CS 543 - Massive Data Storage and Retrieval

Intermediate Report
Date - 10/23/2019
Team Members: Nidhi Harwani (nh417) & Adhish Shrivastava (as3003)

Most Valuable Player Prediction using IPL Dataset

PROJECT GOALS:

- To perform some preliminary analysis on the dataset so as to identify features that can be used in calculating player index.
- To come up with a way to quantify the performance of players, i.e. both Batsman and Bowlers, on the basis of their performance in the past seasons. We plan to create an index to do the same. Finally we plan to find the "Top 5 Most Valuable Players" based on the index calculated.
- To be able to predict the outcome of a match based on the input and the previous outcomes of all the matches across all the seasons. The input would be the playing teams.

PROBLEM DESCRIPTION:

Indian Premier League (IPL) is a cricket tournament which was founded by Board of Control for Cricket in India (BCCI), 12 years ago in 2008. The tournament consists of players from within India as well as players from other cricket playing countries. Every year an auction of players is done in which each team gets a chance to bid for players in the talent pool, the highest bidder get the player for their team. There are a total of 8 teams participating in the tournament.

Every year heavy analysis is done before the auction, taking into account the performance of players during the past IPL seasons as well as their performance in other tournaments between IPL seasons. This is done so as to be able to predict which player could be a good match along with the other players of the team and in general, a good fit for the team. This is basically done to create a team which will end up winning the tournament.

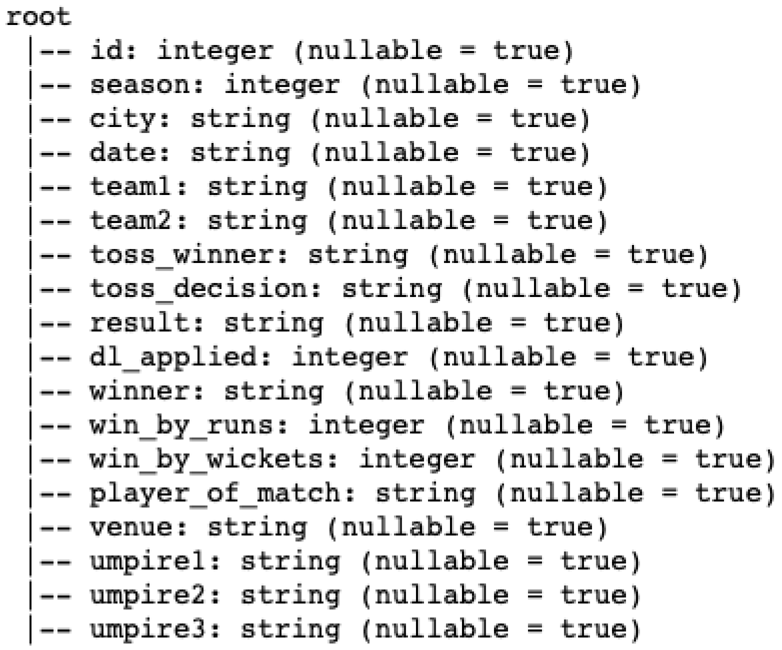
MOTIVATION:

Indian Premier League is a cricket tournament that has been played in India since 2008 during the summer. It has won the hearts of people all over the world. To analyze the data from this tournament is quite exciting. While watching any tournament we often think as to who would win the series. This thought was where the idea of the project came from. The analysis of the past matches and teams and finding out which team would win would give the IPL team owners if the team selected will help them win.

Secondly, the idea of finding out who are the top players of the tournament would help the team owners to bid and make a strong team. Also, it would give them an idea as to what their strengths are and what areas the team needs to focus more upon. Also, the top players list would help the players find their position in the game and how they rank with their competition.

DATA COLLECTION:

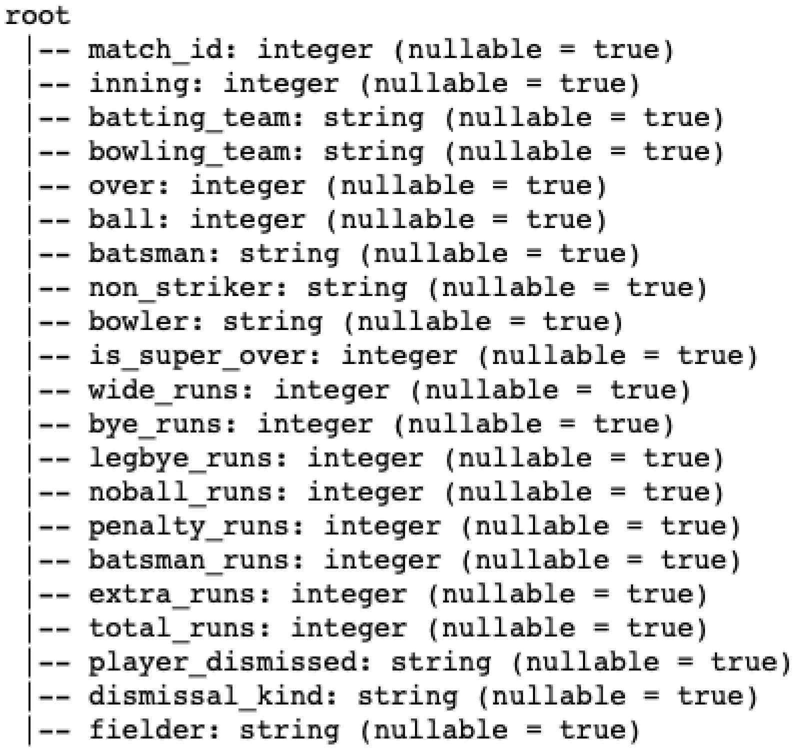
We got the dataset from kaggle. The data has the information of all the matches (approx. 750 matches) played across 12 seasons. The data consists of two csv files. The first file contains the results of all the matches played in the Indian Premier League from year 2008 to 2019. This file has the following schema:



```
root
 |-- id: integer (nullable = true)
 |-- season: integer (nullable = true)
 |-- city: string (nullable = true)
 |-- date: string (nullable = true)
 |-- team1: string (nullable = true)
 |-- team2: string (nullable = true)
 |-- toss_winner: string (nullable = true)
 |-- toss_decision: string (nullable = true)
 |-- result: string (nullable = true)
 |-- dl_applied: integer (nullable = true)
 |-- winner: string (nullable = true)
 |-- win_by_runs: integer (nullable = true)
 |-- win_by_wickets: integer (nullable = true)
 |-- player_of_match: string (nullable = true)
 |-- venue: string (nullable = true)
 |-- umpire1: string (nullable = true)
 |-- umpire2: string (nullable = true)
 |-- umpire3: string (nullable = true)
```

Figure 1: Schema for matches.csv

The second file contains information about delivery by delivery information along with the bowler and the batsman on strike. This file has the following schema:



```
root
 |-- match_id: integer (nullable = true)
 |-- inning: integer (nullable = true)
 |-- batting_team: string (nullable = true)
 |-- bowling_team: string (nullable = true)
 |-- over: integer (nullable = true)
 |-- ball: integer (nullable = true)
 |-- batsman: string (nullable = true)
 |-- non_striker: string (nullable = true)
 |-- bowler: string (nullable = true)
 |-- is_super_over: integer (nullable = true)
 |-- wide_runs: integer (nullable = true)
 |-- bye_runs: integer (nullable = true)
 |-- legbye_runs: integer (nullable = true)
 |-- noball_runs: integer (nullable = true)
 |-- penalty_runs: integer (nullable = true)
 |-- batsman_runs: integer (nullable = true)
 |-- extra_runs: integer (nullable = true)
 |-- total_runs: integer (nullable = true)
 |-- player_dismissed: string (nullable = true)
 |-- dismissal_kind: string (nullable = true)
 |-- fielder: string (nullable = true)
```

Figure 2: Schema for deliveries.csv

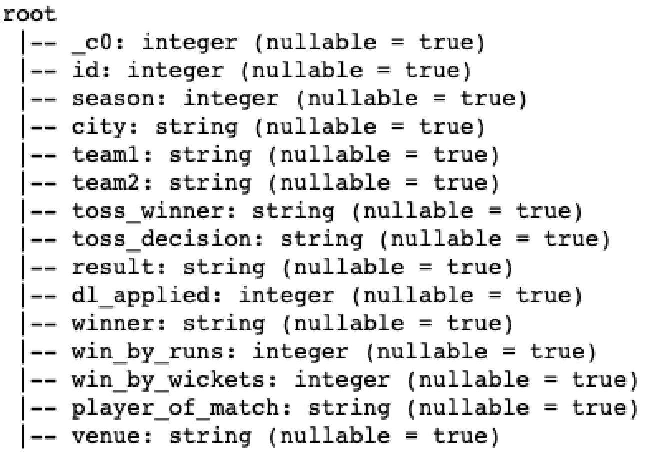
This is the schema of the original data as we got from kaggle.

We plan to clean the data by dropping columns that will not be needed in the analysis or prediction. Also, modifying values like Team names which might be erroneous.

DATA PRE-PROCESSING:

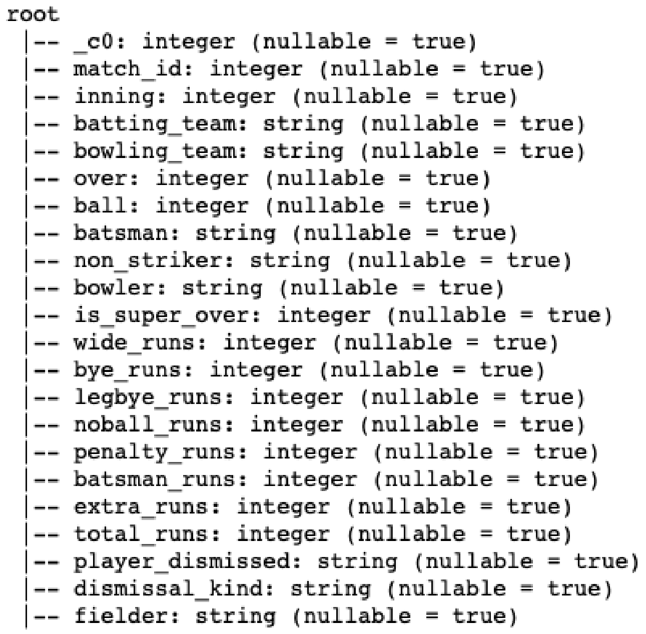
- Find all nan or empty values and handle them accordingly. We replaced all int value columns with -1 and all string value columns with None.
- While going through the data we found that some team names were not correct, so we cleaned that up.
- Some teams during the course of 12 years have changed their names, so we changed all the team names to the current team names.
- For the task of predicting which team will win on the basis of their past played matches, we plan to drop columns like date, umpire1, umpire2 and umpire3 from the matches.csv.

The schema of the data after pre-processing is as follows:



```
root
 |-- _c0: integer (nullable = true)
 |-- id: integer (nullable = true)
 |-- season: integer (nullable = true)
 |-- city: string (nullable = true)
 |-- team1: string (nullable = true)
 |-- team2: string (nullable = true)
 |-- toss_winner: string (nullable = true)
 |-- toss_decision: string (nullable = true)
 |-- result: string (nullable = true)
 |-- dl_applied: integer (nullable = true)
 |-- winner: string (nullable = true)
 |-- win_by_runs: integer (nullable = true)
 |-- win_by_wickets: integer (nullable = true)
 |-- player_of_match: string (nullable = true)
 |-- venue: string (nullable = true)
```

Figure 3: Schema for matches.csv after data processing



```
root
 |-- _c0: integer (nullable = true)
 |-- match_id: integer (nullable = true)
 |-- inning: integer (nullable = true)
 |-- batting_team: string (nullable = true)
 |-- bowling_team: string (nullable = true)
 |-- over: integer (nullable = true)
 |-- ball: integer (nullable = true)
 |-- batsman: string (nullable = true)
 |-- non_striker: string (nullable = true)
 |-- bowler: string (nullable = true)
 |-- is_super_over: integer (nullable = true)
 |-- wide_runs: integer (nullable = true)
 |-- bye_runs: integer (nullable = true)
 |-- legbye_runs: integer (nullable = true)
 |-- noball_runs: integer (nullable = true)
 |-- penalty_runs: integer (nullable = true)
 |-- batsman_runs: integer (nullable = true)
 |-- extra_runs: integer (nullable = true)
 |-- total_runs: integer (nullable = true)
 |-- player_dismissed: string (nullable = true)
 |-- dismissal_kind: string (nullable = true)
 |-- fielder: string (nullable = true)
```

Figure 4: Schema for deliveries.csv after data processing

BASIC ANALYSIS:

Number of matches played = 756
Number of deliveries = 179078
Number of Batsmen = 516
Number of Bowlers = 405
Number of Player = 559

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

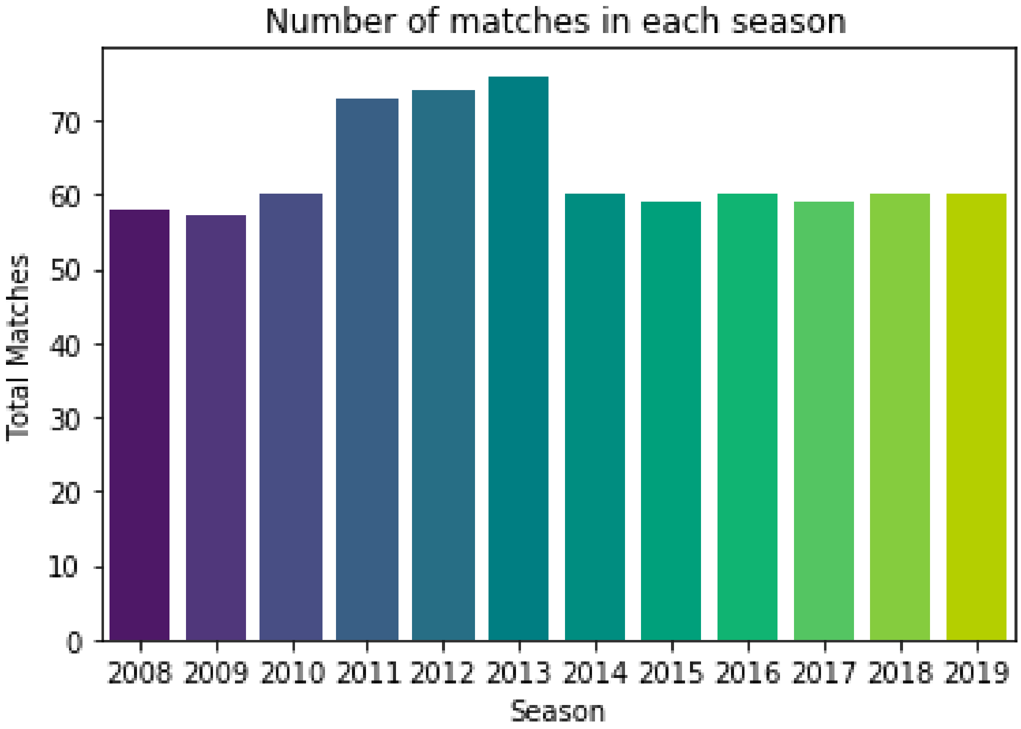
Name of various venues:
'Rajiv Gandhi International Stadium, Uppal', 'Maharashtra Cricket Association Stadium', 'Saurashtra Cricket Association Stadium', 'Holkar Cricket Stadium', 'M Chinnaswamy Stadium', 'Wankhede Stadium', 'Eden Gardens', 'Feroz Shah Kotla', 'Punjab Cricket Association IS Bindra Stadium,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

Name of cities matches were played at:
'Hyderabad', 'Pune', 'Rajkot', 'Indore', 'Bangalore', 'Mumbai', 'Kolkata', 'Delhi', 'Chandigarh', 'Kanpur', 'Jaipur', 'Chennai', 'Cape Town', 'Port Elizabeth', 'Durban', 'Centurion', 'East London', 'Johannesburg', 'Kimberley', 'Bloemfontein', 'Ahmedabad', 'Cuttack', 'Nagpur', 'Dharamsala', 'Kochi', 'Visakhapatnam', 'Raipur', 'Ranchi', 'Abu Dhabi', 'Sharjah', 'Mohali', 'Bengaluru'

ANALYSIS USING SPARK:

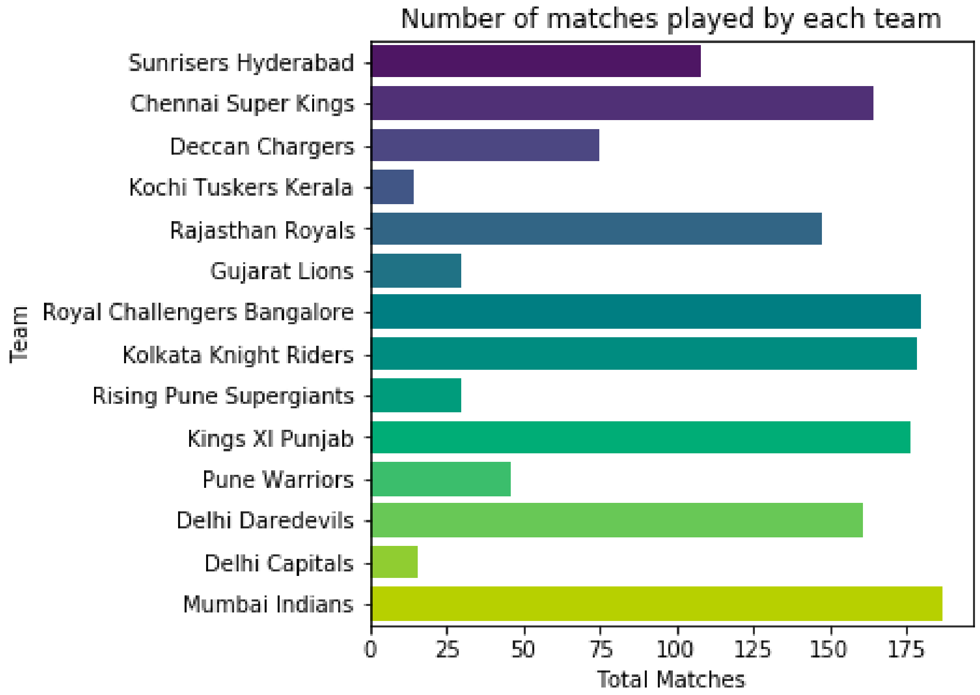
Here we have plotted graphs of some features and derived some insights from it.

1. Number of Matches in each Season:



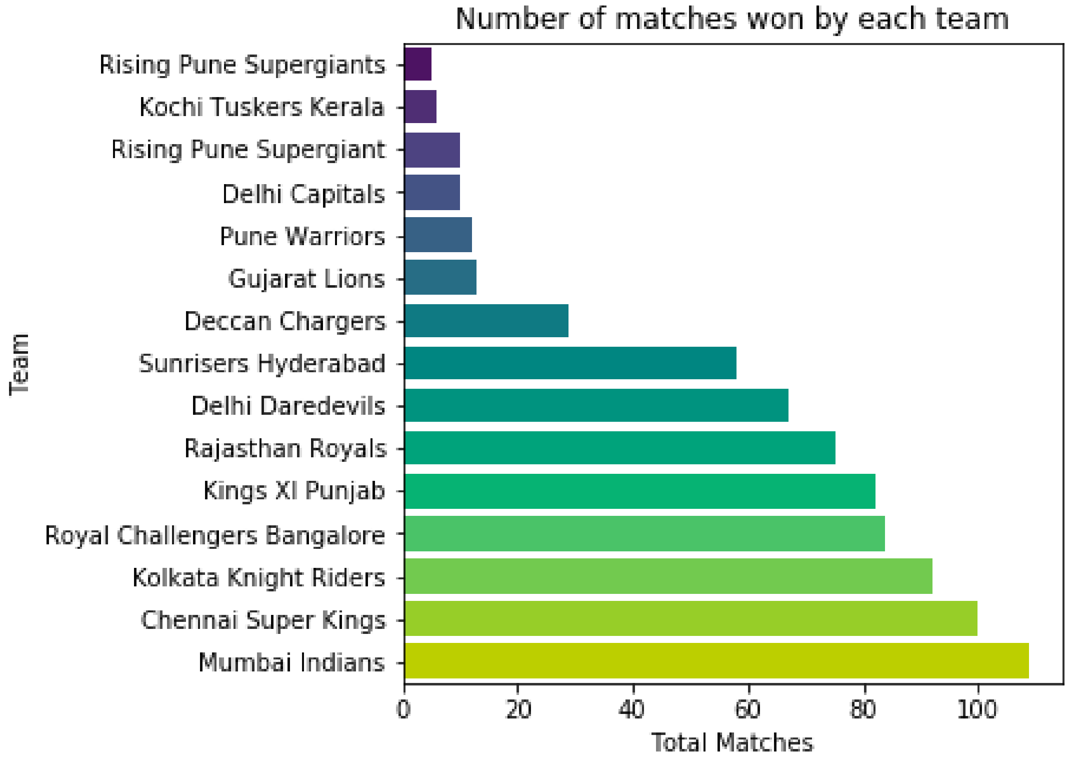
- The above graph shows the total number of matches held in each season from the year 2008 up till 2019.
- We can see that the maximum matches were held during the 2013 season. This must be because the number of teams participating was higher than usual that year. Also, the average number of matches per season was 63 with the lowest being 57 (2009) and the highest being 76 (2013).

2. Number of Matches played by each team:

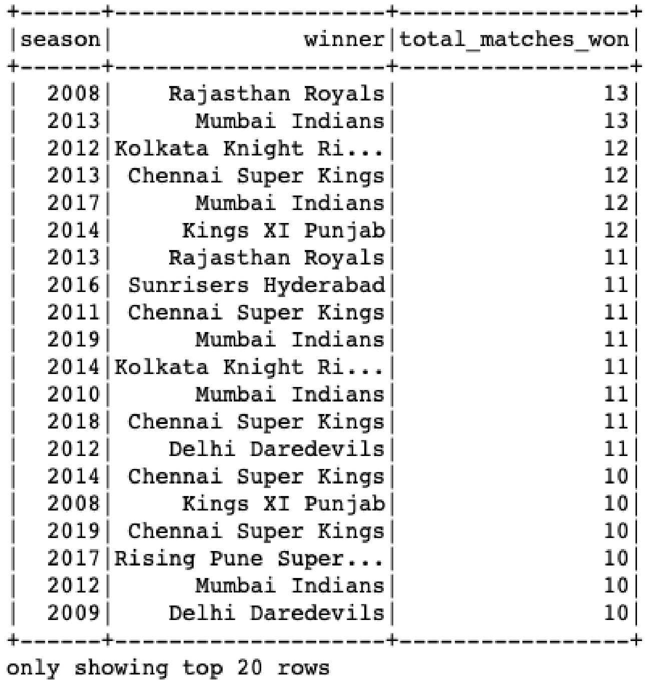


From the above plot, we can see that Mumbai Indians has the max number of matches played overall in the league. Also, we see that Delhi Capitals has the least number of matches. This can be justified as Delhi Daredevils was renamed to Delhi Capitals recently. Also, there are teams like Gujarat Lions, Kochi Tuskers Kerala which played only for some seasons.

3. Number of matches won by each team:

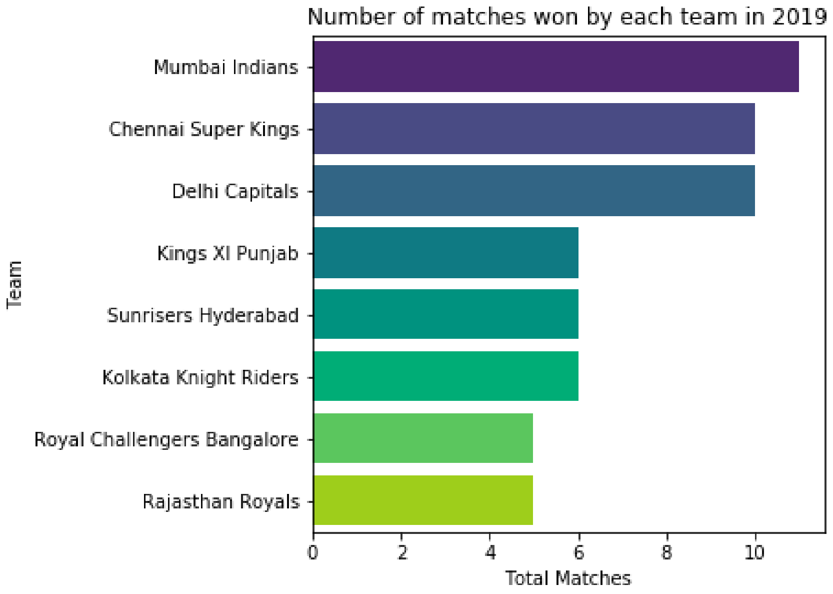


The above graph tells you the total matches won by a team during all seasons. As the team Mumbai Indians have the highest number of total matches played and the highest number of total matches won, the probability of it winning some seasons become higher.



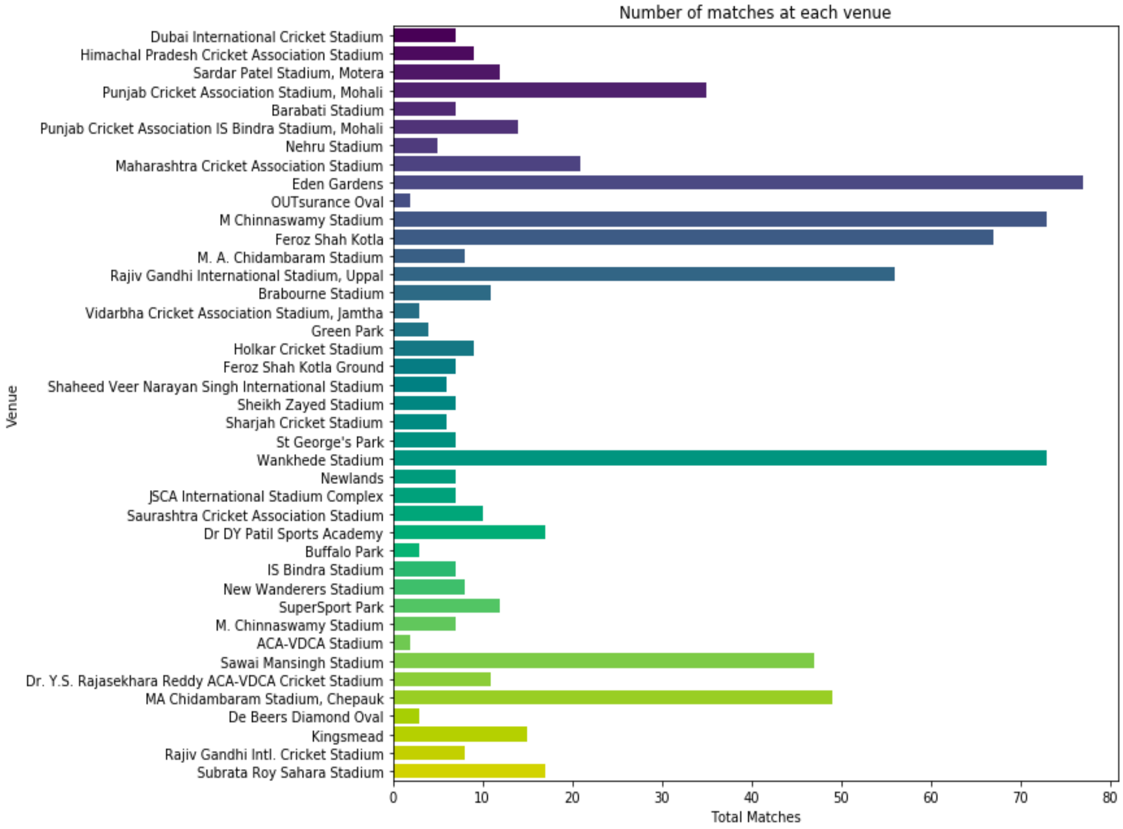
```
+------+--------------------+-----------------+
|season|              winner|total_matches_won|
+------+--------------------+-----------------+
|  2008|    Rajasthan Royals|               13|
|  2013|      Mumbai Indians|               13|
|  2012|Kolkata Knight Ri...|               12|
|  2013| Chennai Super Kings|               12|
|  2017|      Mumbai Indians|               12|
|  2014|     Kings XI Punjab|               12|
|  2013|    Rajasthan Royals|               11|
|  2016| Sunrisers Hyderabad|               11|
|  2011| Chennai Super Kings|               11|
|  2019|      Mumbai Indians|               11|
|  2014|Kolkata Knight Ri...|               11|
|  2010|      Mumbai Indians|               11|
|  2018| Chennai Super Kings|               11|
|  2012|    Delhi Daredevils|               11|
|  2014| Chennai Super Kings|               10|
|  2008|     Kings XI Punjab|               10|
|  2019| Chennai Super Kings|               10|
|  2017|Rising Pune Super...|               10|
|  2012|      Mumbai Indians|               10|
|  2009|    Delhi Daredevils|               10|
+------+--------------------+-----------------+
only showing top 20 rows
```

The above table shows the number of matches won by a team in each season. If the total is higher, it means that the team was closer to winning the season. Now, we could look at a particular season to determine who won the series in that year. Let's look at an example of year 2019.

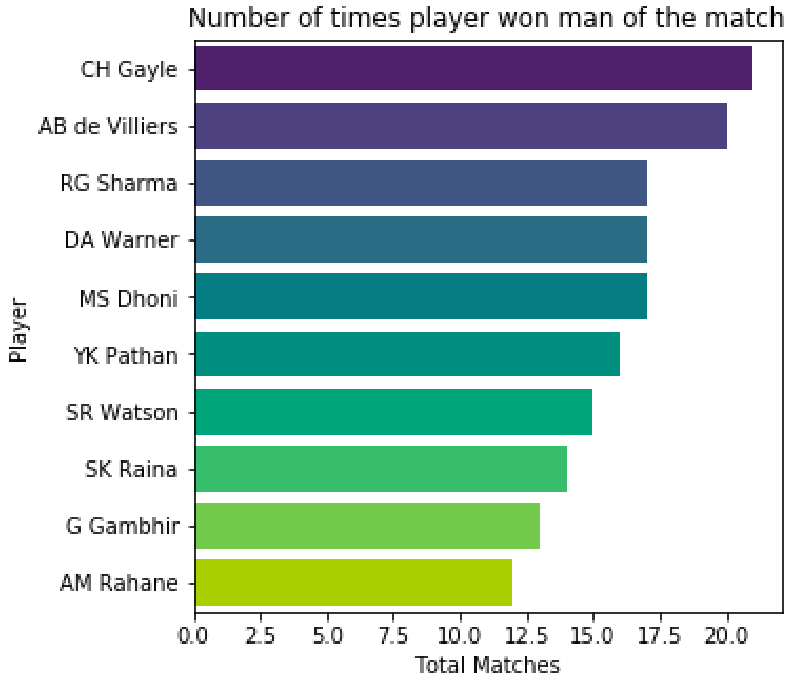


We can see that Mumbai Indians won the max number of matches in 2019. So, we can definitely say that Mumbai Indians was the winner of the series in 2019. Next, we could see that Chennai Super Kings and Delhi Capitals have the same number of matches won. We can say that these two played against each other and one team went to play the final against Mumbai Indians and then lost against them.

4. Number of matches at each venue:

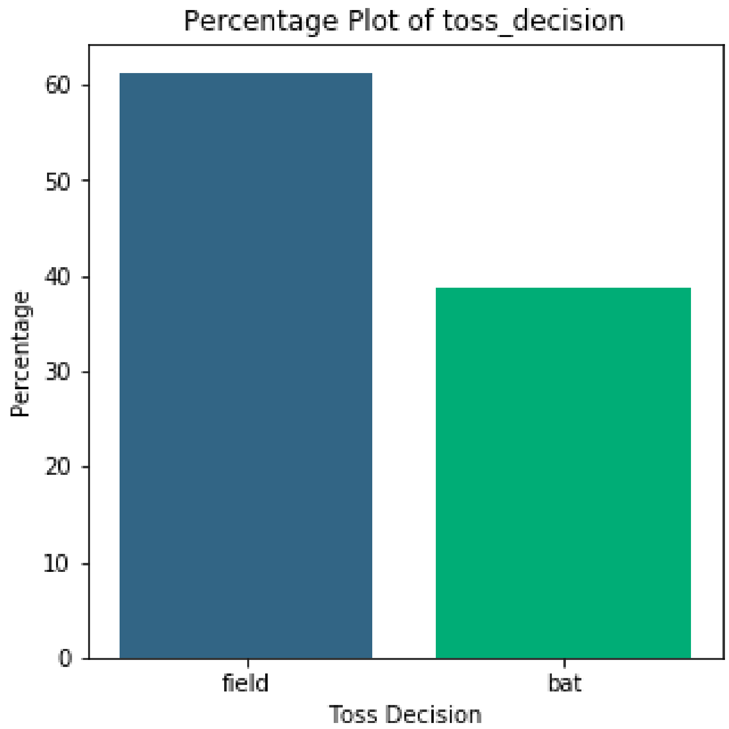


5. Number of times a player won man of the match:



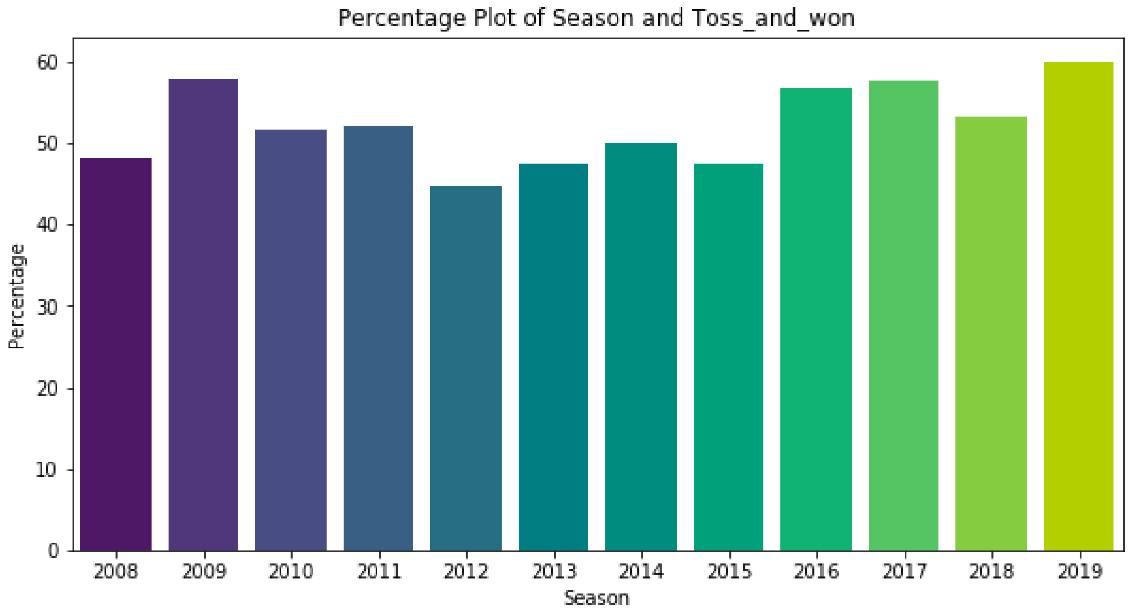
This plot shows the number of times the players won man of the match. It shows the details about the first 10 players with the highest number of man of the match awards. This can be taken into consideration when a team is selecting players for itself.

6. Toss Decision:



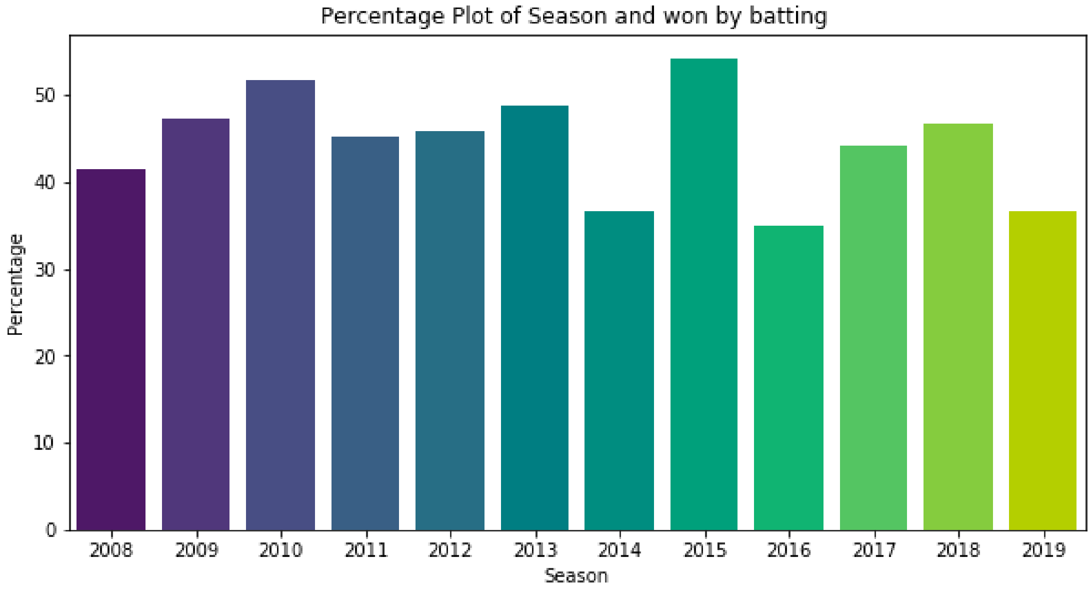
The above plot tells us that when a team wins the toss, it is more likely to choose bowling/fielding in the first inning then batting.

7. Toss win and match win by same team per season:

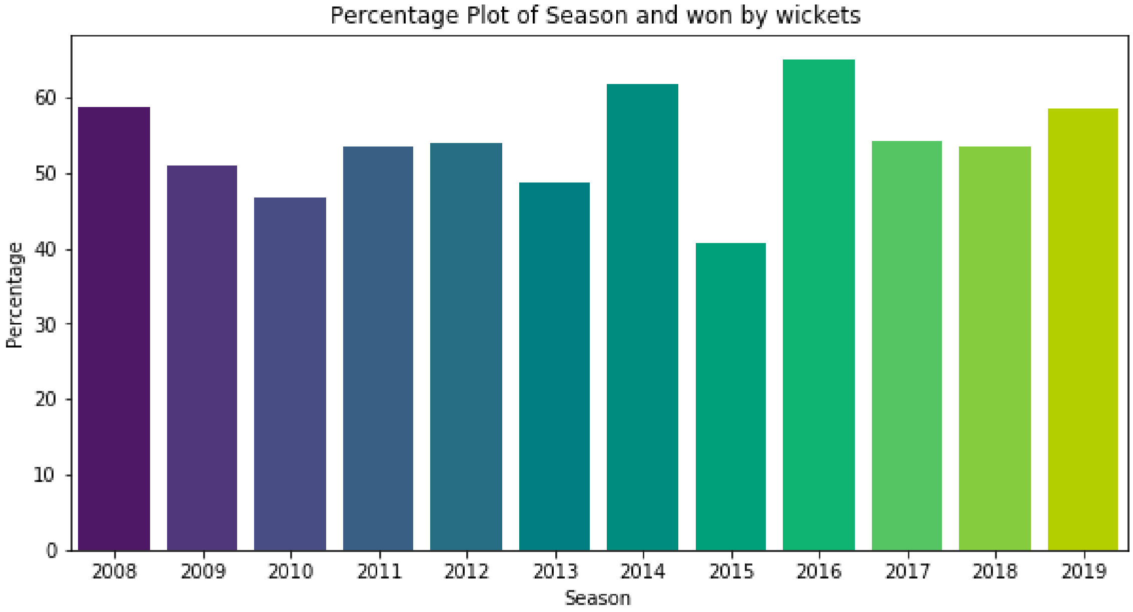


The above graph shows that when a team wins the toss and decides either to bat or bowl, what is the likelihood that the team wins the match after taking the decision i.e. what percentage of teams win both the toss and the match. For example, in 2019 around 60% time the teams on the toss and then went ahead and won the match too.

8. Percentage plot of teams won by batting first:



9. Percentage plot of teams won by bowling first:



Tasks for the next stage of the Project:

- To use the information gathered during this stage of analysis to put in a mechanism to calculate the index which can be used to rank the players. Also, plan to use Most Valuable Player Index (MVPI) technique to get an index for a player on the basis of his performance as a batsman or a bowler. We will then use this index to rank players.

References:

- https://en.wikipedia.org/wiki/Indian_Premier_League#Current_teams
- https://www.kaggle.com/nowke9/ipldata