# Big Data Analysis on Indian Premier League

Suhas Chikkanaravangala Vijayakumar
Mihir Manjrekar
Gautam Gadipudi

## ABSTRACT

Indian Premier League(IPL) is the most viewed cricket league in the world. The Sport analytics helps franchise increase their revenue, buy suitable players, improve player's performance and increase team's quality. The amount of data that is being generated from IPL is enormous, when used in a right way could yield immensely useful information that could help teams stay ahead of the opponents in this extremely competitive league. In this paper ball by ball IPL data is analyzed 1) To create Batting tiers amongst the players, based on their batting performances over the years with the aid of clustering. 2) To Mine most successful batting partnerships over the years. 3) To visualize the attributes over the phases of the game that directly effect the momentum of the game.

## 1. INTRODUCTION

Sports Analysis is game changer in the world of sports, instead of relying on institution and experience, we can now use different stats acquired by monitoring player performance and compare it in different game situations for which a player with particular stats will be ideal. We plan on using different analysis techniques on a game called Cricket, we plan on doing this analysis from data obtained from Indian Premier League(IPL) a professional cricket league with 8 teams from 8 different cities. This analysis can be used for building a team around core players and also it can give insights on features influencing IPL business.

The analysis on player's performance could be used to find his/her role in the match, metrics like bowling and batting are most basic metrics used to analysis player's performance, for example considering a batting metric which include finishing ability, hard-hitting ability, counter attacking ability, etc, such measures can be used to determine the role of player, a player with a good hard hitting ability indicates that the player may be able to recover runs in a match which his team is losing and a player with counter attacking ability can take the pressure off from other players. Based on this data analysis team owners are able to decide what their team lacks in and bid for players which can fill up that gap, in addition to that teams can also analyze other teams players to decide what kind of strategy they will bring in to the game and different roles of a each player.

we plan to combine two different datasets, The first data set consists ball by ball statistics of 756 matches played across all IPL seasons in yaml format[1].The second data set contains Player and club statistics of all seasons in csv format[2]. Our analysis focuses to mine two things, one of them is to mine the most successful batting partnerships, partnerships which have scored more runs and partnerships which have finished the games as this analysis could help IPL clubs buy players who have better understanding in the field which is instrumental for winning big matches. The second data mining technique focuses on clustering players into different Tiers based on their performances, which could help clubs manage their budget in auction.

Along with data mining techniques mentioned above, we also plan to visualise three attributes, runs scored, wickets taken and boundaries scored over 12 IPL seasons, through a time series plot, as this visualization could help improve IPL business, which they could use to check weather these attributes has a direct impact on Television Rating Point(TRP), number of tickets sold in an IPL season. At last we plan to do a pair wise comparison on teams performance in league stage and in playoffs. Performance is calculated based on different attributes like runs scored in power play and in death overs etc.

Section 2 describes the dataset section 3 presents a motivation for this project, and 4 discusses the implementation, 5 discusses our analysis and 7presents the results and describe the current state of the project and what else could be done in the future.

## 2. DATASET

There are primarily 2 datasets, one is based on matches[1] which consists 181,440 instances and the other is based on players[2] which consists of 562 instances. The players dataset initially consisted of basic player information such as age, batting style and bowling skill, it was then updated with new attributes like runs scored and balls faced extracted from matches dataset[1]. Two new datasets partnerships and team progression were extracted from primary sources[1][2]. Partnerships data consists of batting partnership attributes such as runs scored by each

partner, number of balls faced by each partner,strike rate of each partner and total runs scored by the partnership between 2 players. Team progression data consists of progression of teams score at each over in the match, it also includes other attributes like home team, away team and season. The Partnerships data consists of 9925 instances which can be clustered using total runs and strike rate of the partnership from which player-pairs can be clustered into different tiers. Similarly team progression dataset which consists 28654 instances can be clustered based on teams score and overs.

# 3. MOTIVATION

Application of Data Analysis in Sports growing at a rapid speed. Sport Analysis is widely used in many sports industries like Basketball, Soccer, football, Cricket etc. There have been many advances of data analysis in sports industries over the years, players activities are tracked and monitoring to generate useful data to analysis players performance, for example in basketball no of basket scored, no of three pointers scored etc can be used to evaluate players performance and can detect area's that the player lacks, similarly a same type of analysis on teams can be done to predict if the teams going to win or no. Various Data mining techniques can be applied to such data to give and more accurate summary in players/teams performance. We aim to apply these data mining technique for sports analysis of a sport called cricket, we will be conducting this analysis on cricket league in India called the Indian premier league(IPL).

# 4. IMPLEMENTATION

## 4.1 Clustering

Clustering is done on batting statistics per match for every player for all all seasons. The attributes used to cluster are:

- runs
- strike_rate
- sixes
- fours

### 4.1.1 Pre-processing

Since the dataset has ball-by-ball data, we first need to extract the above batting stats for each player per match.

### 4.1.2 Normalization

Then, we normalize the attributes so that all the fields are in the range of 0 and 1. The player with highest value of the attribute will have a normalized value of 1, and the player with the least value of the attribute will have a normalized value of 0. The formula used to calculate the normalized value for each field is:

$$A_{i\_normalized} = \frac{A_i - A_{min}}{A_{max} - A_{min}} \quad (1)$$

where, $A_i$ is the attribute value for a player, $A_{max}$ is the maximum value for that attribute and $A_{min}$ is the minimum value for that attribute.

This way, we normalized the values for $runs$, $strike\_rate$, $fours$ and $sixes$.

### 4.1.3 Select initial centroids

For a particular value of k, select a random sample of k items from the data to be clustered. These will be the centroid points for the first iteration of the algorithm.

### 4.1.4 Assign points to clusters

Now, for each item in the dataset, calculate the distance to every other centroid and assign that item to the closest centroid. In calculating the distance, use the Euclidean distance in four dimension:

$$\text{distance} = \sqrt{(runs_i - runs_c)^2 + (strike\_rate_i - strike\_rate_c)^2 + (sixes_i - sixes_c)^2 (fours_i - fours_c)^2} \quad (2)$$

### 4.1.5 Recalculate the centroids

Calculate the new centroid for each cluster by taking a mean of all the data points / items under that cluster.

$$\text{centroid}_{new} = (\textstyle\sum_{i=1}^{n} runs_i/n, \sum_{i=1}^{n} strike\_rate_i/n, \sum_{i=1}^{n} sixes_i/n, \sum_{i=1}^{n} fours_i/n) \quad (3)$$

where, $n$ is the number of elements in the cluster.

### 4.1.6 Further iterations

Now, do section 4.1.4 and section 4.1.5 one after the other for a given number of iterations or until the centroids converge.

### 4.1.7 Selecting the ideal k

For a range of k, do section 4.1.3, section 4.1.4, section 4.1.5 and section 4.1.6 and calculate the Sum of Squared Errors (SSE) for each k using:

$$\text{SSE} = \textstyle\sum_{i=1}^{k} \sum_{j=1}^{n} (runs_i - runs_j)^2 + (strike\_rate_i - strike\_rate_j)^2 + (sixes_i - sixes_j)^2 + (fours_i - fours_j)^2 \quad (4)$$

where, $k$ is the number of clusters and $n$ is the number of elements in that cluster.

Now plot k vs SSE graph for all the range of k values calculated above.

**Select the k that is at the tip of the elbow**.

### 4.1.8 Final clusters

Finally, do section 4.1.3, section 4.1.4, section 4.1.5 and section 4.1.6 with the k value from section 4.1.7.

## 4.2 Association

Item set Mining is performed on partnership data set which consists of players and venue.The algorithm is pruned at 2 steps to reduce the time-complexity.

### 4.2.1 Pre-processing

Since the dataset has ball-by-ball data, we are extracting partnerships and its attributes such as strike rate, runs scored etc, along with these attributes we are also extracting common attributes like venues, opponent etc.

### 4.2.2 filter Partnerships

At first the we filter those partnerships with runs less than 30 and the data is fed for 3 levels of iteration with min support($min_sup$) set to 15 This filtered data is considered to be at $\text{level}_0$.

### 4.2.3 $Level_i$ candidate generation

candidates for $level_i$ is generated from the data of previous level, this step provides all possible combinations from $\text{level}_{i-1}$.

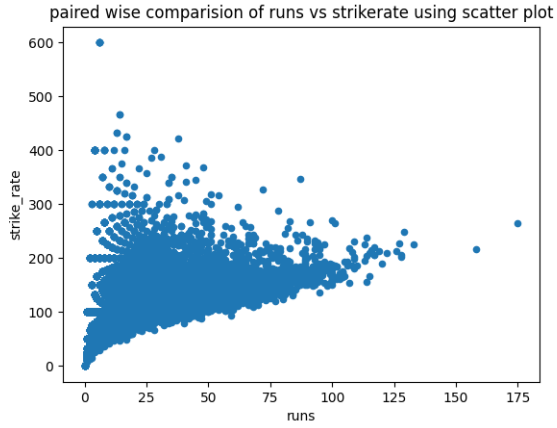


**Figure 1: scatter plot of runs scored vs strike rate**

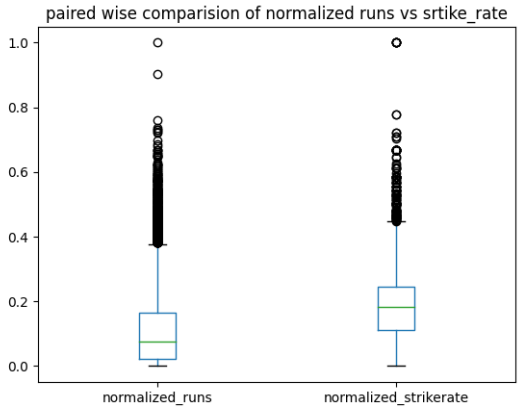


**Figure 2: pairwise comparison of runs scored vs strike rate**

#### 4.2.4 *Prune invalid candidates at level$_i$*

Since the section 4.2.3 generates every possible candidate, some of these items do not occur in the real partnerships data. These candidates are pruned.

#### 4.2.5 *Calculate the frequency of items at level$_i$*

The frequency remaining valid candidates are calculated using hash map.

#### 4.2.6 *Prune candidates with support less than $min_sup$*

Now prune all the candidates which fails to meet minimum support. save the remaining candidates as level$_i$ items, stop the algorithm if -> $i = 3$., else -> go to 4.2.3.

## 5. ANALYSIS

### 5.1 Attribute Analysis

Prior to the implementation of any data mining technique, it is important to identify principal attributes which can provide meaningful information. Here's the Fig1 showing runs scored vs batting strike rate of all the batting innings of all the players across 12 IPL seasons. The graph 1 shows that the as the number of runs increase the relationship between strike rate and runs scored becomes highly linear. Herein the Fig. 2, from the box plot of normalized runs scored vs normalized strike rate it can be seen that mean of strikerate is slightly higher than that of runs scored and shows similarity after third quartile. Another Attribute pair which shows high variance is number of fours and number of sixes, it can be seen in Fig.7 that the number of fours are higher than sixes through out. All of these four attributes are chosen for clustering.

### 5.2 Clustering Analysis

k-means is one of the popular data mining algorithm, the algorithms creates k number of clusters and each data point is assigned to the cluster with nearest mean(centroid). The algorithm requires number of clusters to be specified in advance. There are various techniques to determine the number of clusters [3]. one of the most popular technique is identifying elbow point in k vs sum of squared(SSE) plot, the plot can be obtained by implementing k-means clustering iteratively by changing the k value, SSE is calculated at the end of each iteration. The plot here in Fig.3 shows sudden drop in SSE at initial k values and starts to saturates between 4 and 6 forming a elbow shaped graph. k-means clustering with k=5 was used to cluster the batting innings across all 12 seasons. A batting inning consists of runs scored, strikerate,sixes,fours ,strikerate indicates the rate at which the runs were scored, sixes represents number of sixes hit during the inning and fours represent the number of fours hit during the inning. All these attributes together helps in determining the innings quality. All the attributes were normalized prior to clustering. euclidean distance was used as a distance metric for calculating the distance between the data points. After clustering evaluation of cluster quality becomes important, while there are many techniques for determining cluster quality Silhouette coefficient is the most popular.silhouette co-efficient is obtained based on similarity and dissimilarity of each point by calculating the average distance from each point to every other point in the same cluster as well as average distance from each point to every other point in other clusters,it ranges from -1 to 1. Silhouette co-efficient was used in our implementation to determine quality of the clusters formed. The silhouette co-efficient was 0.41, the good clusters usually have silhouette co-efficient ranging from 0.5-1, considering how dense our data points are 0.41 is fairly a good result. The figure 11 shows the data points at different clusters.

### 5.3 Association Mining Analysis

Apriori is an Association Mining Algorithm, the goal of association mining is to find meaningful relations between data items in dataset. We have used Association Mining on player partnership and venue to find meaningful relations between them. In Cricket there two batsmen playing, one of them is a striker and the other is a non-strike, this is known as partnership. The partnership comes to an end when one of the batsmen is dismissed. Batting partnership is very crucial in cricket, it helps build the foundation of your game, it helps you concentrate and think of your next move freely. In bating partnership trust and communication

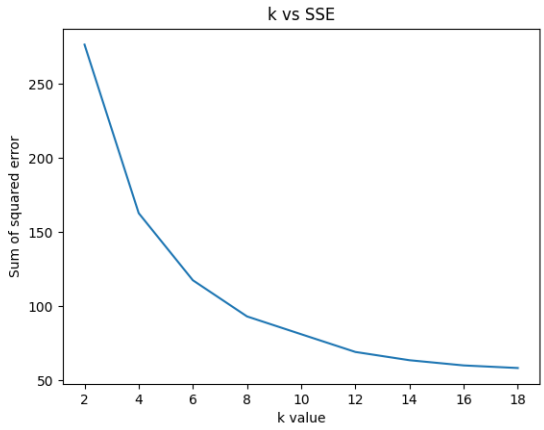


**Figure 3: K vs sum of squared errors**

is the key, the two players need to think fast and make quick decision sometimes, not all players are like that towards each other, some players may perform better in a presence of a player they maybe trust more or players with whom he has good communication with. In cricket even the venue you play on holds importance, for example players who play at their home team venue usually perform better than the away team cause the home team is familiar with the venue and have practices many times there. We have extracted some of this partnerships of players and also added venue to check which partnership can be beneficial in which venue.

We divided them in three levels:

- Level 1: At this level we get players who regardless of their partner and venue are cap to score 30+ runs in a match. Shown in Fig 4
- Level 2: At this level we get two types of relations , one is player-player and the other is player-venue. player-player pairs in this level show player partnership who are able to score 30+ runs regardless of which venue they are playing on, while player-venue pair gives us information about player which regardless of their partner are able to score 30+ runs in that venue, this shows us in what venue the player is comfortable playing in. Shown in Fig 5
- Level 3: At this level we get relations between player partnership and venue. This level gives us an item-set containing two players and venue where they seem to show good performance. Shown in Fig 6

## 5.4 Visualization

The number of runs scored or number of runs conceded during the particular phase of game is crucial in T-20(Twenty Twenty) cricket, any abrupt changes in these attributes indicates the shift in momentum of the game. The phase in cricket is referred to the overs. The Visualization of these attributes could help teams identify where they had lost/gained momentum during the game. Since franchises change players once in every three years the following points should be considered to obtain a meaningful visualizations 1) these values should only be

level1

| | count Int32 | item1 String |
|---|---|---|
| 21 | 92 | "Punjab Cricket Association I |
| 22 | 91 | "AT Rayudu" |
| 23 | 84 | "SR Watson" |
| 24 | 84 | "Maharashtra Cricket Associat |
| 25 | 82 | "BB McCullum" |
| 26 | 79 | "PA Patel" |
| 27 | 77 | "MK Pandey" |
| 28 | 76 | "KA Pollard" |
| 29 | 73 | "Yuvraj Singh" |
| 30 | 73 | "YK Pathan" |

**Figure 4: Level 1 of Association Mining**

level2

| | count Int32 | item1 String | item2 String |
|---|---|---|---|
| 1 | 34 | "AB de Villiers" | "M Chinnaswamy Stadium" |
| 2 | 33 | "CH Gayle" | "M Chinnaswamy Stadium" |
| 3 | 33 | "CH Gayle" | "V Kohli" |
| 4 | 33 | "MA Chidambaram Stadium, Chep | "MS Dhoni" |
| 5 | 31 | "KA Pollard" | "Wankhede Stadium" |
| 6 | 28 | "Rajiv Gandhi International S | "S Dhawan" |
| 7 | 27 | "AB de Villiers" | "V Kohli" |
| 8 | 27 | "DA Warner" | "S Dhawan" |
| 9 | 26 | "G Gambhir" | "RV Uthappa" |
| 10 | 23 | "MS Dhoni" | "SK Raina" |
| 11 | 22 | "M Vijay" | "MA Chidambaram Stadium, Chep |

**Figure 5: Level 2 of Association Mining**

level3

| | count Int32 | item1 String | item2 String | item3 String |
|---|---|---|---|---|
| 1 | 14 | "CH Gayle" | "M Chinnaswamy Stadium" | "V Kohli" |
| 2 | 14 | "Eden Gardens" | "G Gambhir" | "RV Uthappa" |
| 3 | 13 | "KA Pollard" | "RG Sharma" | "Wankhede Stadium" |
| 4 | 12 | "AB de Villiers" | "M Chinnaswamy Stadium" | "V Kohli" |
| 5 | 8 | "M Vijay" | "MA Chidambaram Stadium, Chep | "MEK Hussey" |
| 6 | 7 | "Feroz Shah Kotla" | "RR Pant" | "SS Iyer" |
| 7 | 7 | "DA Warner" | "Rajiv Gandhi International S | "S Dhawan" |
| 8 | 7 | "MA Chidambaram Stadium, Chep | "MS Dhoni" | "RA Jadeja" |
| 9 | 7 | "MA Chidambaram Stadium, Chep | "MEK Hussey" | "SK Raina" |
| 10 | 7 | "AT Rayudu" | "RG Sharma" | "Wankhede Stadium" |
| 11 | 7 | "AM Rahane" | "R Dravid" | "Sawai Mansingh Stadium" |
| 12 | 6 | "F du Plessis" | "MA Chidambaram Stadium, Chep | "SK Raina" |
| 13 | 6 | "MA Chidambaram Stadium, Chep | "MS Dhoni" | "SK Raina" |

**Figure 6: Level 3 of Association Mining**

visualized over a range of season/seasons not across all seasons. 2) these values should be plotted during the particular phase of the game (between overs(1-20)). 3) these values should be based on type of innings(1 or 2), since the chasing team often tends to score at the rate that is required to get the target set by opponents. By taking all these points into consideration the visualization was implemented in a interactive way. To calculate average runs scored by teams between any overs in a particular innings and across the specified seasons. we first match the overs and seasons given as the input, then we group the data by batting teams and compute average runs scored during the selected overs . Herein Fig. 8, the plot shows the average runs scored by IPL teams during the powerplay across latest two seasons(2018, 2019). Powerplay is played between overs 1-6 of an innings, where only two fielders are allowed outside the inner ring(30-yards). Teams often sees this situation as a scoring opportunity and tries to accelerate the scoring rate. This plot in Fig.8 helps teams identify their the scoring rate during the powerplay and also compare their average with opponents ahead of the game and make necessary changes to the team/tactics prior to the game. This is one example of how we can utilize the visualization, similarly teams can choose overs between (6-16) if they feel they are losing momentum in middle overs or overs between(16-20) if they feel they are losing momentum in the death overs. At the end of the season teams can identify the phases in which they lack behind other teams buy players who tends to score big during these phase of the game. It Can be seen from the graph in Fig.8 that some teams tend to score really low during the first two overs and scoring anything below 6 runs per over is below par and it is considered as a bad over from batting perspective.Visualization of runs conceded is also equally important since the couple of expensive overs during the innings could change the game entirely upside down. If the bowling team is defending a below par or a just par target, teams cannot offer to concede too many runs. This Fig. 9 plot was implemented similar to the runs scored plot in Fig.8, but the runs is grouped by bowling team instead of batting team. Fig.9 shows the average run conceded by the teams during the death overs(between 16-20) in last two seasons(2018,2019). Teams tend to double their scoring rate in these overs, as a bowling unit during this phase conceding less runs could help them chase a much lesser target than the projected score while bowling first. While defending these overs decides the outcome of the game and is very crucial. It can be seen from the plot 9 that conceding anything less than 8 runs per over during death overs is considered to be a good over from a bowling perspective. Here in Fig.10, shows the number of sixes hit across all 12 IPL seasons. Number of sixes is the entertainment factor of T20 Cricket, the crowd tends to involve more in the game when batsmen hit a six. This Fig.10 can be compared with number of television viewers of IPL across all seasons.

## 6. RESULTS

This section briefly discusses the outcome of data mining techniques implemented. The k-means clustering was implemented to cluster the the batting innings with the k = 5. The 3D plot in Fig.11 shows the innings was divided into 5 different clusters. Once the cluster is assigned to

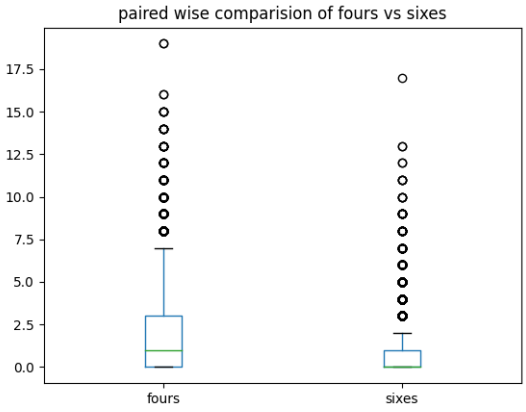


**Figure 7: Pairwise comparison of number of fours vs sixes**

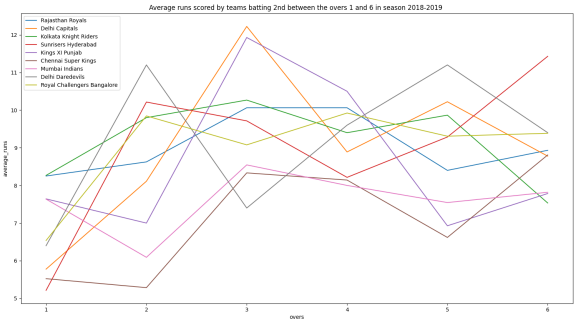


**Figure 8: Runs Scored by teams during the Power play in last two season**

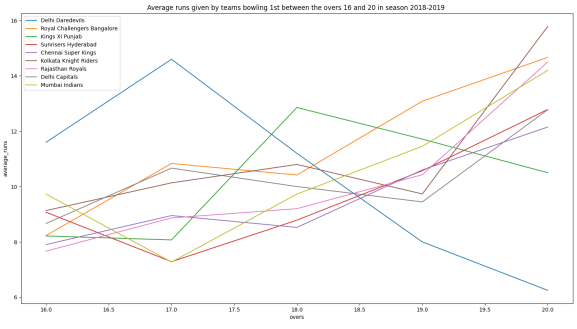


**Figure 9: Runs Conceded by teams during the death overs in last two season**

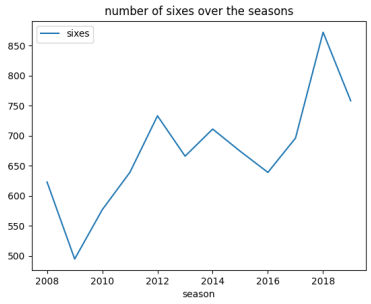


**Figure 10: sixes hit over the 12 IPL seasons**

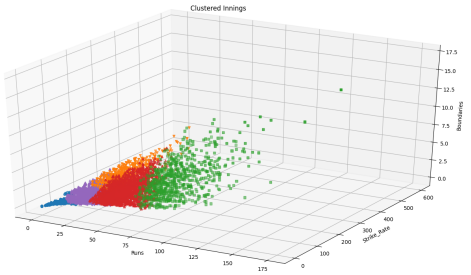


**Figure 11: Clustered batting innings**

each innings, the clusters are then ranked from 1 to 5 based on the average value of attributes in this order 1)runs 2)strikerate 3)sixes 4)fours. Now each rank will be associated with points called batting points. The points are assigned based on ranks, cluster with rank-1 is worth 10 points, cluster with rank-2 is worth 8 points, similarly cluster with rank-5 gets 2 point. every rank is separated by 2 points difference. A player called SK Raina has played 187 innings across all season spread over all 5 clusters, here the figure 12 is an example of how the batting points for a particular player is calculated. As shown in the figure 12 SK has played 21 innings which was ranked-1, so the number of innings in each cluster is multiplied by cluster worth ( for ranked-1 cluster -21*10=210 points). The total points obtained by SK Raina is 1068. Batting Tiers(1-5) are created based on these points. The figure 13 shows the players with highest batting points and their batting tier. The second data mining technique implemented was item-set Mining on batting partnerships using apriori, The Level 3 of item set mining provides player-player-venue item sets, These items represents most successful player pairs at a particular venue. These pairs are highly capable of building 30+ run partnership at a given venue. Here the figure 6 shows the most frequent partnerships across all seasons.

## 7. FUTURE WORK AND CONCLUSION

Players have been divided into different tiers based on batting points obtained with the help of clustering. These tiers help IPL franchises buy and distinguish batsmen of all quality. Item set mining gave the most frequent batting

| | rank | innings | multiplier | total |
|---|---|---|---|---|
| 2 | 1 | 21 | 10 | 210 |
| 3 | 2 | 61 | 8 | 488 |
| 4 | 3 | 10 | 6 | 60 |
| 5 | 4 | 60 | 4 | 240 |
| 6 | 5 | 35 | 2 | 70 |
| 7 | | | | **1068** |
| 8 | | | | |

**Figure 12: Calculation of batting points of a player**

players

| | name String | batting_points Int32 | batting_tier Int32 |
|---|---|---|---|
| 1 | "SK Raina" | 1068 | 1 |
| 2 | "V Kohli" | 1012 | 1 |
| 3 | "RG Sharma" | 970 | 1 |
| 4 | "S Dhawan" | 940 | 1 |
| 5 | "RV Uthappa" | 936 | 1 |
| 6 | "MS Dhoni" | 912 | 1 |
| 7 | "G Gambhir" | 856 | 1 |
| 8 | "KD Karthik" | 828 | 2 |
| 9 | "DA Warner" | 814 | 2 |
| 10 | "AB de Villiers" | 808 | 2 |

**Figure 13: player's batting points with tier**

player pair with 30+ run partnership. This could help teams buy players in pair, who could have immediate an impact while playing together. Efforts has been made to visualize the attributes like runs scored, runs conceded, number of sixes that would shift the momentum of the game. Visualization was implemented in such way that the teams can choose the particular phases during the game to analyze their weaknesses and buy/substitute players who can perform during those phases. Apart from runs conceded all the analysis conducted were based on either batsmen or batting. Similar to the clustering performed here to obtain batting points and tiers, there is scope for performing clustering on bowling attributes like wickets taken, runs conceded in an over, bowling strike rate and average to create bowling tiers.

## 8. REFERENCES

[1] Indian Premier League data. `https://cricsheet.org/downloads/`. Accessed: 2020-06-08.

[2] IPL Player stats. `https://data.world/raghu543/ipl-data-till-2016-set-of-csv-files`. Accessed: 2020-06-08.

[3] X. Wang, Y. Jiao, and S. Fei. Estimation of clusters number and initial centers of k-means algorithm using watershed method. In *2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)*, pages 505–508, 2015.