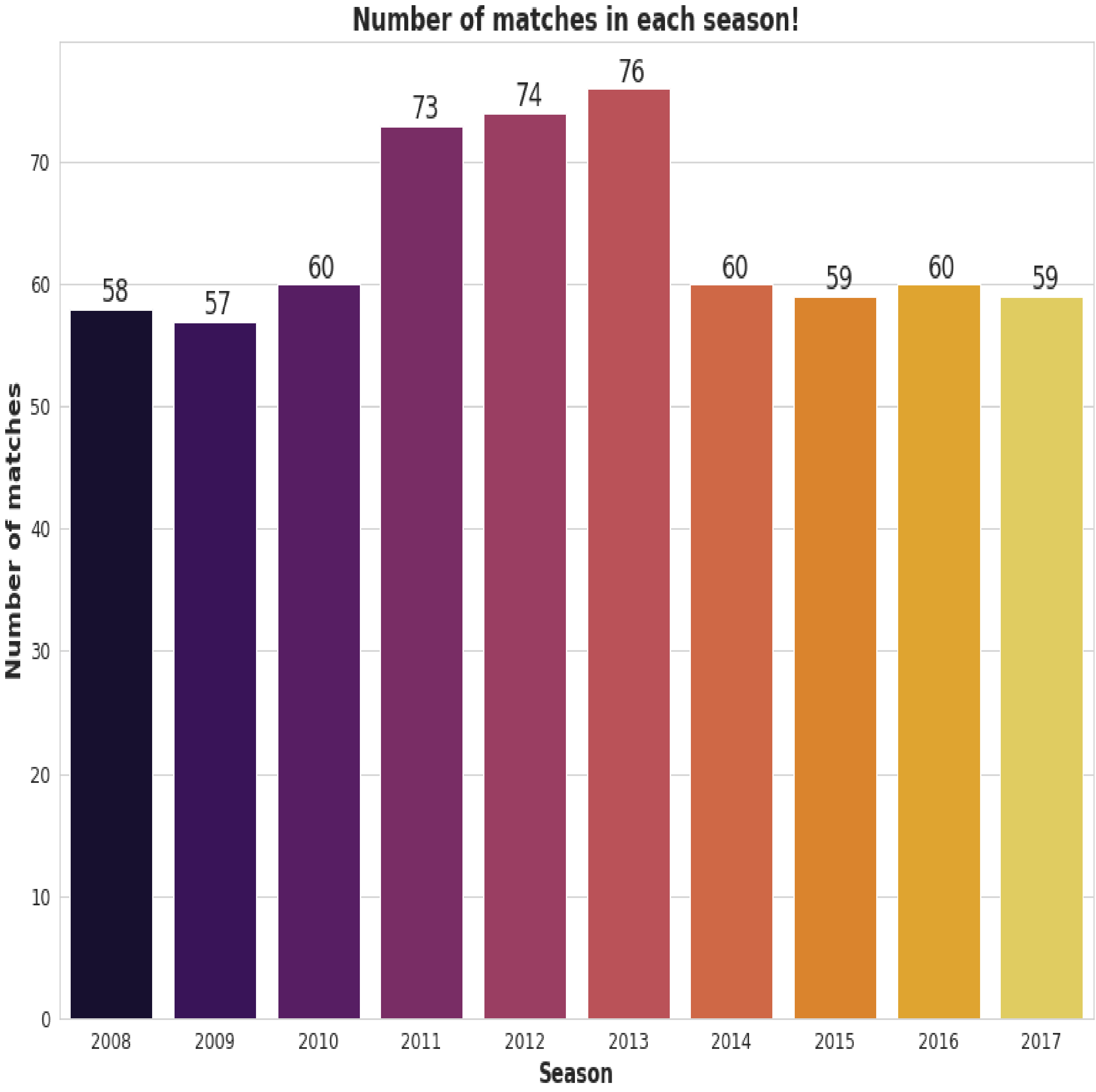
Total umpires used in IPL till now : 45

Detailed Analysis Using Spark:

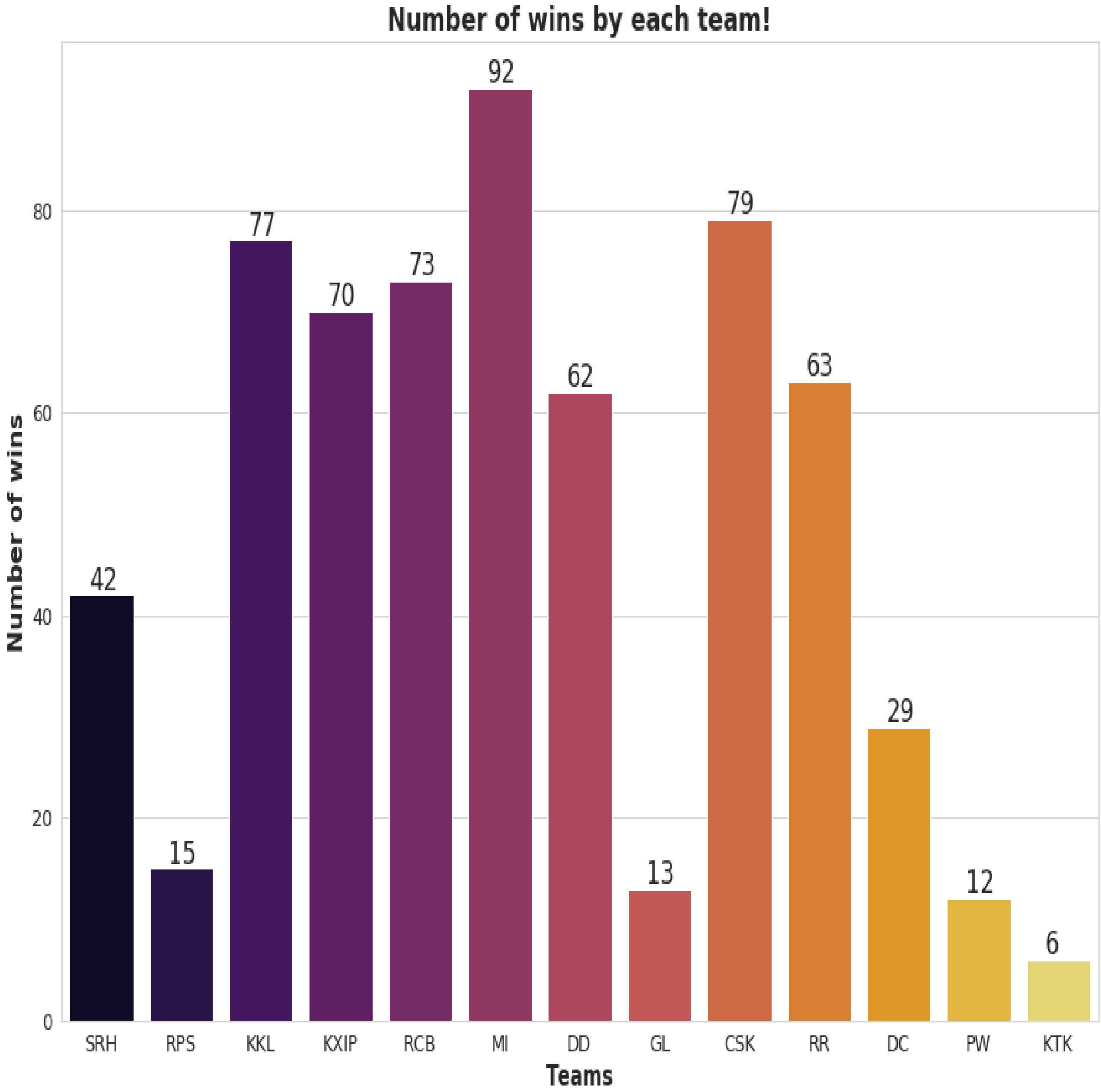
Using pyspark, analysis has been carried out and visualized as specified below :

1. Number of matches played at a venue.
2. Win percentage at a venue batting first.
3. Win percentage at a venue bowling first.
4. Teamwise comparative analysis of win/loss/tie if a team bats first.
5. Percentage wins by a team when it wins toss
6. Percentage wins by a team when it loses toss
7. Teams most and least affected by the event of winning/losing toss.

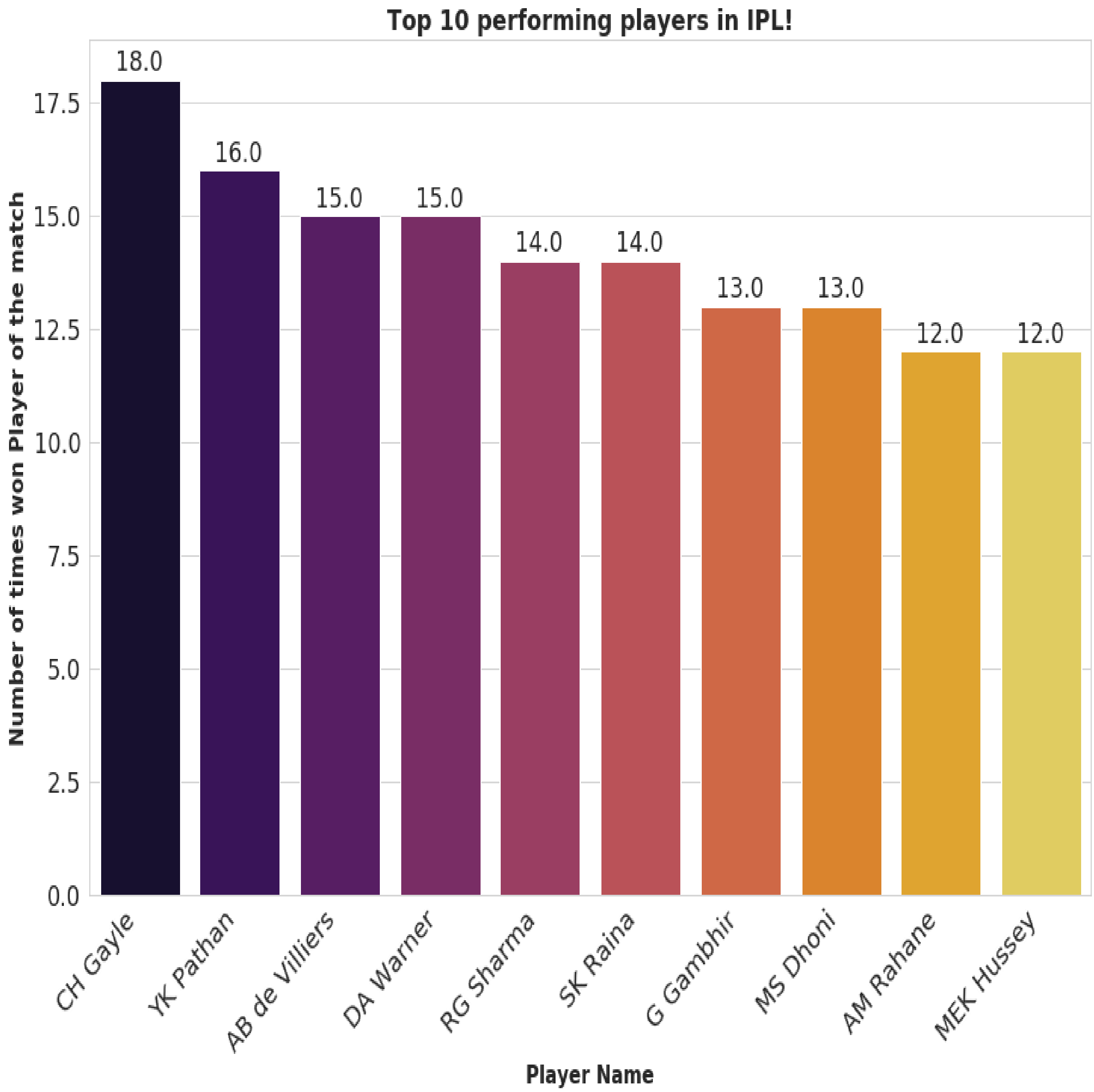
Please find sample screenshots for the analysis below. Kindly refer to the attached notebook and pdf file for detailed visualization of the analysis.



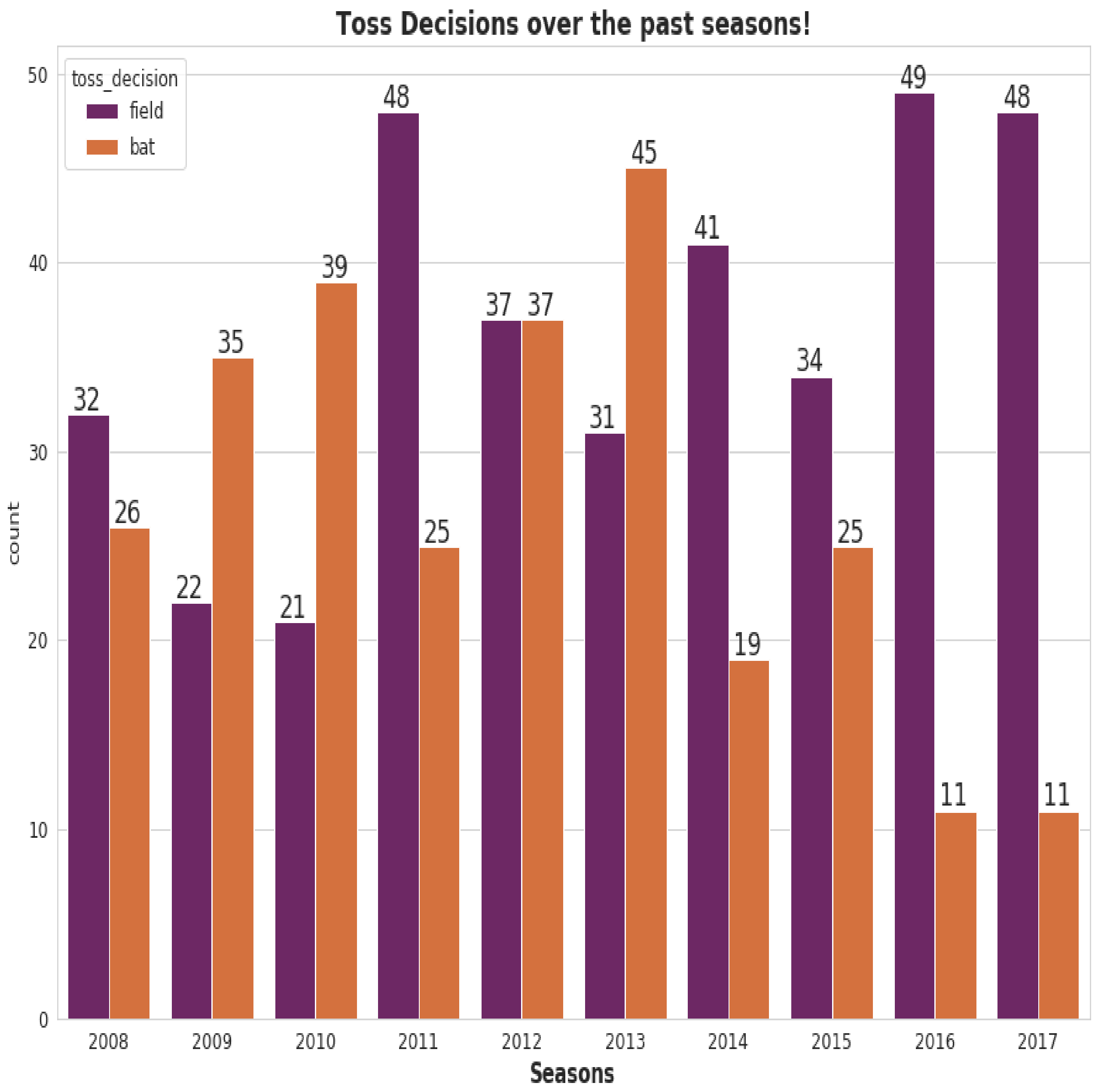
- 2013 was the season with most number of matches.
- 2011, 2012, 2013 were the seasons with most number of matches, this is mainly because the number of teams participating in IPL in those seasons increased from 8 to 10.
- 2014 onwards, again the number of teams were 8. Hence, the reduction in the number of matches.



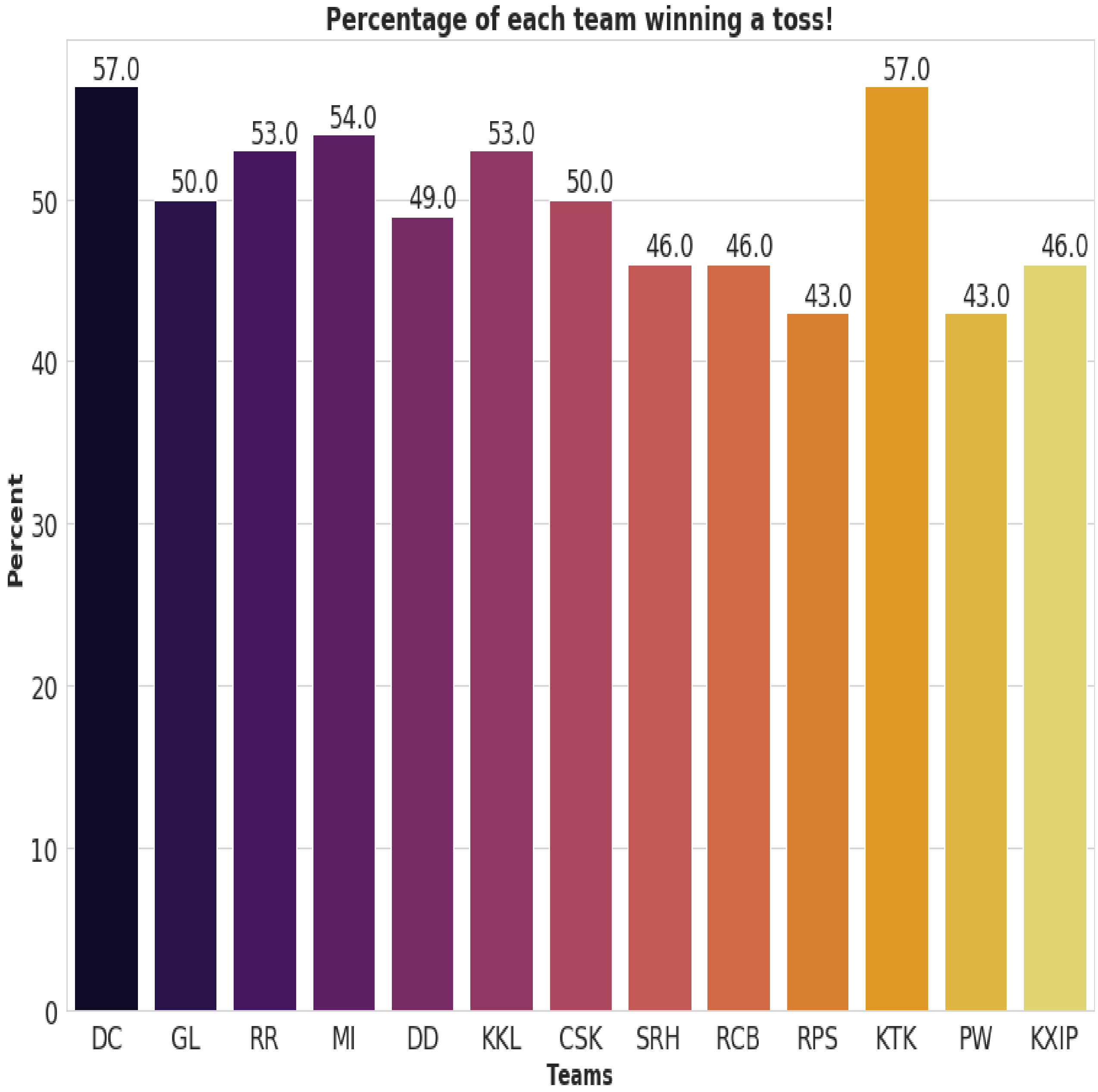
- Mumbai Indians(MI) seems to be the most successfull IPL team over all the seasons with most number of wins, followed by Chennai Super Kings(CSK) and Kolkata Knight Riders(KKR).
- Kochi Tuskers Kerala (KTK) has the least number of wins in IPL, mainly because it just participated in 1 season, i.e. 2011.
- Pune Warriors (PW), Gujrat Lion (GL) and Rising Pune Supergiants (RPS) also seem to have a bad record. As mainly, these took part only in 2 seasons of IPL



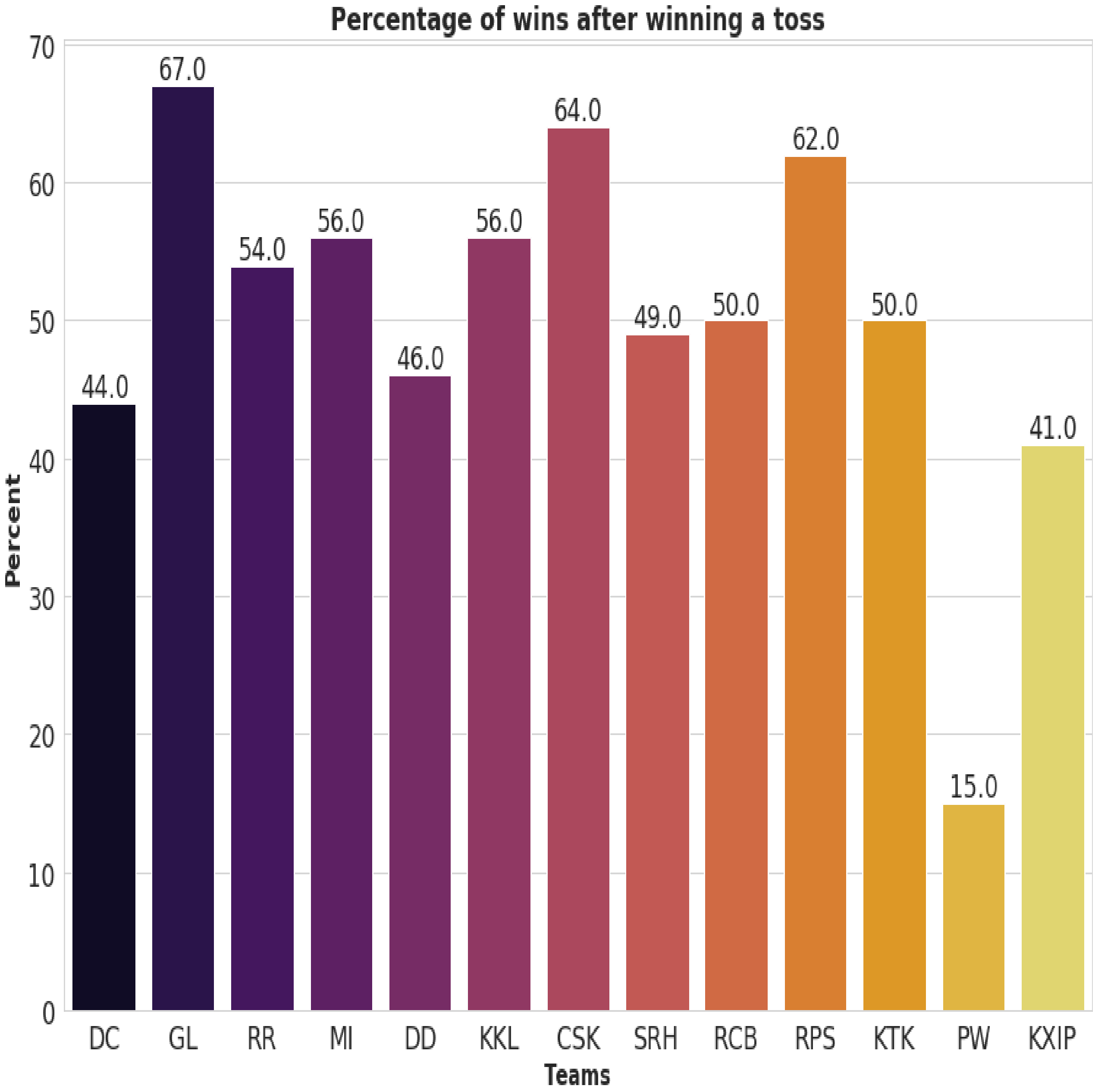
- This analysis is mainly based on the number of times each Player has won the 'Player Of the Match' award
- Chris Gayle leads the list of top 10 performing players with 18 'Player of the Match' awards.



- It is very clear from this graph that every season, teams prefer fielding first rather than batting.
- 2012 is the only season where the number of toss decisions is the same for both bat and field.

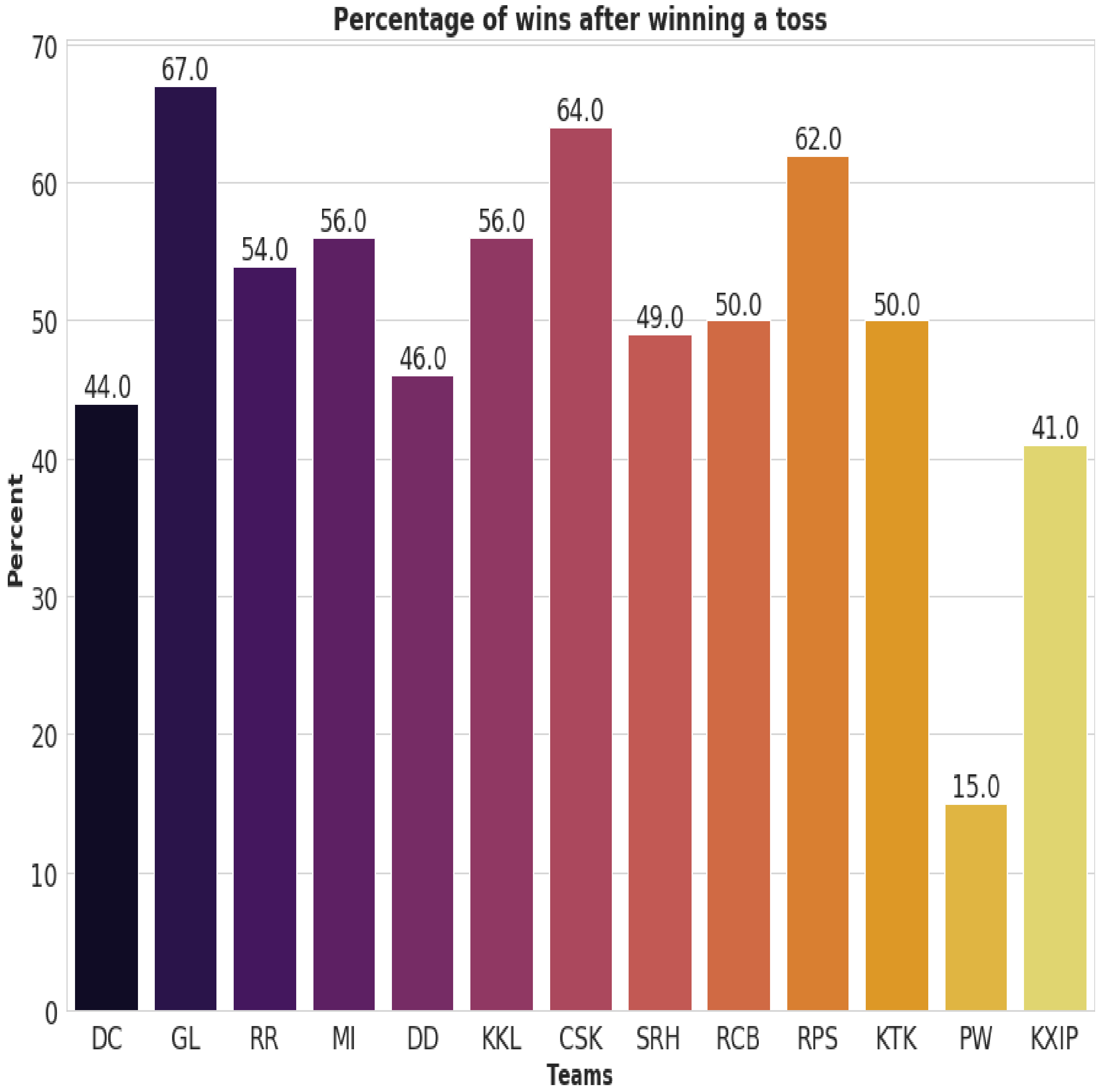


- Every teams luck factor seems to move towards 50, with the number of matches they play.
- Although, Deccan Chargers(DC) and Kochi Tuskers Kerala (KTK) seem to be the most luckiest compared to the rest of the teams

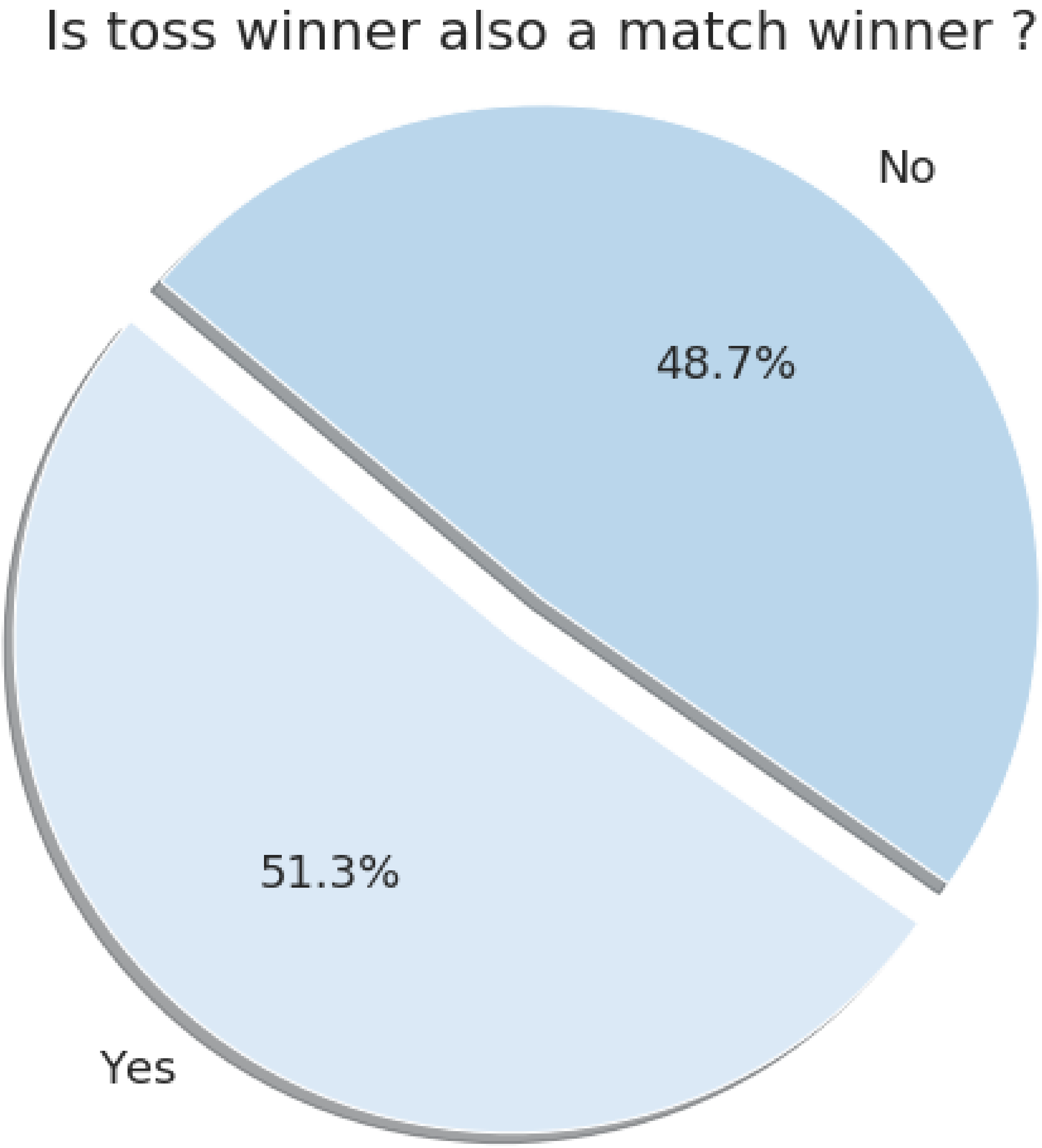


- Gujrat Lions (GL) seems to be the team which takes most advantage of winning a toss, followed by Chennai Super Kings (CSK) and Rising Pune Supergiants (RPS). They definitely must be sticking to a particular stratergy and pulling it off really well.

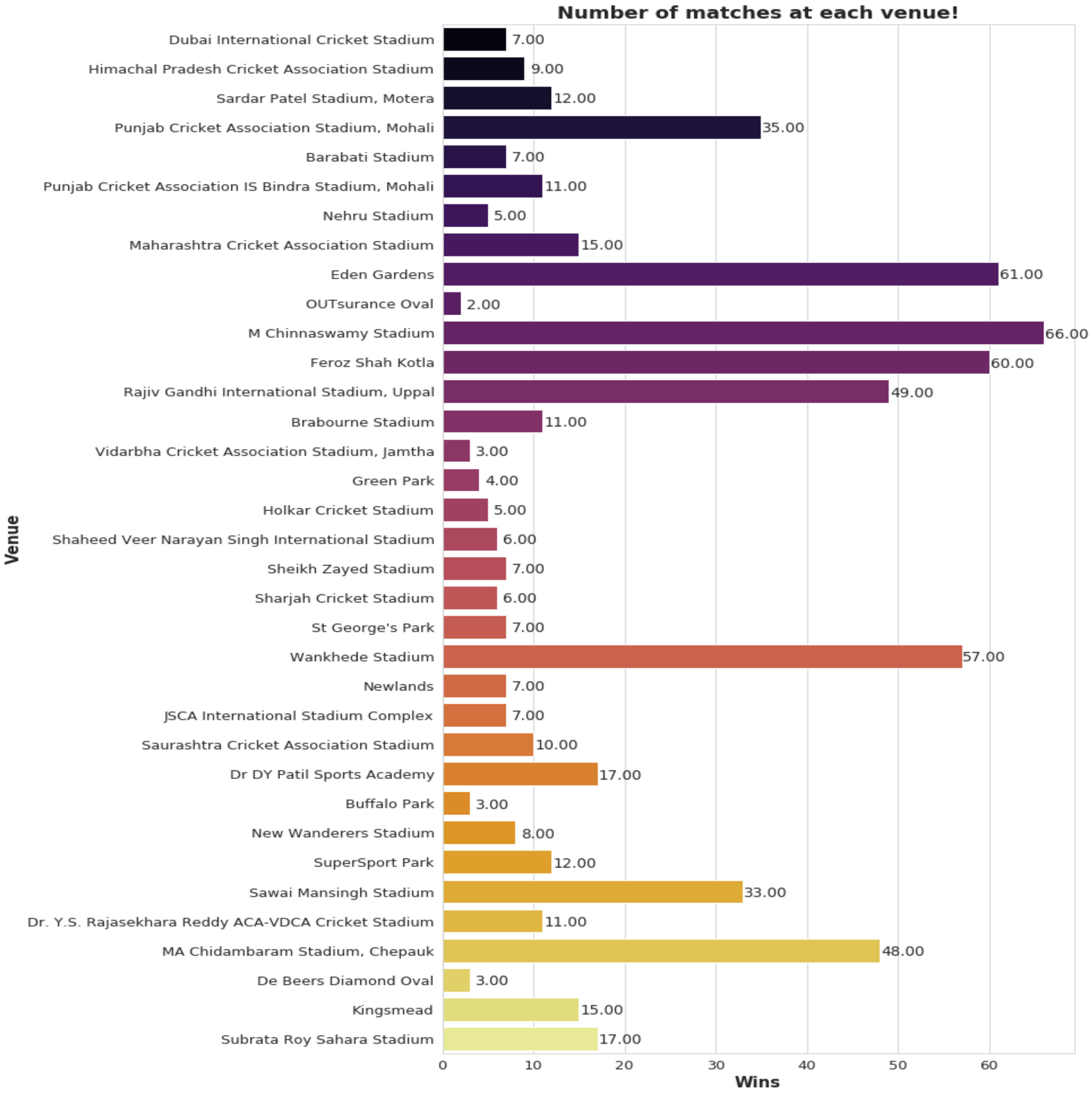
- Pune Warriors(PW) is the one which doesn't really make the best of winning a toss and has the worst percentage of winning ratio even after winning the toss



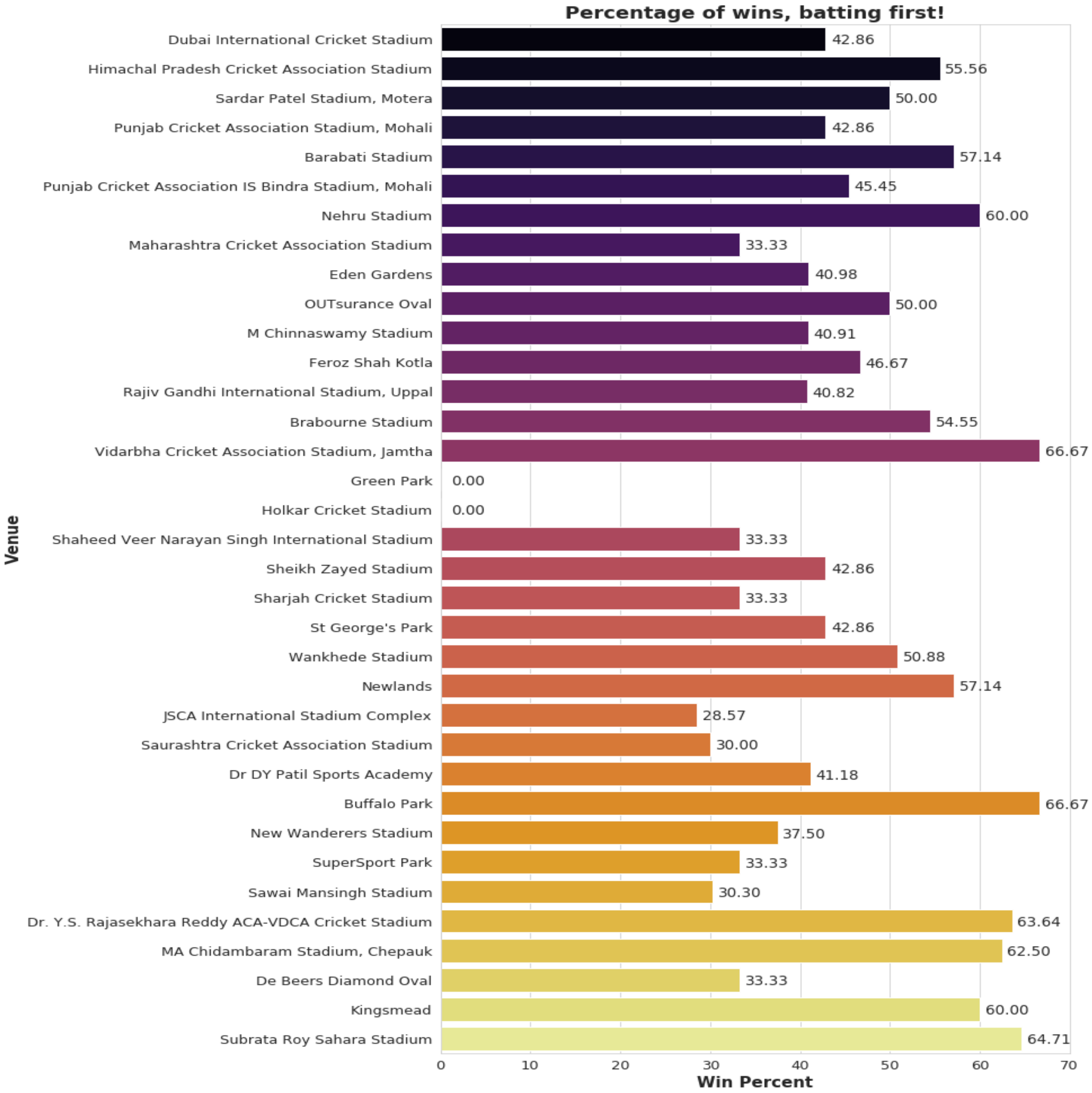
- Mumbai Indians (MI) and Sunrisers Hyderabad (SRH) seem to be the strongest of teams as their ratop of winning the matches even after loosing the toss is the most.
- Gujrat Lion (GL) seems to be the one whose game really gets affected if they don't win the toss



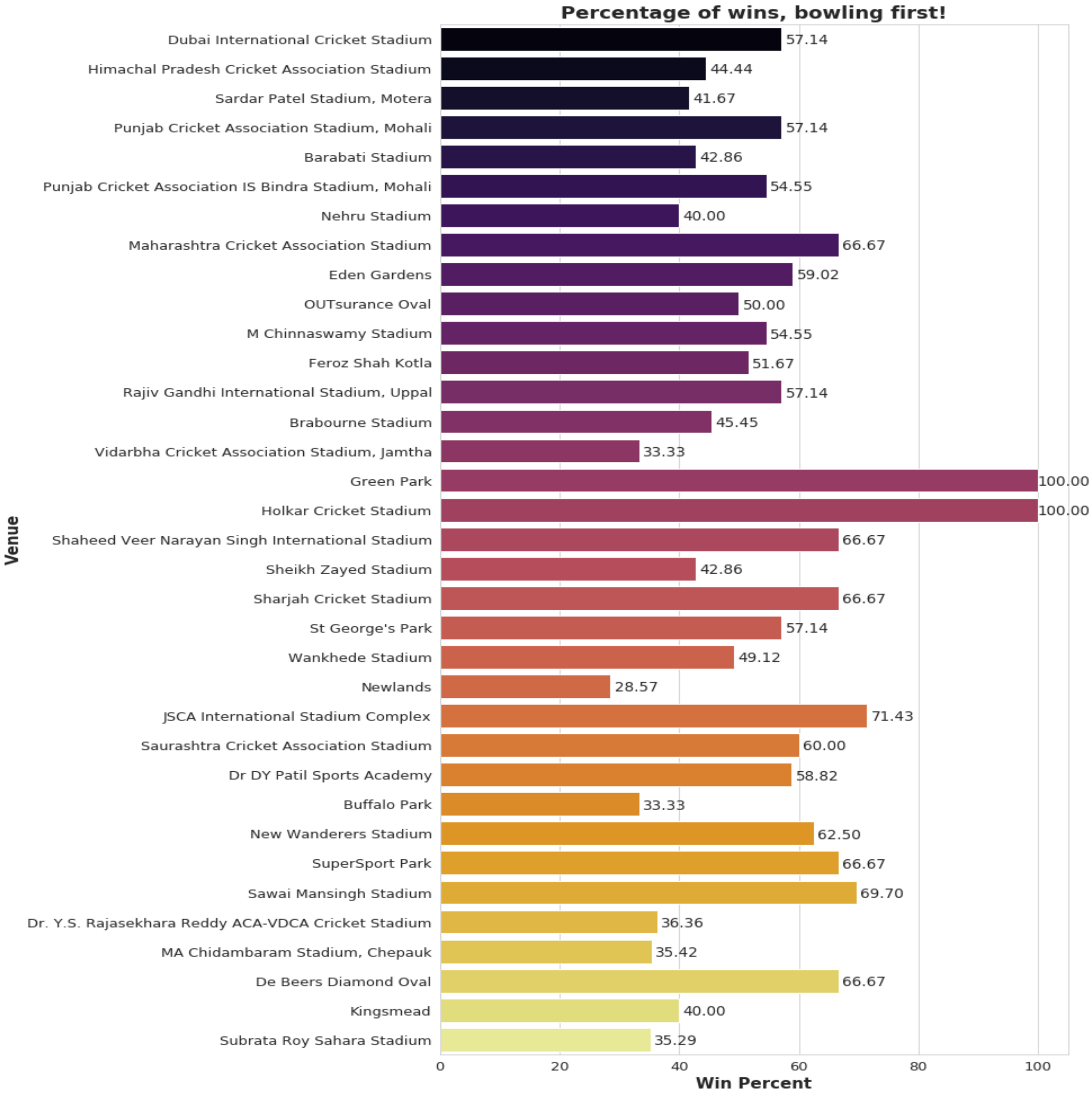
- On an overall scenario, winning a toss doesn't seem to affect the final result of the match.
- This shows the preparation of each team, because on an overall basis, they are not letting loosing the toss affect their play and still seem to pull off a win.



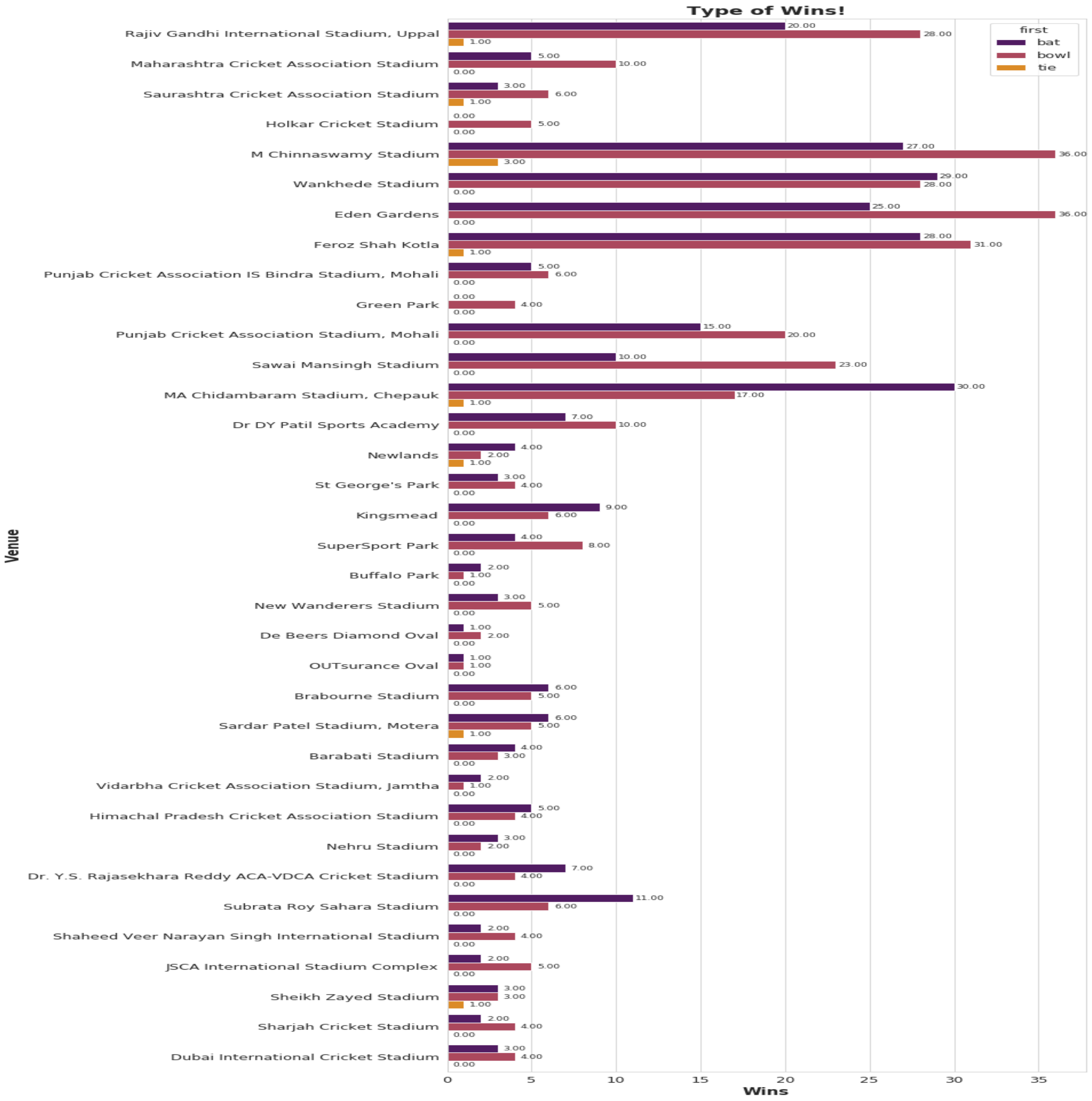
- M Chinnaswami is the most favourite ground as most of the matches are played there, followed by Eden Gardens and Feroz Shah Kotla



- Vidarbha Cricket Ass and Buffalo Park seem to be most favourable for batting first.



- Green Park and Holkar Cricket Stadium seem to be most favourable for bowling first as they have a 100 percent ratio of wins for all the teams who have chosen to Field first there.


Type of Wins!
first
bat
bowl
tie
Venue
Rajiv Gandhi International Stadium, Uppal
20.00
28.00
1.00
Maharashtra Cricket Association Stadium
5.00
10.00
0.00
Saurashtra Cricket Association Stadium
3.00
6.00
1.00
Holkar Cricket Stadium
0.00
5.00
0.00
M Chinnaswamy Stadium
27.00
36.00
3.00
Wankhede Stadium
29.00
28.00
0.00
Eden Gardens
25.00
36.00
0.00
Feroz Shah Kotla
28.00
31.00
1.00
Punjab Cricket Association IS Bindra Stadium, Mohali
5.00
6.00
0.00
Green Park
0.00
4.00
0.00
Punjab Cricket Association Stadium, Mohali
15.00
20.00
0.00
Sawai Mansingh Stadium
10.00
23.00
0.00
MA Chidambaram Stadium, Chepauk
30.00
17.00
1.00
Dr DY Patil Sports Academy
7.00
10.00
0.00
Newlands
4.00
2.00
1.00
St George's Park
3.00
4.00
0.00
Kingsmead
9.00
6.00
0.00
SuperSport Park
4.00
8.00
0.00
Buffalo Park
2.00
1.00
0.00
New Wanderers Stadium
3.00
5.00
0.00
De Beers Diamond Oval
1.00
2.00
0.00
OUTsurance Oval
1.00
1.00
0.00
Brabourne Stadium
6.00
5.00
0.00
Sardar Patel Stadium, Motera
6.00
5.00
1.00
Barabati Stadium
4.00
3.00
0.00
Vidarbha Cricket Association Stadium, Jamtha
2.00
1.00
0.00
Himachal Pradesh Cricket Association Stadium
5.00
4.00
0.00
Nehru Stadium
3.00
2.00
0.00
Dr. Y.S. Rajasekhara Reddy ACA-VDCA Cricket Stadium
7.00
4.00
0.00
Subrata Roy Sahara Stadium
11.00
6.00
0.00
Shaheed Veer Narayan Singh International Stadium
2.00
4.00
0.00
JSCA International Stadium Complex
2.00
5.00
0.00
Sheikh Zayed Stadium
3.00
3.00
1.00
Sharjah Cricket Stadium
2.00
4.00
0.00
Dubai International Cricket Stadium
3.00
4.00
0.00
0
5
10
15
20
25
30
35
Wins

Calculation of Most Valuable Player Index (MVPI):

- Calculated scores for following 5 feature of batsman:
  1. Fast scoring ability
  2. Consistency
  3. Hard hitting ability
  4. Running between the wickets
  5. Finishing ability
- Calculated scores for following 5 feature of bowler:
  1. Economy
  2. Wicket taking ability
  3. Consistency
  4. Cricial Wicket taking ability
  5. Short performance index

As the first step, the above 10 feature scores have been calculated. These scores are normalized to remove any metric bias, using the formula:

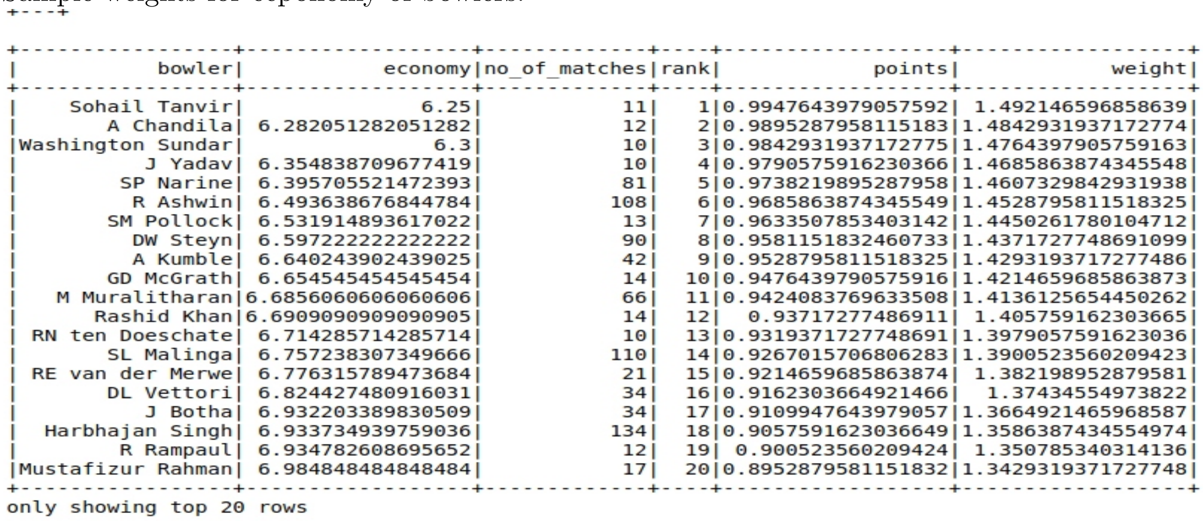
Score for a Feature = (Players Count Rank in that Feature / Players Count)

In the next step, calculation of respective weights for each of the selected features has been calculated. This has been done using the Recursive Feature Elimination technique, which retains features by recursively narrowing down to smaller and smaller sets of attributes.
Now its time to train the model. The model is first trained by feeding an initial set of attributes followed by computation of relative importance of each attribute. Then, the least important attributes are eliminated from the present set of attributes. This approach is then repeated on the reduced set each time, until the desired number of attributes to be selected is eventually reached.

In this way, weights for each feature takes into account the importance of that feature in the match. A high weight is assigned to the important features while relatively smaller weight is assined to the features which carry less importance from the match impact perspective. For example, Fast scoring ability and hard hitting ability carry a high weight for batting and wicket taking ability and crucial wicket taking ability carry a high weight for bowling.
Now, as the weights have been calculated, its time to multiply these weights with the respective feature score to get the points. These points calculated for individual feature are now added to get the total points for a particular batsman and bowler. These players, Players get sorted according to the pints they get in decreasing order. By this way, top 10 batsman and top 10 bowlers are computed.

The computed points for batsman and bowlers are merged to get the total points for a team. This makes us enable to find out the teams with total points. We can typically expect the teams with higher total points to fare well and be the table toppers for the tournament.

Sample weights for ecponomy of bowlers:



```
+---+

+------------------+------------------+-------------+----+------------------+------------------+
|            bowler|           economy|no_of_matches|rank|            points|            weight|
+------------------+------------------+-------------+----+------------------+------------------+
|     Sohail Tanvir|              6.25|           11|   1|0.9947643979057592| 1.492146596858639|
|        A Chandila| 6.282051282051282|           12|   2|0.9895287958115183|1.4842931937172774|
|Washington Sundar|               6.3|           10|   3|0.9842931937172775|1.4764397905759163|
|           J Yadav| 6.354838709677419|           10|   4|0.9790575916230366|1.4685863874345548|
|         SP Narine| 6.395705521472393|           81|   5|0.9738219895287958|1.4607329842931938|
|          R Ashwin| 6.493638676844784|          108|   6|0.9685863874345549|1.4528795811518325|
|        SM Pollock| 6.531914893617022|           13|   7|0.9633507853403142|1.4450261780104712|
|          DW Steyn| 6.597222222222222|           90|   8|0.9581151832460733|1.4371727748691099|
|          A Kumble| 6.640243902439025|           42|   9|0.9528795811518325|1.4293193717277486|
|        GD McGrath| 6.654545454545454|           14|  10|0.9476439790575916|1.4214659685863873|
|    M Muralitharan|6.6856060606060606|           66|  11|0.9424083769633508|1.4136125654450262|
|       Rashid Khan|6.6909090909090905|           14|  12|  0.93717277486911| 1.405759162303665|
|  RN ten Doeschate| 6.714285714285714|           10|  13|0.9319371727748691|1.3979057591623036|
|        SL Malinga| 6.757238307349666|          110|  14|0.9267015706806283|1.3900523560209423|
|  RE van der Merwe| 6.776315789473684|           21|  15|0.9214659685863874| 1.382198952879581|
|        DL Vettori| 6.824427480916031|           34|  16|0.9162303664921466|  1.37434554973822|
|           J Botha| 6.932203389830509|           34|  17|0.9109947643979057|1.3664921465968587|
|   Harbhajan Singh| 6.933734939759036|          134|  18|0.9057591623036649|1.3586387434554974|
|         R Rampaul| 6.934782608695652|           12|  19| 0.900523560209424| 1.350785340314136|
|Mustafizur Rahman| 6.984848484848484|           17|  20|0.8952879581151832|1.3429319371727748|
+------------------+------------------+-------------+----+------------------+------------------+
only showing top 20 rows
```

Applying Machine Learning:

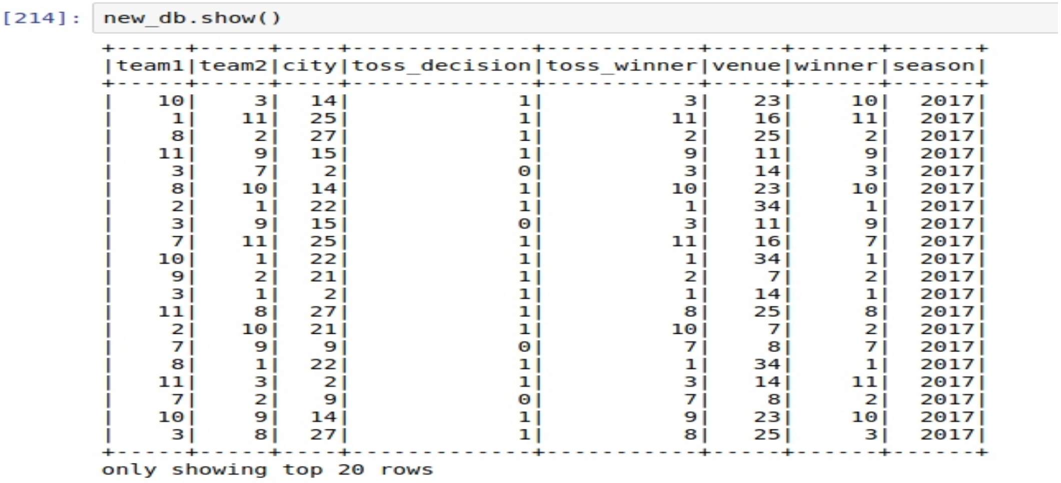
The goal of the machine learning section is to predict the match winner based on many different features associated with each match. Initially, we applied logistic regression model to the data without batting or bowling weights. i.e. we predicted the match winner based on team1, team2 and toss_winner. It gave us moderate accuracy and cross-validation score. This was giving us the accuracy of 79% and cross validation score of 44.538%. We felt that this accuracy and cross-validation score could be increased by adding few new features.
So, we felt that, if we could take the already calculated weights for each team (which we had already calculated as a part of the above MVPI function), it would definitely increase the probability of increase in accuracy and cross-validation score.

These features include :

- Team1
- Team2
- Toss_winner

Hence, we added the weights to the training data to check if we could get more accurate predictions. We started it by loading players.csv data to the players dataframe. Then team_weights were added to the dataframe. As most of the columns from this dataset were categorical, we carried out label encoding on the columns so that input could be fed to logistic regression model.

Below is the 'Encoded data' consisting of all the features:

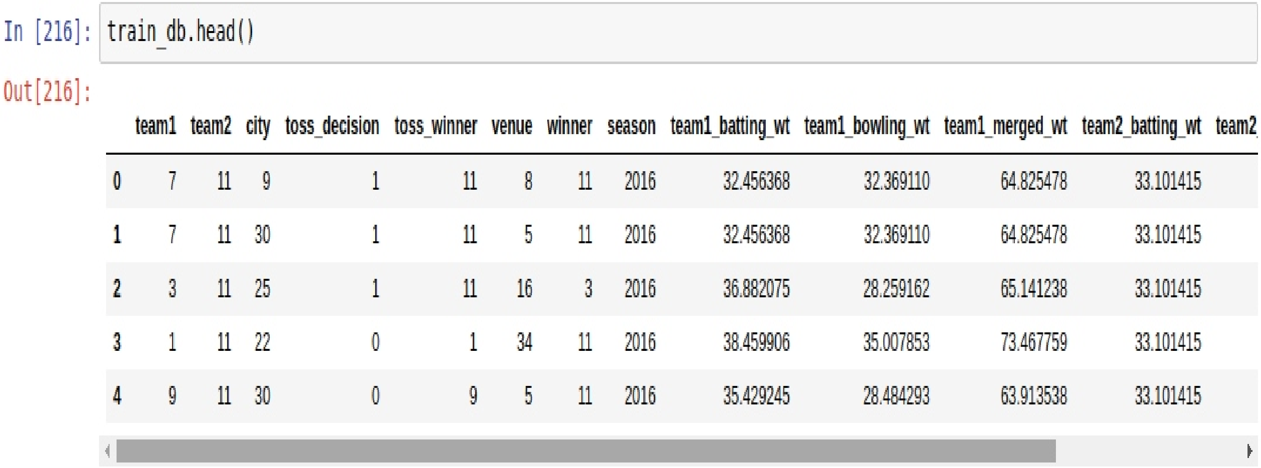


```
[214]: new_db.show()

+-----+-----+----+-------------+-----------+-----+------+------+
|team1|team2|city|toss_decision|toss_winner|venue|winner|season|
+-----+-----+----+-------------+-----------+-----+------+------+
|   10|    3|  14|            1|          3|   23|    10|  2017|
|    1|   11|  25|            1|         11|   16|    11|  2017|
|    8|    2|  27|            1|          2|   25|     2|  2017|
|   11|    9|  15|            1|          9|   11|     9|  2017|
|    3|    7|   2|            0|          3|   14|     3|  2017|
|    8|   10|  14|            1|         10|   23|    10|  2017|
|    2|    1|  22|            1|          1|   34|     1|  2017|
|    3|    9|  15|            0|          3|   11|     9|  2017|
|    7|   11|  25|            1|         11|   16|     7|  2017|
|   10|    1|  22|            1|          1|   34|     1|  2017|
|    9|    2|  21|            1|          2|    7|     2|  2017|
|    3|    1|   2|            1|          1|   14|     1|  2017|
|   11|    8|  27|            1|          8|   25|     8|  2017|
|    2|   10|  21|            1|         10|    7|     2|  2017|
|    7|    9|   9|            0|          7|    8|     7|  2017|
|    8|    1|  22|            1|          1|   34|     1|  2017|
|   11|    3|   2|            1|          3|   14|    11|  2017|
|    7|    2|   9|            0|          7|    8|     2|  2017|
|   10|    9|  14|            1|          9|   23|    10|  2017|
|    3|    8|  27|            1|          8|   25|     3|  2017|
+-----+-----+----+-------------+-----------+-----+------+------+
only showing top 20 rows
```

Now the updated features for the training data :

- Team1
- Team2
- Toss_winner
- Team1_batting_weight
- Team1_bowling_weight
- Team1_merged_weight
- Team2_batting_weight
- Team2_bowling_weight
- Team2_merged_weight



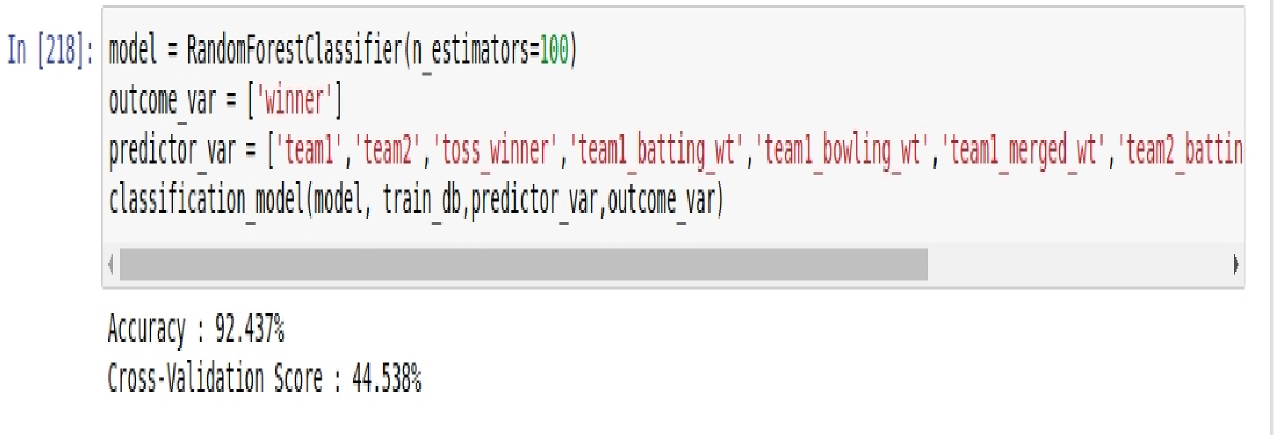
```
In [216]: train_db.head()
```

Out[216]:

	team1	team2	city	toss_decision	toss_winner	venue	winner	season	team1_batting_wt	team1_bowling_wt	team1_merged_wt	team2_batting_wt	team2
0	7	11	9	1	11	8	11	2016	32.456368	32.369110	64.825478	33.101415	
1	7	11	30	1	11	5	11	2016	32.456368	32.369110	64.825478	33.101415	
2	3	11	25	1	11	16	3	2016	36.882075	28.259162	65.141238	33.101415	
3	1	11	22	0	1	34	11	2016	38.459906	35.007853	73.467759	33.101415	
4	9	11	30	0	9	5	11	2016	35.429245	28.484293	63.913538	33.101415	

The model was trained on the above train data and the results were classified using random forest classifier. Accuracy and cross-validation scores were calculated. We were able to achieve accuracy of 92.43% and cross-validation score of 49.377%. So , it was clearly evident that by adding the weights of teams (i.e. batting and bowling weights for individual teams), accuracy and cross-validation scores were increased. Therefore, this new model with weights really fitted well for the data and brought the expected outcome.

Below is the accuracy which we were able to achieve:



```
In [218]: model = RandomForestClassifier(n_estimators=100)
          outcome_var = ['winner']
          predictor_var = ['team1','team2','toss_winner','team1_batting_wt','team1_bowling_wt','team1_merged_wt','team2_battin
          classification_model(model, train_db,predictor_var,outcome_var)

          Accuracy : 92.437%
          Cross-Validation Score : 44.538%
```

Conclusion:

We were able to carry out decent analysis and visulizations which depicted how the teams and players fared over the whole tournament. Not only the most important 10 players were calculated but a model was fitted and it successfully predicted the match winner.

Insights gained:

We gained understanding about the following points :

- Toss is actually not a big factor in a cricket match though it is considered to be one. The analysis clearly showed that a team losing the toss has almost the same chances of winning the match as of the team winning it. (51.3% of the matches have been won by the team winning the toss while 48.7% of the matches have been won by the team losing it.)
- In the seasons 2016 and 2017, a surprisingly large number of times teams have opted for field first after winning the toss. This clearly shows that as T20 cricket is evolving day by day, teams want a definite target that they want to achieve. This is exactly opposite the traditional approach of win the toss and bat first !!
- This analysis gave us the information about certain teams which have performed really well at their home ground. These teams are: Royal Chanllengers Banglore (RCB), Mumbai Indians (MI) and Chennai Super Kings (CSK). All the other teams were good irrespective of their home ground.

Future scope:

It is possible to take more number of features into consideration than what we have taken in this project. We have currently considered 5 batsman and 5 bowlers features. This number could be increased to 10. Similarly weights could be calculated out more accurately. This is ensure the correct assignment of total weight to every player. The training data can be made larger in the future for better model learning and classification. Various other data analysis techniques and a better machine learning algorithm can also be used for improving the accuracy of the model.

Attachments:

IPL.ipynb

References

[1] Indian Premier League, https://en.wikipedia.org/wiki/Indian_Premier_League.

[2] https://www.kaggle.com/manasgarg/ipldeliveries.csv

[3] Parker, David, Phil Burns, and Harish Natarajan. ”Player valuations in the indian premier league.” Frontier Economics 116 (2008)

[4] https://www.researchgate.net/publication/309225601_Data_Analytics_based_Deep_Mayo_Predictor_for_IPL 9