**CS 543 - Massive Data Storage and Retrieval**

Final Report

Group No.: 19

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.

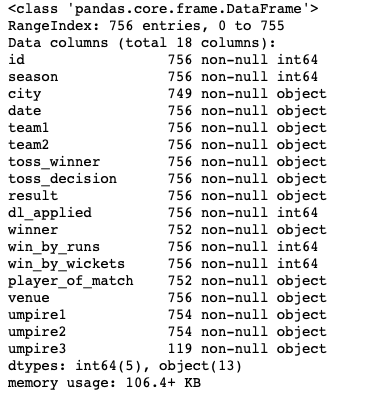
**Motivation:**

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.

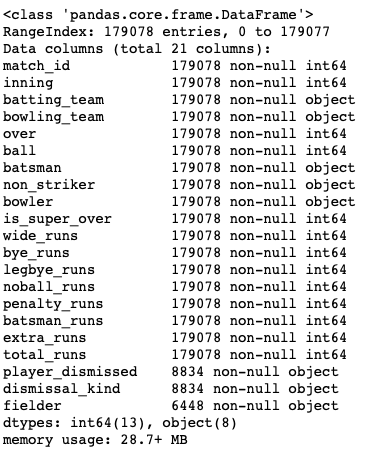
**Data Collection:**

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:



**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:



**Figure 2: Schema for deliveries.csv**

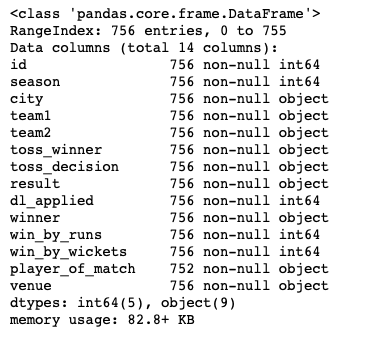
This is the schema of the original data as we got from kaggle.

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.

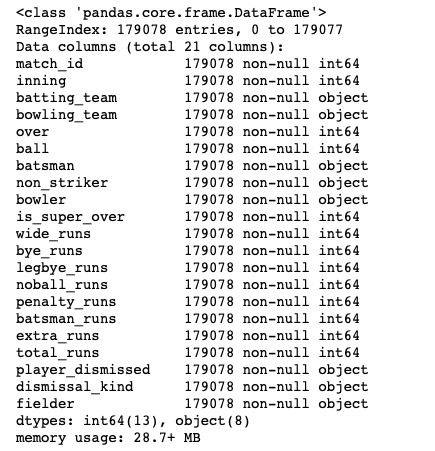
**Data Pre-processing:**

* 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.
* In the ‘winner’ column, all the places where there was null value (no result), we changed if to ‘Draw’.
* It was found that all the matches that took place in Dubai, had the city column missing in them. We used the name of the stadium to fill the city columns.
* 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:



**Figure 3: Schema for matches.csv after data processing**



**Figure 4: Schema for deliveries.csv after data processing**

**Basic Analysis:**

* **All the teams that have played in the league since 2008:**

Total number of teams participated so far: 14

Teams participated so far:

* 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
* Delhi Capitals
* **The venues where the matches were held: 41**

'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'

* **Total number of cities where the matches were played : 33**

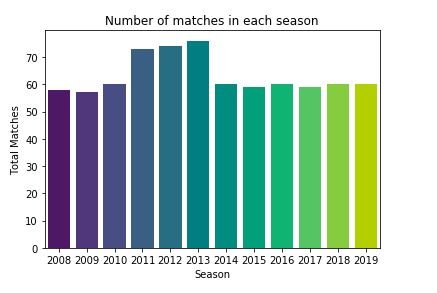
'Hyderabad', 'Pune', 'Rajkot', 'Indore', 'Bangalore', 'Mumbai', 'Kolkata', 'Delhi', 'Chandigarh', 'Kanpur', 'Jaipur', 'Chennai', 'Cape Town', 'Port Elizabeth', 'Durban', 'Centurion', 'East London', 'Johannesburg', 'Kimberley', 'Bloemfontein', 'Ahmedabad', 'Cuttack', 'Nagpur', 'Dharamsala', 'Kochi', 'Visakhapatnam',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

* **Number of bowlers : 405**
* **Number of batsmen: 516**
* **Total players that have played in the league since 2008: 559**
* **Total matches played: 756**
* **Number of deliveries: 179078**

**Analysis Using Spark:**

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

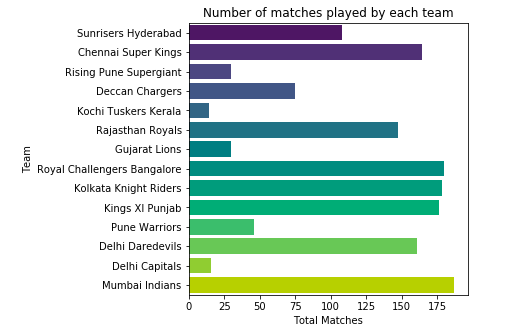
1. **Number of Matches in each Season:**



The above graph shows the total number of matches held in each season from the year 2008 up till 2019.

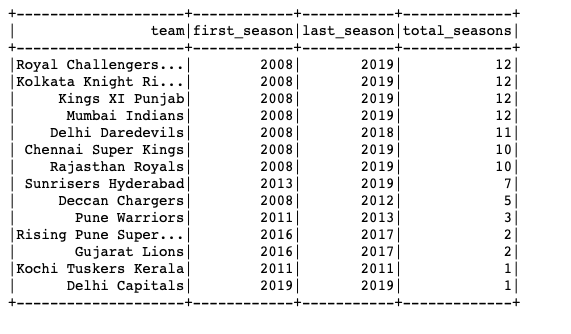
We can see that the maximum matches were held during the 2013 season. This must be because the number of teams participating was higher than usual that year. Also, the average number of matches per season was 63 with the lowest being 57 (2009) and the highest being 76 (2013).

1. **Number of Matches played by each team:**



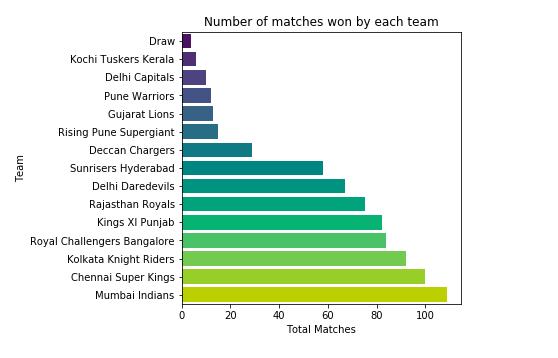
From the above plot, we can see that Mumbai Indians has the max number of matches played overall in the league. Also, we see that Delhi Capitals has the least number of matches. This can be justified as Delhi Daredevils was renamed to Delhi Capitals recently. Also, there are teams like Gujarat Lions,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

1. **Total season in which teams have played:**

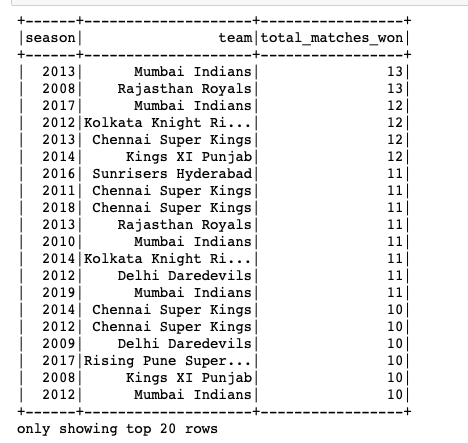


The above graph tells us about the total seasons each team has played in the league. The team Kochi Tuskers Kerala played only in the year 2011. Whereas 4 teams have played in all the seasons up till now. Also, as mentioned above, Delhi Daredevils was changed to Delhi Capitals in 2019 as the owner of the team was changed. This makes Delhi Capitals a new team with only one season under its belt.

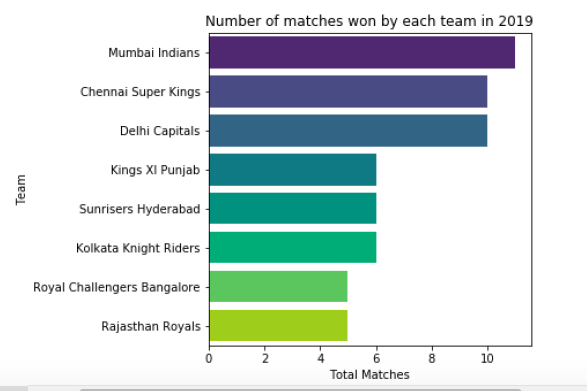
1. **Number of matches won by each team:**



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.

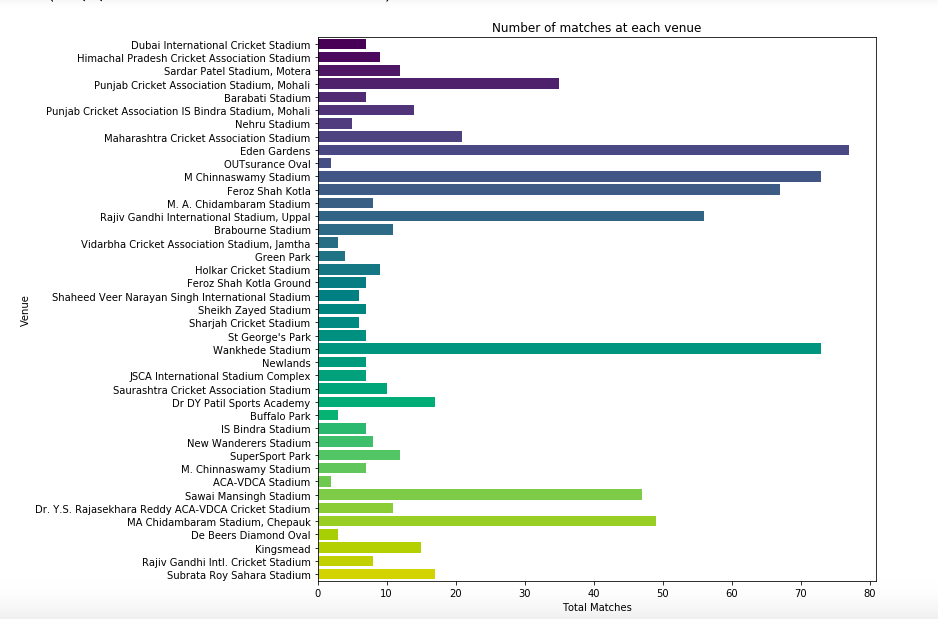


The above table shows the number of matches won by a team in each season. If the total is higher, it means that the team was closer to winning the season. Now, we could look at a particular season to determine who won the series in that year. Let’s look at an example of year 2019.

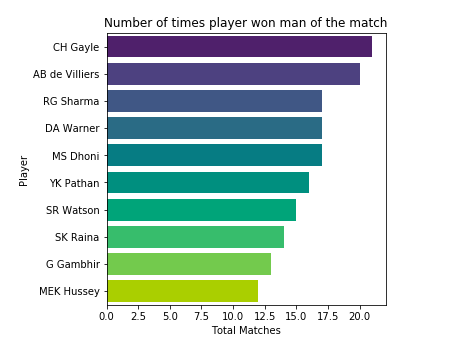


We can see that Mumbai Indians won the max number of matches in 2019. So, we can definitely say that Mumbai Indians was the winner of the series in 2019. Next, we could see that Chennai Super Kings and Delhi Capitals have the same number of matches won. We can say that these two played against each other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

1. **Number of matches at each venue:**

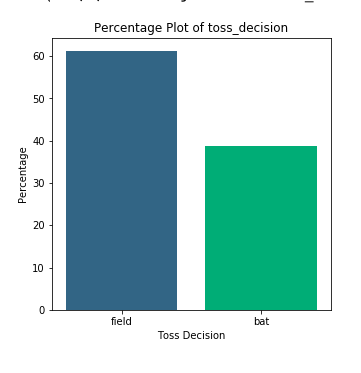


1. **Number of times a player won man of the match:**



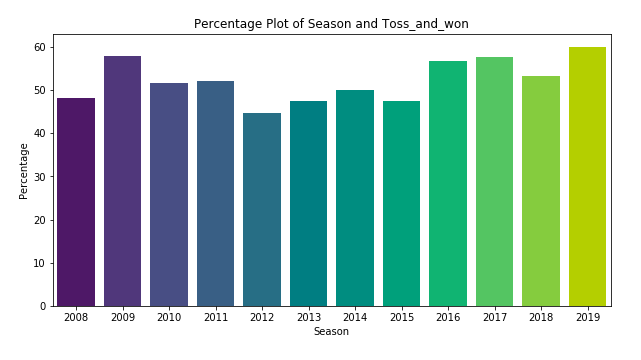
This plot shows the number of times the players won man of the match. It shows the details about the first 10 players with the highest number of man of the match awards. This can be taken into consideration when a team is selecting players for itself.

1. **Toss Decision:**



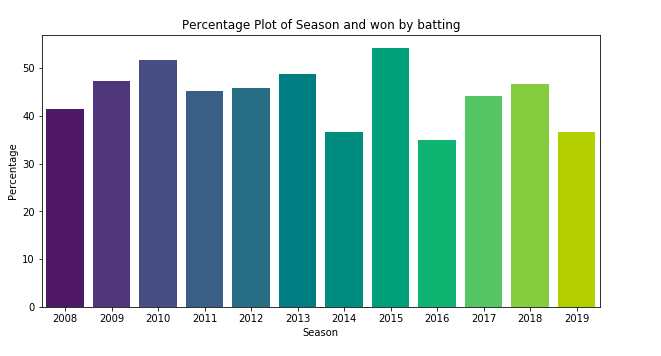
The above plot tells us that when a team wins the toss, it is more likely to choose bowling/fielding in the first inning then batting.

1. **Toss win and match win by same team per season:**

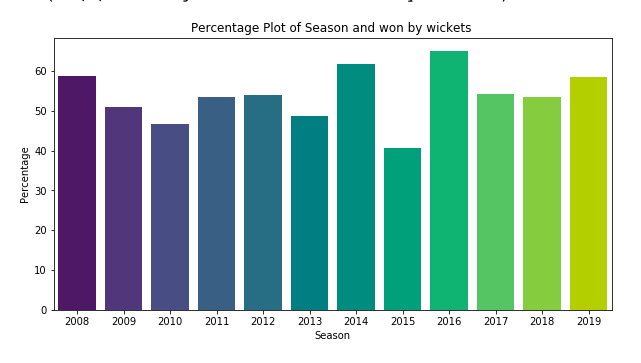


The above graph shows that when a team wins the toss and decides either to bat or bowl, what is the likelihood that the team wins the match after taking the decision i.e. what percentage of teams win both the toss and the match. For example, in 2019 around 60% time the teams won the toss and then went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

1. **Percentage plot of teams won by batting first:**



1. **Percentage plot of teams won by bowling first:**

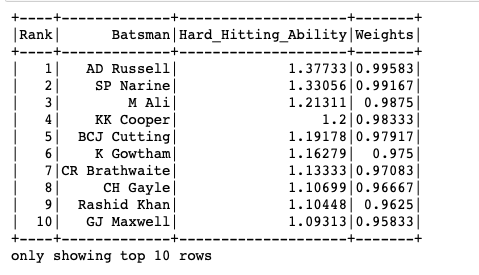


**Final Analysis**

For Batsmen as well as for the bowlers we calculated some parameters based on which we can measure their impact in their respective teams.

The parameters we calculated for Batsmen are:

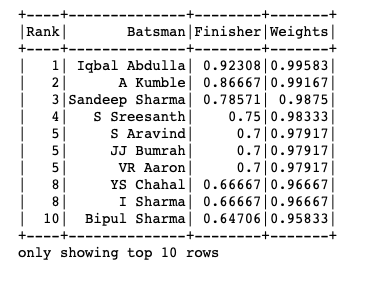
* **Hard Hitting Ability**



Hard hitting ability is the total runs made by scoring 4’s and 6’s over the total balls played by that batsman. This score is a measure for how many fours and sixes a player can make in the least number of balls. The higher the hard hitting ability, the better batsman the player can be.

The score is a weighted average for each batsman. This can be a major aspect for a team owner in selecting the batsmen for the team. The above table shows the top 10 players with the highest hard hitting ability score.

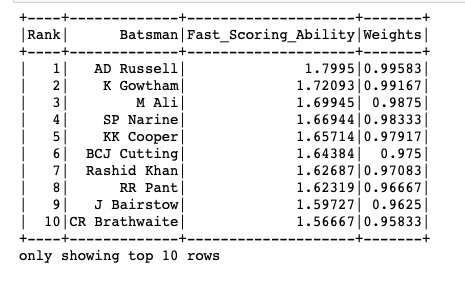
* **Finisher**



Finisher score tells us if the batsman can race to the bottom or not. It is calculated as follows:

This is also a weighted score. The finishing score is a pivotal impact on the chances of the team winning the match. Iqbal Abdulla has the highest finishing score of 0.92. The above table represents the top 10 finishers in the league.

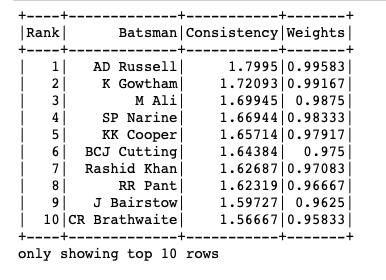
* **Fast Scoring Ability**



Fast scoring ability is the ability of the batsman to score maximum number of runs in less number of balls. In T20 matches, each ball is important and wasting one ball can bring the downfall of the team in that match. This score is calculated as follows:

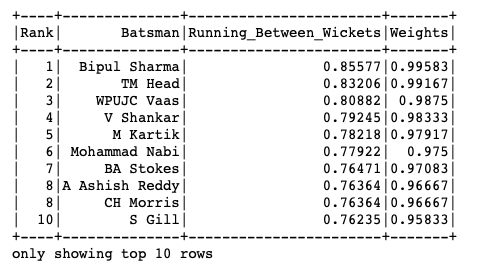
AD Russell has the highest fast scoring ability which makes him a good candidate for any team. The above table shows the top 10 fast scoring batsman in the league.

* **Consistency**

****

Any team looks for a player who is consistent in his game. Consistency is calculated here as the average number of runs made by the player based on the number of times he was out by the other team. The above table displays the top 10 consistent players in the league.

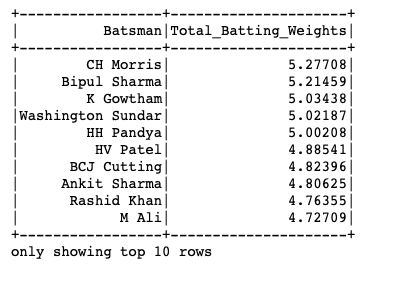
* **Running Between Wickets**



Running between wickets is the rate of runs made by the batsman by running between the wickets. In T20 matches, runs made by running between the wickets give the players an edge above the others. We find the total number of runs by running by subtracting the boundaries from the total runs. The running between the wickets score is calculated as follows:

The top 10 players with the highest running between the wickets score are shown above.

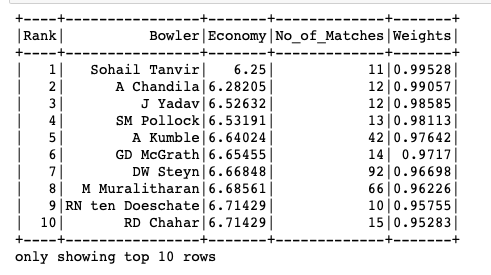
**Total Batting Weights**



The total batting weight is the sum of the weights of the 5 batting matrices i.e. hard hitting ability, finishing score, fast scoring ability, consistency and running between wickets. This is the final metric for batting. The above table shows the top 10 batsman with the maximum batting weights.

The parameters we calculated for Bowlers are:

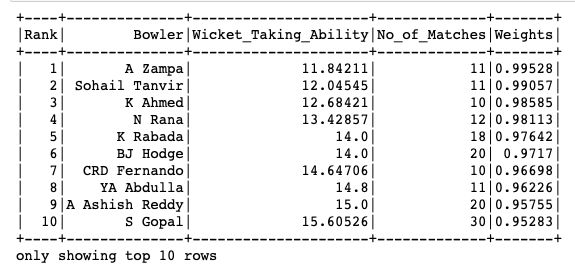
* **Economy**

****

Economy score is a bowling metric. A team needs a good economic bowler to ensure that the opposition team is not able to score more runs which would then lead to giving a chance to the other bowlers to put pressure on the batsmen.  Economy score is calculated as follows:

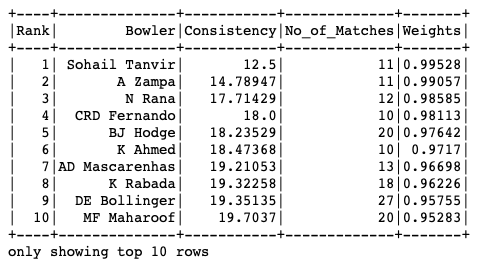
The total overs can be found by total balls bowled by the bowler over 6.  The above table shows the top 10 economic bowlers in the league.

* **Wicket Taking Ability**



The wicket taking ability is a very crucial metric for a bowler. The wicket taking ability tells about the average number of balls it takes for a number of wickets. A wicket is important in a T20 match for 2 specific reasons. Each wicket that the bowling team takes adds pressure on the batting team. Also, the wicket helps in slowing down the run rate. Thus, a team owner looks for a bowler with higher wicket taking ability as it would help the team score a win. The above table displays the ranking of the bowlers with high wicket taking ability.

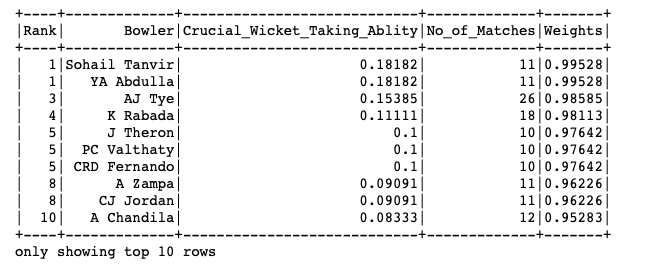
* **Consistency**



A consistent bowler is the one who takes more wickets while keeping the runs at bay. The consistency of a bowler is calculated as

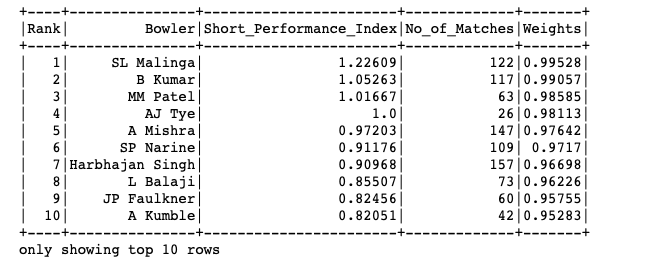
The above table shows the top 10 bowlers that have been consistent during the league.

* **Crucial Wicket Taking Ability**

****

Crucial wicket taking ability is a higher metric of the wicket taking ability. In this bowling metric, we look for the bowlers who take 4 - 5 wickets in an inning. The crucial wicket taking bowlers are the strong pillars of a team. This metric is calculated as follows: number of times the bowler has taken 4-5 wickets / Total innings played by the bowler. As can be seen, it is very difficult for any bowler to take 4-5 wickets in one inning. Thus, the highest score in this metric is 0.18 which is very low as compared to the other bowling metrics.

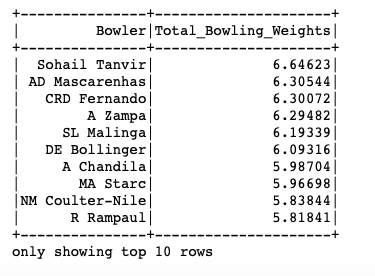
* **Short Performance Index**



Short performance index is the complement of the crucial wicket taking ability. This means that in the short performance index, we look into the bowlers who take less than 4 wickets in one inning. So, this index is calculated as follows:

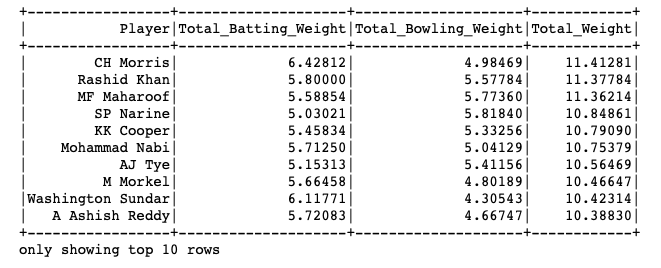
This index factors the good bowling periods that a bowler demonstrated during the league. We can see that SL Malinga has the highest short performance index which means that he takes less than 4 wickets on average in an inning.

**Total Bowling Weights**



The total bowling weight is the sum of the 5 bowling metric weights i.e. economic score, wicket taking ability, consistency, crucial wicket taking ability and the short performance index. This is the final bowling metric for any bowler. The table tells us the top 10 bowlers with the highest bowling metrics.

**Total Player Weights**



The above metric is for finding about the players who are good at both batting and bowling. It gives us top 10 overall players in the league by using the basic analysis. Here, it may happen that a player is a really good batsman but not a good bowler. Such a player would not be too high in the list. But, a player who is above average in both bowling and batting would be much higher in the list.

From the above analysis, we have found out the top 5 batsmen, top 5 bowlers and the top 5 all-rounder players in the league.

**I. Top 5 Batsmen of the league are:**

1. **CH Morris**
2. **Bipul Sharma**
3. **K Gowtham**
4. **Washington Sundar**
5. **HH Pandya**

**II. Top 5 Bowlers of the league are:**

1. **Sohail Tanvir**
2. **AD Mascarenhas**
3. **CRD Fernando**
4. **A Zampa**
5. **SL Malinga**

**III. Top 5 All-rounder players of the league are:**

1. **CH Morris**
2. **Rashid Khan**
3. **MF Maharoof**
4. **SP Narine**
5. **KK Cooper**

This is the first part of our project where we determine the top players of the league. These lists help the team owners in identifying which players would be more beneficial for the team to win the series. The next part of the project focuses on predicting which team would win the match.

**Model Training**

To train a model on this data, we first need to encode all the values into numerical values. We have encoded all the team names. Another value ‘Draw’ which means that the match was a draw has also been encoded. We have use a very basic encoding system which is shown as below.

For columns like city, venue and toss decision, we have used the sklearn library of the python language. We have used the LabelEncoder function to encode the values in these columns.

For training the machine learning model, we have used two approaches.

* **Using Random Forest Classifier on the Data**

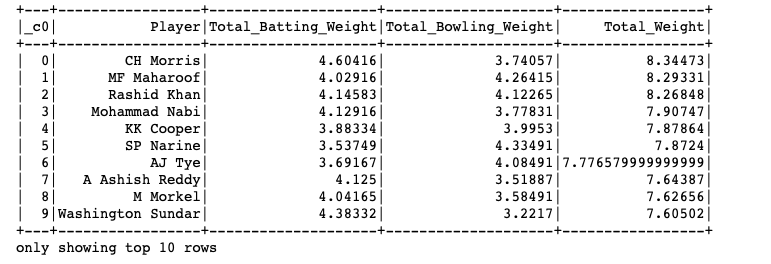
In this approach, we have taken the columns 'team1', 'team2', 'venue', 'toss\_winner', 'city' and 'toss\_decision' as the input and ‘winner’ column as the output value. In random forest, the model is created using random samples from the data which are converted into decision trees which are then used for prediction. Here, we have used the random forest classifier of the sklearn library of the python language. We trained the model by building the forest of trees on the training data. Then we predict the winner based on the trained model and check the accuracy of it as well. For the selected features we got an accuracy of 85.8%. Next, we cross-validated the data using the K-fold model of the sklearn library which gave us a score of 50.663%. This was a simple approach to create a model to predict which team would win the match. To increase the accuracy of the model, we used the second approach.

* **Using weighted data for Random Forest Classifier**

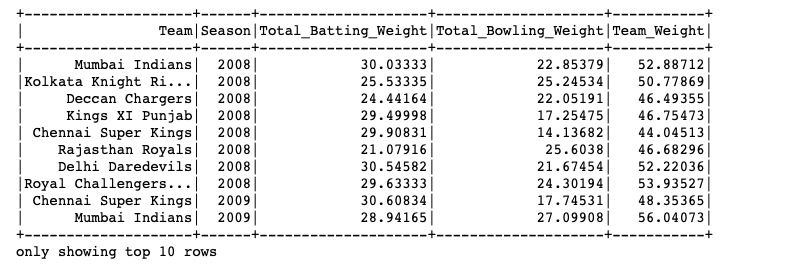
In this approach, we have first created another table called the players. In this dataframe, we have the information about which player was playing for which team during a season.



This table above shows the players with respect to the team and the season. A player can occur multiple times as it may happen that the player was auctioned by a different team in multiple seasons. We calculated the weights of the players in the analysis part of the project.



Now, based on the players each team had in a season, the batting and bowling weights for each team was calculated.



As you can see Royal Challengers Bangalore and Mumbai Indians had the highest total team weights in 2008 season.

Now, we encode this data using the simple encoding technique shown above.

This is the encoded version of the team weights that we calculated based on the seasons. Now, to train the model, we inculcate the above data with other inputs.

The input for the model is [ team1, team2, city, toss decision, toss winner, venue, season, bowling weights of each team, batting weights of each team, the total weights of each team] and winner as the output.

Now, we use the random forest classifier on this set of data and train the model. By using this data, we get an increased accuracy and we get a total accuracy of 88% and cross validation of 49.46%. Thus, we can see that by using the weights for the input we get a better accuracy of the model.

Now, we can use this model to predict which team would win the match given the two teams and who won the toss and what season it is.

A simple example for predicting the team that wins the match is shown below. The teams playing the match are Mumbai Indians and Sunrisers Hyderabad where Mumbai Indians win the toss. The model predicts that Mumbai Indians win the match. While going through the data, we can see that the prediction is correct based on the inputs that were given.

Summary / Conclusion (Please add something if needed):

This project was carried out in 4 major portions.

In the first part, we downloaded the data and cleaned it so as to remove irrelevant columns and fill out the null values wherever needed. This step is a crucial one as all the data needs to be pre-processed before we begin any analysis on it.

The second part of our project included basic analysis of the data. To do this, we used PySpark in Jupyter notebook. Here, we found out insights about the players who won the man of the match multiple times. Looking at the graphs plotted above, we can also say that batting first or bowling first does not play a huge role in deciding if the team would win the match. Also, we can see that if a team wins the toss then the likelihood of the team winning the match is around 50%. This tells us that winning the toss may not be a good source for judging the outcome of the match.

The third step in our project was detailed analysis about the players who played in the league so far. A player can either be a good batsman or a good bowler or both. Here, we took into consideration 5 metrics for a good batsman and 5 metrics for a good bowler. For a team to select a player in the auction, the team owner can look into the lists of these metrics and get players who specialize in either one or more of them. To find out the top 5 batsmen, we then aggregated the batting metrics and calculated a final batting weight for each player. We did a similar approach for finding the top 5 bowlers. Each of the 10 metrics shown above are crucial and needs to be taken into consideration while finding the top 5 all-rounder players. Thus, by adding the total batting and bowling scores of each player, we found out the top 5 players who were good in batting as well as bowling.

This is the last part of our project. Here, we used machine learning techniques to build a model to try and predict which team would win the match. For this, we first tried the basic approach by training a model on just the team names, venue and toss decision and winner. Then, to increase the accuracy of the model, we incorporated the weights of the team based on the players in the team as an input to the model. This helped us to increase the accuracy to 88%.

**Future Scope: (Dont know)**

**References:**

https://en.wikipedia.org/wiki/Indian\_Premier\_League#Current\_teams

https://www.kaggle.com/nowke9/ipldata

https://www.formcept.com/blog/360-player-analysis-now-at-your-fingertips-explore-how-formcept-is-transforming-cricket-into-the-new-face-of-big-data/