Business Understanding Document
IPL Analytics

Abhinav Chawla IMT2013002
Aditya Naidu IMT2013003
Shivam Kumar IMT2013042
Siddartha Padhi IMT2013043

16 October 2016

1 Problem Area Description

Coaches and scouts since long have been using intuition and gut feeling to select players in the squad and to decide the strength and the weakness of a player. If we are able to analyse the sports data, coaches will have analytics computed by us supporting or contradicting this intuition which will help in better understanding of the qualities of a player. As a big cricketing nation, we have chosen to take cricket as the sport to analyse upon.

Our project broadly aims at analysing the previous IPL seasons to find out various weaknesses, strengths and abilities of various players. We aim to find insights about which players have a mutual understanding with each others, which player is strong and weak against which type of bowler, how does a player perform under pressure etc.

2 Current Existing Solution

Most of the high level teams have hired a data analytics team who help them in this task. We want to emulate them and if possible, try to find new correlations.

3 Business Objectives Definition

Our project aims at coming up with an analytics driven solution which assists the team management with player selection and devising game play strategies based on several factors such as individual and team performance of a player against the opponent team and/or players. The business objectives we hope to achieve with our solution are implementation of better player selection methods than just intuition/experience and amplification of team's success rate.

4 Business Success Criteria

We, as the end output, want to predict which squad should be chosen by a specific team for a particular match, decide the batting order and decide which bowler to bowl to exploit the opponents's weakness and provide the best results.

5 Situation Assessment

We will describe the dataset in this section. We are provided with 578 Excel files, one for each IPL match since 2008. Each excel file includes metadata as description of the matchfor first 20 rows which includes the teams contesting the match, date, venue, toss result, umpire name, man of the match, the winner of the match and other extra details. The rows after that include the actual data. Each row consists information of each delivery which includes bowler name, batsman on strike, batsman off strike and runs scored on that ball.

6 Resource Inventory

We will be using data set provided by "cricsheet.org" for a ball by ball detail of every IPL match conducted. We would be scraping data from "3rd party websites to gather basic information about playing styles of every player. We have a competent team of 4 members with the required skill set to carry out the analysis. We have the required computing power i.e. Personal Laptops to carry out the data mining tasks.

7 Requirements, Assumptions and Constraints

There are no restrictions in legal or economic restrictions. The dataset from "cricsheet.org" is publicly available and free to use. Our main assumption is the validity of data. We are assuming that the data is correct and up to date. The data is limited to only IPL matches. A players performance is judged only by the way he performs in these matches. Performance in other formats of cricket matches such as International T20, ODI and Test Cricket are not being considered.

8 Risks and contingencies

The data set is limited to only IPL matches. We have not considered a players performance in other formats of the game which might have an important role in analysing a players performance accurately.

9 Cost/Benefit Analysis

There is no economic factor associated with this project. The data is publicly available and free to use. The softwares (Tableau) being used have an academic liscence and are free to use for this project. Out insights can help coaches form the optimum squad to play against a team and increase their chances of winning the match.

10 Project Plan

Phase	Time	Duration
Business Understanding	1 week	9/10/2016 - 16/10/2016
Data Understanding	1 week	17/10/2016 - 24/10/2016
Data Preparation	1 week	25/10/2016 - 1/11/2016
Data Modeling	2 weeks	2/11/2016 - 16/11/2016
Evaluation	1 week	17/11/2016 - 24/11/2016
Deployment	1 week	25/11/2016 - 2/12/2016

Cricket Insights

1. Venue vs Elected To

Venue vs Batting Order Victory

Venue Vs Batting Order Victory-Season-Wise

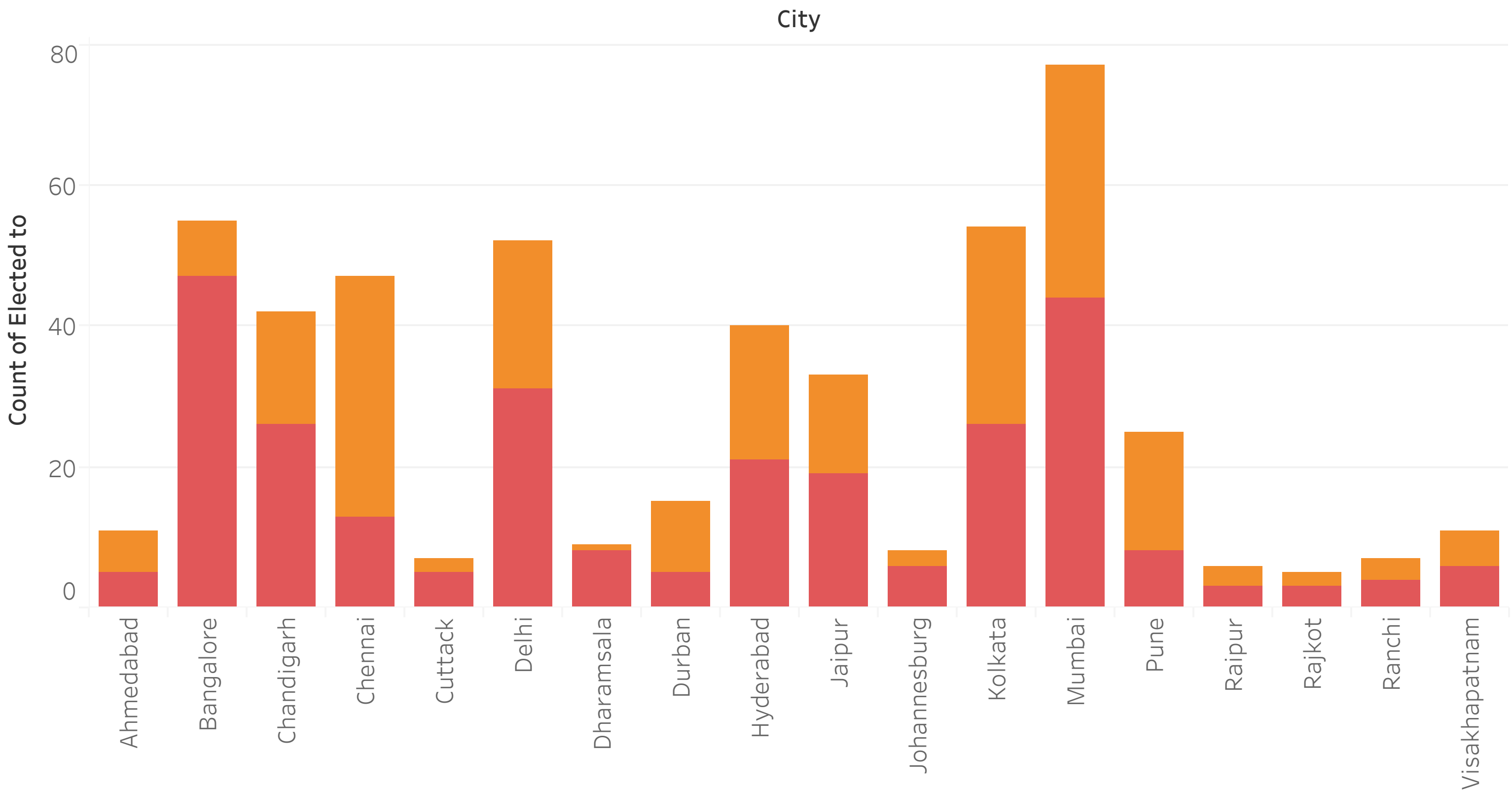
RCB winning partnership

Team vs Opening Partnership

Average Run rate vs Over

Batting Style Wise Average Against Bowling Style

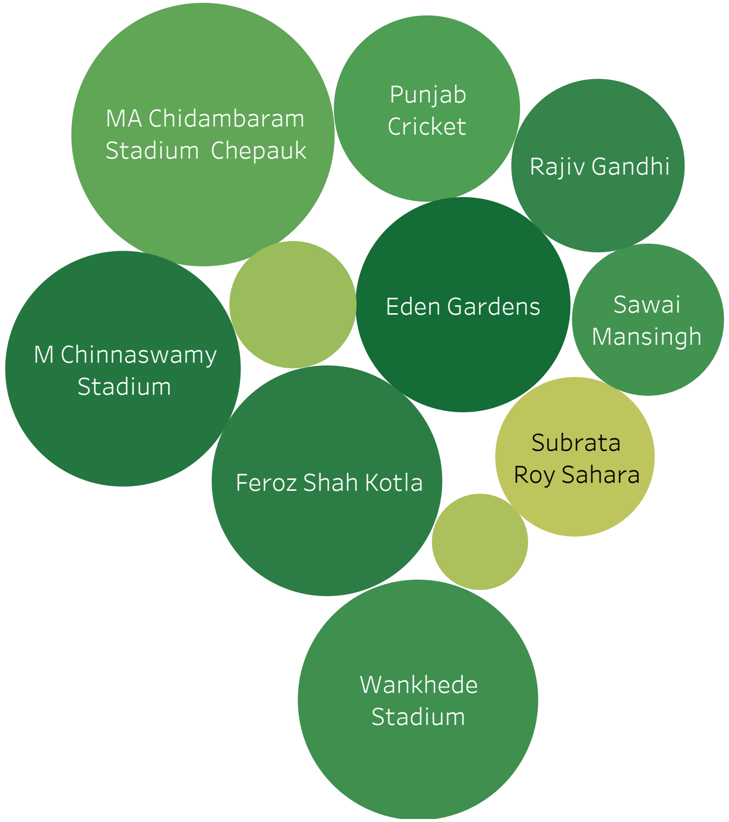
Batting Style Wise Strike Rate Vs Bo..



This is a graph depicting choice of toss winner for every venue. Teams tend to choose to field first in Bangalore and tend to choose to bat first in Chennai.

Cricket Insights

- 1. Venue vs Elected To
- Venue vs Batting Order Victory
- Venue Vs Batting Order Victory-Season-Wise
- RCB winning partnership
- Team vs Opening Partnership
- Average Run rate vs Over
- Batting Style Wise Average Against Bowling Style
- Batting Style Wise Strike Rate Vs Bo..



Batting_second_wins

0.00 34.00

This is a bubble graph where a larger size describes batting first victory while a darker colour describes batting second victory.

Cricket Insights

1. Venue vs Elected To

Venue vs Batting Order Victory

Venue Vs Batting Order Victory-Season-Wise

RCB winning partnership

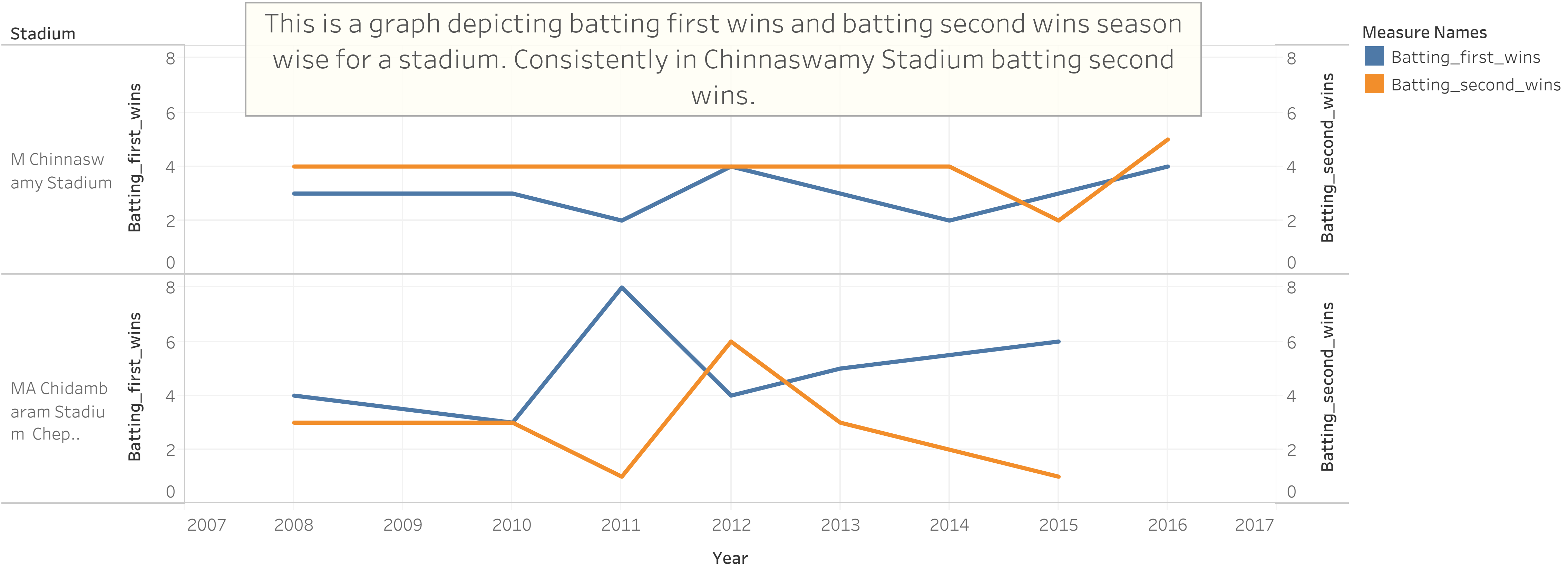
Team vs Opening Partnership

Average Run rate vs Over

Batting Style Wise Average Against Bowling Style

Batting Style Wise Strike Rate Vs Bo..

This is a graph depicting batting first wins and batting second wins season wise for a stadium. Consistently in Chinnaswamy Stadium batting second wins.



Cricket Insights

1. Venue vs Elected To

Venue vs Batting Order Victory

Venue Vs Batting Order Victory-Season-Wise

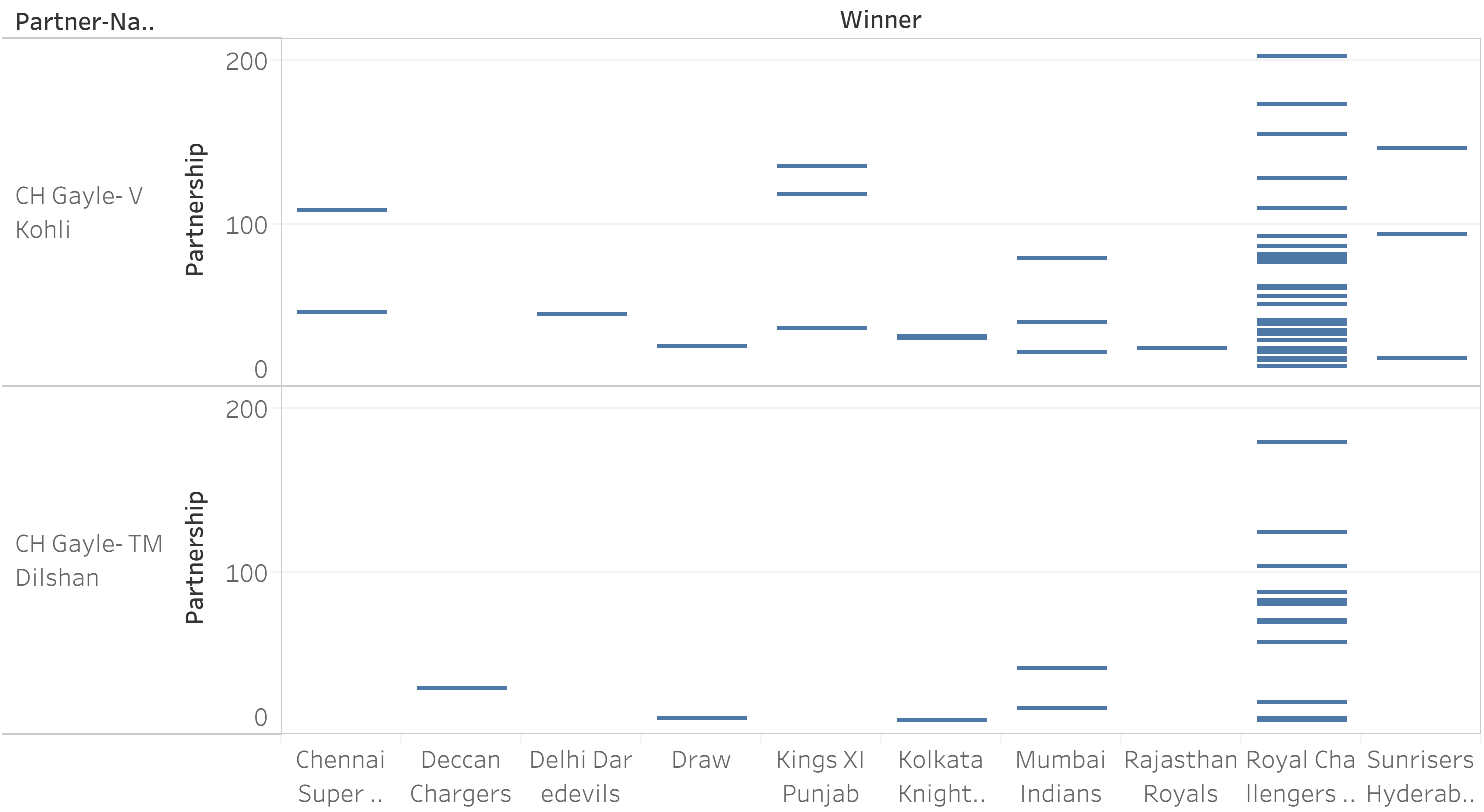
RCB winning partnership

Team vs Opening Partnership

Average Run rate vs Over

Batting Style Wise Average Against Bowling Style

Batting Style Wise Strike Rate Vs Bo..



This is a chart depiciting partnership and winning team. The winning mantra for RCB is for Gayle and Dilshan to score more than 50 partnership.

Cricket Insights

1. V enu e v..

Venue vs Batting Order Victory

Venue Vs Batting Order Victory-Season-Wise

RCB winning partnership

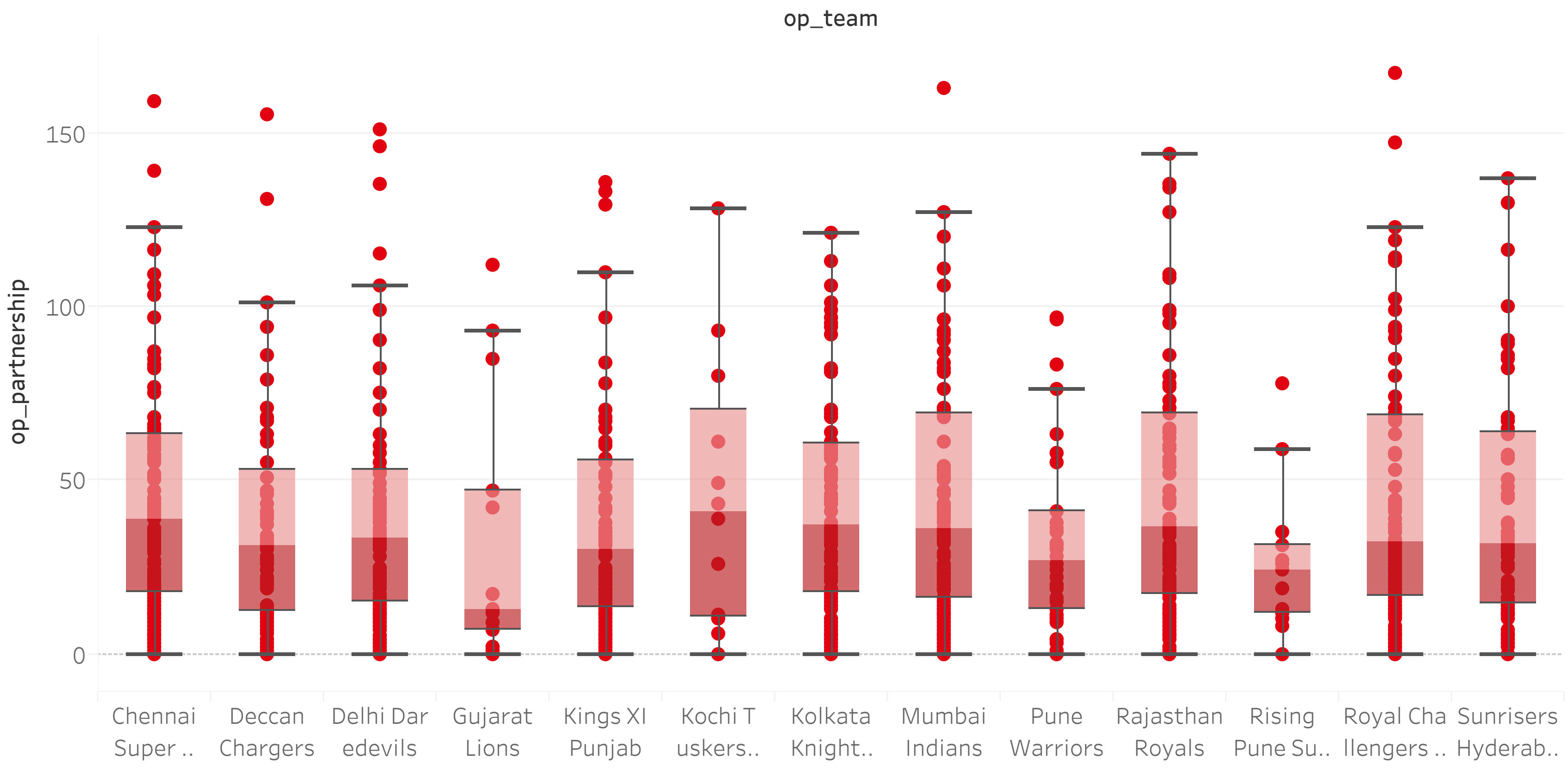
Team vs Opening Partnership

Average Run rate vs Over

Batting Style Wise Average Against Bowling Style

Batting Style Wise Strike Rate Vs Bowling Style

Venue Wise PErfo..



This is a box plot depicting opening partnerships for every team. Most of the IPL winners have an excellent opening partnership.

Cricket Insights

Ven ue v s B..

Venue Vs Batting Order Victory-Season-Wise

RCB winning partnership

Team vs Opening Partnership

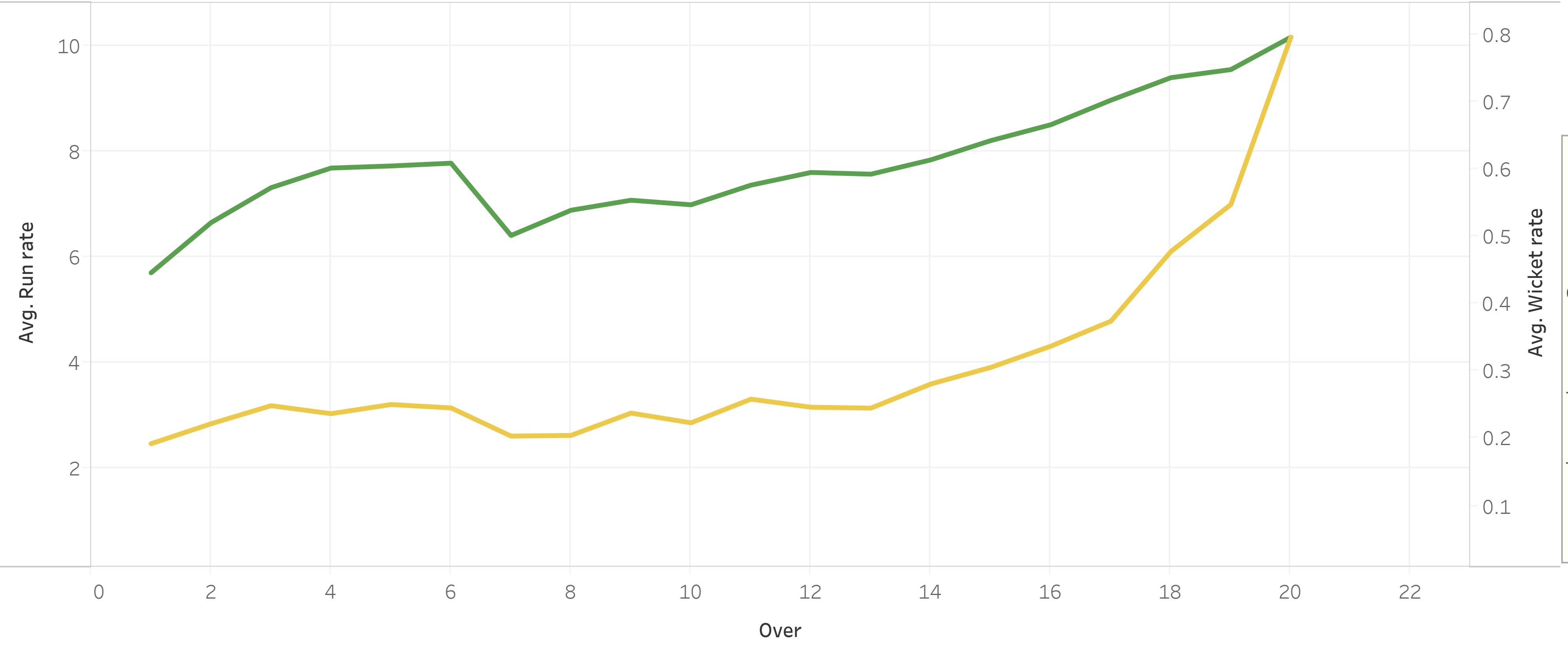
Average Run rate vs Over

Batting Style Wise Average Against Bowling Style

Batting Style Wise Strike Rate Vs Bowling Style

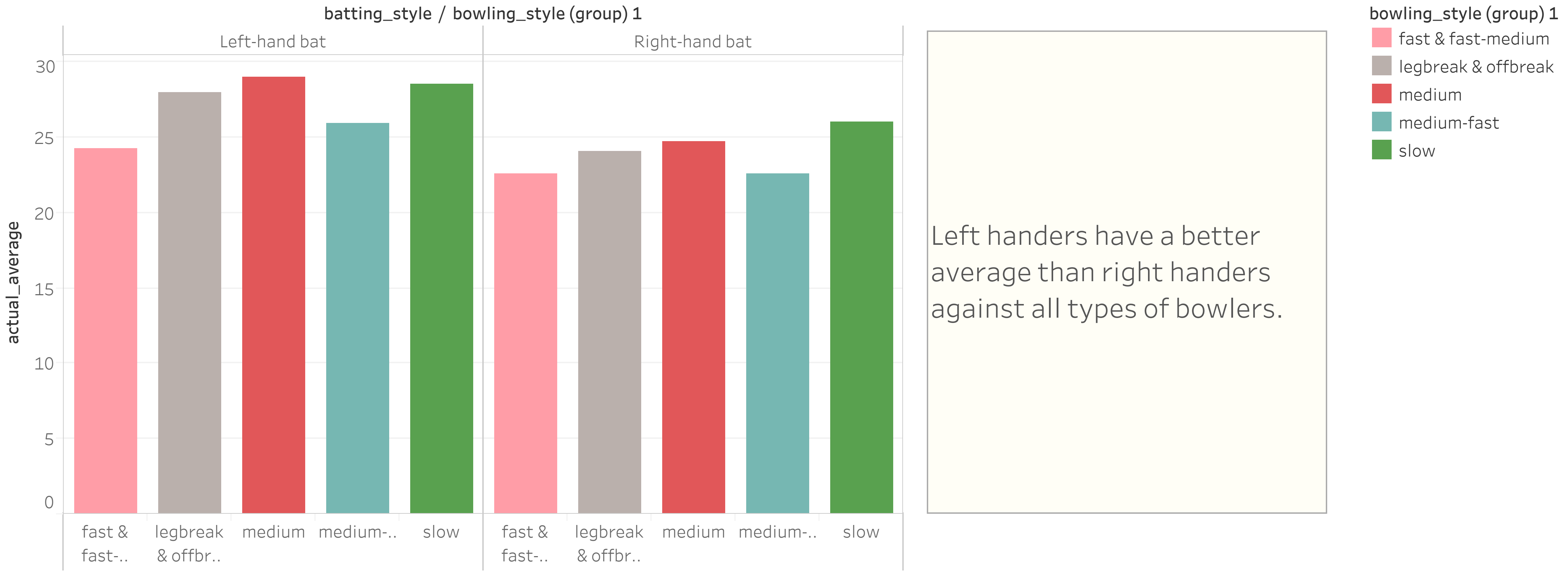
Venue Wise PErformance of KKR

Bowlin g Style Wise ..



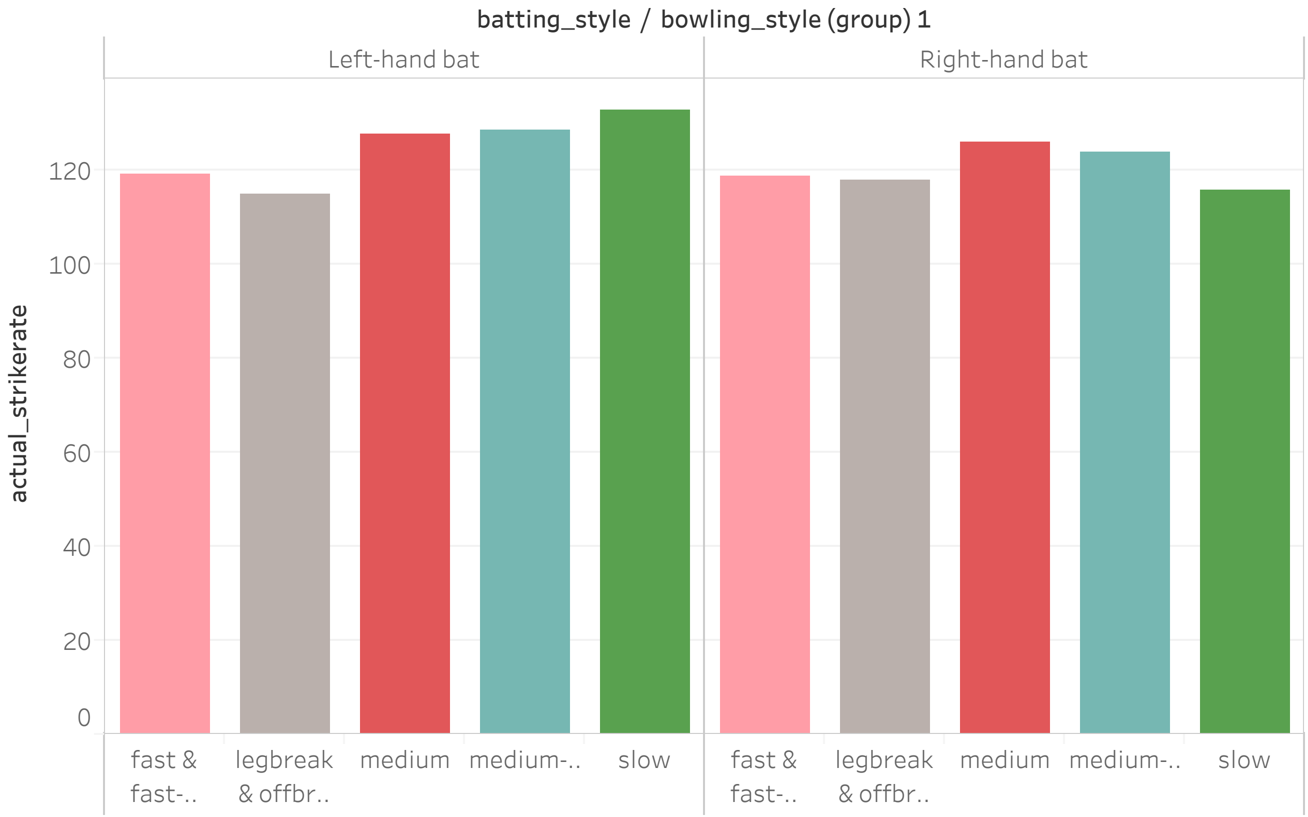
The graph depictc average runs and wickets taken per over. The average run rate, wicket taken tends to decrease at the end of power play and begin to increase toward the end of the innings.

Cricket Insights



Left handers have a better average than right handers against all types of bowlers.

Cricket Insights



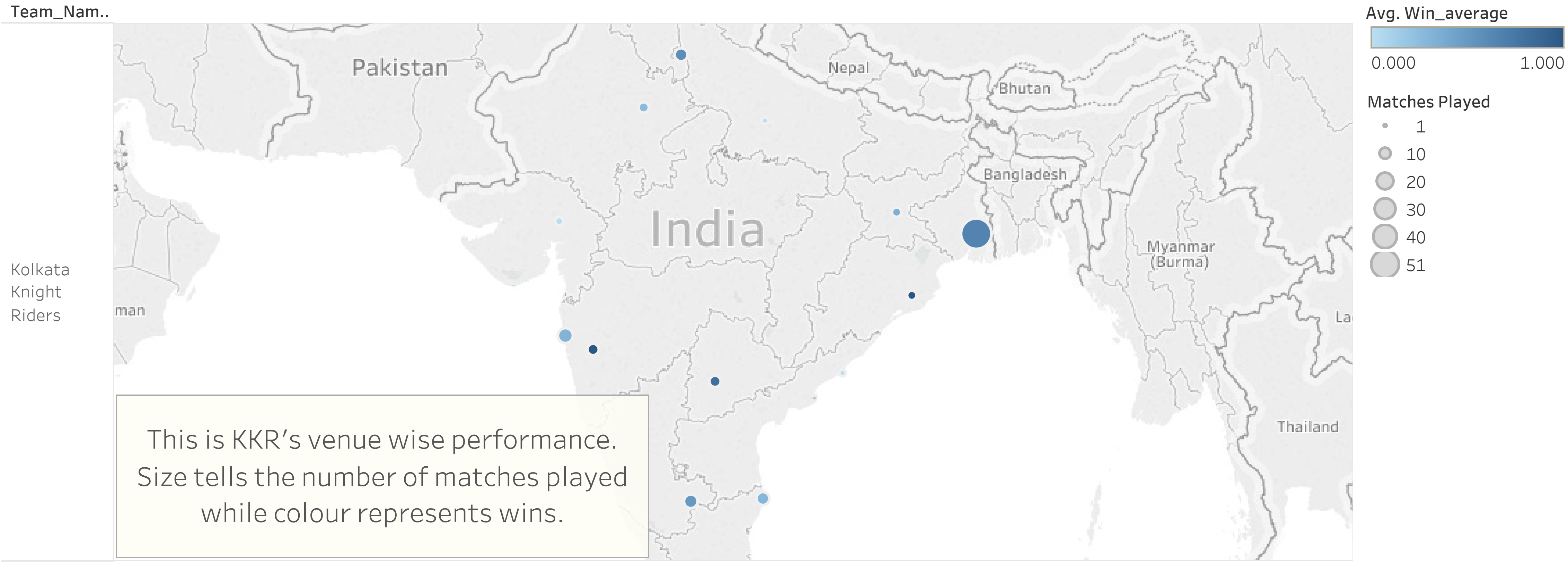
Left handers have a better strike rate than right handers against all types of bowlers.

bowling_style (group) 1

- 🟥 fast & fast-medium
- ▨ legbreak & offbreak
- 🟥 medium
- 🟦 medium-fast
- 🟩 slow

Cricket Insights

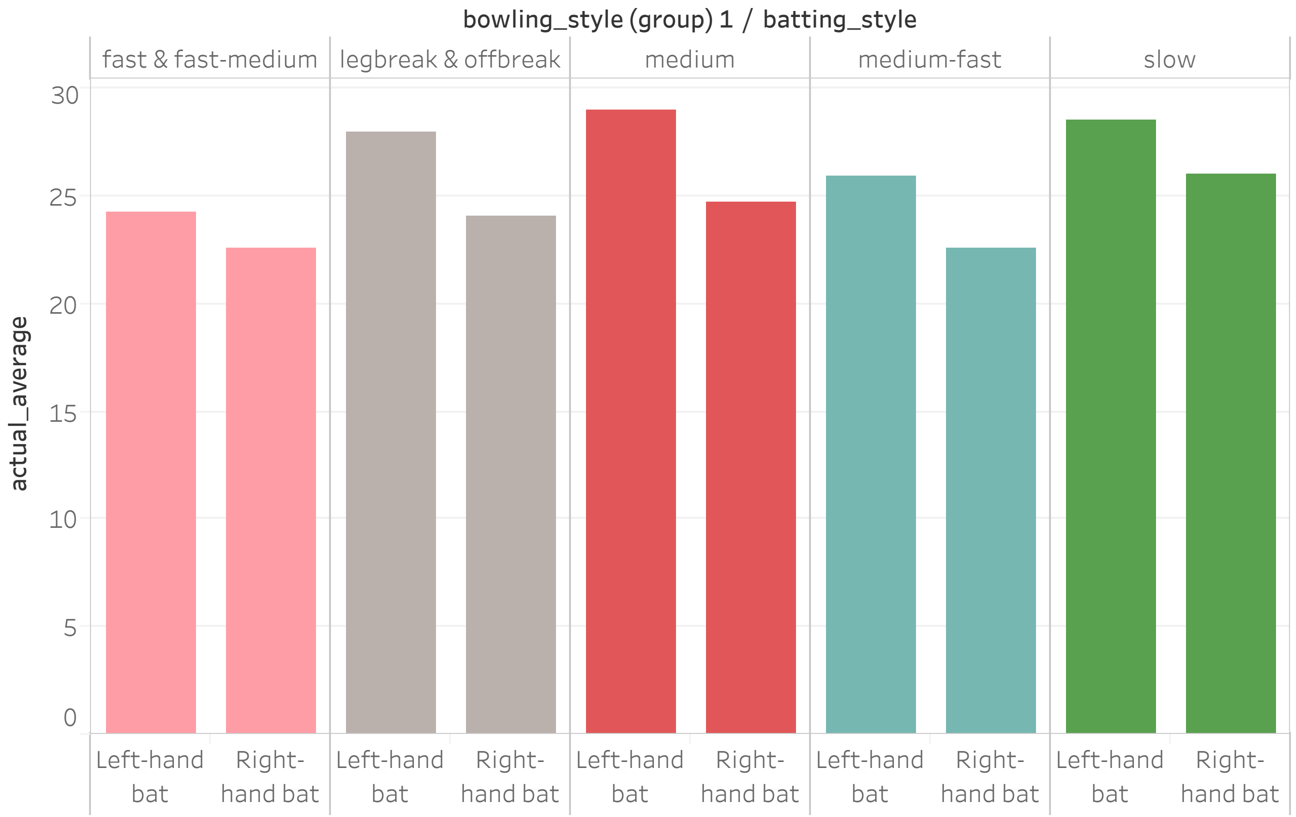
Team vs Opening Partnership | Average Run rate vs Over | Batting Style Wise Average Against Bowling Style | Batting Style Wise Strike Rate Vs Bowling Style | Venue Wise PErformance of KKR | Bowling Style Wise Average Against Bating Style | Bowling Style Wise Strike Rate Vs Batting Style | Average Runs Vs Strike Rate



This is KKR’s venue wise performance. Size tells the number of matches played while colour represents wins.

Cricket Insights

- Team vs Opening Partnership
- Average Run rate vs Over
- Batting Style Wise Average Against Bowling Style
- Batting Style Wise Strike Rate Vs Bowling Style
- Venue Wise PErformance of KKR
- **Bowling Style Wise Average Against Bating Style**
- Bowling Style Wise Strike Rate Vs Batting Style
- Average Runs Vs Strike Rate



This shows how Left-handed and Right-handed batsmen perform against various bowlers by comparing average of the batsmen. From the graph it can be concluded that Right-handed batsmen fare better against Legbreak and Offbreak bowlers only.

Cricket Insights

Team vs Opening Partnership

Average Run rate vs Over

Batting Style Wise Average Against Bowling Style

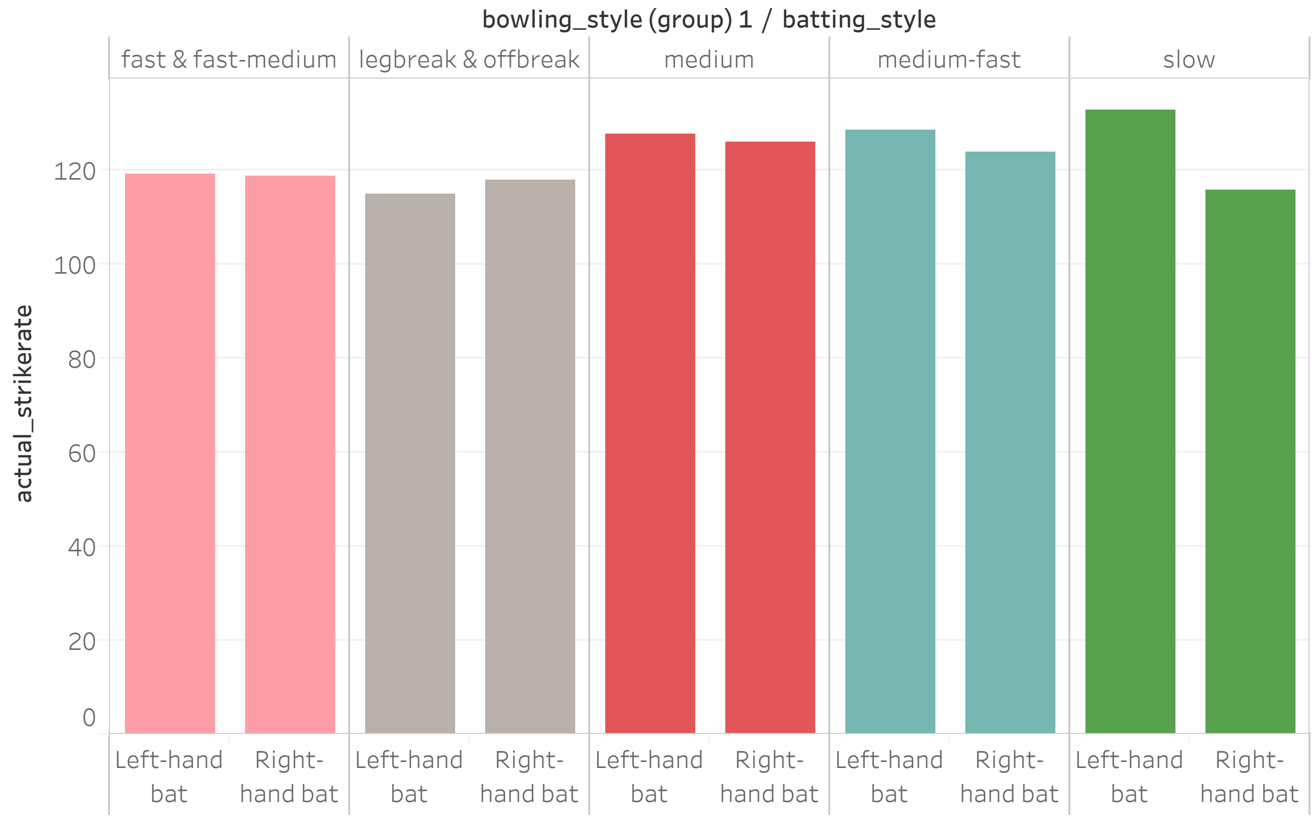
Batting Style Wise Strike Rate Vs Bowling Style

Venue Wise PErformance of KKR

Bowling Style Wise Average Against Bating Style

Bowling Style Wise Strike Rate Vs Batting Style

Average Runs Vs Strike Rate



This shows how Left-handed and Right-handed batsmen perform against various bowlers by comparing strike rate of the batsmen. From the graph it can be concluded that Right-handed batsmen fare better against Legbreak and Offbreak bo..

Cricket Insights

Team vs Opening Partnership

Average Run rate vs Over

Batting Style Wise Average Against Bowling Style

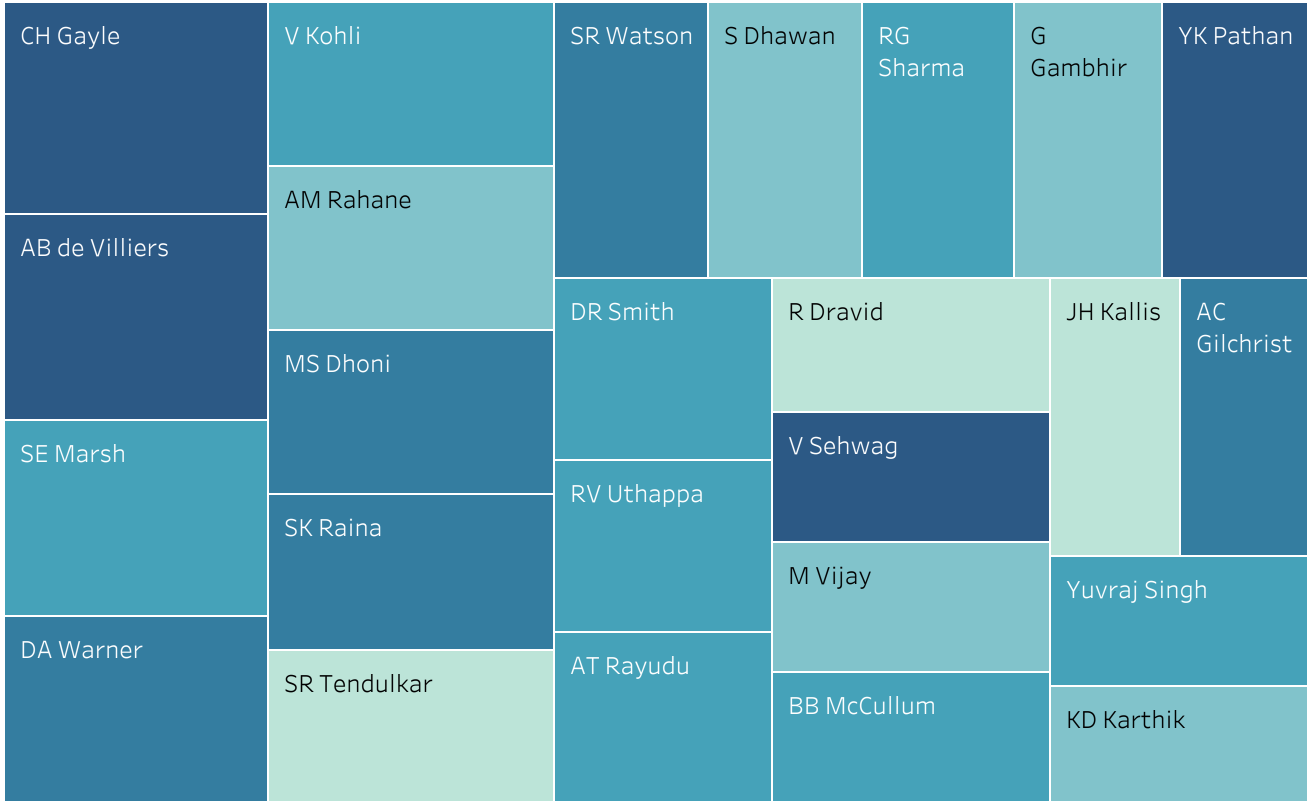
Batting Style Wise Strike Rate Vs Bowling Style

Venue Wise PErformance of KKR

Bowling Style Wise Average Against Bating Style

Bowling Style Wise Strike Rate Vs Batting Style

Average Runs Vs Strike Rate



This is a tree map where the size represents the average runs scored and the color represents strike rate of a batsman. We have taken the top batsmen as for the visualization, the size of the cells are somewhat comparable but players like Gayle, De Villiers and Pathan are known for big hitting and that is evident from the color of their cells.

Cricket-Insights Project Deliverable
Predictive Analytics using Classification

Abhinav Chawla(IMT2013002)

Aditya Naidu(IMT2013003)

Shivam Kumar(IMT2013042)

Siddartha Padhi(IMT2013043)

1 Introduction

This document presents the results that we obtained for predictive analysis using classification on our dataset. We used three main classification algorithms on our dataset with several parameters and analyzed the accuracy for each using a confusion matrix. The algorithms that we experimented with are as follows:

- Random Forest
- Decision Trees
- Support Vector Machine

Overall, average accuracy that we received for all the algorithms was around 75% with SVM reaching the maximum accuracy of 82.69%. In the subsequent sections, we present the details of the dataset used for training and testing these algorithms and the results that we obtained.

2 Dataset

Our raw data consists of 577 yaml files where each file contains ball-by-ball details of IPL matches. The data was scraped from CricInfo for all the IPL seasons i.e. 2008-2016. We decided to divide the data in every match into three sections namely segment-one(0-6 overs), segment-two(7-13 overs) and segment-three(14-20) overs. For each segment, one csv file was maintained containing the following data extracted using python scripts:

- Match_ID: Unique identifier for a match
- Venue: Location of play
- Team1: Team which had to bat first
- Team2: Team which had to bat second
- MR: Runs scored by Team1 that segment
- OR: Runs scored by Team2 in that segment
- MRN: Run rate of Team1 in that segment
- ORN: Run rate of Team2 in that segment
- MW: Wickets lost by Team1 in that segment
- OW: Wickets lost by Team2 in that segment
- Toss_win: Winner of toss for the match
- Win_Bat_First: A binary column, where a 1 corresponds to victory of team batting first and 0 to victory of that batting second

For our classification and prediction purposes, we used the third segment and used all the above mentioned attributes.

2.1 Training Data

We start by randomly shuffling the data of third segment in R. Then we pick top 70% of the records for training and the rest 30% is reserved for testing. Before using any of the algorithms mentioned in the previous section, we have scaled the quantitative attributes MR,OR,MRN,ORN,MW,OW using scale function in R. Scale function in R by default uses Z-score method to normalize each column. After normalization, we assigned the following weights to attributes:

- MR: 0.65
- OR: 0.65
- MW: 0.35
- OW: 0.35
- MRN: 0.65
- ORN: 0.65

By assigning weight for example of 0.65 to MR, we mean that we replaced the column MR by 0.65*MR in the data frame. We arrived at these weights by a series of random trials to check which combination of parameters gives better accuracy.

2.2 Test Data

The 30% of the data reserved for testing the model is stripped off the Win_Bat_First column because that is the one that we want to predict. The attributes which were normalized in training data are also normalized in test data before giving it as input to the model for prediction. In the subsequent section, we describe the results that we received for three different models.

Aim Our aim for predictive analysis was to predict which team won the match based on their performance in the third segment i.e. 14-20 overs.

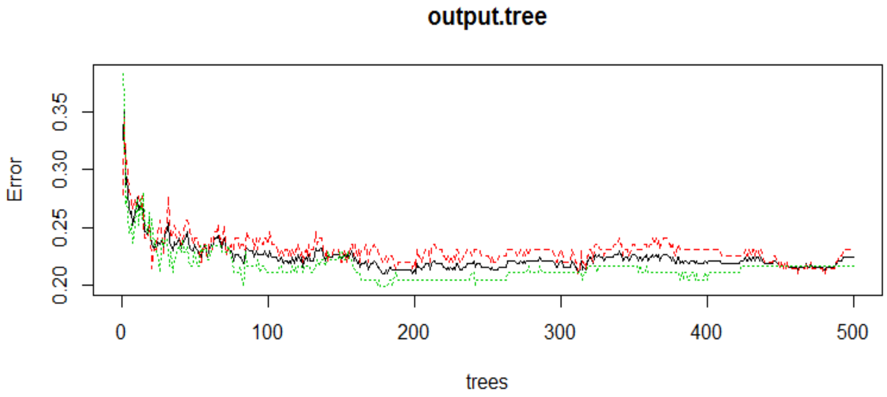
3 Results

In this section, we describe the results and the plots that we got for each of the algorithms. We mention the accuracy(Number of correctly classified rows/Total rows in test data) and the model plot for each.

3.1 Random Forest

- Accuracy: 76.92%
- Confusion Matrix

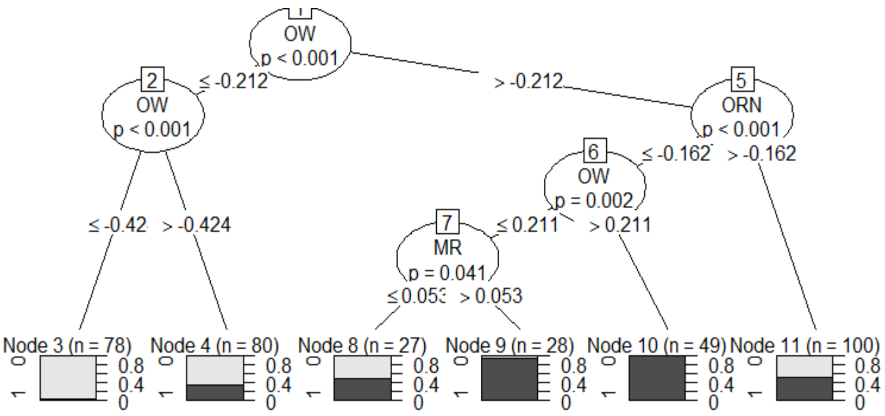
	Reference 0	Reference 1
Predicted 0	59	16
Predicted 1	20	61



3.2 Decision Trees

- Accuracy: 77.5641%
- Confusion Matrix

	Reference 0	Reference 1
Predicted 0	57	13
Predicted 1	22	64



3.3 Support Vector Machines

- Accuracy: 82.692%
- Confusion Matrix

	Reference 0	Reference 1
Predicted 0	67	15
Predicted 1	12	62

4 Conclusion

Support vector machine algorithm gave us the highest accuracy our test dataset. On an average, the accuracy for different training and test data sets obtained by random shuffling vary from 72%-83%. When we increase our train data ratio from 70% onwards, the accuracy for all the models was noticed to fall down. Scaling the attributes and assigning them weights played a crucial role in improving the model accuracy as without normalization, the average accuracy was noticed to be around 61%.

Cricket-Insights Project Deliverable Predictive Analytics using Association Rules

Abhinav Chawla(IMT2013002) Aditya Naidu(IMT2013003)

Shivam Kumar(IMT2013042) Siddartha Padhi(IMT2013043)

1 Introduction

This document presents the results that we obtained for predictive analysis using association rules on our dataset. We used the apriori algorithm using the apriori function call in R from arules library. We analyzed the rules obtained by varying parameters such as minimum length, minimum support and/or minimum confidence. We also tried to obtain association rules by forcing certain constraints on L.H.S and/or R.H.S for the rules. In the subsequent sections, we describe the data preparation phase, the rules that we obtained for our dataset and their interpretations.

2 Data Preparation

For our previous deliverables, we had already extracted both quantitative and categorical data from our raw data using several scripts written in R and python. Further processing was required before we could use the available data for deriving the association rules. In case of columns with categorical data, processing was trivial as the values in such columns was just prepended with the column name itself. For quantitative data however, processing was carried out in the following two phases:

1. **Discretization**: Quantitative columns were converted to columns with discrete values of HIGH, MEDIUM or LOW. Initially, we used the default discretize function call in R to make our data discrete but it did not give useful association rules in many of the cases. We then discretized the columns based on our domain knowledge. For example, An individual score of 30 and above in IPL was categorized as HIGH while below that was categorized as LOW.

2. **Prefixation**: Discretized values were then prepended with the column name as was done for categorical data.

3 Results

After data preparation, we generated association rules and picked those whose confidence was atleast 0.6. Generated rules were sorted in a descending order based on lift for analysis. The rules and their interpretations are presented below.

Rule	Support	Confidence	Lift
op_partner_name=CH Gayle-TM Dilshan, partnership_label=HIGH ⇒ op_winner=Royal Challengers Bangalore	0.008591065	0.8333333	7.028986
op_partner_name=CH Gayle-TM Dilshan ⇒ op_winner=Royal Challengers Bangalore	0.012027491	0.7368421	6.2151030
op_partner_name=CH Gayle-TM Dilshan ⇒ partnership_label=HIGH	0.010309278	0.6315789	1.6520402
Interpretation: If Gayle and Dilshan open for a match, they tend to have a high partnership and if they do, RCB will most likely win.			

Rule	Support	Confidence	Lift
op_partner_name=CH Gayle-V Kohli, partnership_label=HIGH ⇒ op_winner=Royal Challengers Bangalore	0.007731959	0.75	6.3260870
op_partner_name=CH Gayle-V Kohli ⇒ op_winner=Royal Challengers Bangalore	0.012886598	0.60	5.0608696
Interpretation: If Gayle and Kohli open for a match and have a high partnership, RCB will most likely win.			

Rule	Support	Confidence	Lift
Toss=Toss_Gujarat Lions ⇒ Elected_to=field	0.01441441	1	1.843854
Interpretation: If Gujarat Lions win the toss, they always elect to field.			

Rule	Support	Confidence	Lift
Toss=Toss_Royal Challengers Bangalore,Elected_to=field, City=Bangalore ⇒ Win_team=win_team_Royal Challengers Bangalore	0.01981982	0.6111111	5.062189
City=Bangalore ⇒ Elected_to=field	0.08468468	0.8545455	1.575657
Interpretation 1: If RCB wins the toss in Bangalore and elects to field, they are likely to win.			
Interpretation 2: Teams who win the toss at Bangalore usually elect to field first.			

Rule	Support	Confidence	Lift
Toss=Toss_Pune Warriors,Elected_to=field ⇒ loser=loser_Pune Warriors	0.01621622	1.0000000	16.818182
Interpretation: If Pune Warriors win the toss and elect to field, they always lose the match.			

4 Conclusion

We were able to relate many of the association rules to the trends that we derived in the descriptive analysis phase. Association rules are a useful tool in deriving trends from the available data and can be used for predictive analysis.